دانشکده مهندسی کامپیوتر
رشته مهندسی کامپیوتر، گرایش هوش مصنوعی

رساله دکتری

به‌کارگیری محتوای مکانی در مدل‌سازی موضوعی احتمالاتی

نگارنده: مرضیه رحیمی

استاد راهنما:
دکتر مرتضی زاهدی

استاد مشاور:
دکتر هدی مشایخی

آذر ماه ۱۳۹۶

پیوست شماره ۲

دانشکده : مهندسی کامپیوتر

گروه : هوش مصنوعی

رساله دکتری خانم مرضیه رحیمی

تحت عنوان: به‌کارگیری محتوای مکانی در مدل‌سازی موضوعی احتمالاتی

در تاریخ **۱۳۹۶/۰۹/۲۹** توسط کمیته تخصصی زیر جهت اخذ مدرک رساله دکتری ارزیابی گردید و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:		نام و نام خانوادگی:	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:			
نام و نام خانوادگی:			

به نام خدا

در اینجا لازم می‌دانم که پس از سپاس از خداوند بزرگ به خاطر تمام آنچه به من عطا فرموده و پدر و مادر عزیزم که راهنمایی‌ها و پشتیبانی‌هایشان همیشه باعث دلگرمی و آرامشم بوده است، از استاد راهنمای عزیز و بزرگوارم جناب آقای دکتر مرتضی زاهدی به خاطر راهنمایی‌های بی‌دریغ و زحمات دلسوزانه‌شان در تمامی دوران تحصیلم در دانشگاه صنعتی شاهرود و همه آنچه از ایشان آموختم و نیز از استاد محترم مشاورم در این پایان‌نامه خانم دکتر مهدی مشایخی که صبورانه از هیچ کمکی در این مسیر کوتاهی نکردند، تشکر و قدردانی نمایم. همچنین از استادان گرانقدرم آقای دکتر علی اکبر پویان، دکتر حمید حسن پور و دکتر علیرضا احمدی فرد که به دلیل همه آنچه از ایشان آموختم همیشه مدیون ایشان خواهم بود، از خواهر خوبم به خاطر همه‌ی همراهی‌ها و همفکری‌هایش و همه دوستانم در آزمایشگاه وب کاوی و شناسایی الگو، تشکرم.

تعهد نامه

اینجانب **مرضیه رحیمی** دانشجوی دوره دکتری رشته **مهندسی کامپیوتر** دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **به‌کارگیری محتوای مکانی در مدل‌سازی موضوعی احتمالاتی** تحت راهنمائی **دکتر مرتضی زاهدی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهش‌های محققان دیگر، به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- *کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.*
- *استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.*

چکیده

در تحلیل داده‌های متنی، مدل‌سازی موضوعی آماری نقش بسیار جدی و پویایی را بازی می‌کند. زیرا در این حیطه، به دلیل اینکه متون بر اساس کلماتشان توصیف می‌شوند، با داده‌هایی با ابعاد بالا و تقریبا بی‌ساختار روبرو هستیم. بنابراین، ارائه مدلی موضوعی که داده‌های متنی را به جای کلمات، بر مبنای موضوعات توصیف کند، می‌تواند ضمن کاهش ابعاد داده، وجوه معنادارتری از متون را نیز برجسته نماید. در ارائه چنین مدلی معمولاً مبنای تصمیم‌گیری، آمار همرخدادی کلمات در محدوده یک سند است. یکی از مشکلات بسیاری از مدل‌های موضوعی که مبتنی بر همرخدادی کلمات در محدوده یک سند هستند، عدم توانایی در بهره‌گیری از روابط محلی یا مکانی کلمات است. برخی از محققان سعی کرده‌اند با ترکیب مدل‌های موضوعی و مدل‌های زبانی موجود، این مشکل را حل نمایند. اما وارد کردن ترتیب دقیق کلمات، مشکل تنکی را با خود به همراه می‌آورد. از طرفی، در بسیاری از کاربردها، ترتیب کلمات نقش حیاتی بازی نمی‌کند.

در این رساله، هدف ما طراحی مدلی است که بتواند از اطلاعات نسبی مکانی و ارتباطات محلی کلمات بهره برده و در عین حال، مشکل تنکی را برجسته ننماید. برای این منظور به‌جای تکیه بر ترتیب دقیق کلمات، از موقعیت نسبی آنها بهره‌گرفته‌ایم. یعنی برای هر کلمه از متن، پنجره‌ای بر روی کلمات اطرافش تصور شده است که از آن، برای کد کردن اطلاعات مربوط به موقعیت نسبی کلمات، بهره برده‌ایم.

برای ارزیابی، از سه معیار پرپلکسیتی، انسجام و قابلیت مدل در کاربرد خوشه‌بندی، بر روی داده‌های واقعی استفاده کرده‌ایم. نتایج این آزمایشات، همچنین با برخی از مدل‌های مرتبط معرفی‌شده در سال‌های اخیر، مقایسه شده و بهبودهای چشمگیری را در هر سه معیار، نشان داده‌اند. مدل‌های پیشنهادی در بهترین حالت، قادرند موضوعاتی را تولید نمایند که نسبت به موضوعات تولیدشده توسط روش‌های موجود، ۲۸٪ منسجم‌ترند.

کلمات کلیدی: مدل‌سازی موضوعی احتمالاتی، تحلیل متن، تخصیص پنهان دریکله، مدل‌های گرافیکی، نمونه‌برداری گیبس، مدل‌های مولد، ارتباطات محلی کلمات، همرخدادی

لیست مقالات مستخرج

مقالات پذیرفته یا چاپ شده:

Rahimi, Marziea, Morteza Zahedi, and Hoda Mashayekhi. “Incorporating local word relationships into probabilistic topic models”. In *Information and Knowledge Technology (IKT), 2015 7th Conference on*, pp. 1-4. IEEE, 2015.

Rahimi, Marziea, Morteza Zahedi, and Hoda Mashayekhi. “A two level probabilistic topic model” In *Electrical Engineering (ICEE), 2016 24th Iranian Conference on*, pp. 108-112. IEEE, 2016.

Rahimi, Marziea, Morteza Zahedi, and Hoda Mashayekhi. “A Probabilistic Topic Model based on an Arbitrary-Length Co-occurrence Window.” In *International Journal of Information & Communication Technology Research (IJICTR), 2017, v. 9, no. 2, pp.19-27.*

مرضیه رحیمی، مرتضی زاهدی و هدی مشایخی. ”یک مدل موضوعی آماری جهت درنظر گرفتن وابستگی محلی کلمات“ در ۲۱امین کنفرانس انجمن کامپیوتر ایران- اسفند ۹۴

مرضیه رحیمی، مرتضی زاهدی و هدی مشایخی. ” استخراج روابط وابسته به ترتیب کلمات با استفاده از یک مدل سلسله‌مراتبی بیز“ در مجله مهندسی برق دانشگاه تبریز.

سایر مقالات:

Rahimi, Marziea, Morteza Zahedi, and Hoda Mashayekhi. “Two-Level Latent Dirichlet Allocation.” *Machine learning*: under-review, since 1 Nov 2017.

Rahimi, Marziea, Morteza Zahedi, and Hoda Mashayekhi. “A Probabilistic Topic Model based on Short Distance Co-occurrences.” *Knowledge and Information Systems*: under-review, since 7 Aug 2017.

مرضیه رحیمی، مرتضی زاهدی و هدی مشایخی. ”یک مدل موضوعی احتمالاتی مبتنی بر روابط محلی کلمات در پنجره‌های همپوشان“ در مجله پردازش علائم و داده‌ها. تحت داوری از ۳۰ اردیبهشت ۹۶.

فهرست مطالب

فهرست جداول

فهرست تصاویر

فهرست اختصارات

20 Newsgroups (20NG)

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Co-Clustering (LDCC)

Latent Semantic Analysis (LSA)

Probabilistic Latent Semantic Analysis (PLSA)

Reuters-R8 (R8)

Singular Value Decomposition (SVD)

لغت‌نامه فارسی به انگلیسی

اتحاد معنایی	Semantic association
انسجام	Coherence
امیدریاضی-بیشینه‌سازی	Expectation-maximization
ایست‌واژه	Stopword
باناظر	Supervised
بی‌درنگ	Real-time
بیش‌برازش	Over-fitting
بی‌ناظر	Unsupervised
پسین	Posterior
پیشین	Prior
تجزیه مقادیر تکین	Singular Value Decomposition (SVD)
تحلیل احتمالاتی معنای پنهان	Probabilistic Latent Semantic Analysis (PLSA)
تحلیل معنای پنهان	Latent Semantic Analysis (LSA)
تخصیص پنهان دریکله	Latent Dirichlet Allocation (LDA)
تشخیص مولف	Authorship attribution
تقلیل مطلق	Absolute discounting
تنکی	Sparseness
توسعه درخواست	Query expansion

توصیه‌گر	Recommender
جابه‌جایی پذیری	Exchangeability
چندمعنایی	Polysemy
خارجی	Extrinsic
خوشه‌بندی سخت	Hard clustering
خوشه‌بندی نرم	Soft clustering
خوشه‌بندی همزمان دریکله	Latent Dirichlet Co-Clustering (LDCC)
درخت وابستگی	Dependency tree
درستنمایی	Likelihood
غیرعملی	Intractable
فاصله تغییرات اطلاعات	Variation of Information (VI) distance
فراوانی‌گرا	Frequentist
کلمات دستوری	Function words
کلمات شاخص	Descriptor words
کلمات معنایی	Content words
کلمات منفرد	Unique words / Vocabulary
کیسه کلمات	Bag of words
کیسه قطعات	Bag of segments
گستره	Span
ماهوی	Intrinsic

Mixture	مخلوط
Conjugate	مزدوج
Aspect model	مدل وجهی
Topic	موضوع
Generative	مولد

لغت‌نامه انگلیسی به فارسی

Aspect model	مدل وجهی
Authorship attribution	تشخیص مولف
Bag of segments	کیسه قطعات
Bag of words	کیسه کلمات
Coherence	انسجام
Conjugate	مزدوج
Content words	کلمات معنایی
Dependency tree	درخت وابستگی
Descriptor words	کلمات شاخص
Exchangeability	جابه‌جایی پذیری
Expectation-maximization	امیدریاضی-بیشینه‌سازی
Extrinsic	خارجی
Frequentist	فراوانی‌گرا
Function words	کلمات دستوری
Generative	مولد
Hard clustering	خوشه‌بندی سخت
Intractable	غیرعملی
Intrinsic	ماهوی

Latent Dirichlet Allocation (LDA)	تخصیص پنهان دریکله
Latent Dirichlet Co-Clustering (LDCC)	خوشه‌بندی همزمان دریکله
Latent Semantic Analysis (LSA)	تحلیل معنای پنهان
Likelihood	درستنمایی
Mixture	مخلوط
Over-fitting	بیش‌برازش
Prior	پیشین
Probabilistic Latent Semantic Analysis (PLSA)	تحلیل احتمالاتی معنای پنهان
Polysemy	چندمعنایی
Posterior	پسین
Real-time	بی‌درنگ
Query expansion	توسعه درخواست
Recommender	توصیه‌گر
Semantic association	اتحاد معنایی
Singular Value Decomposition (SVD)	تجزیه مقادیر تکین
Soft clustering	خوشه‌بندی نرم
Span	گستره
Sparseness	تنکی
Stopword	ایست‌واژه
Supervised	باناظر

Topic	موضوع
Unique words	کلمات منفرد
Unsupervised	بی‌ناظر
Variation of Information (VI) distance	فاصله تغییرات اطلاعات

۱

مقدمه

۱- مقدمه

بسیاری از سیستم‌های پیچیده را می‌توان با استفاده از قواعد و مفاهیم آماری تحلیل و بررسی نمود. هدف نهایی مدل‌سازی احتمالاتی، توصیف داده‌های مشاهده‌شده با استفاده از تئوری احتمالات است. برای اینکه چنین توصیفی عملی باشد، معمولا نیازمندیم تا فرض‌هایی را برای ساده‌سازی سیستم تولیدکننده داده‌های مورد توصیف درنظر بگیریم. در حیطه تحلیل متن، یکی از فرض‌های رایجی که برای ساده‌سازی در نظر گرفته می‌شود، فرض جابه‌جای‌پذیری[۱] [۱, ۲] کلمات یک متن یا سند است. یعنی برای نمایش متن از مدل کیسه کلمات[۲] استفاده می‌شود. نمایش کیسه کلمات به دلیل سادگی، نادیده گرفتن ترتیب در متن و تکیه بر کلمات منفرد متن[۳]، روش کارآمدی است ولی تعداد ابعاد فضای برداری حاصل که هریک متناظر با یکی از کلمات منفرد متن است، بسیار زیاد است. زمانی که از نمایش کیسه کلمات استفاده می‌نماییم، یکی از چالش‌هایی که با آن روبرو می‌شویم این است که یک مفهوم خاص را می‌توان با کلمات مختلفی بیان کرد (هم‌معنایی[۴]) و در عین حال از یک کلمه می‌توان در بیان مفاهیم مختلفی استفاده نمود (چندمعنایی[۵]) یعنی یک کلمه بسته به زمینه مورد استفاده ممکن است معانی متفاوتی بیابد. بنابراین زمانی که هدف ما تحلیل یک متن زبان طبیعی است تکیه بر کلمات به صورت منفرد منطقی نیست. کاهش ابعاد داده‌ها بدین صورت که ابعاد فضای داده به‌جای کلمات منفرد، مفاهیم و یا به عبارت دیگر، موضوعات باشند که بر مبنای مجموعه‌ای از کلمات بیان شوند، مسئله را حل خواهد نمود.

---

[1]- Exchangeability
[2]- Bag of words
[3]- Vocabulary
[4]- Synonymy
[5]- Polysemy

بحث فوق پایه اولیه شکل‌گیری روش‌های مدل‌سازی موضوعی است. هرچند مدل‌سازی موضوعی بر مبنای کاهش ابعاد داده‌ها آغاز شده است، ولی امروزه موضوعاتی که با روش‌های مطرح در این حیطه تعیین می‌شوند در کاربردهای گوناگونی به کار گرفته می‌شوند که روز به روز در حال گسترش و نو شدن هستند. در ادامه به چند کاربرد متفاوت آنها در سال‌های اخیر اشاره می‌نماییم که می‌توانند بسیار الهام‌بخش باشند. در مخازن نرم‌افزاری[۶]، درک پیغام‌های مربوط به تغییرات نرم‌افزار[۷]، بسیار مهم است. دسته‌بندی آنها به گونه‌ای که دسته‌ها بتوانند مثلا نماینده نوع تغییرات باشند، می‌تواند در مسیر درک این پیغام‌ها مفید باشد. از LDA می‌توان برای دسته‌بندی چنین پیغام‌هایی بصورت نیمه‌نظارتی[۸] [۴] استفاده نمود [۳]. ارزیابی کیفیت تصاویر، کاربرد دیگری است که در آن از LDA بهره گرفته شده است [۵]. به این شکل که هر تصویر، به صورت یک سند در نظر گرفته می‌شود. کلمات این سند، مجموعه‌ای از ویژگی‌های از پیش تعیین‌شده و موثر در ارزیابی کیفیت تصویر هستند. به این ترتیب موضوعات تعیین‌شده در واقع نماینده انواع خرابی‌های ممکن در تصویر خواهند بود و می‌توان برای هر تصویر تعیین کرد که هر خرابی به چه میزان در آن تصویر اثر گذاشته است و در نهایت، بر این اساس، امتیازی را به عنوان کیفیت تصویر محاسبه نمود. می‌توان از مدل‌های موضوعی آماری جهت ارائه یک نمایش چندسطحی از متون استفاده کرد [۶]. به این ترتیب که تعداد موضوعات مختلفی در نظر گرفته می‌شود. هر تعداد موضوع، یک سطح نمایش خاص از سند را ارائه می‌دهد که ارائه‌کنندگان روش، مدعی هستند، می‌تواند نوع خاصی از دانش را در مورد سند نمایش دهد. از مدل‌های موضوعی، همچنین در جهت تشخیص مولف[۹] متون استفاده شده است [۷]. گاهی اوقات دو یا چند نفر ادعای نگارش یک متن را دارند یا گاهی متنی در اختیار داریم که هیچ یک از افراد مطرح مسئولیت نگارش آن را نمی‌پذیرد. هر فرد سبک نگارش مخصوص به خود را دارد.

---

[6]- Repository
[7]-Change message
[8] -Semi-supervised
[9]- Authorship attribution

به‌خصوص مجموعه کلماتی که هر فرد استفاده می‌کند، تا حد زیادی با دیگران متفاوت است. برای تشخیص نویسنده یک متن، موضوعات متن مورد بررسی، با مجموعه متون دیگری که توسط فرد مورد نظر نوشته شده است، مقایسه می‌گردد و نتیجه این مقایسه، به عنوان امکان نگارش آن متن توسط فرد مذکور محسوب می‌گردد.

کاربردهای فوق، وسعت و تنوع حیطه‌هایی را که مدل‌های موضوعی می‌توانند در آنها مفید باشند و در نتیجه اهمیت این مدل‌ها را نشان می‌دهد. تمرکز ما در متن حاضر، بر روی کاربرد این روش‌ها در تحلیل متن است. مدل‌های موضوعی را عمدتا مدل‌های آماری-احتمالی بی‌ناظر تشکیل می‌دهند. امروزه گسترش روزافزون منابع متنی دیجیتال امکان استفاده از روش‌های آماری را در حیطه تحلیل متن فراهم آورده است و در عین حال نیاز به استفاده از آنها را نیز افزایش داده است. در روش‌های آماری، وابسته به ویژگی‌های ماهیتی و قواعد حاکم بر داده مورد نظر نیستیم. به عنوان مثال، زبان پیکره مورد استفاده محدودیتی برای مدل ایجاد نخواهد کرد. از آنجا که زبان‌های گوناگون دارای قواعد و ماهیت‌های گاه کاملا متفاوت هستند، در روش‌های غیرآماری مثلا روش‌های مبتنی بر قانون[10] [۸, ۹] ناچار خواهیم بود با قواعد زبان دست‌وپنجه نرم کنیم و روش‌های ارائه شده محدود به زبان داده‌های موجود هستند در صورتی که روش‌های آماری قابل اعمال بر هر زبانی خواهند بود.

یک مدل آماری، یک توصیف ریاضی از روابط بین متغیرهاست. این رابطه می‌تواند به صورت احتمالی بیان شود یعنی $p(x,y)$ و یا به صورت یک رابطه معین[11] مثل $y=f(x)$. با این حال در اغلب موارد، مدل‌های آماری، احتمالی هستند [۱۰]. یک مدل احتمالی معمولا به صورت یک زوج (D,P) بیان می‌شود که در آن D مجموعه‌ای از مشاهدات و P مجموعه‌ای از توزیع‌های احتمال بر روی مقادیر D است. فرض بر این است که عنصری در P وجود دارد که تولید کننده مشاهدات است و هدف ما این است که آن عنصر

[10]- Rule based
[11]- Deterministic

را با دقت مناسبی بیابیم[۱۱]. لازم به ذکر است که آمار و احتمال علی‌رغم ارتباط تنگاتنگ، دو حیطه کاملا متفاوت، با تفاوتهای کاملا آشکار هستند. ولی زمانی که صحبت از مدل‌سازی است، این تفاوت‌ها تا حد زیادی محو شده و در بسیاری موارد به‌ویژه حیطه مدل‌سازی موضوعی که مورد بحث متن حاضر است، هرگاه صحبت از مدل‌سازی آماری به میان می‌آید منظور همان مدل‌سازی احتمالی است چنان‌که در متون دیگر این حیطه نیز رایج است مانند [۱۲, ۱۳]. به همین دلیل، در متن حاضر، این دو واژه را به‌جای یکدیگر به کار خواهیم برد.

امروزه، بیشتر مدل‌های موضوعی، برگرفته از مدل‌های وجهی هستند. مدل‌های وجهی بر مبنای آمار همرخداد تصمیم‌گیری می‌کنند ولی نحوه به‌کارگیری مفهوم همرخدادی در آنها بیشتر ضمنی است. استفاده ضمنی به این معنی است که هرچند داده ورودی مستقیما یک ماتریس همرخدادی نیست ولی این روش‌ها عملا بر مبنای همرخدادی کلمات کار می‌کنند [۱۶, ۱۷]. بسیاری از روش‌های آماری مدل‌سازی موضوعی مطرح، جزو این دسته قرار می‌گیرند. به عنوان یک نمونه بسیار مطرح و پرکاربرد می‌توانیم از روش تخصیص پنهان دریکله[۱۲] (LDA) [۱۸] یاد نماییم. در این روش، به‌جای آنکه هر سند را به عنوان هیستوگرامی بر روی مجموعه کلمات نشان دهیم، خوشه‌هایی از کلمات را ایجاد می‌کنیم و هر سند را هیستوگرام یا توزیعی بر روی این خوشه‌ها محسوب می‌نماییم. هریک از این خوشه‌ها، یک موضوع[۱۳] نامیده می‌شود که خود توزیعی بر روی کلمات است. بدین‌ترتیب علاوه بر کاهش ابعاد داده و حل مشکل تنکی[۱۴] مفهوم آشنای موضوعات را نیز وارد دنیای بازیابی اطلاعات [۱۹-۲۲] و تحلیل متن نموده‌ایم.

[12] -Latent Dirichlet Allocation (LDA)
[13] -Topic
[14] -Sparseness

۱-۱- تعریف مسئله

مبنای اصلی تصمیم‌گیری در LDA و بسیاری از مدل‌های موضوعی کنونی، همرخدادی کلمات در محدوده یک سند است. در این مدل‌ها، استفاده از سایر تعاریف رایج برای همرخدادی؛ مثل همرخدادی در یک پنجره یا همرخدادی رابطه که ادعا می‌شود، معمولا نتایج دقیق‌تری را در خوشه‌بندی صریح موجب می‌شوند [۲۴, ۲۵]، ممکن نیست. منطقی است، اگر بیندیشیم که در روش‌هایی مشابه LDA، که به صورت ضمنی مبتنی بر همرخدادی کلماتند نیز، استفاده از یک تعریف دقیق‌تر بتواند، نتایج بهتری را تولید نماید.

با تعریف یک مدل موضوعی، می‌توانیم، ارزش هر کلمه در اسناد را در هر موضوع را تعیین نماییم. در بخش‌های بعد خواهیم دید که احتمال انتساب هر کلمه w_i از سند d_i به هر موضوع مانند z_i به صورت زیر محاسبه خواهد شد که در آن θ_{d_i} نماینده هیستوگرام سند d_i و φ_{w_i,z_i} نماینده توزیع موضوع مربوطه است. :

$$p\left(z_i, w_i \mid \varphi_{w_i,z_i}, \theta_{d_i}\right) = p\left(z_i \mid \theta_{d_i}\right) p\left(w_i \mid z_i, \varphi_{w_i,z_i}\right) \quad (۱-۱)$$

چنان‌که در این رابطه (۱-۱) می‌بینیم، برای تمام تکرارهای یک کلمه مقادیر بدست آمده یکسان خواهند بود و به این ترتیب تمام تکرارهای یک کلمه در یک سند به احتمال زیاد به یک موضوع اختصاص خواهند یافت. اما در یک سند ممکن است یک کلمه در چندین موضوع به‌کار گرفته شود که تعیین آن بستگی به کلمات اطرافش دارد. در رابطه فوق هیچ توجهی به این مسئله نشده است.

در تعریف مدل‌های پایه موضوعی، چندین شرط برای ساده‌سازی مسئله در نظر گرفته شده است. یکی از این شروط، جابه‌جایی‌پذیری[۱۵] در یک سند است که بر اساس آن هیچ ترتیبی برای کلمات متن در نظر گرفته نمی‌شود. یعنی اگر ترتیب کلمات متن را به هر صورتی تغییر دهیم، در آمار بدست آمده و نتیجه

[15]- Exchangeability

حاصل، تغییری ایجاد نخواهد شد. این شرط که ما به عنوان یک انسان، می‌دانیم کاملا غیرواقعی است، یکی از دلایلی است که منجر به مسئله فوق شده است. در تعیین موضوعات یک مجموعه متن، در واقع هدف ما این است که در هر موضوع به کلماتی که دارای ارتباط معنایی بیشتری متناسب با آن موضوع هستند، امتیاز بالاتری نسبت دهیم. اما معنای کلمات در زبان‌های انسانی کاملا وابسته به اطلاعات محلی، یعنی کلمات اطراف و ویژگی‌های آنها نسبت به کلمه مورد بررسی، است. به همین دلیل، یکی از حیطه‌های تحقیق در مدل‌های وجهی، تلاش برای رفع این شرط است. در بسیاری از این روش‌ها، هر کلمه وابسته به دنباله دقیقا مرتب کلمات اطرافش فرض شده است و برای اجتناب از تنکی معمولا این دنباله را یک‌کلمه‌ای فرض می‌کنند یعنی تنها، کلمه بلافاصله قبل از کلمه جاری را در نظر می‌گیرند.

فرض ما این است که برخلاف بسیاری از مدل‌های پیشنهادی که به عنوان توسعه‌ای بر مدل اولیه‌ی LDA صورت گرفته‌اند، لزوما فقط کلمه بلافاصله قبل نیست که در مفهوم کلمه جاری موثر است. چنان‌که در بسیاری موارد، کلمه قبل می‌تواند یک کلمه کاملا عام باشد، مثل یک صفت، و نتواند اطلاعات زیادی را در مورد موضوع کلمه بعد، در اختیار ما بگذارد. این در حالی است که هرکدام از کلمات قبل یا اطراف کلمه جاری ممکن است در موضوع آن موثر باشند. در فصل‌های بعد، شواهدی را به نقل از محققان این زمینه بیان خواهیم کرد که نشان می‌دهد، کلمات اطراف و حتی فاصله آنها با کلمه جاری، برکلمه جاری موثر هستند. همچنین، موقعیت کلمات در متن نیز می‌تواند، بر میزان ارتباط آنها با یکدیگر موثر باشد. چنین ویژگی‌های محلی‌ای را که وابسته به موقعیت کلمات در متن و فاصله نسبی آنها از یکدیگر است، در این پایان‌نامه، با عنوان "محتوای مکانی"، یاد خواهیم نمود. سعی ما این است، مدلی موضوعی ارائه دهیم که قادر باشد، این اطلاعات محلی را به کار گیرد. در ادامه خواهیم دید که بیشتر مدل‌های موضوعی‌ای که سعی در به‌کارگیری اطلاعات محلی دارند، الهام‌گرفته از مدل‌های زبانی n-gram و در نتیجه وابسته به ترتیب دقیق کلمات هستند. چنین رویکردی موجب بروز مشکل تنکی می‌گردد. این درحالی است که در

بسیاری کاربردهای مدل‌های موضوعی مثل بازیابی اطلاعات، ترتیب دقیق کلمات نقش اساسی بازی نمی‌کند. در این کاربردها، بیش از ترتیب کلمات، موقعیت نسبی آنها و ارتباطات محلی‌شان دارای اهمیت است. بنابراین در این پایان‌نامه با به‌کارگیری مفهوم محتوای مکانی در مدل‌های موضوعی، سعی داریم با اجتناب از تشدید مشکل تنکی موضوعاتی منسجم‌تر استخراج نماییم.

در این راستا، مدلی ارائه شده است که در آن یک پنجره بر روی هر کلمه در نظر گرفته می‌شود به گونه‌ای که کلمه مذکور (کلمه جاری)، آخرین کلمه این پنجره است. فرض می‌گردد که کلمه جاری، وابسته به یکی از کلمات موجود در این پنجره است. این کلمه را "مولد" کلمه جاری می‌نامیم. در مدل مذکور، هر موضوع، دیگر یک توزیع منفرد بر روی کلمات نیست بلکه مجموعه‌ای از توزیع‌ها بر روی کلمات است که هریک متناظر با یک کلمه مولد می‌باشد. مشکل این مدل این است که تعداد این توزیع‌ها بسیار زیاد است و درنتیجه مدل با وجود برخورداری از قابلیت‌های مفید، قادر به تولید موضوع‌هایی بهتر از LDA نیست. برای رفع این مشکل، دو مدل مختلف ارائه شده‌اند که سعی می‌کنند تعداد این توزیع‌ها را کاهش دهند. یکی از این مدل‌ها که مدل "همپوشان" نامیده خواهد شد، با اجتناب از فرض صریح وابستگی بین کلمات، این کار را انجام می‌دهد و دیگری که مدل "دولایه" نامیده می‌شود، با تبدیل توزیع‌های متناظر با کلمات مولد، به توزیع‌هایی برای موضوعات خاص‌تر و در نتیجه کاهش تعداد آنها.

در مدل همپوشان، از پنجره‌های همپوشان برای کد کردن موقعیت نسبی کلمات استفاده شده است. در این مدل، همانند مدل اول، ترتیب کلمات نادیده گرفته نشده است ولی وابسته به ترتیب دقیق کلمات نیز نیستیم. بنابراین مشکل تنکی در آنها برجسته نیست. در هرسه‌ی این مدل‌ها، همرخدادی را در حیطه محدودتری از کل یک سند، در نظر گرفته‌ایم. مدل دولایه اما، برخلاف دو مدل دیگر، محدودیت را به موقعیت کلمات اعمال نمی‌نماید بلکه یک سلسله‌مراتب دو سطحی از موضوعات را تشکیل داده و

محدودیت را بر دسته‌بندی این موضوعات اعمال می‌نماید. در ادامه خواهیم دید که هرسه این روش‌ها بهبودهای چشمگیری را نسبت به مدل‌های پایه نشان داده‌اند.

۱-۲- مفروضات

بر اساس آنچه تاکنون گفته شد، مفروضات مسئله را می‌توانیم به صورت زیر خلاصه نماییم:

- ✓ مدل‌های موضوعی احتمالاتی بر مبنای همرخدادی کلمات عمل می‌کنند.
- ✓ بیشتر مدل‌های موجود مبتنی بر همرخدادی کلمات در یک سند هستند.
- ✓ یک متن زبان طبیعی یک رخداد تصادفی است.
- ✓ در تولید یک متن زبان طبیعی، ترتیب دقیق کلمات، نقش اساسی بازی نمی‌کند.
- ✓ کلمات اطراف یک کلمه در تولید (رخداد) آن کلمه موثرند.

۱-۳- ساختار مطالب

در ادامه متن، پس از مروری بر ادبیات موضوع در بخش دو، مقدمات مبحث را به صورت مختصر در بخش سه بیان خواهیم نمود. در این بخش همچنین به بررسی پیشینه و تعاریف موجود برای همرخدادی کلمات، عوامل موثر در آن و کاربردهای رایج آن در حوزه متن می‌پردازیم و عوامل دیده نشده جدیدی را معرفی می‌نماییم. سپس در بخش‌های چهار، پنج و شش، به معرفی سه مدل آماری متفاوت می‌پردازیم که در جهت مباحث مطرح شده پیشنهاد شده‌اند. همچنین این مدل‌ها را در آزمایش‌های مختلف ارزیابی کرده و با چند مدل پایه مقایسه می‌نماییم. نهایتا در بخش هفت به جمع‌بندی موضوع و ارائه پیشنهاداتی برای ادامه کار می‌پردازیم.

۲

مروری بر ادبیات موضوع

۲- مروری بر ادبیات موضوع

بحث تقسیم کلمات موجود در یک مجموعه از اسناد به تعدادی موضوع به معنای خوشه‌های مجزایی که کلمات منتسب به هر یک از آنها دارای مشابهت معنایی باشند، بحث جدیدی در پردازش متن محسوب نمی‌شود. به عنوان مثال تحلیل معنای پنهان[۱] (LSA) [۲۶] یک روش معین است که از تجزیه مقادیر تکین[۲] (SVD) [۲۷] برای نگاشت خطی داده‌هایی که در فضای TF-IDF[۳] [۲۸] نمایش داده شده‌اند به فضایی با ابعاد کمتر استفاده می‌کند. در این فضای جدید که فضای معنای پنهان یا فضای موضوعات خوانده می‌شود، روابط معنایی بین کلمات، آشکار می‌شود و کلمات و اسنادی که دارای ارتباط معنایی هستند نزدیکر به یکدیگر نمایش داده می‌شوند. در واقع کلمات و اسناد در فضای جدید خوشه‌بندی (نرم[۴]) می‌شوند. این مدل در کاربردهای مختلفی به منظور کاهش ابعاد بر اساس روابط معنایی کلمات، با موفقیت، به کار گرفته شده است.

ولی در سال‌های اخیر نگاه جدیدی به مفهوم "موضوع" شده است که حاصل آن مجموعه‌ای از مدل‌های آماری به نام مدل‌های وجهی است. تفاوت اصلی مدل‌های وجهی با روش‌های خوشه‌بندی در این است که در آنها به جای تشکیل موضوعات به عنوان خوشه‌های مجزا سعی می‌کنیم که ارزش هر کلمه را در هر خوشه یا موضوع مشخص نماییم. به این ترتیب هر موضوع در واقع وجه مفهومی متفاوتی از مجموعه کلمات را برجسته می‌نماید.

شروع مبحث مدل‌های وجهی را می‌توان مقاله [۱۲] دانست که در سال ۱۹۹۸ منتشر شده است. در این مقاله یک خانواده جدید از مدل‌های وجهی معرفی می‌شوند که قادرند داده‌های همرخداد را در تعدادی

[1] - Latent Semantic Analysis (LSA)
[2] - Singular Value Decomposition (SVD)
[3] - Term Frequency-Inverse Document Frequency (TF-IDF)
[4] - Soft clustering

وجه[5] مشترک نمایش دهند. با الهام از مفهوم مدل‌های وجهی و همچنین LSA مدل تحلیل احتمالاتی معنای پنهان[6] (PLSA) [۲۹] معرفی شده است. در LSA، ابعاد جدید معرفی شده مفهوم قابل درکی برای انسان ندارند و به این ترتیب نمی‌توان با اطمینان توضیح داد که چرا در فضای جدید، موضوعات متن برجسته شده‌اند. مدل PLSA این مشکل را با معرفی یک مدل مولد آماری حل می‌کند که در آن ابعاد فضای جدید می‌تواند به عنوان لیستی از کلمات قابل تفسیر توسط انسان نمایش داده شوند. در این مدل، هر سند یک مخلوط[7] از موضوعات پنهان متن است و هر موضوع توزیعی بر روی کلمات. بنابراین موضوعات که ابعاد فضای جدید را تشکیل می‌دهند، به صورت مستقیم برای انسان قابل تفسیر هستند. در این مدل، هیچ راهکاری برای تعیین نسبت موضوعات در اسناد ارائه نمی‌شود. در واقع مدلی برای سند ارائه نمی‌شود. بنابراین نمی‌توان برای اسناد جدید، احتمالات لازم را محاسبه نمود. برای رفع این مشکل، LDA معرفی شده است. به طور کلی مدل‌های موضوعی مبتنی بر این ایده هستند که هرسند مخلوطی از موضوعات است و هر موضوع توزیعی بر روی کلمات. مدل ارائه شده در LDA نیز مستثنی از این قاعده نبوده و یک مدل مولد آماری است که بر مبنای فرض فوق، یک فرایند ساده آماری را برای تولید یک سند ارائه می‌کند. مدل‌های مولد آماری در واقع مبتنی بر قواعد ساده نمونه‌گیری آماری هستند که نحوه تولید کلمات یک سند بر مبنای متغیرهای پنهان را مشخص می‌نمایند. در اینجا هدف این است که مجموعه متغیرهای پنهان را به گونه‌ای بیابیم که بهترین توصیف برای داده‌های مشاهده شده (غیرپنهان) باشند، با این فرض که داده‌ها توسط همان مدل تولید شده‌اند. بعد از معرفی LDA در سال ۲۰۰۳، بیشتر تحقیقات انجام شده در زمینه مدل‌های موضوعی شامل توسعه این مدل بوده‌اند. رشد و توسعه این مدل در چندین جهت مختلف صورت می‌گیرد که در ادامه به برخی از آنها می‌پردازیم.

[5]- Aspect
[6] -Probabilistic Latent Semantic Analysis (PLSA)
[7] - Mixture

یکی از این جهات، تخفیف اثر شرطهای غیرواقعی در نظر گرفته شده در مدل پایه مثل شرط جابه‌جایی‌پذیری اسناد [۳۰] است که در برخی کاربردها قابل فرض نیست. مثلا زمانی که بخواهیم روند تغییرات پدید آمده در یک حیطه را در طول زمان بسنجیم یا در متونی مثل چت که ترتیب زمانی تولید پیغام‌ها کاملا در محتوای متون موثر است. نمونه دیگری از این شروط، جابه‌جایی‌پذیری کلمات یک متن است که در ادامه به آن خواهیم پرداخت.

ارائه روش‌هایی برای افزایش دقت یا سرعت محاسبه پارامترها مثل [۳۱, ۳۲] مسیر تحقیقاتی دیگری در این زمینه است. یکی دیگر از حیطه‌های مورد توجه وارد کردن مواردی از دانش اولیه در مورد داده‌ها [۳۳, ۳۴] در مدل است مثل اینکه مثلا در مورد برخی کلمات می‌دانیم که احتمال اینکه مربوط به یک موضوع باشند بیشتر است (مثلا بر مبنای یک فرهنگ لغت[8]). این کار بیشتر از طریق تغییر احتمالات اولیه مدل نجام می‌شود یعنی احتمالات اولیه را بر اساس این دانش اولیه تعیین می‌کنند.

مسیر دیگر نیز مربوط به بهره گرفتن از مدل‌های موضوعی در کاربردهای جدید است. مدل‌های موضوعی به سرعت در حیطه‌های مختلف مورد توجه و استفاده قرار گرفته‌اند و دائما کاربردهای جدید پیدا می‌کنند که به نمونه‌هایی از آنها در مقدمه اشاره نمودیم.

برخی از محققان سعی کرده‌اند مدل‌های موضوعی را به‌گونه‌ای تغییر دهند که قابل اعمال بر روی داده‌های دارای برچسب بوده و بتوانند از اطلاعات اضافه‌ای که برچسب‌ها فراهم می‌کنند بهره بگیرند. این مدل‌ها را مدل‌های موضوعی باناظر [۳۵, ۳۶] می‌نامند.

یکی از مشکلات LDA این است که به دلیل مشکل تنکی، بر روی متون کوتاه خوب عمل نمی‌کند. بر همین اساس شاخه‌ای از توسعه مدل‌های موضوعی تمرکز بر ارائه مدل‌هایی دارند که بتوانند بر روی جملات یک سند یا متون کوتاه، عمل کنند. به عنوان مثال چنگ و همکاران روشی [۳۷] ارائه داده‌اند که در آن سعی شده است تا امکان اعمال مدل‌های موضوعی بر روی متون کوتاه فراهم شود. در این مقاله،

[8]- Thesaurus

داده‌های ورودی، دیگر، اسناد مجموعه داده نیستند و به‌جای مدل کردن اسناد، زوج کلمات موجود در مجموعه داده مدل می‌شوند. به این ترتیب که ابتدا تمام زوج کلماتی را که در پنجره‌هایی با اندازه معین رخ داده‌اند، استخراج کرده و سپس با نادیده گرفتن اسناد، این زوج کلمات را با استفاده از تعدادی موضوع مدل می‌نماید. در این مدل فرض می‌شود که هر دو عضو یک زوج کلمه منتسب به یک موضوع یکسان هستند با این تصور که زوج کلماتی که به فراوانی در یک محدوده مشخص با یکدیگر اتفاق می‌افتند، به احتمال زیاد متعلق به یک موضوع هستند. به این ترتیب، با در نظر گرفتن زوج کلمات به صورت سراسری و در کل مجموعه به جای در نظر گرفتن کلمات در اسناد کوتاه مشکل تنکی حل می‌شود. در این مدل با استخراج زوج کلمات به شکلی که معادل همرخدادی مرتبه اول آنهاست، اطلاعات همرخدادی‌های مراتب بالاتر را از دست داده‌ایم. در واقع این روش را می‌توان یک مصالحه بین از دست دادن بخشی از اطلاعات و کاهش اثر تنکی در متون کوتاه هنگام استفاده از روش‌های استخراج موضوع دانست. کاربرد این روش و روش‌های مشابه [۳۸] این است که احتمال استخراج موضوعات در متون کوتاه را فراهم نمایند یا به عبارت دیگر امکان مدل کردن متون کوتاه با استفاده از موضوعات را. هدف ما در این پایان‌نامه این است که از موقعیت مکانی یا ارتباطات محلی کلمات که وابسته به موقعیت نسبی آنها در کنار یکدیگر است برای بهبود مدل‌های موضوعی بهره بگیریم. به عبارت دیگر، هدف استخراج موضوعات برمبنای همرخدادی‌های سختگیرانه‌تر کلمات به گونه‌ای است که قادر باشیم ارتباطات دقیق‌تر و قابل‌اعتمادتری را بین کلمات استخراج نماییم. بنابراین هدف این مدل‌ها کاملا متفاوت با هدف روش پیشنهادی است.

چنان‌که ذکر شد یکی از مسیرهای تحقیقات در زمینه مدل‌های موضوعی آماری تلاش در جهت حذف شرط جابه‌جایی‌پذیری کلمات است. این جهت از تحقیقات مورد توجه خاص ما در این پژوهش است بنابراین در ادامه به بررسی مفصل روش‌های ارائه شده در این جهت می‌پردازیم.

بیشتر این مدل‌ها مبتنی بر مدل‌های زبانی n-gram هستند. این دسته البته متعلق به دسته بزرگتری از مدل‌های موضوعی است که در استخراج موضوعات بر ارتباط محلی کلمات تکیه کرده و سعی می‌کنند همرخدادی را در بازه‌ای محدودتر از کل یک سند (چنان‌که در LDA انجام می‌شود) در نظر گیرند. اولین نمونه این مدل‌ها را می‌توان مدل موضوعی بایگرم[9] (BTM) [۳۹] دانست. در این مدل هنگام تعیین موضوعات در یک سند، ترتیب کلمات نیز در نظر گرفته می‌شود یا به عبارتی مدل موضوعی بر مبنای یک مدل زبانی بایگرم عمل می‌کند که البته قابل گسترش به مدل‌های ترایگرم[10] و بالاتر نیز هست. برای این کار از یک مدل زبانی بایگرم [۴۰] که در برمبنای توزیع سلسله‌مراتبی دریکله ارائه گردیده، استفاده شده است. در مدل BTM هر موضوع، مجموعه‌ای از N توزیع بر روی کلمات است و هر توزیع نماینده یکی از کلمات ممکن به عنوان کلمه پیشین می‌باشد. مدل BTM الهام‌بخش بسیاری از مدل‌های این دسته است. این مدل نیز به تفصیل در بخش بعد بررسی شده است. در مدل BTM فرض شده است که هر کلمه علاوه بر موضوع خود، وابسته به کلمه‌ی پیشین خود نیز هست.

مدل مشابه دیگری [۴۱] توسط *باربیری* و همکارانش پیشنهاد شده است که همین فرض را در نظر می‌گیرد. در این مدل، سعی شده است شرط عدم اهمیت ترتیب کلمات سند به سه شکل مختلف شکسته شود به گونه‌ای که در هر یک از این اشکال، تولید کلمه جاری وابسته به اطلاعات مربوط به کلمه قبل نیز باشد. برای این منظور سه مدل مختلف ارائه شده‌اند. در اولین مدل، مشابه BTM، هرکلمه به کلمه بلافاصله قبل از خودش وابسته است و در دومی، هر کلمه علاوه بر کلمه قبلی و موضوع آن کلمه نیز وابسته خواهد بود. در سومین مدل، انتخاب کلمه فعلی به موضوع کلمه قبلی وابسته است. هر یک از این مدل‌ها در یک سیستم توصیه‌گر[11] به کار گرفته شده‌اند و هر یک بهبودهایی را در معیارهای دقت[12] و

[9]- Bigram Topic Model (BTM)
[10] - Trigram
[11] - Recommender
[12] - Precision

فراخوانی مجدد[۱۳] [۴۲] نشان داده‌اند. *گریفیس* و همکارانش مدلی [۴۳] را پیشنهاد کرده‌اند که فرض می‌کند هر دو کلمه پشت سر هم می‌توانند یک "ترکیب" را تشکیل دهند. یعنی هر کلمه یا توسط یک موضوع تولید می‌شود و یا توسط کلمه پیشینش. در مدل مربوطه که LDA-Collocation نامیده می‌شود، برای انتخاب یکی از این دو حالت از یک متغیر برنولی استفاده شده است. ونگ و همکارانش تعمیمی [۴۴] بر LDA-Collocation ارائه کرده‌اند که در آن هر کلمه بر مبنای موضوع خود می‌تواند تصمیم بگیرد که آیا با کلمه قبلی یک ترکیب را تشکیل می‌دهد یا خیر. ینگ و همکارنش [۴۵] فرض مشابهی را در نظر می‌گیرند و علاوه بر آن فرض می‌کنند که یک سلسله‌مراتب از موضوعات وجود دارد و هر کلمه مسیری مشخص را در این سلسله مراتب طی می‌کند تا توسط یک موضوع خاص تولید شود. جمیل و همکارانش [۴۶] یک مدل موضوعی باناظر را در ترکیب با مدل زبانی بایگرم ارائه داده‌اند. هرچند بسیاری از این مدل‌ها قابل تعمیم به n-gramهای بالاتر هستند ولی عموم آنها فقط یک کلمه قبل را در نظر می‌گیرند. دلیل این امر چنان‌که توضیح داده شد مسئله تنکی است.

در مقاله دیگری [۴۷] الگوریتمی ارائه می‌شود که قصد آن برقراری ارتباط بین وابستگی‌های دوربرد و کوتاه‌برد در یک مدل موضوعی است. وابستگی‌های دوربرد بیشتر نماینده‌ی ارتباطات معنایی هستند؛ از آنجا که جملات یک متن از نظر معنایی با یکدیگر مرتبط هستند بنابراین کلمات موجود در جملات مختلف با یکدیگر وابستگی معنایی دارند. در حالی که وابستگی‌های کوتاه‌برد بیشتر در یک جمله قابل تعریف بوده و در واقع بیان‌کننده ارتباطات دستوری هستند. مدل ارائه شده کلمات را به دو نوع کلمات معنایی[۱۴] و کلمات دستوری[۱۵] تقسیم می‌نماید. مدل ارائه شده دارای دو مولفه است؛ یک مدل پنهان مارکوف (HMM)[۱۶] که نماینده کلمات دستوری است و دیگری یک مدل موضوعی که نماینده کلمات

[13] - Recall
[14]- Content word
[15]- Function word
[16] -Hidden Markov Model (HMM)

معنایی است. مدل مذکور می‌تواند در ساختار مدل موضوعی، نقش دستوری کلمات را نیز لحاظ نماید. بدین ترتیب در مدل مذکور، برخلاف مدل LDA، نیازی به حذف ایست‌واژه‌ها[17] نیست و به مدل به خوبی قادر است آنها را نیز بر اساس نقششان دسته بندی نماید.

دسته دیگر که زیر دسته‌ای از مدل‌های فوق هستند، از فرایند پیتمن-یور (HPY)[18] [۴۸] استفاده می‌کنند. این دسته از روش‌ها که عموما برگرفته از [۴۸] هستند، برخلاف دسته قبل، فقط محدود به یک کلمه قبل نبوده‌اند و نتایج آن‌ها برای ترکیباتی با طول‌های مختلف گزارش شده است. مدل‌های این دسته نیز با توجه به مشکل تنکی باید بر روی مجموعه‌داده‌های بسیار بزرگ آموزش داده شوند تا نتایج آنها قابل اعتماد باشد. اما این مدل‌ها با توجه به بار محاسباتی سنگین، بر روی مجموعه‌داده‌های بزرگ بسیار پرهزینه و غیرعملی هستند. [۴۹] و [۵۰] نمونه‌هایی از مدل‌های این دسته‌اند.

دسته دیگری از مدل‌ها که می‌توان آن را مدل‌های موضوعی مبتنی بر قطعه (به جای سند کامل) نام نهاد، برخلاف مدل‌های دسته قبل، ترتیب کلمات را نادیده می‌گیرند ولی لایه‌ی دیگری از جزئیات را به مدل می‌افزایند. مدل‌های این دسته عمدتا الهام گرفته از مدل خوشه‌بندی همزمان پنهان دریکله[19] (LDCC) [۵۱] هستند. در این مدل‌ها فرض می‌شود که هر سند مجموعه‌ای از قطعات جابه‌جایی‌پذیر است. این قطعات می‌تواند یک پاراگراف، یک جمله [۵۲] یا هر قطعه دیگری از متن [۵۳] باشد که البته بیش از اندازه کوچک نباشد که مشکل تنکی را غیرقابل تحمل نماید. معمولا در این نوع روش‌ها دو نوع موضوع در نظر گرفته می‌شود: قطعه‌موضوع و کلمه‌موضوع. در LDCC هر قطعه‌موضوع به قطعات و کلمه‌موضوع که وابسته به قطعه‌موضوع خود است به کلمات آن قطعه اختصاص می‌یابد. از آنجا که یک سطح از موضوعات در این گونه مدل‌ها بر مبنای همرخدادی در قطعاتی از اسناد شکل می‌گیرد، می‌توان گفت که این

[17]- Stop word
[18] -Hierarchical Pitman-Yor (HPY) process
[19]- Latent Dirichlet Co-Clustering (LDCC)

روش‌ها، نسبت به مدل‌های مبتنی بر همرخدادی در سند مانند LDA، موضوعات را بر اساس همرخدادی‌های محلی‌تری استخراج می‌نمایند.

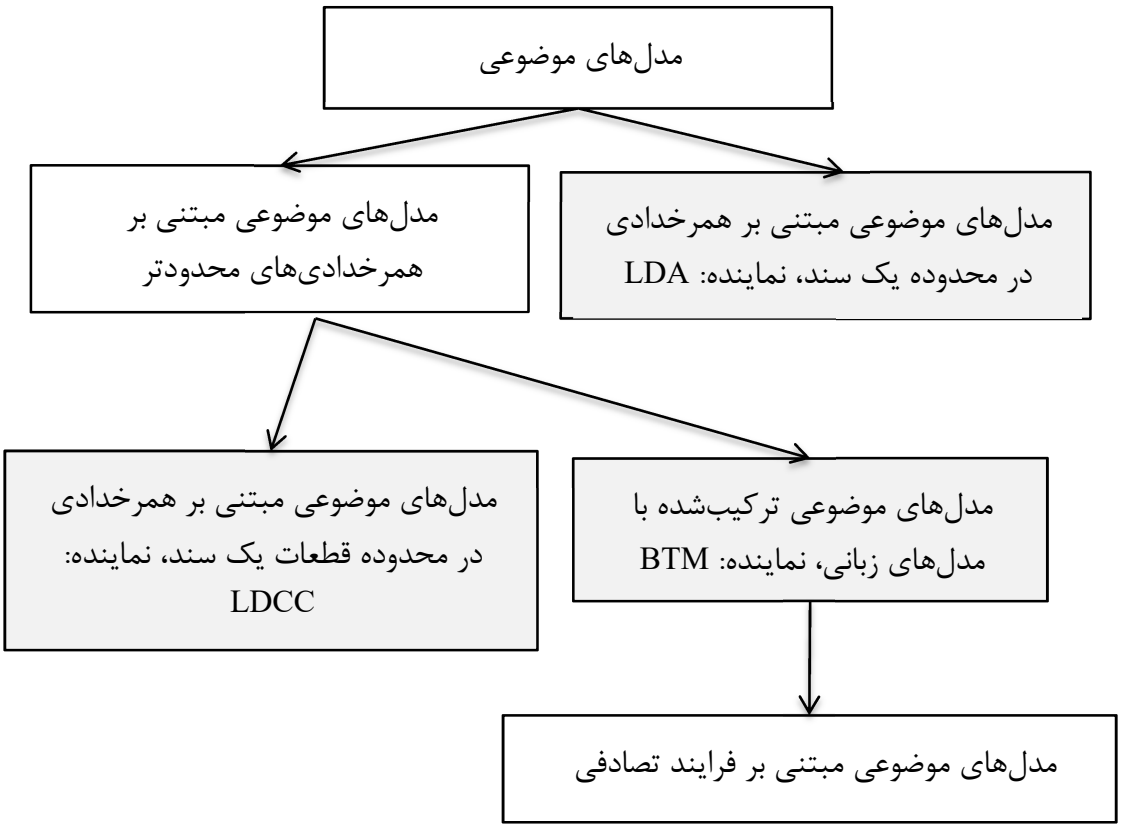
مدل‌های مختلف دیگری هم هستند که ارتباطات محلی‌تر را در نظر می‌گیرند ولی به طور خاص برای یک کاربرد یا نوع داده‌ای مانند تصویر تعریف شده‌اند. هرچند مدل‌های موضوعی عموما برای داده‌های متنی پیشنهاد شده‌اند ولی اکثر آنها قابل تعمیم به داده‌های غیرمتنی نیز هستند. برخلاف این، مدل‌هایی که به طور خاص برای تصویر پیشنهاد می‌شوند لزوما قابل تعمیم به داده متنی نیستند. به عنوان مثال در مقاله [۵۴] مدلی برای قطعه‌بندی تصویر استفاده شده است که در آن هر تصویر مجموعه‌ای از اسناد همپوشان است؛ فرضی که برای اسناد متنی قابل اعمال نیست. [۵۵] و [۵۶] نمونه‌های دیگری از چنین‌مدل‌هایی هستند.

۲-۱- مدل‌های پایه

بنابر آنچه در بخش قبل بیان شد، از دیدگاه این پایان‌نامه، مدل‌های موضوعی را می‌توان به دو دسته تقسیم نمود. دسته اول مدل‌هایی هستند که همرخدادی کلمات را در سطح یک سند در نظر می‌گیرند و در نتیجه مبتنی بر ارتباطات محلی کلمات نیستند. . نماینده این دسته را می‌توان LDA دانست که به تفصیل در بخش بعد بررسی شده است. دسته دوم، مدل‌هایی هستند که سعی در استخراج موضوعات بر مبنای ارتباطات محلی‌تر کلمات دارند. دسته دوم خود از دو زیر دسته تشکیل می‌شود. دسته اول مدل‌هایی هستند که با ترکیب مدل‌های n-gram با مدل‌های موضوعی، ترتیب کلمات و به‌همراه آن ارتباطات محلی آنها را وارد مدل‌های موضوعی می‌کنند. به دلیل وابستگی به ترتیب کلمات این مدل‌ها به شدت از مشکل تنکی رنج می‌برند. دسته دوم مدل‌هایی هستند که موضوعات را در دو سطح مختلف از دقت مطرح کرده و یکی از این سطوح را بر مبنای همرخدادی کلمات در قطعاتی از اسناد ونه اسناد کامل استخراج می‌کنند. از آنها با نام مدل‌های مبتنی بر قطعه یاد می‌کنیم. هدف این گونه مدل‌ها عمدتا وارد

کردن ساختار متن (پاراگراف یا جمله) در فرایند تولید اسناد است. در بخش‌های بعد مدل پیشنهادی را با مدل‌های LDA، BTM و LDCC به نمایندگی از به ترتیب، مدل‌هایی که از ارتباطات محلی بهره نمی‌برند، مدل‌هایی که برای بهره‌گیری از ارتباطات محلی از مدل‌های n-gram استفاده می‌نمایند و مدل‌هایی که همرخدادی را در بازه‌ای کوچکتر از یک سند کامل بررسی می‌نمایند، مقایسه خواهیم نمود و از آنها با عنوان مدل‌های پایه یاد خواهیم کرد.

بر اساس مطالب ذکرشده، مدل‌های موضوعی را از دیدگاه این رساله می‌توان در سلسله‌مراتبی که در شکل ۲-۱ آمده است، دسته‌بندی کرد. روش‌هایی که هایلایت شده‌اند چنان‌که ذکر شد، برای مقایسه مدل‌های پیشنهادی مورد استفاده قرار خواهند گرفت.



شکل ۲-۱- سلسله‌مراتب مدل‌های موضوعی

۳
پیش‌زمینه‌ها

۳- پیش‌زمینه‌ها

موضوع اصلی مورد بحث ما در این متن مدل‌های موضوعی آماری یا همان احتمالی هستند. این مدل‌ها مولد و غیرپارامتریک بوده و به صورت بی‌ناظر یادگرفته می‌شوند. ابزار مناسب برای نمایش چنین مدل‌هایی که در بسیاری موارد، مدل‌های پیچیده‌ای هستند، مدل‌های گرافیکی [۵۷] و به‌ویژه شبکه‌های جهت‌دار بیزین می‌باشند. به این ترتیب در این بخش به معرفی مختصر مقدمات مربوطه برای درک چنین مدل‌هایی می‌پردازیم.

چنانچه توصیف شد، در مدل‌سازی هدف ما این است که تعیین کنیم، کدام یک از عناصر مجموعه P تولیدکننده‌ی داده‌های مشاهده شده است. در روش‌های پارامتریک فراوانی‌گرا[۱]، با بهینه کردن معیارهای مثل درستنمایی[۲] سعی می‌کنیم مقادیر خاصی را برای یک پارامتر بیابیم، در مقابل در روش‌های بیزین، توزیع‌های پیشینی[۳] بر روی پارامترها تعریف می‌شود و سپس با استفاده از نظریه بیز سعی می‌شود تا توزیع پسین[۴] مناسب به شرط داده‌های مشاهده شده محاسبه می‌شود [۵۸]. در چنین حالتی توزیع‌های پیشین مزدوج[۵] نقشی بسیار اساسی بازی می‌کنند. استفاده از چنین توزیع‌های پیشینی باعث می‌شود تا توزیع پسین فرم تابعی مشابه با توزیع پیشین داشته باشد.

یکی از محدودیت‌های روش‌های پارامتریک این است که فرم تابعی خاصی را با تعداد پارامترهای ثابت در نظر می‌گیرند که ممکن است برای برخی کاربردها مناسب نباشد. راهکار دیگر، روش‌های غیر پارامتریک هستند که در آنها فرم خاصی برای توزیع در نظر گرفته نمی‌شود یا تعدادی پارامترها ثابت نبوده و وابسته

[1]- Frequentist
[2]- Likelihood
[3]- Prior
[4]- Posterior
[5]- Conjugate

به اندازه داده‌های موجود است [۵۹]. چنین مدل‌هایی همچنان دارای پارامتر هستند، ولی این پارامترها بیشتر کنترل کننده‌ی پیچیدگی مدل هستند تا فرم تابعی توزیع‌های آن [۶۰].

۳-۱- متغیرهای دوجمله‌ای

متغیر یک بعدی و باینری $x \in \{0,1\}$ را در نظر بگیرید که مثلا می‌تواند توصیف‌کننده نتایج پرتاب یک سکه باشد؛ $x=1$ نشان دهنده‌ی شیر و $x=0$ نشان دهنده‌ی خط. احتمال رخداد شیر را می‌توانیم با پارامتر μ نشان دهیم:

$$p(x=1|\mu)=\mu \qquad (۱-۳)$$

که در آن $0 \le \mu \le 1$ و طبیعتا $p(x=0|\mu)=1-\mu$. به این ترتیب توزیع احتمال $p(x|\mu)$ را می‌توانیم با $\mu^x(1-\mu)^{1-x}$ مشخص نماییم که یک توزیع برنولی است. حال فرض کنیم مجموعه داده‌ی $D=\{x_1,x_2,\ldots,x_N\}$ شامل مجموعه مقادیر مشاهده شده‌ی x است. تابع درستنمایی را می‌توان به صورت زیر محاسبه کرد.

$$p(D|\mu)=\prod_{n=1}^{N}p(x_n|\mu)=\prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n} \qquad (۲-۳)$$

به این ترتیب مقدار μ با حداکثر کردن تابع فوق یا معمولا لگاریتم آن، به صورت زیر قابل محاسبه است. لازم به ذکر است، اگر تعداد شیرهای رخداده در مجموعه D برابر m باشد آنگاه می‌توانیم بنویسیم:

$$p(D|\mu)=\mu^m(1-\mu)^{N-m} \qquad (۳-۳)$$

که اگر در ضریب نرمالسازی $\binom{N}{m}$ که نشان دهنده‌ی تعداد راه‌های انتخاب m عنصر از میان N عنصر است، ضرب شود، یک توزیع دوجمله‌ای مثل $p(m|\mu)$ است. مقدار μ با حداکثر کردن تابع فوق یا معمولا لگاریتم آن به صورت زیر قابل محاسبه است.

$$ln\,p(D|\mu)=\sum_{n=1}^{N} ln\,p(x_n|\mu)=\sum_{n=1}^{M} x_n\,ln\,\mu+(1-x_n)ln(1-\mu) \qquad (۴-۳)$$

برای محاسبه حداکثر تابع فوق می‌توان مشتق آن را نسبت به μ بدست آورد و آن را برابر صفر قرار داد که در نتیجه خواهیم داشت:

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N} x_n \qquad (۵-۳)$$

که در واقع همان میانگین نمونه است. اگر تعداد مقادیر $x_n=1$ در نمونه مشاهده شده را برابر با m در نظر بگیریم آنگاه $\mu_{ML}=\frac{m}{N}$. به این ترتیب، اگر تمام مقادیر مشاهده شده در تعداد اندکی آزمایش، شیر ($x_n=1$) باشد آنگاه $\mu_{ML}=1$ که به این ترتیب چون $p(x_n=1)=\mu$ مقدار مشاهدات بعدی نیز، ۱ پیش‌بینی خواهد شد. روشن است چنین پیش‌بینی‌ای، به‌ویژه زمانی که تعداد مشاهدات کم باشد، به هیچ وجه منطقی نبوده و مصداق بیش‌برازش[۶] است. راه حل این مشکل، از دید تئوری بیز، تعریف یک توزیع پیشین بر روی μ است.

به خاطر بیاورید که تابع درستنمایی دارای فرم $\mu^m(1-\mu)^{N-m}$ است. اگر توزیع پیشینی انتخاب نماییم که آن نیز حاصل ضرب توانهایی از μ و $(1-\mu)$ باشد، آنگاه، احتمال پسین نیز فرمی مشابه تابع پیشین خواهد داشت. چنانچه پیش از این اشاره شد، این ویژگی را تزویج می‌نامیم. توزیع پیشینی که معمولا برای توابع برنولی انتخاب می‌شود، توزیع بتا است.

$$beta(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(a)}\mu^{a-1}(1-\mu)^{b-1} \qquad (۶-۳)$$

a و b را فراپارامتر می‌نامند. زیرا پارامترهایی هستند که توزیع μ را که خود یک پارامتر است، کنترل می‌کنند. ضریب $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(a)}$ برای اطمینان حاصل کردن از نرمال بودن توزیع فوق است، یعنی اینکه:

6_ Over-fitting

$$\int_0^1 bete(\mu \mid a,b)\, d\mu = 1 \qquad (۷-۳)$$

حال می‌توان توزیع پسین را از حاصل ضرب توزیع پیشین و تابع درستنمایی دوجمله‌ای بدست آورد. اگر تنها جملات وابسته به μ را نگه داریم آنگاه می‌توانیم بگوییم :

$$p(\mu \mid m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+a-1} \qquad (۸-۳)$$

که در آن $l = N - m$. تناسب فوق نشان دهنده همان ویژگی تزویج است. فرم کامل به صورت زیر است که یک توزیع بتا با پارامترهای $l+a-1$ و $m+b-1$ به جای a و b می‌باشد. به این ترتیب اگر داده جدیدی مشاهده شود، احتمال پسین قبلی، می‌تواند به عنوان احتمال پیشین برای محاسبه پسین جدید در نظر گرفت و به این ترتیب یک روش یادگیری دنباله‌ای[۷] را در اختیار می‌گذارد که برای کار با داده‌های بزرگ و کاربردهای بی‌درنگ[۸] بسیار کارآمد است.

$$p(\mu \mid m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+a-1} \qquad (۹-۳)$$

۳-۲- متغیرهای چندجمله‌ای

در بخش قبل، متغیرهای مورد بحث، تنها می‌توانستند یکی از دو مقدار ۰ یا ۱ را بپذیرند ولی در بسیاری موارد متغیرها ممکن است یکی از K مقدار را بپذیرند. یک راه برای نمایش چنین متغیرهایی این است که آن را به صورت یک بردار K تایی نشان دهیم که هر بار فقط یکی از عناصر آن ۱ و بقیه صفر است. به عنوان مثال اگر $K = 4$ باشد یکی از مقادیر ممکن می‌تواند به صورت زیر باشد.

$$\boldsymbol{x} = [0,0,1,0]^T \qquad (۱۰-۳)$$

[7]- Sequential
[8]- Real-time

روشن است که همواره $\sum_{k=1}^{K} x_k = 1$. اگر احتمال ۱ بودن x_k را با μ_k نشان دهیم یعنی $p(x_k = 1) = \mu_k$ توزیع مقادیر را می‌توانیم به صورت زیر بیان نماییم.

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad (۱۱-۳)$$

که در آن $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]^T$ و پارامترهای μ_k باید دارای شرایط $\mu_k \geq 0$ و $\sum_k \mu_k = 1$ باشند. این توزیع را می‌توانیم به عنوان تعمیمی بر توزیع برنولی در نظر بگیریم و به همان ترتیب، تابع درستنمایی آن به صورت زیر محاسبه می‌گردد.

$$p(D \mid \boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{n,k}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{n,k}} = \prod_{k=1}^{K} \mu_k^{m_k} \qquad (۱۲-۳)$$

که در آن m_k برابر است با تعداد دفعات ظاهر شدن مقدار ۱ برای متغیرهای x_k در مجموعه داده D و باز مشابه توزیع دوجمله‌ای در اینجا نیز با افزودن یک ضریب نرمال‌سازی $\binom{N}{m_1 m_2 \ldots m_K}$ می‌توانیم احتمال فوق را به فرم یک توزیع چندجمله‌ای بنویسیم.

$$p(\boldsymbol{m} \mid \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} = mul(\boldsymbol{m} \mid \boldsymbol{\mu}, N) \qquad (۱۳-۳)$$

و باز با همان توجیه قبل به روش بیزین رو آورده و یک توزیع پیشین مزدوج، بر روی مقادیر پارامتر تعریف می‌نماییم. توزیع مزدوج برای توزیع چندجمله‌ای، توزیع دریکله[9] است که در بخش بعد به توصیف آن می‌پردازیم.

[9]- Dirichlet

۳-۳- توزیع دریکله

با توجه به فرم تابعی توزیع چندجمله‌ای یک توزیع اولیه مزدوج برای آن باید دارای فرم $\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$ باشد که در آن α_kها پارامتهای توزیع بوده و بردار $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\ldots,\alpha_K]^T$ را تشکیل می‌دهند. بر همین اساس، توزیع دریکله $dir(\boldsymbol{\mu}|\boldsymbol{\alpha})$ به عنوان توزیع پیشین مزدوج برای توزیع چندجمله‌ای تعریف می‌گردد.

$$dir(\boldsymbol{\mu}|\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \qquad (۱۴-۳)$$

که در آن $\Gamma(.)$ تابع گاما و $\alpha_0=\sum_{k=1}^{K}\alpha_k$. توزیع پسین را می‌توان از حاصل ضرب توزیع پیشین و درستنمایی بدست آورد.

$$p(\boldsymbol{\mu}|D,\boldsymbol{\alpha})\propto p(D|\boldsymbol{\mu})\,p(\boldsymbol{\mu}|\boldsymbol{\alpha})\propto\prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1} \qquad (۱۵-۳)$$

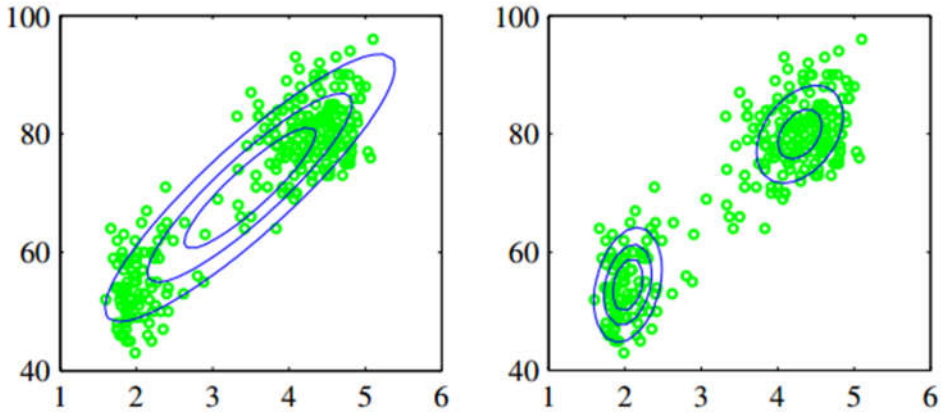
و چنان‌که در بخش‌های قبل هم دیدیم، $p(\boldsymbol{\mu}|D,\boldsymbol{\alpha})=dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})$ که در آن $\boldsymbol{m}=[m_1,m_2,\ldots,m_K]^T$ و در نتیجه می‌توانیم رابطه فوق را به صورت زیر بنویسیم.

$$p(\boldsymbol{\mu}|D,\boldsymbol{\alpha})=dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})=\frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\ldots\Gamma(\alpha_K+m_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1} \qquad (۱۶-۳)$$

۴-۳- مدل‌های مخلوط

داده‌های واقعی، معمولا قابل مدل کردن با یک توزیع احتمال واحد نیستند بلکه ترکیبی از چندین دسته هستند که اگر هریک را به طور جداگانه با یک توزیع مدل کنیم قادر خواهیم بود داه‌ها را با دقت بالاتری تحلیل نماییم و البته این توزیع‌ها از یک خانواده ولی با پارامترهای متفاوت هستند. معمولا چنین داده‌هایی را به صورت ترکیبی خطی از توزیع‌های مذکور مدل می‌کنند. هر یک از این توزیع‌ها را یک مولفه می‌نامند. شکل ۳-۲ نمونه‌ای از داده‌ها را نشان می‌دهد که قابل مدل کردن با یک توزیع واحد

گوسی نیستند ولی اگر آنها را به صورت ترکیبی خطی از دو توزیع گوسی نشان دهیم به دقت بالاتری دست خواهیم یافت.

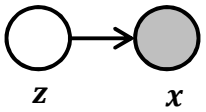


شکل ۳-۱- نمونه‌ای از داده‌ها که با یک توزیع منفرد گوسی قابل مدل کردن نیستند[۶۰]

در بسیاری از موارد دانش اولیه‌ای درباره اینکه احتمال رخداد هریک از این مولفه‌ها چقدر است وجود دارد و بنابراین معمولا برای هر مولفه مانند k وزنی در نظر گرفته می‌شود که آن را با π_k نشان می‌دهیم. به این ترتیب می‌توانیم رابطه زیر رابنویسیم که در آن $p(\boldsymbol{x}|\mu_k)$ توزیعی با پارامتر μ_k است. روشن است که $\sum_k \pi_k = 1$ و در نتیجه باید $0 \le \pi_k \le 1$.

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\mu_k) \qquad (۱۷-۳)$$

حال مدل گرافیکی شکل ۳-۲ را در نظر بگیرید که در آن $\boldsymbol{z}$ برداری است به طول K که هربار فقط یکی از مولفه‌های آن مقدار ۱ داشته و بقیه، ۰ هستند یعنی $\boldsymbol{z}$ یک متغیر چندجمله‌ای است.



شکل ۳-۲ - مدل گرافیکی از یک مدل مخلوط برای متغیرهای منفرد

اگر بخوهیم احتمال توام آن یعنی $p(\boldsymbol{x},\boldsymbol{z})$ را بدست آوریم، خواهیم داشت:

$$p(\boldsymbol{x},z)=p(z)\,p(\boldsymbol{x}\mid z) \tag{۱۸-۳}$$

و به این ترتیب، اگر بخواهیم احتمال حاشیه‌ای $\boldsymbol{x}$ یعنی $p(\boldsymbol{x})$ را محاسبه نماییم، با توجه به اینکه z یک متغیر گسسته است، باید حاصل‌جمع $p(\boldsymbol{x},z)$ را به ازای تمام مقادیر ممکن z بدست آوریم:

$$p(\boldsymbol{x})=\sum_{z}p(\boldsymbol{x},z)=\sum_{z}p(z)\,p(\boldsymbol{x}\mid z) \tag{۱۹-۳}$$

فرض کنیم که $p(z_k=1)=\pi_k$. در این صورت می‌توانیم بنویسیم:

$$p(z)=\prod_{k=1}^{K}\pi_k^{z_k} \tag{۲۰-۳}$$

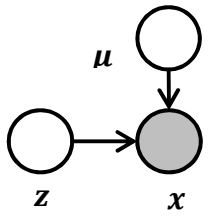
همچنین فرض کنیم $p(\boldsymbol{x}\mid z_k=1)$ دارای توزیعی با پارامتر μ_k باشدکه در این صورت نمایش گرافی آن مانند شکل ۳-۳ است و می‌توانیم بنویسیم:

$$p(\boldsymbol{x}\mid z)=\prod_{k=1}^{K}p(\boldsymbol{x}\mid \mu_k)^{z_k} \tag{۲۱-۳}$$

احتمال توام به صورت زیر محاسبه خواهد شد:

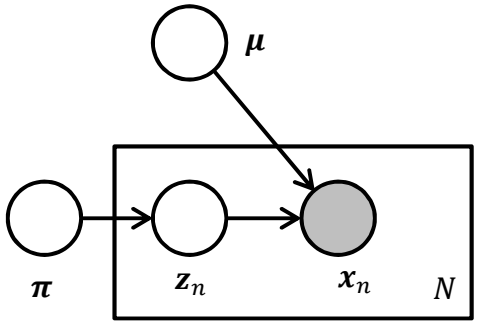
$$p(\boldsymbol{x})=\sum_{z}p(z)\,p(\boldsymbol{x}\mid z)=\sum_{k=1}^{K}\pi_k\,p(\boldsymbol{x}\mid \mu_k) \tag{۲۲-۳}$$

بنابراین احتمال حاشیه‌ای $\boldsymbol{x}$ یعنی $p(\boldsymbol{x})$ دارای فرم یک مدل مخلوط است.



شکل ۳-۳- مدل گرافیکی از یک مدل مخلوط برای متغیرهای منفرد با فرض توزیعی با پارامتر μ_k

حال فرض کنیم که مجموعه داده $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ را مشاهده نموده‌ایم. به این ترتیب اگر بخواهیم همان مدل مخلوط را نمایش دهیم باید برای هر متغیر $\boldsymbol{x}_n$ یک متغیر $\boldsymbol{z}_n$ را در نظر بگیریم و مدل مورد نظر ما بر اساس قواعد مدل‌های گرافیکی، به صورت شکل ۳-۴، نمایش داده خواهد شد. در این شکل، کادر مستطیلی نماینده تکرار است. یعنی عناصر داخل کادر در تکرارهای مستقل رخ داده‌اند.



شکل ۳-۴- مدل گرافیکی از یک مدل مخلوط برای مجموعه‌ای $\boldsymbol{n}$تایی از مشاهدات

بر اساس مدل گرافیکی شکل ۳-۴ می‌توانیم روابط زیر رابنویسیم:

$$\begin{aligned} p(D\mid\boldsymbol{\pi},\boldsymbol{\mu}) &= \sum_{z} p(D,\boldsymbol{z}\mid\boldsymbol{\pi},\boldsymbol{\mu}) = \sum_{z}\prod_{n=1}^{N} p(x_n,z_n\mid\boldsymbol{\pi},\boldsymbol{\mu}) \\ &= \sum_{z}\prod_{n=1}^{N} p(z_n\mid\boldsymbol{\pi})\,p(x_n\mid z_n,\boldsymbol{\mu}) = \sum_{z_1}\sum_{z_2}\ldots\sum_{z_N}\prod_{n=1}^{N} p(z_n\mid\boldsymbol{\pi})\,p(x_n\mid z_n,\boldsymbol{\mu}) \\ \Rightarrow \prod_{n=1}^{N}\sum_{z_n} p(z_n\mid\boldsymbol{\pi})\,p(x_n\mid z_n,\boldsymbol{\mu}) &= \prod_{n=1}^{N}\sum_{k=1}^{K} p(z_{nk}=1\mid\pi_k)\,p(x_n\mid z_{nk},\mu_k) \end{aligned} \qquad (۲۳-۳)$$

که در آن جمله اول زیر سیگما همان π_k بوده و جمله دوم دارای توزیع $p(x_n\mid\mu_k)$ است بنابراین خواهیم داشت:

$$p(D\mid\boldsymbol{\pi},\boldsymbol{\mu}) = \prod_{n=1}^{N}\sum_{k=1}^{K}\pi_k\,p(x_n\mid\mu_k) \qquad (۲۴-۳)$$

بنابراین می‌توانیم لگاریتم تابع درستنمایی را به صورت زیر محاسبه نماییم:

$$\ln p(D\mid\boldsymbol{\pi},\boldsymbol{\mu}) = \sum_{n=1}^{N}\ln\left(\sum_{k=1}^{K}\pi_k\,p(x_n\mid\mu_k)\right) \qquad (۲۵-۳)$$

محاسبه حداکثر مقدار فوق، به دلیل وجود جمع در زیر لگاریتم که به دلیل وجود چندین مولفه در مدل، شکل گرفته است، قابل محاسبه مستقیم نمی‌باشد. ولی وجود متغیر پنهان z این امکان را برای ما فراهم می‌آورد تا با استفاده از الگوریتم امید ریاضی- بیشینه‌سازی[10] (EM) پاسخ مسئله را با دقت مناسبی تخمین بزنیم.

امید ریاضی- بیشینه‌سازی یا EM الگوریتمی است که برای محاسبه بیشینه درستنمایی یا بیشینه احتمال پسین، به روش تکراری، هنگام تخمین پارامترها در مدل‌های آماری و زمانی که مدل وابسته به متغیرهای پنهان است، مورد استفاده قرار می‌گیرد. این الگوریتم شامل دو گام امید ریاضی و بیشینه‌سازی است که در گام اول امید ریاضی لگاریتم تابع درستنمایی، بر اساس تخمین فعلی پارامترها ارزیابی می‌گردد و در مرحله دوم با حداکثر کردن تابع درستنمایی بدست آمده در مرحله قبل، تخمین دیگری از پارامترها بدست می‌آید که توقع داریم، دقیق‌تر از تخمین قبل باشد. برای اطلاعات بیشتر می‌توانید به [۶۰, ۶۱] مراجعه نمایید.

۳-۵- مدل‌های مولد

طبق آنچه تاکنون بیان شد، یک مدل احتمالی به فرمی که در شکل ۳-۴ آمده است را می‌توان با استفاده از یک مدل مخلوط بیان و مشخص نمود. این مدل، یک مدل مولد است. مدل‌های مولد که در مقابل مدل‌های تمایزی[11] قرار می‌گیرند، مدل‌هایی هستند که برای تولید تصادفی مشاهدات، معمولا به شرط تعدادی متغیر پنهان معرفی می‌شوند.

مدل‌های مولد برخلاف مدل‌های تمایزی، یک مدل کامل احتمالی از تمام متغیرهای مسئله هستند، در حالی که مدل‌های تمایزی فقط متغیرهای هدف را با توجه به مشاهدات مدل می‌کنند. یعنی در مدل‌های

[10]- Expectation-Maximization (EM)
[11] - Discriminative

مولد $p(x,z)$ را مدل می‌کنیم سپس با استفاده از قوانین بیزین $p(x|z)$ را بدست می‌آوریم، در حالی که در روش‌های تمایزی مستقیما $p(x|z)$ را مدل می‌نماییم. نمونه‌هایی از مدل‌های مولد عبارتند از: مدل‌های مخلوط، مدل پنهان مارکوف و نایو بیز[۱۲].

زمانی که با داده‌های بزرگ و دارای برچسب روبه‌رو هستیم، یعنی قادر به استفاده از روش‌های دارای ناظر[۱۳] هستیم، تقریبا همیشه روش‌های تمایزی ترجیح داده می‌شوند، اما زمانی که داده‌ها کوچک هستند یا داده‌های برچسب‌زده در اختیار ما نیستند و نیازمند استفاده از روش‌های بدون ناظر هستیم، عموما روش‌های مولد ترجیح داده می‌شوند. برای اطلاعات بیشتر می‌توانید به [۶۱] مراجعه نمایید.

۳-۶- مدل‌های گرافیکی

مدل‌های گرافیکی، ابزاری برای نمایش مدل‌های احتمالی هستند. در این مدل‌ها، متغیرهای تصادفی مسئله و روابط بین آنها یعنی وابستگی و استقلال بین متغیرها نمایش داده می‌شود. به عبارت دیگر یک مدل گرافیکی را می‌توان مانند یک دیتابیس احتمالی دانست که می‌تواند درخواست‌هایی را با توجه به مقادیر مجموعه‌ای از متغیرهای تصادفی پاسخ گوید. پاسخ این درخواست‌ها توسط قواعد نظریه احتمال تولید می‌شود. مجموعه متغیرهای تصادفی $X=\{X_1,X_2,\ldots,X_N\}$ را در نظر بگیرید. اگر مدل گرافیکی ما تشکیل شده از این متغیرها باشد، درخواست‌های مطرح شده عموما مربوط به روابط بین متغیرها و یا زیرمجموعه‌ای از آنها و محاسبه احتمالات شرطی بر روی آنها می‌باشد، به عنوان مثال، $p(X_2|X_1)$.

مدل‌های گرافیکی به دو دسته‌ی کلی جهت‌دار و بدون جهت تقسیم می‌شوند که دسته مورد نظر ما در متن حاضر، دسته جهت‌دار است. بنابراین در ادامه بحث به معرفی این دسته از مدل‌های گرافیکی می‌پردازیم.

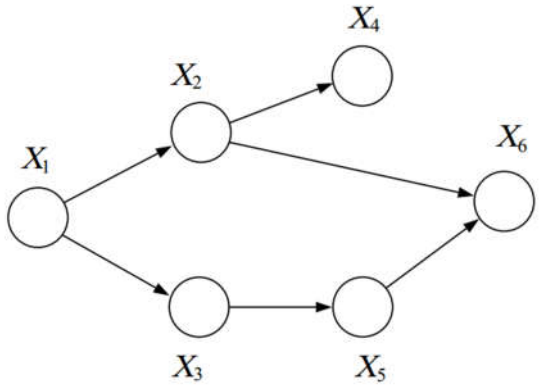
---

[12] - Naïve Bayes
[13]- Supervised

۳-۶-۱- گراف‌های جهت‌دار و احتمال توام

یک گراف جهت‌دار را به صورت یک زوج (G,E) تعریف می‌نماییم که در آن نماینده G راس‌ها و E مجموعه لبه‌های گراف است. گراف‌های مورد نظر ما در این متن گراف‌های بدون دور هستند. هر نود از گراف متناظر با یک متغیر تصادفی است. هر نود از این گراف در یک مدل گرافیکی، متناظر با یک متغیر تصادفی است و با X_i نمایش داده می‌شود. مجموعه نودهای پدر نود X_i را با X_{π_i} نشان می‌دهیم.

به عنوان مثال شکل ۳-۵ را در نظر بگیرید. این شکل نماینده یک مدل گرافیکی جهت‌دار شامل ۶ متغیر تصادفی $X=\{X_1,X_2,X_3,X_4,X_5,X_6\}$ است. مجموعه نودهای پدر نود X_6، مجموعه $X_{\pi_6}=\{X_2,X_5\}$ است.



شکل ۳-۵- نمونه‌ای از یک مدل گرافیکی

در یک مدل گرافیکی به طور کلی می‌توان احتمال توام $p(x_1,x_2,\ldots,x_6)$ را به صورت زیر تعریف کرد که در آن x_i نماینده مقدار واقعی متغیر تصادفی X_i است. بنابراین احتمال توام گراف زیر را می‌توان به صورت حاصل ضربی از احتمال‌های شرطی محلی مانند $p(x_i \mid x_{\pi_i})$ شکست.

$$p(x_1,x_2,\ldots,x_N)=\prod_{i=1}^{N}p(x_i \mid x_{\pi_i}) \qquad (۲۶-۳)$$

برای مجموعه متغیرهای $p(x_1,x_2,\ldots,x_6)$ را در نظر بگیرید. بر اساس قانون زنجیره‌ای می‌توان احتمال توام کامل این متغیرها را بر اساس ترتیب این متغیرها در ساختار گراف شکل ۳-۵ به صورت زیر تعریف نماییم. در این رابطه اگر فرض بر استقلال شرطی برخی از این متغیرها باشد، می‌توان برخی از عبارات را حذف نمود. چنین فرض‌هایی را می‌توان با عدم وجود نودهای بین متغیرها نمایش داد.

$$p(x_1,x_2,x_3,x_4,x_5,x_6) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_2,x_1)\,p(x_4|x_3,x_2,x_1)\,p(x_5|x_4,x_3,x_2,x_1)\,p(x_6|x_5,x_4,x_3,x_2,x_1) \quad (۲۷-۳)$$

رابطه فوق بدون در نظر گرفتن فرض خاصی در رابطه با استقلال متغیرها است. اما اگر به گراف شکل ۳-۵ نگاه کنیم بین برخی از زوج متغیرها نودی وجود ندارد. بر اساس ساختار این گراف رابطه فوق به صورت زیر نوشته می‌شود.

$$p(x_1,x_2,x_3,x_4,x_5,x_6) = p(x_1)\,p(x_2\,|\,x_1)\,p(x_3\,|\,x_1)\,p(x_4\,|\,x_2)\,p(x_5\,|\,x_3)\,p(x_6\,|\,x_5,x_2) \quad (۲۸-۳)$$

به طور کلی مجموعه متغیر X_A را مستقل یا به طور حاشیه‌ای مستقل از متغیر X_B می‌نامیم اگر

$$p(x_A,x_B) = p(x_A)\,p(x_B) \quad (۲۹-۳)$$

این استقلال حاشیه‌ای را با $X_A \perp X_B$ نشان می‌دهیم و همچنین متغیر X_A را به شرط X_C مستقل از X_B می‌نامیم اگر

$$p(x_A\,|\,x_B,x_C) = p(x_A\,|\,x_C) \quad (۳۰-۳)$$

یا

$$p(x_A,x_B\,|\,x_C) = p(x_A\,|\,x_C)\,p(x_B\,|\,x_C) \quad (۳۱-۳)$$

این استقلال شرطی را به صورت $X_A \perp X_B | X_C$ نمایش می‌دهیم. در یک مدل گرافیکی جهت‌دار، اگر مجموعه X_{v_i} نماینده تمامی نودهای قبل از نود X_i در مجموعه نودها باشد، آنگاه

$$X_i \perp X_{v_i} | X_{\pi_i} \quad (۳۲-۳)$$

به عنوان مثال برای شکل ۳-۵ می‌توان استقلال‌های شرطی زیر را تعریف نمود.

$$\begin{aligned} &X_3 \perp X_2 \mid X_1 \\ &X_4 \perp \{X_1, X_3\} \mid X_2 \\ &X_5 \perp \{X_1, X_2, X_4\} \mid X_3 \\ &X_6 \perp \{X_1, X_3, X_4\} \mid \{X_2, X_5\} \end{aligned} \quad (۳۳-۳)$$

برای اطلاعات بیشتر می‌توانید به [۶۲] مراجعه نمایید.

۳-۷- همرخدادی

تمرکز ما در این پایان‌نامه بر روی مدل‌های موضوعی احتمالاتی است که اساس کار آنها بر همرخدادی کلمات است. مفهوم محتوای مکانی را بر اساس تعاریف موجود همرخدادی و عوامل استفاده نشده در آنها وارد مدل‌های موضوعی خواهیم نمود. در این بخش به معرفی مفهوم همرخدادی، تعاریف موجود برای آن و عوامل دیده نشده در این تعاریف می‌پردازیم.

بسیاری از مدل‌های آماری، مبتنی بر داده‌های همرخداد[۱۴] (COD) هستند و تحلیل و مدل کردن آماری چنین داده‌هایی، یک چالش جدی در بسیاری از کاربردهاست. برای اینکه بتوانیم بحث را ادامه دهیم، ابتدا باید مشخص نماییم که منظور از داده‌های همرخداد به طور کلی چیست. دو مجموعه متناهی $X = \{x_1, x_2, \ldots, x_N\}$ و $Y = \{y_1, y_2, \ldots, y_M\}$ را در نظر بگیرید. زوج (x_i, y_j) یک رخداد مشترک (همرخدادی) شی x_i با شی y_j است.

[14]- Co-occurrence data

انواع COD در کاربردهای مختلفی قابل مشاهده‌اند. یکی از حالات بسیار معروف COD هیستوگرام‌ها هستند که در آن هر عنصر x_i به صورت توزیعی بر روی y_jها نمایش داده می‌شود. مثلا در بینایی ماشین که X متناظر با پیکسل‌ها یا مجموعه‌ای از پیکسل‌ها است و Y ویژگی‌های آنها باشد. یا مثلا در بازیابی اطلاعات که در آن X یک مجموعه از اسناد[۱۵] و Y مجموعه‌ای از کلمات کلیدی است. یا در پردازش زبان‌های طبیعی[۱۶] که در آن، دو مجموعه، متناظر با مجموعه کلماتی هستند که در یک ساختار گرامری مثل اسم و صفت یا فعل و فاعل و مانند آنها ظاهر شده‌اند. گستردگی این کاربردها نشان می‌دهد تحلیل داده‌های همرخداد در واقع یک مسئله رایج و پایه‌ای، در یادگیری بدون ناظر[۱۷] است.

۳-۷-۱- همرخدادی کلمات

همرخدادی کلمات یکی از رایجترین و پرکاربردترین انواع همرخدادی در حیطه پردازش متن و پردازش زبان‌های طبیعی است. علت آن فرضیه‌ای است که از آن به عنوان "فرضیه همرخدادی" [۶۳] یاد می‌شود. این فرضیه به این قرار است: "همرخدادی کلمات در متن می‌تواند بیانگر ارتباط معنایی آنها باشد". اما خود واژه همرخدادی هم می‌تواند تعاریف مختلفی داشته باشد که در ادامه به برخی از آنها اشاره می‌نماییم.

۳-۷-۱-۱- همرخدادی موقعیت

می‌توان مصداق‌های همرخدادی کلمات را به دو نوع همرخدادی "موقعیت"[۱۸] و "رابطه"[۱۹] تقسیم نمود. همرخدادی موقعیت که دارای پیشینه‌ی طولانی‌تری است برگرفته از تعریفی است که توسط *سینکلر*[۲۰] در سال ۱۹۹۱ [۶۴] مطرح شد. بر اساس این تعریف، همرخدادی دو کلمه وابسته به موقعیت آن دو نسبت به

[15]- Documents
[16]- Natural language processing
[17]- Unsupervised
[18]- Positional
[19]- Relational
[20]- Sinclair

یکدیگر است. یعنی دو کلمه زمانی همرخدادند که درون فاصله معینی از یکدیگر ظاهر شده باشند. منظور از فاصله در اینجا تعداد کلمات مابین است و از آن با عنوان گستره[۲۱] یاد می‌شود. به این ترتیب که اگر گستره همرخدادی برابر با ۴- و ۴ باشد، معنی آن این است که همرخدادی در محدوده ۴ کلمه در هریک از دو طرف کلمه مورد بررسی تعریف می‌گردد. این نوع از تعریف همرخدادی، رایجترین نوع تعریف آن است و البته اشکال گوناگونی از آن در مقالات و پژوهش‌های مختلف به کار گرفته شده است که در ادامه به برخی از آنها اشاره می‌نماییم.

۳-۷-۱-۱-۱- همرخدادی در یک سند

اگر دو کلمه w و w' در یک سند یکسان آمده باشند این دو کلمه معمولا از نظر موضوعی با یکدیگر مرتبط هستند. برای محاسبه همرخدادی دو کلمه در این تعریف دو راه وجود دارد که در ادامه آمده‌اند .

[۶۵] اگر میزان همرخدادی دو کلمه را با $f_{ww'}$ نشان دهیم، آنگاه خواهیم داشت:

۱- $f_{ww'}$ برابر است با تعداد سندهایی که حاوی هر دو کلمه w و w' هستند.

۲- تعداد همرخدادی کلمات در هر سند برابر است با تعداد تمام رخدادهای جفت کلمه w و w'. یعنی اگر در یک سند کلمه w، ۳ بار و w'، ۲ بار آمده باشند، تعداد همرخدادی آنها در آن سند برابر است با ۶. $f_{ww'}$ برابر است با مجموع تمام همرخدادی‌های آنها در سندهای مجموعه داده.

هر چند که این تعریف در بسیاری مواقع می‌تواند موثر باشد ولی تعریف دقیقی نیست زیرا در بسیاری مواقع موضوع مورد بحث در یک سند ثابت نیست و در طول سند تغییر می‌کند. بنابراین نیازمند خواهیم بود که گستره همرخدادی را کوچکتر از یک سند در نظر بگیریم.

[21]- Span

۳-۷-۱-۱-۲- همرخدادی در یک جمله

در این تعریف، دو کلمه همرخداد هستند اگر هر دو در یک جمله یکسان ظاهر شوند. فراوانی‌ها را هم می‌توان به همان ترتیبی که در تعریف فوق شرح داده شد، محاسبه نمود با این تفاوت که محدوده مورد نظر در این تعریف یک جمله از سند است.

۳-۷-۱-۱-۳- همرخدادی در یک پنجره

در این تعریف، دو کلمه همرخداد هستند اگر هر دو در یک پنجره یکسان ظاهر شوند. در این تعریف یک پنجره با طول ثابت در نظر گرفته می‌شود و در یک سند حرکت می‌کند و تعداد همرخدادی کلمات در پنجره‌ها شمرده می‌شود.

۳-۷-۱-۲- همرخدادی رابطه

نوع دیگری از همرخدادی، همرخدادیِ رابطه [۲۵] است به این معنی که دو کلمه همرخدادند اگر دارای یک رابطه دستوری با یکدیگر باشند. این رابطه می‌تواند "اسم+فعل" مثل غذا خوردن، "اسم + صفت" مثل "دشت پهناور" و یا روابط دیگری از این دست باشد. در واقع این تعریف، همرخدادی را در سطح یک جمله در نظر می‌گیرد با این تفاوت که کلمات ظاهر شده در یک جمله همرخداد هستند اگر دارای رابطه گرامری نیز باشند. مثلا اسم‌های "آب"، "چای" و "شربت" همگی اسم‌هایی هستند که همراه فعل "نوشیدن" می‌آیند. برای شمردن چنین همرخدادی‌هایی، معمولا نیازمندیم که یک درخت وابستگی[۲۲] برای هر جمله رسم نماییم و در این صورت همرخدادی فقط می‌تواند بین هر هسته[۲۳] با وابسته‌هایش[۲۴] رخ دهد. برای تعیین این نوع همرخدادی نیازمند این خواهیم بود که ابتدا نقش‌های دستوری کلمات متن

[22]- Dependency tree
[23]- Head
[24]- Modifier

را مشخص نماییم که طبیعتا برای زبان‌های مختلف متفاوت خواهد بود. در ضمن دقت کاربردهای بعدی علاوه بر دقت تعیین کلمات همرخداد، وابسته به دقت تعیین نقش‌های کلمات متن نیز خواهد بود.

۳-۷-۲- عوامل موثر بر همرخدادی

دو عامل اصلی در تفاوت تعاریف مختلفی که تاکنون برای همرخدادی ذکر کرده‌ایم، نقش بازی کرده‌اند. این عوامل عبارتند از:

- ✓ اندازه بازه همرخدادی
- ✓ نقش دستوری کلمات همرخداد

سوالی که در اینجا مطرح است این است که فرضیه همرخدادی تا چه حد درست است و از طرفی عوامل فوق چه اثری روی دقت نتایج حاصله دارند یا به عبارتی ارتباط معنایی کلماتی که بر اساس تعاریف مختلف مبتنی بر عوامل فوق بدست می‌آید، تا چه حد قوی است. در سال‌های اخیر تحقیقاتی برای یافتن این پاسخ، صورت گرفته است. در برخی از این کارها، از مفهومی با عنوان اتحاد معنایی[۲۵] [۶۶, ۶۷] بهره گرفته شده است. یک اتحاد معنایی، نماینده مجموعه کلماتی است که با ذکر یک کلمه که آن را محرک[۲۶] می‌نامند، به ذهن متبادر می‌شوند. هر یک از این کلمات را یک پاسخ[۲۷] می‌نامیم. مجموعه داده‌هایی، به خصوص در زبان انگلیسی، ایجاد شده‌اند که شامل تعداد زیادی زوج کلمات محرک-پاسخ هستند. مراحل تهیه این مجموعه‌ها به این ترتیب است که ابتدا تعدادی کلمات کلیدی که معمولا کلمات پرتکرار هستند، به عنوان کلمات محرک تعیین می‌شوند و سپس از شرکت‌کنندگان خواسته می‌شود تا اولین کلمه‌ای را که در مواجهه با کلمه محرک به ذهنشان خطور می‌کند، بیان کنند. برای هر کلمه محرک ممکن است، پاسخ‌های متفاوتی توسط شرکت‌کنندگان بیان شود که تعداد تکرار هر کدام قوت زوج محرک-پاسخ را

[25]- Semantic association
[26]- Stimulus
[27]- Response

بیان می‌کند. در تحقیقات مذکور، از چنین مجموعه داده‌ای استفاده می‌شود و بررسی می‌گردد که توزیع چنین کلماتی در تعاریف مختلف محتوی چگونه است. به عنوان مثال در [۶۶] یک مجموعه داده حاوی یک میلیون کلمه انگلیسی استفاده شده است. زوج‌های محرک-پاسخ انتخاب شده و جمعا شامل ۴۷ کلمه پرتکرار از نوع اسم و قوی‌ترین اسم ذکر شده به عنوان پاسخ آنها هستند. سپس با استفاده از یک پنجره ۵۰ کلمه‌ای، نشان داده شده که یک رابطه معکوس، مابین فاصله زوج‌های همرخداد و قوت اتحاد آنها وجود دارد. یعنی اتحادهای قوی‌تر، در فاصله نزدیکتر رخ می‌دهد.

هر چه قوت اتحاد بیشتر باشد، میزان همرخدادی آنها نیز بیشتر است. همچنین بررسی‌ها و نتایج ذکر شده آنها نشان می‌دهد که فراوانی همرخدادی کلمات مرتبط، به طور چشمگیری بیش از همرخدادی کلمات نامرتبط است که تاییدی بر فرضیه همرخدادی است.

در مقاله دیگری [۶۳] که در سال ۲۰۰۸ منتشر شده است بررسی جامعی درباره همرخدادی زوج‌های محرک-پاسخ در زبان آلمانی انجام شده است. در این تحقیق از ۱۵۷۸۸ زوج محرک-پاسخ استفاده شده است که در آنها تمامی محرک‌ها فعل هستند. سپس با استفاده از این مجموعه به پرسشهایی از جمله پرسشهای زیر در یک مجموعه داده از متون خبری شامل ۲۰۰ میلیون کلمه پاسخ داده شده است:

- ✓ چه کسری از پاسخ‌ها با محرک خود همرخدادند؟ برای پاسخ به این پرسش تمامی مجموعه متون برای یافتن زوج‌های محرک-پاسخ جستجو شده است. سپس برای هرکلمه محرک یافت شده، در پنجره‌هایی به طول حداقل ۱ و حداکثر ۲۵ قبل و بعد از آن به دنبال پاسخ‌های مربوطه می‌گردند. نتایج آزمایش نشان می‌دهد که زوج‌های محرک-پاسخ با همرخدادی حداقل ۱ در پنجره‌ای با اندازه حداکثر ۲۵ یافت می‌شوند و همچنین هر چه اندازه پنجره بزرگتر می‌شود تعداد زوج‌های قوی جدیدی که مشاهده می‌شوند در حالی که در پنجره‌های کوچکتر یافت نشده‌اند کاهش پیدا می‌کند. حتی زمانی که برای همرخدادی مقدار سختگیرانه‌ی ۲۰ به عنوان حداقل در نظر گرفته

شده باز هم بیش از ۵۰ درصد زوج‌ها در پنجره‌ای به طول حداکثر ۲۵ یافت خواهد شد. به این ترتیب به نظر می‌رسد نه‌تنها پنجره‌های کوچکتر نماینده خوبی برای ارتباط معنایی کلمات هستند. بلکه با انتخاب یک اندازه مناسب، می‌توانند شواهد کافی را برای روش‌های آماری نیز فراهم کنند. طبق این تحقیق هر چند پنجره‌هایی با طول کوچک مناسب‌ترند ولی پنجره‌هایی با طول بزرگ نیز همچنان می‌توانند مناسب باشند. اما حداکثر طول در نظر گرفته‌شده، ۲۵ است که همچنان پنجره کوچکی محسوب می‌شود و همچنین مقایسه‌ای با پنجره‌هایی به طول کل یک سند نیز که بسیار رایج است انجام نگرفته است.

- ✓ همرخدادی زوج‌های نامرتبط چگونه است؟ نتایج حاصل نشان می‌دهد که میزان همرخدادی زوج‌های نامرتبط در پنجره‌های کوچکتر به طور چشمگیری کمتر از همرخدادی کلمات مرتبط در این پنجره‌هاست. و با بزرگتر شدن پنجره میزان همرخدادی آنها با سرعت زیادی افزایش می‌یابد.
- ✓ نتیجه جالب دیگری که در نتیجه آزمایشات این پژوهش بدست آمده است این است که اثر فاصله برای کلمات مختلف می‌تواند متفاوت باشد یعنی برخی کلمات در فاصله نزدیکتر و برخی در فاصله‌های دورتر همرخدادند.

۳-۷-۲-۱- عوامل دیده نشده

همانطور که ذکر شد، مفهوم همرخدادی در ضمن اینکه بسیار ساده است کاربردهای فراوانی داشته و نتایج چشمگیری را نیز تولید کرده است ولی، نکاتی هستند که در بیشتر تعاریف ذکرشده، نادید گرفته شده‌اند.

۱- در تعاریف فوق، تعداد تکرار همرخدادی شمرده می‌شود. ممکن است تعداد رخداد یک کلمه مثل w_1 کم باشد ولی در هر رخداد، در حوالی کلمه دیگری مثل w_2 آمده باشد. در چنین حالتی،

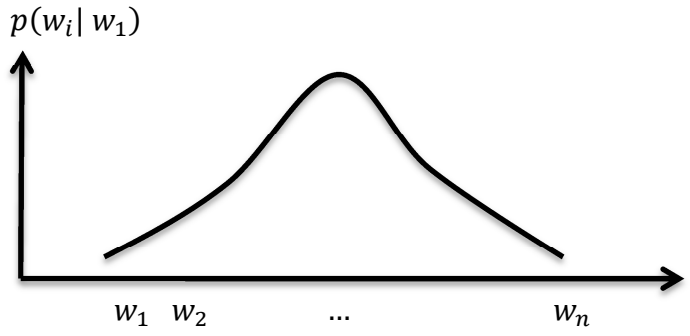
باید همرخدادی w_1 نسبت به w_2 از امتیاز بالایی برخوردار باشد که در بیشتر تعاریف فوق بدلیل تعداد رخداد پایین چنین نیست.

۲- ممکن است کلمه w_1 در تمام تکرارهای خود در حوالی w_2 رخ داده باشد ولی w_2 در بسیاری از رخدادهای خود با کلمات دیگری آمده باشد. بنابراین نباید میزان همرخدادی این دو نسبت به یکدیگر یکسان در نظر گرفته شود. به عبارت دیگر مفهوم همرخدادی کلمات ذاتا یک مفهوم نامتقارن است. ولی در تعاریفی که تا کنون برای همرخدادی کلمات ارائه شده‌اند، این مفهوم متقارن درنظر گرفته می‌شود. مثلا در بازیابی اطلاعات دو کلمه "درخت" و "هرا" را در نظر بگیرید. اگر در سندی کلمه "درخت" ظاهر شود، لزوما آن سند مرتبط با درخواستی که کلمه "هرا" در آن آمده است نیست یا به عبارتی احتمال رخداد کلمه "هرا" در آن سند چندان بالا نخواهد بود. چون "درخت" یک کلمه بسیار رایج زبان است و ممکن است در مفاهیم بسیار متفاوت و مختلفی به کار رود. ولی اگر در یک سند کلمه "هرا" رخ داده باشد، این سند به احتمال بسیار زیاد مرتبط با درخواستی است که حاوی کلمه "درخت" باشد یا به عبارت دیگر احتمال مشاهده کلمه "درخت" در سند مربوطه بسیار زیاد خواهد بود. البته مثال فوق با فرض کنار گذاشتن سایر کلمات سند دارای مفهوم است. سایر کلمات نیز می‌توانند نقش خود را در امتیاز نهایی بازی کنند.

۳- حتی اگر همرخدادی را در یک پنجره درنظر بگیریم، بعضی از کلمات، نزدیکتر به هم رخ می‌دهند و بعضی دورتر. آیا همرخدادی این کلمات یکسان است؟ در واقع همرخدادی به فاصله وابسته است. یعنی برخی کلمات معمولا در فاصله نزدیک همرخدادند و برخی در فواصل دورتر. به عنوان مثال، می‌توانیم همان کلمه "درخت" را در نظر بگیریم. دو کلمه "درخت " و "هرا" اگر همرخداد باشند، معمولا در مجاورت هم و بی‌فاصله رخ می‌دهند. ولی کلمه "خاک" معمولا در فاصله

بیشتری از "درخت" ظاهر می‌شود. مثال دیگری که می‌توان برای این نکته ذکر کرد، همرخدادی کلماتی با نقش‌های دستوری خاص است. مثلا در زبان فارسی، فعل و مفعول در فاصله کمتری رخ می‌دهند تا فعل و فاعل. به عبارت دیگر شاید فاصله بتواند تا حدی نماینده نقش کلمات نیز باشد.

برای لحاظ کردن نکته اول و دوم می‌توانیم به جای در نظر گرفتن $p(w_1, w_2)$ از $p(w_1 | w_2)$ و $p(w_2 | w_1)$ استفاده نماییم، که به این ترتیب هر کلمه عبارت خواهد بود از یک توزیع بر روی سایر کلمات مجموعه و هردو کلمه به صورت شرطی مستقل از یکدیگر خواهند بود. شکل ۳-۶، مثالی فرضی از چنین بیانی است.



شکل ۳-۶- توزیع کلمات

مفهوم محتوای مکانی را می‌توان با استفاده از تعریف محدودتر همرخدادی، یعنی همرخدادی در یک پنجره و همچنین عوامل دیده‌نشده در تعاریف موجود که در بالا به آن اشاره شد، تجسم بخشید. یعنی مدل‌های موضوعی را به‌گونه‌ای تعریف کنیم که به‌جای استفاده از همرخدادی در یک سند، از تعاریفی از همرخدادی استفاده نماید که نکات مذکور را لحاظ کند.

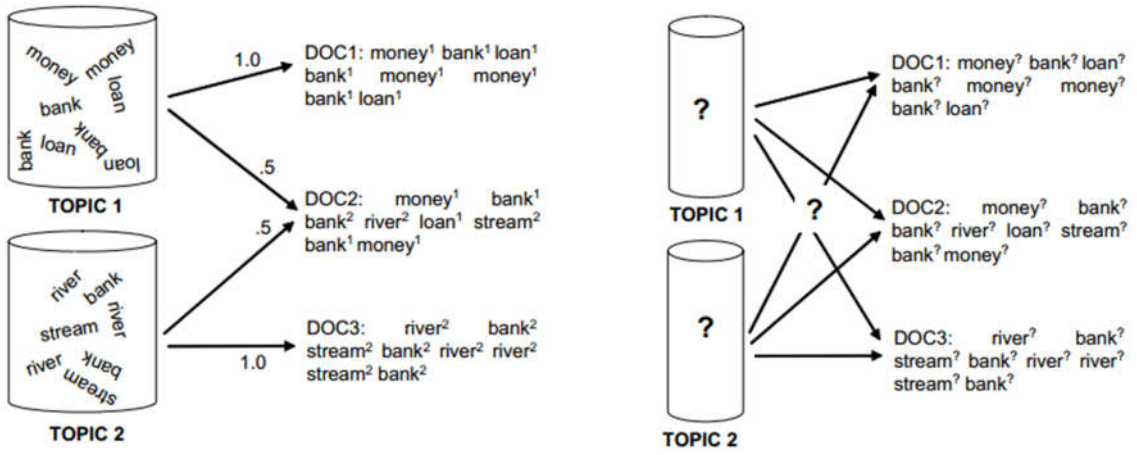
۳-۸- مدل‌های موضوعی

در فصل دوم، به معرفی مختصر انواع مختلفی از مدل‌های موضوعی پرداختیم. برخی از این مدل‌ها در درک مفاهیم مطرح در این پایان‌نامه نقش کلیدی بازی می‌کنند و همچنین برای ارزیابی مدل‌های پیشنهادی مورد استفاده قرار گرفته‌اند. بنابراین نیازمندیم تاین روش‌ها را با تفصیل و جزئیات بیشتری برررسی نماییم. به همین دلیل در این بخش به معرفی جزئی روش‌های LDA، BTM و LDCC که پیش از این از آنها با نام مدل‌های پایه یاد کردیم، می‌پردازیم.

۳-۸-۱- مدل تخصیص پنهان دریکله (LDA)

در LDA، هر سند ترکیبی از موضوعات است و هر موضوع خود، توزیعی بر روی کلمات. روش LDA یک مدل مولد برای تولید سندهاست که در آن، هر سند طی یک فرایند احتمالی ساخته می‌شود. برای ساخت یک سند جدید، یک توزیع بر روی موضوعات در نظر گرفته می‌شود سپس برای تولید کلمات، هر بار بطور تصادفی، یک موضوع را انتخاب کرده و یک کلمه را از آن استخراج می‌نماییم. با استفاده از روش‌های آماری می‌توان این فرایند را معکوس کرد که در نتیجه آن قادر خواهیم بود تا موضوعات سازنده یک سند را بدست آوریم. در اینجا، آنچه قابل مشاهده است سندها و کلمات سازنده آنها هستند و موضوعات را می-توان متغیرهای پنهانی در نظر گرفت که باید بدست آیند.

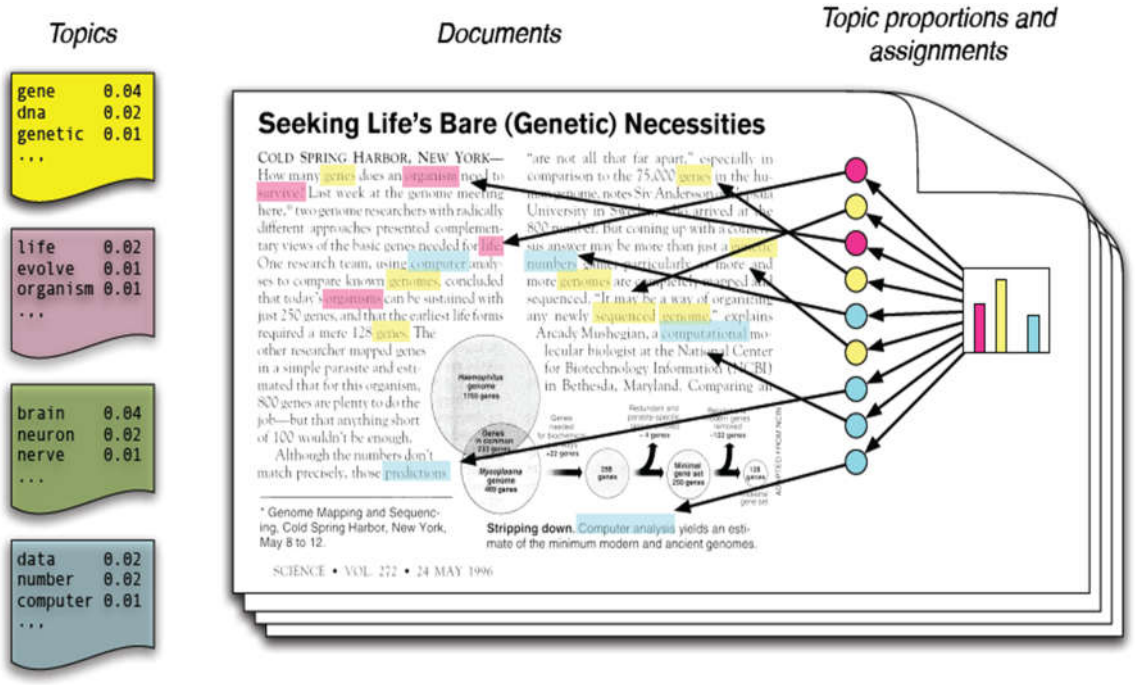
شکل ۳-۷ می‌تواند فرایند ذکر شده را روشن‌تر نماید. این شکل نشان می‌دهد که هر سند و کلمات آن معلوم هستند و موضوع‌ها نامعلوم و بعد از اجرای فرایند مدل‌سازی موضوعی آماری، موضوعات تشکیل شده و شرکت هر یک از آنها در هر سند مشخص شده است. هر کلمه از سند هم با یک وزن به یک موضوع مشخص تخصیص داده شده است.



شکل ۳-۷: فرایند مولد [۱۷]

به طور کلی اگر بخواهیم مدل‌های موضوعی آماری را بر مبنای LDA توصیف کنیم، می‌توانیم آن را با شکل ۳-۸ توصیف نماییم. ورودی این مدل، مجموعه‌ای از اسناد است که به صورت کیسه کلمات نمایش داده می‌شوند. سپس برای مجموعه‌ی اسناد تعدادی موضوع ساخته می‌شود که تعداد آن‌ها از قبل مشخص است. در شکل ۳-۸ تعداد این موضوعات، چهارتا در نظر گرفته شده است که با چهار رنگ نمایش داده شده‌اند. هر موضوع با سه کلمه از آن که دارای بیشترین امتیاز در آن موضوع هستند، نمایش داده شده است. اگر به هریک از این موضوعات دقت کنیم، از دید ناظر انسانی نیز نماینده یک موضوع خاص هستند؛ به عنوان مثال موضوع زرد درباره‌ی ژن‌ها است و موضوع سبز در مورد سیستم عصبی انسان.

نمونه سند نمایش‌داده‌شده در این شکل، در این باره صحبت می‌نماید که یک ارگانیسم ساده، حداقل به چند ژن احتیاج دارد، تا بتواند به حیات خود ادامه دهد. مطالب موجود در سند، شامل سه‌تا از موضوعات ذکرشده است که در هیستوگرامی (توزیع احتمالی) در سمت راست تصویر نمایش داده شده است. برای برخی از کلمات متن نیز، تخصیص آنها به موضوعات ذکرشده مشخص شده است.



شکل ۳-۸- مثالی از عناصر مدل LDA [18]

مثال دیگری از خروجی و عناصر این مدل که روی یک داده نمونه توسط نرم‌افزار ملت[۲۸] [۲۳] تولید شده است، در شکل ۳-۹ قابل مشاهده می‌باشد. تعداد موضوعات استخراج‌شده در این مثال، ۲۰ عدد است. برای سند شماره صفر، توزیع موضوعات، تخصیص برخی کلمات سند به موضوعات و موضوع شماره ۵ که تمام تکرارهای کلمه school در سند صفر، به این موضوع اختصاص یافته‌اند، با پروزن‌ترین ۲۰ کلمه اولش نمایش داده شده است. چنین کلماتی را شاخص[۲۹] می‌نامیم. در روش LDA، می‌توان موضوع‌ها را خوشه‌هایی همپوشان از کلمات دانست. هر کلمه در یک سند به موضوعی تعلق می‌گیرد که سایر کلمات آن سند نیز مرتبط با موضوع مذکور باشند. به این ترتیب، یک کلمه می‌تواند در سند جاری متعلق به یک موضوع باشد و در سندی دیگر، متعلق به موضوعی دیگر.

[28] - Mallet
[29] - Descriptor

توزیع موضوعات در یک سند نمونه (۱۰ موضوع مهم در سند)

5	1	12	19	18	16	6	7	11	10
0.341	0.258	0.084	0.033	0.03	0.026	0.026	0.023	20 0.	0.020

بخشی از سند نمونه (سند شماره ۰) و انتساب کلمات آن به موضوعات مختلف

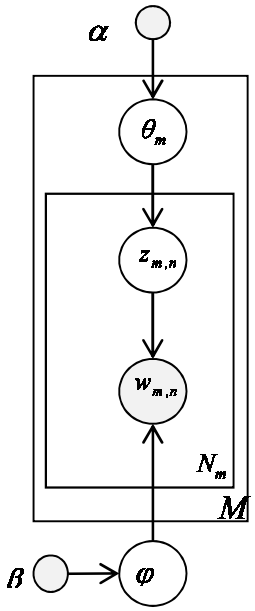
ear 5 0 NA 0 0
0 NA 1 1 student 5
2 2 private 13 0 NA
0 NA 3 3 baptist 5
0 NA 4 4 school 5
0 NA 5 5 allegedly 19
6 killed 15 0 NA 6
her 5 a 0 NA 7 7 t
0 NA 8 8 wounded 15
0 NA 9 9 firing 15
0 NA 10 10 filled 12
om 5 0 NA 11 11 classr
0 NA 12 12 apparently 18
0 NA 13 13 snapped 5
0 NA 14 4 school 5
0 NA 15 14 pastor 6
0 NA 16 15 don 12
0 NA 17 16 happened 12
0 NA 18 17 george 10
0 NA 19 18 sweet 5
0 NA 20 14 pastor 6
0 NA 21 19 atlantic 7
0 NA 22 20 shores 18
0 NA 23 3 baptist 5

پرتکرارترین ۲۰ کلمه اول موضوع ۵

Topics
year
mrs
family
school
children
life
students
ms
wife
women
years
university
woman
son
home
left
mother
parents
college

شکل ۳-۹- مثال تولید شده با نرم‌افزار ملت

در LDA، نقش اساسی را برای تعیین موضوعات، همرخدادی کلمات بازی می‌کند [۱۷]. در این مدل هر سند d_m یک توزیع بر روی موضوعات است که با θ_m نمایش داده می‌شود یعنی $\theta_m = p(k|\theta_m)$ و $p(\theta_m)$ دارای توزیع دریکله با پارامتر α است. هر موضوع نیز یک توزیع چندجمله‌ای بر روی کلمات است که با φ_k نمایش داده می‌شود یعنی $\varphi_k = p(v|k,\varphi_k)$ و $p(\varphi_k)$ از یک توزیع دریکله با پارامتر β پیروی می‌کند. شکل ۳-۱۰ نمودار گرافی مدل را نمایش می‌دهد.



شکل ۳-۱۰- نمودار گرافی مدل LDA

همچنین برای LDA فرایند مولد زیر در نظر گرفته شده است. در این فرایند روشن است که از دیدگاه LDA هر کلمه از توزیع مربوط به موضوع منتسب به آن و بدون توجه به کلمات اطرافش استخراج می‌گردد.

- برای هر سند d_m در مجموعه‌داده D
  - توزیع $\theta_m \sim Dirichlet(\alpha)$ را انتخاب کن.
  - برای هر کلمه w_{mn} در سند d_m

- یک موضوع مانند $z_{mn} \sim multinomial(\theta_m)$ را برای آن انتخاب کن.
- کلمه w_{mn} را با توجه توزیع مربوط به آن موضوع انتخاب کن یعنی:

$$w_{mn} \sim multinomial(\varphi_{z_{mn}})$$

۳-۸-۲- مدل موضوعی بایگرم (BTM)

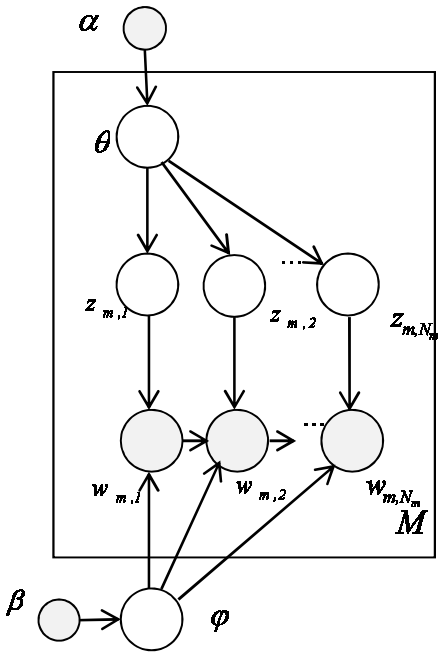
در این مدل نیز هر سند d_m یک توزیع بر روی موضوعات است که با θ_m نمایش داده می‌شود یعنی $\theta_m = p(k|\theta_m)$ و $p(\theta_m)$ دارای توزیع دریکله با پارامتر α است. هر موضوع دیگر فقط یک توزیع یگانه بر روی کلمات نیست بلکه هر موضوع از $|V|$ توزیع بر روی کلمات تشکیل شده است که هر کدام با $\varphi_{kv'}$ نمایش داده می‌شود و $\varphi_{kv'} = p(v|k,v',\varphi_{kv'})$. همچنین $p(\varphi_{kv'})$ از یک توزیع دریکله با پارامتر β پیروی می‌کند.

شکل ۳-۱۱ نمودار گرافی مدل را نمایش می‌دهد. برای BTM فرایند مولد زیر در نظر گرفته شده است. در این فرایند می‌توان دید که از دیدگاه BTM هر کلمه از توزیع مربوط به موضوع منتسب به آن و همچنین کلمه قبلیش استخراج می‌گردد. به این ترتیب در BTM ارتباط محلی کلمات در یک پنجره دوتایی در نظر گرفته شده است. ولی در بخش‌های بعد خواهیم دید که چون آمار چنین زوج کلماتی تنک خواهد بود، نتایج بدست آمده چندان قابل اعتماد نیستند.

- برای هر سند d_m در مجموعه‌داده D:
  - توزیع $\theta_m \sim Dirichlet(\alpha)$ را انتخاب کن.
  - برای هر کلمه w_{mn} در سند d_m:
    - یک موضوع مانند $z_{mn} \sim multinomial(\theta_m)$ را برای آن انتخاب کن.
    - با توجه به اینکه کلمه قبل از کلمه جاری w_{mn-1} است:

- کلمه w_{mn} را از توزیع مربوط به آن موضوع انتخاب کن یعنی:

$$w_{mn} \sim multinomial\left(\varphi_{z_{mn},w_{mn-1}}\right)$$



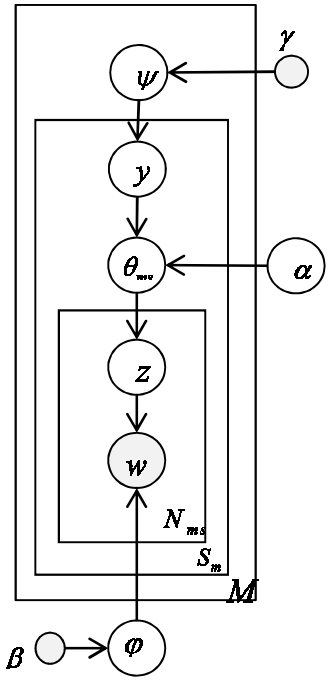
شکل ۳-۱۱- نمودار گرافی مدل BTM

۳-۸-۳- مدل خوشه‌بندی همزمان پنهان دریکله (LDCC)

این مدل که نمایش گرافی آن را در شکل ۳-۱۲ می‌بینیم، سطح جدیدی از جزئیات را برای داده‌ها در نظر می‌گیرد تا بتواند ساختار اسناد را در فرایند تولید آنها لحاظ نماید. در این مدل، هر سند کیسه‌ای از قطعات[۳۰] جابه‌جایی‌پذیر است. یک قطعه می‌تواند یک جمله، یک پاراگراف یا هر قطعه دیگری از سند باشد. هر کدام از این قطعات، نماینده یک موضوع هستند. موضوعات منتسب به قطعات را قطعه‌موضوع می‌نامیم. همچنین هرقطعه کیسه‌ای از کلمات است. هر کلمه از قطعات برگرفته از یک موضوع است. هر

[30] - Bag of segments

یک از موضوعات منتسب به کلمات را کلمه‌موضوع می‌نامیم. به این ترتیب در LDCC دو نوع متفاوت از موضوعات تصور می‌شود.



شکل ۳-۱۲- نمودار گرافی مدل LDCC

هر سند d_m یک توزیع مانند ψ_m از کلمات است. هر قطعه از این سند مانند s_m یک توزیع θ_{mp} بر روی کلمه‌موضوع‌ها است که در آن p نماینده شماره قطعه است. موضوعات منتسب به کلمات را با z و موضوعات منتسب به قطعات را با y نمایش می‌دهیم. توزیع φ همانند قبل نماینده کلمه‌موضوع‌ها است که هرکدام توزیعی بر روی کلمات هستند.

فرایند مولد زیر روند تولید سندها را در LDCC توصیف می‌کند:

- برای هر سند d_m در مجموعه‌داده D
  - توزیعی مانند $\psi_m \sim Dirichlet(\gamma)$ انتخاب کن.
  - برای هر قطعه s_p در سند d_m:

- موضوع y_p را از توزیع $y_p \sim Multinomial(\psi_m)$ استخراج کن.
- توزیع $\theta_{mp} \sim Dirichlet(\alpha, y_p)$ را انتخاب کن.
- برای هر کلمه w_n در قطعه s_p
  - یک موضوع مانند $z_n \sim multinomial(\theta_{mp})$ را برای آن انتخاب کن.
  - کلمه w_{mn} را با توجه به توزیع مربوط به آن موضوع انتخاب کن یعنی:

$$w_n \sim multinomial(\varphi_{z_n})$$

در LDCC هر قطعه‌موضوع توزیعی بر روی کلمه‌موضوع‌ها و هر کلمه‌موضوع توزیعی بر روی کلمات است. تصور این سلسله مراتب به مدل اجازه می‌دهد تا خوشه‌های اسناد را که برگرفته از قطعه‌موضوع‌ها هستند بر مبنای نمایشی کم‌بعد از اسناد تشکیل دهد که توسط کلمه‌موضوع‌ها فراهم شده است. به همین دلیل خوشه‌بندی اسناد دراین مدل معمولا دقیق‌تر از مدل‌های مشابه [۶۸, ۶۹] است. می‌توان گفت که در LDCC، خوشه‌بندی اسناد و کلمات به طور همزمان انجام می‌گیرد و خوشه‌بندی اسناد تنها یک فراورده جانبی از خوشه‌بندی کلمات نیست. به همین دلیل است که این مدل "خوشه‌بندی همزمان پنهان دریکله" نامیده شده است.

چنان‌که ذکر شد، این مدل دو سطح متفاوت از موضوعات را تصور می‌کند و تنها اعمال ساده LDA بر روی قطعات کوچکتر متن نیست. به همین دلیل، می‌توان گفت که این مدل قادر است موضوعات را بر اساس همرخدادی‌های محلی‌تر کلمات استخراج نماید. البته این مدل تنها به اینکه هر کلمه در چه قطعه‌ای از متن ظاهر شده است حساس است ولی فاصله نسبی و ترتیب کلمات را لحاظ نمی‌کند.

۳-۹- جمع‌بندی

در این فصل، به معرفی توزیع‌های چندجمله‌ای و دریکله پرداختیم. همچنین، مدلهای بیزین یا همان مدلهای گرافیکی جهت‌دار را معرفی نموده و بر اساس آن وارد بحث مدلهای سلسله‌مراتبی بیز شده، توضیح دادیم که چگونه می‌توان مدلهای مخلوط را براساس این مفهوم نمایش داده و بررسی نمود. مفاهیم فوق مقدمه ورود به بحث مدل‌های موضوعی هستند که در انتهای این فصل توصیف شده‌اند. این مدل‌ها به طور ضمنی، برمبنای همرخدادی کلمات کار می‌کنند و پیشنهاد رساله نیز با همین خط فکری ارائه شده است. به همین دلیل، بحث همرخدادی نیز در این فصل به تفصیل مورد معرفی و بررسی قرار گرفته است که حاصل آن پیشنهاد مفهوم محتوای مکانی است که در فصل‌های آتی از آن بهره خواهیم برد.

۴

مدل مبتنی بر پنجره همرخدادی

۴- مدل مبتنی بر پنجره همرخدادی

مدل‌های موضوعی با وجود قابلیت‌هایشان، از محدودیت‌هایی رنج می‌برند. یکی از این محدودیت‌ها، عدم امکان بهره‌گیری از روابط محلی کلمات است. اکثر مدل‌های موضوعی برگرفته از LDA هستند. این مدل مبتنی بر همرخدادی کلمات در محدوده یک سند است و همچنین از نمایش "کیسه کلمات" استفاده می‌نماید یعنی کلمات متن را "جابه‌جایی‌پذیر" و مستقل فرض می‌کند. تا کنون روش‌های مختلفی پیشنهاد شده‌اند تا امکان بهره‌گیری از روابط محلی کلمات را برای مدل‌های موضوعی فراهم کنند. بسیاری از این روش‌ها هدف خود را با فرض وابستگی بین کلمات و کنار گذاشتن فرض جابه‌جایی‌پذیری کلمات یا به عبارتی ترکیب مدل‌های زبانی n-gram و مدل‌های موضوعی عملی می‌کنند. الهام‌بخش اینگونه مدل‌ها را می‌توان مدل BTM [۳۹] دانست که در آن فرض می‌شود هر کلمه به کلمه قبلی خود در سند وابسته است.

فرض وابستگی بین کلمات متن در یک مدل موضوعی، یا به عبارتی، در نظر گرفتن ترتیب دقیق کلمات در استخراج موضوعات متن اما، ما را با مشکل تنکی روبرو خواهد کرد. زیرا حتی در ساده‌ترین حالت، (bigram) بسیاری از ترکیبات ممکن کلمات، در مجموعه‌داده مشاهده نخواهند شد و به طور کلی تکرارهای ترکیبات ممکن عموما بسیار اندک خواهد بود که باعث می‌شود نتایج حاصل چندان قابل اعتماد نباشد. هرچه اندازه n در n-gramها بزرگتر باشد این مشکل جدی‌تر خواهد بود.

در این بخش به معرفی مدلی می‌پردازیم که هر کلمه را علاوه بر موضوع آن، برگرفته از کلمات قبلیش فرض می‌کند ولی وابسته به ترتیب دقیق این کلمات نیست. منظور از کلمات قبلی در اینجا، کلماتی هستند که در پنجره با طول مشخص واقع شده‌اند. این پنجره قبل از کلمه جاری قرار گرفته است به گونه‌ای که می‌توان کلمه جاری را به عنوان آخرین کلمه آن فرض کرد. در این مدل هر کلمه خود توزیعی

بر روی کلمات است و هر موضوع، مجموعه‌ای از این توزیعات. کلمه جاری توسط یکی از توزیعات کلمات موجود در پنجره خود تولید می‌شود.

همانطور که گفته شد، در این مدل، هر موضوع، تنها یک توزیع بر روی کلمات نیست بلکه مجموعه‌ای از توزیعات مختلف روی کلمات است به این معنی که بسته به اینکه در اطراف چه کلمه‌ای اتفاق می‌افتد وجوه مختلفی از خود را به نمایش می‌گذارد. برای معرفی این مدل، ابتدا مدلی را معرفی می‌کنیم که بدون در نظر گرفتن موضوعات، توزیع کلمات را استخراج می‌نماید و نشان می‌دهیم که این مدل قادر است با تخفیف مشکل تنکی نسبت به روش‌هایی که ترتیب دقیق کلمات را در نظر می‌گیرند، ارتباطات کلمات را به صورت $p(w_i \mid w_j)$ استخراج نماید. $p(w_i \mid w_j)$ که توزیع مربوط به کلمه w_j را نشان می‌دهد، احتمال رخداد کلمه w_i در محدوده (پنجره) معینی قبل از کلمه w_j است. سپس موضوعات را نیز به این مدل افزوده و آن را با برخی مدل‌های مرتبط مقایسه و ارزیابی می‌نماییم.

۴-۱- مدل زبانی مبتنی بر پنجره همرخدادی

در این بخش، یک مدل سلسله مراتبی بیز برای استخراج روابط محلی کلمات معرفی شده است. این مدل را می‌توان یک مدل برای زبان دانست. مدل‌های احتمالاتی زبانی یا همان مدل‌های زبانی، مکانیزم‌های احتمالاتی برای تولید متن در یک زبان هستند [۷۰]. بر اساس این تعریف، گستره وسیعی از مدل‌های زبانی قابل تصور است. پرکاربردترین این مدل‌ها از ابتدا تا کنون، مدل‌های n-gram بوده‌اند که نقشی اساسی در کاربردهایی از جمله شناسایی صحبت [۷۱]، نویسه‌خوانی نوری [۷۲]، اصلاح املای کلمات [۷۳] و ترجمه ماشینی دارند. خواستگاه مدل‌های n-gram، حوزه شناسایی صحبت است که در آن سعی می‌شود دنباله‌های اصوات با دنباله‌های کلمات نگاشت داده شوند. بنابراین مدل‌های n-gram مبتنی بر ترتیب دقیق کلمات هستند. این مدل‌ها در واقع سعی می‌کنند با توجه به یک دنباله nتایی از کلمات یک

عبارت، کلمه بعدی را پیش‌بینی کنند. به همین دلیل، هموارسازی[1] یکی از مسائل اصلی این مدل‌ها است. یعنی به دلیل اینکه مدل‌های n-gram مبتنی بر دنباله‌های مرتب کلمات هستند، بسیاری از ترکیبات ممکن، در مجموعه داده مشاهده نخواهند شد و بنابراین نیازمندیم که روش‌های هموارسازی را تعیین نماییم که احتمال غیر صفر به چنین ترکیباتی اختصاص دهند. برخی از مهمترین روش‌های هموارسازی که تا کنون معرفی شده‌اند عبارتند از روش کنسر-نای[2] [۷۴]، روش تقلیل مطلق[3] [۷۵]، روش جلینک-مرسر[4] [۷۶] و روش کنسر-نای بهبودیافته (KMN)[5] [۷۷]. برای اطلاعات بیشتر و توصیف این روش‌ها می‌توان به [۷۷] مراجعه نمود. سایر مدل‌های زبانی معرفی شده تاکنون نیز، تا جایی که ما می‌دانیم برمبنای همین مدل‌های n-gram شکل گرفته‌اند ولی با شیوه‌های محاسباتی متفاوت. مانند مدل‌های مبتنی بر شبکه‌های عصبی [۷۸] یا مدل کلمات پنهان [۷۹] که از متغیرهای پنهان و مدل‌های گرافیکی برای محاسبه مدل زبانی استفاده نموده است یا [۸۰] که از روش تخمین تغییراتی برای محاسبه مدل زبانی استفاده نموده است. به طور کلی مدل‌های زبانی موجود را می‌توان به دو دسته تقسیم نمود: مدل‌های بلنددامنه و مدل‌های کوتاه‌دامنه. مدل‌های کوتاه‌دامنه مدل‌هایی هستند که روابط نزدیک کلمات را مدل می‌کنند یا همان مدل‌های n-gram با nهای کوچک. مدل‌های بلنددامنه مدل‌هایی هستند که مبتنی بر روابط دورترند مثل مدل‌هایی که روابط ساختاری را درنظر می‌گیرند [۸۱] یا مدل‌های n-gram با nهای بزرگ. به طورکلی مدل‌های بلنددامنه دقیق‌تر هستند ولی به دلیل مشکل تنکی یا سنگینی محاسبات عموما عملی نیستند [۸۰, ۸۲].

چنان‌که ذکر شد، در مدل‌های زبانی n-gram، به دلیل در نظر گرفتن ترتیب دقیق کلمات با مشکل تنکی روبرو هستیم. در کاربردهایی مانند شناسایی صحبت و ترجمه ماشینی، ترتیب دقیق کلمات دارای

[1] - Smoothing
[2] - Knesser-Ney
[3] - Absolute discounting
[4] - Jelinek-Mercer
[5] - Modified Knesser-Ney (MKN)

نقش کلیدی است ولی در بسیاری از کاربردها، این امر صادق نیست. در دهه‌های اخیر، این مدل‌ها در کاربردهای دیگری مثل بازیابی اطلاعات [۷۰] و استخراج مدل‌های موضوعی [۳۹] نیز به کار گرفته و موثر نیز واقع شده‌اند. در این کاربردها، ترتیب کلمات در برخی موارد مثلا در مورد اصطلاحات یا ترکیبات خاص می‌تواند کارآمد باشد. با این حال، ترتیب دقیق کلمات، در چنین کاربردهایی نقش مستقیم و حیاتی بازی نمی‌کند [۸۳]. باتوجه به اینکه استفاده از مدل‌های n-gram ما را با مشکل تنکی مواجه خواهد نمود، مهم است بتوانیم مدلی از زبان بسازیم که ضمن نادیده نگرفتن ترتیب کلمات بتواند مشکل تنکی را تخفیف دهد. در این بخش، هدف ما معرفی چنین مدلی است. مدل پیشنهادی فرض می‌کند که هر کلمه بر مبنای یکی از کلمات پیش از خودش تولید شده است. این کلمه پیشین می‌تواند در هر موقعیتی، در یک محدوده خاص قبل از کلمه جاری قرار گرفته باشد. به همین دلیل، تغییر ترتیب کلمات متن، کلمات پیشین هر کلمه را تغییر خواهد داد و بنابراین مدل، مستقل از ترتیب کلمات متن نیست. با این حال، از آنجا که دنباله‌های دقیقا مرتب کلمات را در نظر نمی‌گیریم، مشکل تنکی تا حد زیادی تخفیف می‌یابد. هنگامی که انسانی در حال نگارش یک متن است، گذشته از دستور زبان، عوامل گوناگونی ممکن است او را در انتخاب کلمه بعدی متن هدایت نمایند. حداقل دوتا از این عوامل در تحلیل زبان‌های طبیعی مورد توجه قرار گرفته‌اند: مفهوم کلی که برای نگارش متن در ذهن اوست یا همان موضوع متن و سایر کلماتی که پیش از این در متن استفاده کرده است. اولی منجر به پیشنهاد مدل‌های موضوعی و دومی منجر به پیشنهاد مدل n-gram شده است. در این بخش هرچند از مفهوم مدل‌های سلسله‌مراتبی بیز استفاده شده است، اما قصد ما پرداختن به موضوع متن نیست بلکه تکیه ما مستقیما بر رابطه کلمات است. با این حال، ایده این بخش چنان‌که ذکر شد، با مدل n-gram نیز متفاوت است. می‌خواهیم $p(w_i \mid w_j)$ را که احتمال رخداد کلمه w_i است بدست آوریم، به شرط اینکه w_j در یک محدوده مشخص قبل از موقعیت آن رخ داده باشد. در مدل‌های n-gram فرض بر این است که یک کلمه از توزیعی بر روی دنباله مرتب

کلمات قبلیش تولید می‌شود. در مدل پیشنهادی فرض بر این است که هر کلمه از یکی از کلمات قبلیش گرفته شده است ولی از آنجا که فقط به کلمات قبلی نگاه می‌کنیم و موقعیت کلمات دارای اهمیت است می‌توان گفت که ترتیب به کلی نادیده گرفته نشده است بلکه اثر آن تا حد زیادی خفیف شده است. تفاوت این مدل با یک مدل موضوعی مانند LDA این است که در این مدل فرض می‌شود که هر کلمه توسط یکی از کلمات پیشین خود تولید شده است و نه موضوع منتسب به آن. به این ترتیب، در مدل پیشنهادی، هر کلمه خود توزیعی بر روی کلمات است. این مدل ضمن اینکه می‌تواند در کاربردهایی که ترتیب دقیق کلمات در آنها نقش کلیدی بازی نمی‌کند مفید باشد، همچنین به راحتی قابل گنجاندن در یک مدل موضوعی احتمالاتی مانند LDA خواهد بود. مدل پیشنهادی از نظر تواناییش در مدل کردن داده‌ها با استفاده از معیار پرپلکسیتی با مدل n-gram مقایسه شده است و برای پنجره‌هایی با طول‌های مختلف، بهتر از مدل n-gram عمل کرده است.

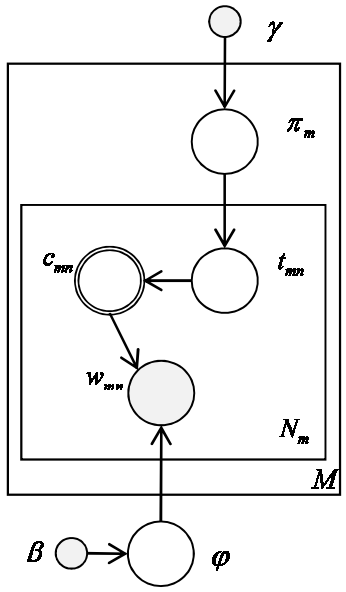
۴-۱-۱- توصیف مدل

چنان‌که پیش از این ذکر شد، در این بخش مدلی معرفی می‌شود که فرض می‌کند هر کلمه خود توزیعی بر روی کلمات است. هر کلمه می‌تواند، از توزیع مربوط به هرکدام از کلمات پیشین خود، در یک پنجره با طول از پیش تعیین شده، تولید شده باشد. در این مدل، هر سند توزیعی بر روی کلمات است.

برای توصیف مدل پیشنهادی فرض می‌کنیم که مجموعه داده D متشکل از M سند d_m است. هرسند d_m متشکل از N_m کلمه w_{mn} است که در آن n، موقعیت کلمه در سند را نشان می‌دهد. مجموعه کلمات منفرد [۶] متن را با $V=\{v_1,v_2,...,v_{|V|}\}$ نشان می‌دهیم. هر کلمه w_{mn} منتسب به یک کلمه c_{mn}

[6] - Vocabulary

است که در پنجره‌ای به طول L قبل از کلمه مربوطه واقع شده است. گراف مربوط به مدل پیشنهادی در شکل ۴-۱ نمایش داده شده است.



شکل ۴-۱- نمایش گرافیکی مدل پیشنهادی

در این شکل، φ حاوی $|V|$ توزیع چندجمله‌ای برروی کلمات است که هریک متناظر با یکی از کلمات منفرد مجموعه‌داده است. پارامترهای این توزیع‌ها از یک توزیع دریکله پیروی می‌کنند با پارامتر β. همچنین در شکل مذکور π_m حاوی یک توزیع چندجمله‌ای بر روی کلمات برای سند d_m است که پارامترهای آن نیز از یک توزیع دریکله با پارامتر γ پیروی می‌کنند. البته از آنجا که اندیس mn در هنگام نگارش باعث شلوغی روابط و کاهش خوانایی آنها می‌گردد به‌جای آن از اندیس i استفاده کرده‌ایم به این شکل که تمام کلمات مجموعه داده به ترتیب ظهورشان از ۱ تا N شماره‌گذاری شده‌اند و اندیس i می‌تواند یکی از مقادیر ۱ تا N را بپذیرد. این درحالی است که تقسیمات فوق همچنان برقرار است ولی

سند حاوی هر کلمه w_i را با d_i نشان می‌دهیم. همینطور پنجره متناظر با آن با s_i نشان داده می‌شود. فرایند مولد مدل پیشنهادی به شرح زیر است:

- برای هر سند d_m در مجموعه‌داده D :
  - ابتدا $\pi_m \sim Dirichlet(\gamma)$ انتخاب کن.
- برای هر کلمه w_i در مجموعه داده:
  - یکی از L موقعیت موجود در پنجره متناظرش s_i را از توزیع $t_i \sim Multinomial(\pi_{d_i})$ انتخاب کن.
  - کلمه موجود در موقعیت t_i را به عنوان c_i قرار بده.
  - کلمه جاری w_i را از توزیع مربوط به $w_i \sim Multinomial(\varphi^{c_i})$ انتخاب کن.

فرایند مولد فوق چگونگی تولید یک سند را از نگاه مدل پیشنهادی، با فرض دانستن پارامترهای مدل، نشان می‌دهد. همانطور که در این فرایند می‌بینیم، برای تولید هر کلمه از متن، یکی از کلمات قبلیش به طور تصادفی انتخاب می‌شود و کلمه هدف از توزیع این کلمه استخراج می‌گردد. برای محاسبه پارامترهای مدل، فرایند مذکور باید معکوس گردد یعنی پارامترهای مدل را به گونه‌ای می‌یابیم که احتمال اسناد برمبنای آنها حداکثر شود یا به بیان دیگر در فرایند تولید هر یک از کلمات متن، منجر به انتخاب بهترین کلمه از بین کلمات قبلیش گردد.

۴-۱-۲- تخمین پارامترهای مدل

همانطور که پیش از این گفته شد، مدل پیشنهادی یک مدل مولد برای اسناد مجموعه داده یا به عبارتی کلمات آن سند است. بر این اساس باید پارامترهای مدل را به گونه‌ای بیابیم که احتمال رخداد این مجموعه را حداکثر نماید. احتمال رخداد یک سند بر مبنای مدل پیشنهادی، با توجه به مدل گرافیکی شکل ۴-۱ به صورتی که در رابطه (۱) آمده است محاسبه می‌گردد.

$$p(D|\varphi) = \prod_{m=1}^{M} \int p(\pi_m) \left(\prod_{n=1}^{N_m} \sum_{t_{mn}} p(t_{mn}|\pi_{mn}) \sum_{c_{mn}} p(c_{mn}|t_{mn})\, p\left(w_{mn}|c_{mn}, \varphi_{c_{mn},w_{mn}}\right) \right) d\pi_m \qquad (۱-۴)$$

انتگرال موجود در رابطه فوق به صورت مستقیم قابل محاسبه نیست بنابراین باید از روش‌های تخمینی برای محاسبه پارامترها استفاده نماییم. برای تخمین پارامترها در مدل‌های مشابه از روش‌های مختلفی [۱۸, ۸۴] استفاده شده است که بهترین عملکرد را روش نمونه‌برداری گیبس داشته است [۸۵]. زمانی که تولید مستقیم نمونه‌های یک توزیع احتمال دشوار یا غیرممکن باشد، از نمونه‌برداری گیبس که یک الگوریتم زنجیره مارکوف مونت کارلو است [۸۶] برای تخمین نمونه مشاهدات [۸۷] استفاده می‌شود. این دنباله از مشاهدات می‌توانند در تخمین توزیع‌های احتمال متغیرهای پنهان یا برخی پارامترهای مدل به کار گرفته شوند. نمونه‌برداری گیبس ابزاری برای استنباط‌های آماری به خصوص استنباط بیزی است. این الگوریتم یک زنجیره مارکوف از نمونه‌ها می‌سازد که در آن، هر حالت، یک نمونه از تمامی متغیرهای موجود است. در هر حالت، هر متغیر به شرط مقادیر سایر متغیرها به‌روز می‌شود و زمانی از یک حالت به حالت دیگر می‌رویم که مقادیر تمامی متغیرها به‌روز شده باشد. در این روش، هر نمونه به نمونه‌های نزدیکش وابسته است بنابراین معمولا از تمام نمونه‌ها استفاده نمی‌شود بلکه مثلا از هر ۱۰۰ نمونه یکی برداشته می‌شود. همچنین تعدادی از اولین نمونه‌های انتخاب شده دور ریخته می‌شوند. برای اینکه نمونه‌برداری گیبس را اعمال نماییم باید $p\left(c_{xy}, t_{xy} | c_{-xy}, t_{-xy}, w\right)$ را محاسبه کنیم که در آن x اندیس سند جاری و y اندیس کلمه جاری است. نماد $-xy$ نیز به معنی همه موقعیت‌ها غیر از موقعیت جاری

یعنی xy است که و c_{xy} کلمه منتسب به کلمه جاری در موقعیت x از سند d_y است و c_{-xy} نماینده کلمات منتسب به سایر کلمات غیر از کلمه جاری است. بازهم برای سادگی در اینجا به جای اندیس xy از i استفاده می‌نماییم. که موقعیت کلمه را در کل مجموعه داده نشان می‌دهد. در این حالت سند حاوی کلمه جاری را با d_i نشان خواهیم داد. فرض مدل این است که t_i موقعیتی از یک پنجره بر روی کلمه w_i را نشان می‌دهد و c_i کلمه موجود در آن موقعیت است. با مشخص شدن مقدار t_i مقدار c_i نیز مشخص است. به این ترتیب با وجود اینکه هر دو متغیر، تصادفی هستند، در هر آزمایش و وابسته به همان آزمایش مقدار یکی با توجه به مقدار دیگری قابل محاسبه است. پس در اینجا به‌جای $p(t_i=l,c_i=v|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i},\boldsymbol{w})$ احتمال $p(t_i=l|\boldsymbol{t}_{-i},\boldsymbol{c},\boldsymbol{w})$ را محاسبه خواهیم نمود و با استخراج t_i از این توزیع، مقدار c_i را نیز خواهیم داشت. این احتمال در الگوریتم شکل ۴-۲ برای به روز کردن نمونه‌ها استفاده می‌شود. احتمال مربوطه به فرم زیر قابل محاسبه است.

$$p(t_i=l|\boldsymbol{t}_{-i},\boldsymbol{c},\boldsymbol{w})=\frac{p(t_i=l,c_i=v,w_i|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i},\boldsymbol{w}_{-i})}{p(w_i,c_i=v|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i},\boldsymbol{w}_{-i})} \qquad (۴-۲)$$

همانطور که در رابطه فوق مشاهده می‌نمایید، مقدار مخرج ثابت خواهد بود و بنابراین قابل حذف است. یعنی داریم:

$$\propto p(t_i=l,c_i=v,w_i|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i},\boldsymbol{w}_{-i})=p(t_i=l,c_i=v|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i})\,p(w_i|c_i=v,\boldsymbol{c}_{-i},\boldsymbol{w}_{-i}) \qquad (۴-۳)$$

فاکتور سوم به دلیل ثابت بودن قابل حذف است. با توجه به مدل گرافیکی شکل ۴-۱، فاکتور اول به صورت زیر قابل محاسبه است:

$$\begin{aligned}&p(t_i=l,c_i=v|\boldsymbol{t}_{-i},\boldsymbol{c}_{-i})=p(c_i=v|t_i=l)\,p(t_i=l|t_{-i})\\&=p(c_i=v|t_i=l)\int p(t_i=l,\pi_m|\boldsymbol{t}_{-i})\,d\pi_m\\&=p(c_i=v|t_i=l)\int p(t_i=l|\pi_m)\,p(\pi_m|t_{-i})\,d\pi_m\end{aligned} \qquad (۴-۴)$$

فرض می‌کنیم که $p(c_i = v \mid t_i = l)$ دارای مقدار ثابت است (مثلا $\frac{1}{|V|}$) و بنابراین:

$$p(t_i = l \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}) \propto \int p(t_i = l \mid \pi_m)\, p(\pi_m \mid t_{-i})\, d\pi_m \qquad (۵-۴)$$

برای زیر انتگرال داریم:

$$p(\pi_{d_i} \mid \boldsymbol{t}_{-i}) \propto p(\pi_{d_i})\, p(t_{-i} \mid \pi_{d_i}) \qquad (۶-۴)$$

از آنجا که $p(\pi_{d_i})$ از $Dir(\gamma)$ پیروی می‌کند و مزدوج $p(t_{-i} \mid \pi_{d_i})$ است بنابراین احتمال پسین $p(\pi_{d_i} \mid \boldsymbol{t}_{-i})$ دارای توزیع $Dir(\gamma + n_{-i,l}^{d_i})$ خواهد بود که در آن $n_{-i,l}^{d_i}$ نماینده تعداد دفعاتی است که l در سند d_i به کلمه‌ای انتساب یافته است بدون در نظر گرفتن موقعیت جاری i. به این ترتیب زیر الگوریتم فرم $\int \pi p(\pi) d\pi$ در می‌آید که برابر $E[\pi]$ است در حالی که π از توزیع $Dir(\gamma + n_{-i,l}^{d_i})$ پیروی می‌کند و میانگین توزیع دریکله مذکور برابر است با $\frac{n_{-i,l}^{d_i} + \gamma}{n_{-i,\cdot}^{d_i} + L\gamma}$ که در آن $n_{-i,\cdot}^{d_i}$ برابر است با تعداد کلمات سند d_i. یعنی:

$$p(t_i = l \mid \boldsymbol{t}_{-i}) \propto \frac{n_{-i,l}^{d_i} + \gamma}{n_{-i,\cdot}^{d_i} + L\gamma} \qquad (۷-۴)$$

بعد از انتخاب t_i مقدار c_i نیز مشخص می‌شود. حال باید $p(c_i = v \mid \boldsymbol{t}, \boldsymbol{c}_{-i}, \boldsymbol{w})$ را محاسبه نماییم که به صورت زیر خواهد بود و به دلیل ثابت بودن مخرج باز هم متناسب است با $p(t_i = l, c_i = v, w_i \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i})$.

$$p(c_i = v \mid \boldsymbol{t}, \boldsymbol{c}_{-i}, \boldsymbol{w}) = \frac{p(t_i = l, c_i = v, w_i \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i})}{p(w_i, t_i = l \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i})} \qquad (۸-۴)$$

در حاصل‌ضرب مربوطه، این بار، $p(t_i = l \mid \boldsymbol{t}_{-i})$ قابل حذف خواهد بود و $p(c_i = v \mid t_i = l)$ را نیز که ثابت فرض نمودیم. بنابراین حاصل‌ضرب در نهایت متناسب است با $p(w_i \mid c_i = v, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i})$ که می‌تواند به صورت زیر محاسبه شود:

$$p(w_i \mid c_i = v, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}) = \int p(w_i, \varphi \mid c_i = v, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}) d\varphi$$
$$= \int p(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}) p(w_i \mid c_i = v, \varphi) d\varphi \quad (۹-۴)$$

برای زیر انتگرال داریم:

$$p(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}) \propto p(\varphi) p(\boldsymbol{w}_{-i} \mid \boldsymbol{c}_{-i}, \varphi) \quad (۱۰-۴)$$

که مشابه قبل می‌توان از آن نتیجه گرفت که احتمال پسین $p(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i})$ از توزیع $Dir(n_{-i,v,w_i} + \beta)$ پیروی می‌کند. بنابراین حاصل انتگرال برابر است با $\frac{n^v_{-i,w_i} + \beta}{n^v_{-i,\cdot} + |V|\beta}$ یعنی:

$$p(w_i \mid c_i = v, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}) \propto \frac{n^v_{-i,w_i} + \beta}{n^v_{-i,\cdot} + |V|\beta} \quad (۱۱-۴)$$

در روابط فوق n^v_{-i,w_i} نماینده تعدد دفعاتی است که کلمه w_i به کلمه v از بین کلمات قبل از خودش، اختصاص یافته است بدون در نظر گرفتن کلمه جاری ($-i$) و $n^v_{-i,\cdot}$ تعداد دفعاتی است کلمه‌ای در سرتاسر مجموعه‌داده به کلمه v اختصاص یافته است، بازهم بدون احتساب موقعیت جاری. پارامترهای مدل برای هر یک از نمونه‌ها، با استفاده از روابط (۴-۱۲) و (۴-۱۳) قابل محاسبه هستند.

$$\pi_{d_i l} = \frac{n^{d_i}_l + \gamma}{n^{d_i}_{\cdot} + |L|\gamma} \quad (۱۲-۴)$$

$$\varphi_{w_i}^{v} = \frac{n_{w_i}^{v} + \beta}{n_{\cdot}^{v} + |V|\beta} \qquad (۱۳-۴)$$

که در آنها $n_l^{d_i}$ نماینده تعداد دفعاتی است که در سند d_i، کلمه‌ای به موقعیت l از پنجره متناظرش انتساب یافته است. $n_l^{d_i}$ تعداد کلمات سند d_i است. $n_{w_i}^{v}$ تعداد دفعاتی است که کلمه w_i در سرتاسر مجموعه داده به کلمه مولد v انتساب یافته است و $n_{w_i}^{v}$ نماینده تعداد دفعاتی است که کلمه‌ای در سرتاسر مجموعه داده به کلمه مولد v انتساب یافته است.

۱- ورودی‌ها: اندازه پنجره L، مقادیر فراپارامترهای β و γ، حداکثر تعداد تکرار $maxIter$
۲- به صورت تصادفی هر کلمه متن را به یکی از کلمات قبلیش که در محدوده پنجره‌ای به طول L واقع شده است، منتسب کن.
۳- مقادیر اولیه شمارنده $n_l^{d_i}$ را که نماینده تعداد کلماتی است که در سند d_i به موقعیت l از پنجره متناظرشان انتساب یافته‌اند، مشخص کن.
۴- مقادیر اولیه شمارنده $n_{\cdot}^{d_i}$ را که نماینده تعداد کلمات در سند d_i است، مشخص کن.
۵- مقادیر اولیه شمارنده $n_{w_i}^{v}$ را که نماینده تعداد کلماتی مانند v_i است که در سرتاسر مجموعه داده به کلمه مولد v انتساب یافته‌اند، مشخص کن.
۶- مقادیر اولیه شمارنده $n_{\cdot}^{v}$ را که نماینده تعداد کل کلماتی است که سرتاسر مجموعه داده به کلمه v انتساب یافته‌اند، مشخص کن.
۷- برای تعداد تکرار ۱ تا $maxIter$
۸- برای هر سند d_i
۹- برای هر موقعیت i در سند d_i
۱۰- آمار موقعیت جاری را از شمارنده‌ها حذف کن.
۱۱- کلمات موجود در پنجره، در موقعیت‌های $i-1$ تا $i-L$ را در محدوده سند در نظر بگیر.
۱۲- یکی از آنها را با توجه به توزیع $p(t_i \mid t_{-i})$ در رابطه ۶ انتخاب کن.
۱۳- موقعیت و کلمه جدید را به موقعیت جاری (کلمه w_i) منسوب کن.
۱۴- شمارنده‌ها را با توجه به موقعیت و کلمه انتسابی جدید به‌روز کن.
۱۵- بعد از پایان تکرارها مقادیر پارامترهای مدل یعنی π و φ را بر اساس روابط (۴-۱۲) و (۴-۱۳) محاسبه کن.

شکل ۴-۲- الگوریتم نمونه‌برداری گیبس مربوط به محاسبه پارامترهای مدل

بر اساس روابط بدست آمده و توضیحات داده شده در نهایت می‌توان الگوریتم نمونه‌برداری گیبس برای مدل را به صورتی که در شکل ۴-۲ آمده است توصیف کرد. چنان‌که در این شکل قابل مشاهده است، پنجره‌ای که در طول توصیف مدل ذکر شد در نمایش گرافیکی ظهور نمی‌یابد ولی در طول محاسبات، فضای حالات را محدود می‌کند. در روابط فوق β و γ پارامترهای توزیع‌های دریکله یا همان فراپارامترهای مدل هستند. در این پایان‌نامه چنان‌که در مدل‌های مشابه مرسوم است، این پارامترها ثابت و متقارن در نظر گرفته شده‌اند.

۴-۱-۳- آزمایشات و نتایج

در مقدمه توضیح داده شد که در این بخش، هدف معرفی مدلی زبانی است که ترتیب کلمات را به طور کلی نادیده نمی‌گیرد و در عین حال وابسته به ترتیب دقیق کلمات نیست زیرا در بسیاری از کاربردها با اینکه ترتیب به طور کلی بی‌اهمیت نیست ولی نقش مستقیم و حیاتی در نتیجه بازی نمی‌کند و به این ترتیب نیازی نیست تا با در نظر گرفتن ترتیب دقیق کلمات با مشکل تنکی دست‌وپنجه نرم کنیم. در این بخش، هدف ما این است که یک مدل اولیه را با ویژگی‌های مذکور معرفی نموده و آن را از نظر قابلیتش در مدل کردن زبان با روش‌های n-gram مقایسه نماییم. بنابراین بررسی و مقایسه مدل پیشنهادی را در کاربردهای ذکر شده به مراحل بعدی کار وامی‌گذاریم.

۴-۱-۳-۱- معیار ارزیابی

در این بخش برای ارزیابی مدل از معیار پرپلکسیتی [۸۸, ۸۹] استفاده شده است. این معیار نشان می‌دهد که مدل تخمین زده‌شده تا چه حد در پیش‌بینی یک نمونه از مدل هدف، خوب است. هر چه مقدار پرپلکسیتی بدست آمده کمتر باشد مدل مربوطه مدل بهتری از داده‌ها است. این معیار به صورت معکوس میانگین هندسی احتمال کلمات مجموعه بر اساس مدل، به صورتی که در رابطه (۴-۱۴) آمده

است محاسبه می‌گردد. در رابطه مذکور w_i یک کلمه در مجموعه‌داده است، $\mathcal{M}$ نماینده مدل و N نماینده تعداد اسناد مجموعه است. N_m تعداد کلمات سند d_m را نشان می‌دهد.

$$ppl = \left(\prod_{i=1}^{N} p(w_i \mid \mathcal{M}) \right)^{-\frac{1}{N}}, N = \sum_{m=1}^{M} N_m \qquad (۱۴-۴)$$

اگر ابتدا لگاریتم رابطه (۱۴-۴) را گرفته و بعد معکوس آن را انجام دهیم خواهیم داشت:

$$ppl = a^{-\frac{\left(\sum_{n=1}^{N} log_a\, p\left(w_{n_{test}} \mid \mathcal{M} \right) \right)}{N}} \qquad (۱۵-۴)$$

در رابطه (۱۵-۴)، a می‌تواند مقادیر مختلفی را بپذیرد ولی از آنجاکه مدل‌های n-gram و همچنین مقدار پرپلکسیتی آنها در این مقاله با استفاده از نرم‌افزار SRILM [۹۰, ۹۱] بدست آمده و در این نرم‌افزار مقدار a برابر ۱۰ فرض شده است، در این بخش نیز مقدار آن را ۱۰ در نظر می‌گیریم.

البته محاسبه این معیار به روش فوق مستلزم جمع روی تمام حالت‌های ممکن انتساب کلمات به موضوعات است که غیرعملی[۷] است. برای حل این مسئله و محاسبه سریعتر پرپلکسیتی چنان‌که در [۹۲] پیشنهاد شده است، از نمونه‌برداری اهمیت[۸] برای تخمین آن استفاده می‌نماییم. این روش به گستردگی در متدهای مونت کارلو مورد استفاده قرار می‌گیرد [۸۵]. در این روش، S نمونه از متغیر پنهان $\mathbf{z}$ (دراینجا بردار کلمات مولد منتسب به کلمات مجموعه داده) از توزیع $p(z \mid w, \mathcal{M})$ استخراج می‌گردند. سپس $p(w \mid \mathcal{M})$ با استفاده از میانگین هارمونیک $p(w \mid z, \mathcal{M})$ بر روی نمونه‌های استخراج شده تخمین زده می‌شود یعنی:

$$P(w \mid \mathcal{M}) \simeq HM\{ P(w \mid z^{(s)}, \mathcal{M}) \}_{s=1}^{S} = \frac{S}{\sum_{s=1}^{S} 1 / p(w \mid z^{(s)}, \mathcal{M})} \qquad (۱۶-۴)$$

[7] -Intractable
[8] -Importance sampling

۴-۱-۳-۲- مجموعه‌های داده

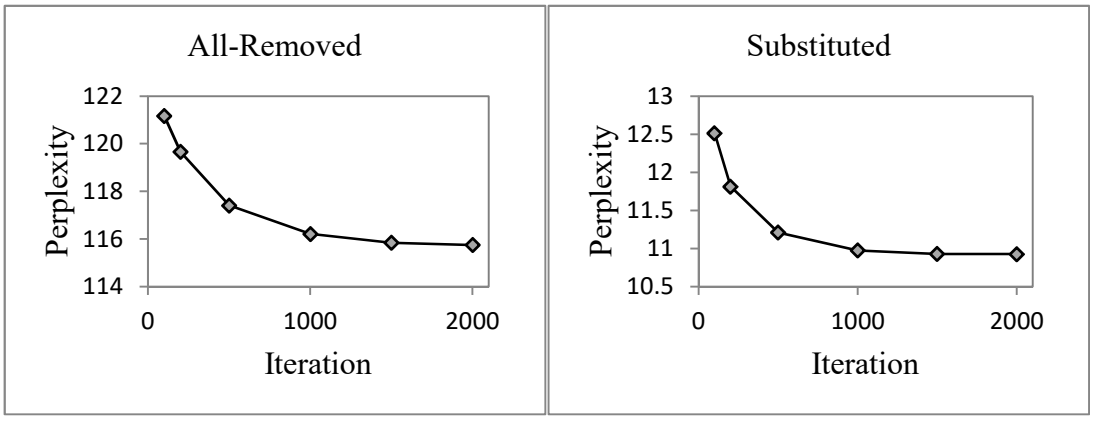
در آزمایشات این مقاله، از مجموعه داده 20 newsgroups [۹۳] استفاده شده است. دو نسخه مختلف از این مجموعه داده تشکیل شده است. برای ایجاد این دو نسخه، ابتدا، ۲۰۰۰ سند برای یادگیری و ۲۰۰۰ سند برای آزمون به صورت تصادفی از مجموعه داده‌ها انتخاب شدند. سپس در نسخه اول که از آن با نام All-Removed یاد خواهیم کرد تمامی ایست‌واژه‌ها، اعداد و علائم حذف شدند. در نسخه دوم که با Substituted نشان داده خواهد شد، ایست‌واژه‌ها، اعداد و علائم هرکدام با یک علامت خاص جایگزین شدند. تفاوت این دو نوع مجموعه‌داده، این است که در اولی دنباله‌های تکراری بسیار کم هستند و بنابراین داده بسیار تنکی محسوب خواهد شد، در حالی که در دومی دنباله‌های تکراری زیادی وجود دارند. در بخش‌های بعد نشان خواهیم داد که در هر دو نوع مجموعه، روش پیشنهادی بهتر از مدل‌های n-gram عمل می‌کند. بعد از مقایسه براساس دو مجموعه داده فوق که نشان می‌دهد که مشکل تنکی در مدل پیشنهادی به بزرگی مدل‌های n-gram نیست، مدل‌های مذکور را بر روی یک مجموعه داده متفاوت نیز مقایسه خواهیم کرد. این مجموعه داده Brown [۹۴] نام دارد و یکی از مجموعه داده‌هایی است که مکررا در مقالات مربوط به مدل‌های زبانی مورد استفاده قرار گرفته است. مجموعه Brown شامل ۵۰۰ متن ۲۰۰۰ کلمه‌ای از ادبیات انگلیسی است. نتایج حاصل در این بخش نیز نشان می‌دهند که مدل پیشنهادی مدل بهتری از مجموعه داده‌ها است.

۴-۱-۳-۳- بررسی نتایج حاصل

در مقالات متعددی که به مقایسه مدل‌های n-gram موجود با روش‌های هموارسازی مختلف پرداخته‌اند [۷۷, ۸۸] ذکر شده است که روش Knesser-Ney بهبودیافته [۷۷] به صورت معناداری بهتر از سایر روش‌های موجود عمل می‌کند و معمولا روش‌های دیگر با این روش مقایسه می‌شوند. بنابراین در این

پایان‌نامه نیز روش پیشنهادی با روش Knesser-Ney بهبودیافته مقایسه شده است که از آن با عنوان MKN یاد خواهیم نمود.

شکل ۴-۳ همگرا شدن مدل را بر روی دو زیرمجموعه کوچک شامل ۲۰۰ سند از هر یک از مجموعه‌داده‌های مذکور را نشان می‌دهد. این آزمایشات با اجرای یک زنجیره مارکوف (نمونه‌برداری گیبس) با ۲۰۰۰ تکرار و پنجره‌ای به طول ۱۰ انجام شده‌اند. همانطور که در شکل‌ها قابل مشاهده است کاهش پرپلکسیتی در ابتدا سریع است و سپس شروع به کند شدن می‌کند. باتوجه به شکل‌ها ۱۰۰۰ تکرار برای حصول نتیجه مناسب می‌تواند کافی باشد.



شکل ۴-۳- همگرا شدن مدل پیشنهادی برروی مجموعه داده‌های مورد استفاده

جدول ۴-۱ نتایج حاصل از اعمال روش MKN برای طول‌های ۱ تا ۵ را روی مجموعه داده‌های مذکور نشان می‌دهد. همچنین این جدول حاوی نتایج حاصل از اعمال روش پیشنهادی برای پنجره‌هایی با طول‌های ۲، ۵، ۱۰ و ۲۰ را نشان می‌دهد. در نظر گرفتن پنجره‌ای به طول یک، به این معنی است که فقط کلمه قبلی می‌تواند به عنوان کلمه سازنده کلمه جاری در نظر گرفته شود. پنجره‌ای به طول پنج یعنی یکی از پنج کلمه‌ای که قبل از کلمه جاری قرار گرفته‌اند می‌تواند سازنده آن باشد. سایر طول‌ها را نیز می‌توان به شکل مشابه تفسیر کرد. برای هر کدام از این آزمایشات، پنج زنجیره مارکوف در ۱۰۰۰ تکرار اجرا شده‌اند و بعد از کنار گذاشتن ۵۰۰ تکرار اول، در هر ۱۰۰ تکرار، یک نمونه برداشته شده است.

چنان‌که ذکر شد، توقع داریم مدل پیشنهادی عملکرد بهتری داشته باشد زیرا قادر است اثر تنکی را تخفیف دهد. به عنوان مثال، تعداد کلمات منفرد در مجموعه داده allRemoved، ۸۳۲۱ است. تعداد عناصر غیرصفر ماتریس φ برای پنجره‌ای به طول ۲۰ برابر ۵۴۸۱۲۶ است که کمتر از یک درصد حالت‌های ممکن است. این در حالی است که تعداد 2-gramهای موجود در مجموعه مذکور ۱۳۸۹۶۴ است که کسر خیلی (تقریبا چهار برابر) کوچکتری از حالت‌های ممکن است. پس ماتریس مربوطه در روش ما، همچنان تنک است ولی اثر تنکی کاهش یافته است. البته باید توجه داشت که کاهش تنکی صرفا به معنی افزایش تعداد عناصر غیرصفر نیست، بلکه به طور کلی فراوانی رخدادها افزایش می‌یابد و باعث می‌شود مدارک محکمتری برای ارتباط کلمات داشته باشیم. از آنجا که ماتریس مربوطه همچنان تنک است، از نظر فضای مصرفی هم دچار مشکل نخواهیم شد. در جدول ۴-۱، می‌توان مشاهده نمود که روش پیشنهادی با هر چهار اندازه پنجره بهتر از مدل n-gram عمل نموده است یعنی مقدار پرپلکسیتی برای روش پیشنهادی کمتر است که نشان می‌دهد روش پیشنهادی هر دو نوع داده را بهتر مدل کرده است. روش‌های مذکور همچنین روی مجموعه داده Brown مقایسه شده‌اند که نتایج مربوطه، در جدول ۴-۲ گزارش شده‌اند. در این جدل نیز مقدار پرپلکسیتی برای روش پیشنهادی کمتر است که نشان می‌دهد روش پیشنهادی داده‌ها را بهتر مدل کرده است.

جدول ۴-۱- مقادیر perplexity برای hها و طول پنجره‌های مختلف روی دو مجموعه داده 20 newsgroups

Models	Substituted	All-removed
2-gram	40.99	3058.85
3-gram	41.35	2827.91
4-gram	42.69	2819.34
5-gram	44.10	2824.92
Proposed-model-L2	11.27	336.38
Proposed-model-L5	11.00	155.10
Proposed-model-L10	11.07	116.44
Proposed-model-L20	11.04	102.53

جدول ۴-۲- مقادیر perplexity برای hها و طول پنجره‌های مختلف روی مجموعه داده Brown

Models	Brown
2-gram	418.18
3-gram	419.99
4-gram	420.88
5-gram	420.90
Proposed-model-L2	379.54
Proposed-model-L5	323.84
Proposed-model-L10	323.27
Proposed-model-L20	318.69

در جدول ۴-۱ می‌بینیم که مقدار پرپلکسیتی مدل‌های n-gram بر روی مجموعه داده substituted، به مراتب کمتر از مجموعه داده all-removed است. این مشاهده، نشان می‌دهد، همان‌طور که توقع داشتیم، اگر در مجموعه‌ای، تعداد دنباله‌های تکراری کمتر و در نتیجه مشکل تنکی برجسته‌تر باشد، پرپلکسیتی مدل بیشتر خواهد بود. در مورد مدل پیشنهادی هم، همین نکته تکرار می‌گردد. در مقایسه دو مدل نیز می‌بینیم که مدل پیشنهادی چنان‌که توقع داشتیم بر مشکل تنکی غلبه کرده و بر روی هر دو مجموعه، پرپلکسیتی را به طور چشمگیری کاهش داده است. این کاهش بر روی مجموعه all-removed چشمگیرتر است زیرا در این مجموعه، مشکل تنکی برجسته‌تر است.

۴-۱-۴- جمع‌بندی

در این بخش، یک مدل زبانی بر مبنای مدل‌های سلسله‌مراتبی بیز معرفی شده است. ایده پایه این مدل این است که در هنگام نگارش یک متن، کلمه جاری بر مبنای یکی از کلماتی که پیش از آن نوشته شده است شکل می‌گیرد یا به عبارتی یکی از کلمات پیشین است که محرک انتخاب کلمه جاری در ذهن نگارنده است. در این مدل چون کلمات قبل از کلمه جاری مورد توجه قرار می‌گیرند ترتیب تا حدی دارای اهمیت است، اما ترتیب دقیق کلمات دخالت مستقیم در مدل ندارد. به همین دلیل مشکل تنکی برای این مدل خفیف‌تر از مدل‌های n-gram است.

مدل پیشنهادی بر روی دو مجموعه داده متفاوت که هر دو از یک زیر مجموعه ۴۰۰۰ سندی از مجموعه 20newsgroups گرفته شده‌اند با مدل Knesser-Ney که معمولا در بین مدل‌های n-gram عملکرد بهتری از نظر پرپلکسیتی دارد مقایسه شده‌اند. دو مجموعه داده فقط از نظر پیش‌پردازش‌های انجام شده بر روی آنها متفاوتند. در یکی از آنها تمام ایست‌واژه‌ها، ارقام و علائم حذف شده‌اند و در دیگری هر دسته با علائم یکسان جایگزین شده است. روشن است که در حالت اول دنباله‌های تکراری بسیار تنک هستند

ولی در دومی دنباله‌های تکراری بیشتری خواهیم داشت. توانایی مدل پیشنهادی در مدل کردن داده‌های هردو مجموعه با استفاده از معیار پرپلکسیتی سنجیده شده است.

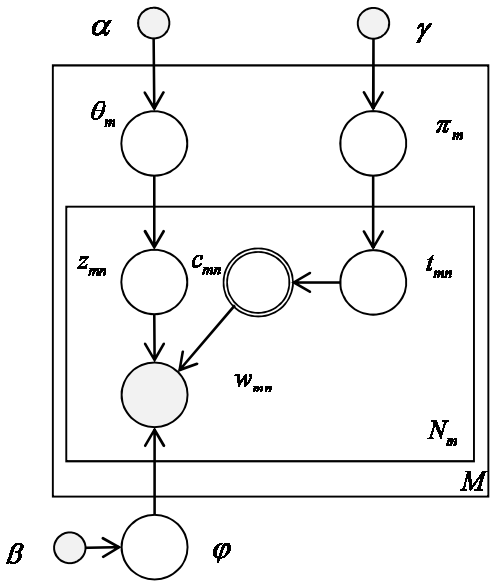
مدل پیشنهادی با طول پنجره‌های متفاوت با مدل زبانی Knesser-Ney بهبودیافته [۷۷] با طول‌های مختلف مقایسه شده است و در تمام موارد عملکرد مدل پیشنهادی به طور چشمگیری بهتر است. تفاوت مدل پیشنهادی با مدل n-gram در مجموعه اول چشمگیرتر به نظر می‌رسد. این نتیجه منطبق با منطق مدل پیشنهادی است یعنی اثر تنکی تخفیف یافته است. در برخی کاربردها مثل بازیابی اطلاعات، در نظر گرفتن ارتباطات محلی کلمات می‌تواند بسیار مفید باشد اما برای مدل کردن این نوع ارتباطات، لزومی به در نظر گرفتن ترتیب دقیق کلمات نیست. مدل پیشنهادی می‌تواند در چنین کاربردهایی کارآمد باشد. در این مقاله نشان داده‌ایم که مدل پیشنهادی اثر تنکی را کاهش داده است و در مدل کردن داده‌ها بهتر از n-gram عمل کرده است، در مراحل بعدی کار قصد داریم که عملکرد مدل پیشنهادی را در کاربردهای مذکور بررسی نماییم.

۴-۲- مدل موضوعی مبتنی بر پنجره همرخدادی

در بخش قبل یک مدل زبانی را توصیف نمودیم که قادر است ارتباطات محلی کلمات را بدون وابستگی به ترتیب دقیق کلمات، مدل نماید. حال این مدل را در کنار موضوعات، در یک مدل موضوعی، به کار می‌گیریم. به این ترتیب هر کلمه علاوه بر موضوعی که بدان انتساب یافته، از کلمات اطرافش نیز، مستقیما تاثیر خواهد پذیرفت. در این مدل موضوعی، هر موضوع، مجموعه‌ای از توزیع‌ها، و نه یک توزیع منفرد، بر روی کلمات است و هر یک از این توزیع‌ها متناظر با یک کلمه مولد از مجموعه کلمات منفرد است. بنابراین در این مدل، برای تولید کلمات متن، علاوه بر موضوع هر کلمه، باید کلمه مولد آن را نیز بدانیم تا مشخص شود که کلمه مربوطه از کدامیک از توزیع‌های کلمات باید انتخاب شود. در ادامه به توصیف مفصل‌تر این مدل می‌پردازیم.

۴-۲-۱- توصیف مدل

شکل ۴-۴ نمایش گرافی مربوط به این ترکیب را نشان می‌دهد. همانطور که در این شکل مشخص است، برای تولید هر کلمه، مدل باید تصمیم بگیرد که آن کلمه از کدام موضوع و همچنین کدامیک از کلمات اطرافش گرفته شده است. منظور از کلمات اطراف، کلماتی است که در پنجره متناظر با آن کلمه قرار گرفته‌اند. چنان‌که پیش از این ذکر شد، پنجره مذکور، قبل از کلمه جاری واقع شده است به شکلی که کلمه جاری آخرین کلمه آن است.



شکل ۴-۴- نمایش گرافی مدل موضوعی مبتنی بر پنجره همرخدادی

فرایند مولد مدل پیشنهادی به شرح زیر است:

- برای هر سند d_m در مجموعه‌داده D :
  - توزیع $\pi_m \sim Dirichlet(\gamma)$ انتخاب کن.

- توزیع $\theta_m \sim Dirichlet(\alpha)$ انتخاب کن.

- برای هر کلمه w_i در مجموعه داده:

  - یکی از K موضوع ممکن را از توزیع $z_i \sim Multinomial(\theta_{d_i})$ انتخاب کن.

  - یکی از L موقعیت موجود در پنجره متناظرش s_i را از توزیع $t_i \sim Multinomial(\pi_{d_i})$ انتخاب کن.

  - کلمه موجود در موقعیت t_i یعنی c_i را مشخص کن.

  - کلمه جاری w_i را از توزیع مربوط به $w_i \sim Multinomial(\varphi^{z_i,c_i})$ انتخاب کن.

فرایند مولد فوق چگونگی تولید یک سند را از نگاه مدل پیشنهادی، با فرض دانستن پارامترهای مدل، نشان می‌دهد. همانطور که در این فرایند می‌بینیم، برای تولید هر کلمه از متن، یک موضوع و یکی از کلمات قبلیش به طور تصادفی انتخاب می‌شود. کلمه انتخاب شده را همانند قبل کلمه مولد می‌خوانیم. کلمه هدف از توزیع متناظر با کلمه مولد در مجموعه توزیع‌های موضوع انتخابی، استخراج می‌گردد. برای محاسبه پارامترهای مدل، فرایند مذکور باید معکوس گردد یعنی پارامترهای مدل را به گونه‌ای می‌یابیم که احتمال اسناد برمبنای آنها حداکثر شود یا به بیان دیگر در فرایند تولید هر یک از کلمات متن، منجر به انتخاب بهترین زوج موضوع و کلمه مولد از بین زوج‌های ممکن گردد.

۴-۲-۲- تخمین پارامترهای مدل

اگر بخواهیم روند نمونه‌برداری گیبس را برای این مدل انجام دهیم به همان شکل مدل قبل، باید احتمال توام زیر را بدست آوریم:

$$p\left(t_i = l, \mid \boldsymbol{t}_{-i}, \boldsymbol{c}, z, \boldsymbol{w}\right) = \frac{p\left(t_i = l, c_i = v, z_i = k, w_i \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, z_{-i},, \boldsymbol{w}_{-i}\right)}{p\left(w_i, c_i = v, z_i = k \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, z_{-i}, \boldsymbol{w}_{-i}\right)} \quad (۱۷-۴)$$

مخرج ثابت و قابل حذف است بنابراین داریم:

$$\begin{aligned} &\propto p\left(t_i = l, c_i = v, z_i = k, w_i \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, z_{-i},, \boldsymbol{w}_{-i}\right) \\ &= p\left(t_i = l \mid \boldsymbol{t}_{-i}\right) p\left(c_i = v \mid t_i = l\right) p\left(z_i = k \mid z_{-i}\right) p\left(w_i \mid c_i = v, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{v}_{-i}, \boldsymbol{w}_{-i}\right) \end{aligned} \quad (۱۸-۴)$$

که فاکتورهای سوم و چهارم ثابت و بنابراین ساده‌شدنی هستند چون مقادیر w_i و $c_i = v$ تغییر نمی‌کنند. فاکتور اول مانند قبل قابل محاسبه است یعنی داریم:

$$p\left(t_i = l \mid \boldsymbol{t}_{-i}\right) \propto \frac{n_{-i,l}^{d_i} + \gamma}{n_{-i,\cdot}^{d_i} + L\gamma} \quad (۱۹-۴)$$

فاکتور دوم را مثل قبل ثابت می‌گیریم. حال با فرض مقادیر $\boldsymbol{t}_i$ و $\boldsymbol{c}_i$ احتمال $p\left(z_i = l, \mid \boldsymbol{t}, \boldsymbol{c}, z_{-i}, \boldsymbol{w}\right)$ را محاسبه می‌نماییم که بازهم متناسب است با $p\left(t_i = l, c_i = v, z_i = k, w_i \mid \boldsymbol{t}_{-i}, \boldsymbol{c}_{-i}, z_{-i},, \boldsymbol{w}_{-i}\right)$ و در حاصل‌ضرب مربوطه، این بار فاکتور اول و دوم قابل حذفند. فاکتور سوم به صورت زیر قابل محاسبه است.

$$p\left(z_i \mid \boldsymbol{z}_{-i}\right) = \int p\left(z_i, \theta_{d_i} \mid \boldsymbol{z}_{-i}\right) d\theta_{d_i} = \int p\left(\theta_{d_i} \mid \boldsymbol{z}_{-i}\right) p\left(z_i \mid \theta_{d_i}\right) d\theta_{d_i} \quad (۲۰-۴)$$

برای زیر انتگرال داریم:

$$p\left(\theta_{d_i} \mid \boldsymbol{z}_{-i}\right) \propto p\left(\theta_{d_i}\right) p\left(\boldsymbol{z}_{-i} \mid \theta_{d_i}\right) \quad (۲۱-۴)$$

از آنجا که $p\left(\theta_{d_i}\right)$ از $Dir(\alpha)$ پیروی می‌کند و مزدوج $p\left(z_{-i} \mid \theta_{d_i}\right)$ است بنابراین احتمال پسین $p\left(\theta_{d_i} \mid \boldsymbol{z}_{-i}\right)$ دارای توزیع $Dir\left(\alpha + n_{-i,k}^{d_i}\right)$ خواهد بود که در آن $n_{-i,k}^{d_i}$ نماینده تعداد دفعاتی است که k در سند d_i به کلمه‌ای انتساب یافته است بدون در نظر گرفتن موقعیت جاری i. به این ترتیب زیر

انتگرال به فرم $\int \theta p(\theta) d\theta$ در می‌آید که برابر $E[\theta]$ است در حالی کـه θ از توزیـع $Dir\left(\theta + n_{-i,k}^{d_i}\right)$ پیروی می‌کند و میانگین توزیع دریکله مذکور برابر است با $\frac{n_{-i,k}^{d_i} + \alpha}{n_{-i,\cdot}^{d_i} + K\alpha}$ که در آن $n_{-i,\cdot}^{d_i}$ برابر است با تعداد کلمات سند d_i. یعنی:

$$p\left(z_i = k \mid \boldsymbol{z}_{-i}\right) \propto \frac{n_{-i,k}^{d_i} + \alpha}{n_{-i,\cdot}^{d_i} + K\alpha} \tag{۲۲-۴}$$

فاکتور چهارم به صورت زیر قابل محاسبه است:

$$\begin{aligned} p\left(w_i \mid c_i = v, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}\right) &= \int p\left(w_i, \varphi \mid c_i = v, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}\right) d\varphi \\ &= \int p\left(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}\right) p\left(w_i \mid c_i = v, z_i = k, \varphi\right) d\varphi \end{aligned} \tag{۲۳-۴}$$

برای زیر انتگرال داریم:

$$p\left(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}\right) \propto p(\varphi)\, p\left(\boldsymbol{w}_{-i} \mid \boldsymbol{c}_{-i}, \boldsymbol{z}_{-i}, \varphi\right) \tag{۲۴-۴}$$

کـه مشـابه قبـل مـی‌تـوان از آن نتیجـه گرفـت کـه احتمـال پسـین $p\left(\varphi \mid \boldsymbol{c}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}\right)$ از توزیـع $Dir\left(n_{-i,w_i}^{k,v} + \beta\right)$ پیروی می‌کند. بنابراین حاصل انتگرال برابر است با $\frac{n_{-i,w_i}^{k,v} + \beta}{n_{-i,\cdot}^{k,v} + |V|\beta}$ یعنی:

$$p\left(w_i \mid c_i = v, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}\right) \propto \frac{n_{-i,w_i}^{k,v} + \beta}{n_{-i,\cdot}^{k,v} + |V|\beta} \tag{۲۵-۴}$$

فرض مدل این است که t_i موقعیتی از یک پنجره بر روی کلمه w_i را نشان می‌دهد و c_i کلمه موجود در آن موقعیت است. بنابراین با وجود اینکه هر دو متغیر تصادفی هستند، در هر آزمایش و وابسته بـه همـان

آزمایش مقدار یکی با توجه به مقدار دیگری قابل محاسبه است. به همین دلیل، در هر مرحله از الگـوریتم گیبس، ابتدا مقدار t_i را با توجه به $p(t_i = l \mid \boldsymbol{t}_{-i}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{w})$ و سپسس مقدار c_i را بر اساس مقدار t_i بدست می‌آوریم. در نهایت پس از پایان تکرارها خواهیم داشت:

$$\pi_{d_i,l} = \frac{n_l^{d_i} + \gamma}{n_\cdot^{d_i} + L\gamma} \tag{۲۶-۴}$$

$$\theta_{d_i,k} \propto \frac{n_k^{d_i} + \alpha}{n_\cdot^{d_i} + K\alpha} \tag{۲۷-۴}$$

$$\varphi_{w_i}^{k,v} \propto \frac{n_{w_i}^{k,v} + \beta}{n_\cdot^{k,v} + |V|\beta} \tag{۲۸-۴}$$

بر اساس روابط بدست آمده و توضیحات داده شده در نهایت می‌توان الگوریتم نمونه‌برداری گیبس را برای مدل، به صورتی که در شکل ۴-۵ آمده است، توصیف کرد. چنان‌که در این شکل مشاهده می‌کنیم، پنجره‌ی دکرشده در توصیف مدل در نمایش گرافیکی ظهور نمی‌یابد ولی در محاسبات، فضای حالات را محدود می‌کند. در روابط فوق β و γ پارامترهای توزیع‌های دریکله یا همان فراپارامترهای مدل هستند. در این مقاله چنان‌که در مدل‌های مشابه مرسوم است، این پارامترها ثابت و متقارن در نظر گرفته شده‌اند. در مدل پیشنهادی می‌خواهیم که تعداد اندکی از کلمات متن در تولید هر سند موثر باشند و همچنین می‌خواهیم که ارتباط هر کلمه با تعداد اندکی از کلمات دیگر مجموعه قوی و با بیشتر آنها ضعیف باشد. بنابراین می‌خواهیم که ماتریسهای φ و θ هردو تنک باشند. این نکته را با انتخاب مقادیر کوچک برای فراپارامترهای مدل، کنترل می‌کنیم.

۱- ورودی‌ها: تعداد موضوعات K، اندازه پنجره L، مقادیر فراپارامترهای α، β و γ، حداکثر تعداد تکرار

$maxIter$.

۲- به به صورت تصادفی هر کلمه متن را به یکی از K موضوع ممکن منتسب کن.

۳- به صورت تصادفی هر کلمه متن را به یکی از کلمات قبلیش که در محدوده پنجره‌ای به طول L واقع شده است، منتسب کن.

۴- مقادیر اولیه شمارنده $n_k^{d_i}$ را که نماینده تعداد کلماتی است که در سند d_i به موضوع k انتساب یافته‌اند، مشخص کن.

۵- مقادیر اولیه شمارنده $n_l^{d_i}$ را که نماینده تعداد کلماتی است که در سند d_i به موقعیت l از پنجره متناظرشان انتساب یافته‌اند، مشخص کن.

۶- مقادیر اولیه شمارنده $n_{\cdot}^{d_i}$ را که نماینده تعداد کلمات در سند d_i است، مشخص کن.

۷- مقادیر اولیه شمارنده $n_{w_i}^{k,v}$ را که نماینده تعداد کلماتی مانند w_i است که در سرتاسر مجموعه داده به زوج موضوع وکلمه مولد (k,v) انتساب یافته‌اند، مشخص کن.

۸- مقادیر اولیه شمارنده $n_{\cdot}^{k,v}$ را که نماینده تعداد کل کلماتی است که سرتاسر مجموعه داده به زوج موضوع وکلمه مولد (k,v) انتساب یافته‌اند، مشخص کن.

۹- برای تعداد تکرار ۱ تا $maxIter$

۱۰- برای هر سند d_i

۱۱- برای هر موقعیت i در سند d_i

۱۲- آمار موقعیت جاری را از شمارنده‌ها حذف کن.

۱۳- یکی از موضوعات را با توجه به توزیع $p(z_i \mid z_{-i})$ در رابطه ۶ انتخاب کن.

۱۴- کلمات موجود در پنجره‌ای در موقعیت $i-1$ تا $i-L$ را مشخص کن.

۱۵- یکی از آنها را با توجه به توزیع $p(t_{mn} \mid t_{-mn})$ در رابطه ۶ انتخاب کن.

۱۶- موقعیت و کلمه جدید را به موقعیت جاری (کلمه w_{mn}) منسوب کن.

۱۷- شمارنده‌ها را با توجه به موضوع، موقعیت و کلمه انتسابی جدید به‌روز کن.

۱۸- در پایان تکرارها مقادیر پارامترهای مدل یعنی π و φ را بر اساس روابط (۴-۱۲) و (۴-۱۳) محاسبه کن.

شکل ۴-۵- الگوریتم نمونه‌برداری گیبس مربوط به محاسبه پارامترهای مدل

۴-۲-۳- آزمایشات و نتایج

معمولا مدل‌های موضوعی را با استفاده از دو نوع معیار ارزیابی مقایسه می‌نمایند: معیارهای ماهوی[۹] و معیارهای خارجی[۱۰]. معیارهای ماهوی مدل را مستقیما و بر اساس ویژگی‌های ماهوی آن ارزیابی می‌نمایند در حالی که معیارهای خارجی عملکرد آن را در یک کاربرد می‌سنجند. بنابراین آزمایشات ما عبارتند از نمایش همگرایی مدل‌ها، مقایسه انسجام آنها و قابلیت آنها در خوشه‌بندی اسناد. برای آزمایشات مذکور به ترتیب از معیارهای پرپلکسیتی، انسجام[۱۱] و فاصله تغییرات اطلاعات[۱۲] (VI) استفاده خواهیم نمود. پرپلکسیتی و انسجام معیارهای ماهوی بوده و فاصله تغییرات اطلاعات در اینجا یک معیار خارجی محسوب می‌گردد چون با استفاده از آن عملکرد مدل را در کاربرد خوشه‌بندی اسناد می‌سنجیم. در ادامه این بخش قبل از پرداختن به نتایج آزمایشات، ابتدا به معرفی مجموعه‌های داده و پیش‌پردازش‌های انجام شده بر روی آنها و سپس به معرفی معیارهای ارزیابی می‌پردازیم.

۴-۲-۳-۱- مجموعه‌های داده

آزمایشات این بخش روی مجموعه داده‌های 20 newsgroups و Reuters-R8 [۹۵, ۹۶] انجام شده‌اند که در این مقاله آنها را به ترتیب 20NG و R8 خواهیم خواند. مجموعه 20NG شامل ۲۰ گروه از پیش تعیین شده می‌باشد که هرکدام یک گروه خبری محسوب شده و به همین دلیل می‌توان هر یک از آنها را یک خوشه از اسناد دانست که توسط کاربر انسانی تشکیل شده است. ایست‌واژه‌ها، علائم و اعداد در پیش‌پردازش حذف شده‌اند. آدرس‌های وب و ایمیل با @ جایگزین و سپس حذف شده‌اند. همچنین کلمات نادر که در کمتر از پنج سند در کل مجموعه تکرار شده‌اند نیز حذف شده‌اند. مجموعه داده دیگر چنان‌که ذکر شد، مجموعه R8 است. پیش‌پردازش‌های انجام شده مانند همان مجموعه 20NG است. این

[9] - Intrinsic
[10] - Extrinsic
[11] - Coherence
[12] - Variation of Information (VI) distance

مجموعه شامل هشت دسته است و هر سند، تنها در یک دسته قرار دارد. در هردو مجموعه از تقسیم آموزش و تست پیش‌فرض خود مجموعه استفاده شده و تغییری در نسبت آنها داده نشده است.

۴-۲-۳-۲- پرپلکسیتی

معیار پرپلکسیتی در بخش قبل معرفی شده است. البته ذکر این نکته در اینجا ضروری است که برخلاف بخش قبل، برای لگاریتم گرفتن در این بخش از پایه e استفاده می‌کنیم، چنان‌که درسال‌های اخیر در ارزیابی مدل‌های موضوعی مرسوم بوده است [۹۷]. به این ترتیب رابطه آن به صورت زیر خواهد بود.

$$perplexity(D_{test}) = exp(-\frac{\log p(w_{test} | \mathcal{M})}{N_{test}}) \quad (۲۹-۴)$$

چنان‌که ذکر شد، برای محاسبه پارامترهای مدل سعی می‌کنیم مقادیر آنها را به‌گونه‌ای تعیین کنیم که احتمال بیشتری به کلمات مجموعه اختصاص یابد. بر این اساس معمولا یکی از روش‌هایی که برای ارزیابی مدل‌های موضوعی مورد استفاده قرار می‌گیرد معیار پرپلکسیتی است. با این حال در سال‌های اخیر ثابت شده است که پرپلکسیتی معیار چندان قابل اعتمادی برای مقایسه و سنجش مدل‌های موضوعی نیست [۹۸].

۴-۲-۳-۳- انسجام موضوعات

از همان ابتدای معرفی مدل‌های موضوعی، یکی از راه‌های ارزیابی این مدل‌ها نمایش موضوعات تولید شده و واگذاری قضاوت در باره کیفیت و انسجام آنها توسط خوانندگان بوده است .ولی تا سال‌های اخیر معیاری برای سنجش کمی انسجام موضوعات وجود نداشته است و ارزیابی ماهوی و کمی مدل‌های موضوعی محدود به سنجش توانایی آنها در مدل کردن داده‌های نادیده[۱۳] با استفاده از معیارهایی مانند پرپلکسیتی

[13] -Unseen

[۱۸] و درستنمایی [۹۸] بوده است. از آنجا که یکی از اهداف معرفی این مدل‌ها استخراج موضوعاتی بوده که از دید انسان معنی‌دار باشند، بسیار مهم است که قضاوت‌های معیار مورد استفاده منطبق بر توانایی مدل در استخراج موضوعات معنی‌دار باشد. بر مبنای برخی تحقیقاتی [۹۹, ۱۰۰] که در سال‌های اخیر صورت گرفته‌اند، اصل مذکور در مورد معیار پرپلکسیتی صادق نیست و در بسیاری موارد مدلی که پرپلکسیتی کمتر(بهتر)ی دارد موضوعات کم‌کیفیت‌تری تولید می‌نماید [۹۹]. بر این مبنا، مدل‌هایی [۱۰۱-۱۰۳] معرفی شده‌اند که مستقیما کیفیت یا به عبارتی انسجام و معنی‌داری موضوعات را می‌سنجند. این معیارها عموما مبتنی بر این ذهنیت هستند که موضوعاتی که کلمات شاخص آنها به فراوانی، به صورت همرخداد با یکدیگر در متون ظاهر شده باشند، با احتمال زیاد، از دید انسان، منسجم و معنی‌دار هستند. در این پایان‌نامه از معیاری [۱۰۳] استفاده شده است که مبتنی بر همرخدادی کلمات شاخص در مجموعه داده مورد استفاده در آموزش مدل است و به همین دلیل که نیاز به دانشی خارج از مجموعه داده آموزش ندارد، به سادگی قابل محاسبه است. همچنین نتایج این معیار منطبق بر نظرات انسان هستند. این معیار توسط رابطه زیر قابل محاسبه است:

$$C(k,T^k)=\sum_{i=2}^{I}\sum_{j=1}^{i-1}log(\frac{N(t_i^k,t_j^k)+1}{N(t_j^k)}) \qquad (۳۰-۴)$$

در این رابطه $C(k,T^k)$ نماینده انسجام موضوع k است که با لیست مرتب T^k از I کلمه شاخصش نمایش داده شده است و $T^k=(t_1^k,t_2^k,...,t_I^k)$. $N(t_i^k,t_j^k)$ تعداد سندهایی است که هردو کلمه شاخص t_i^k و t_j^k در آن‌ها رخ داده‌اند. $N(t_j^k)$ تعداد سندهایی است که حاوی کلمه شاخص t_j^k هستند.

۴-۲-۳-۴- فاصله تغییرات اطلاعات

توقع داریم که مدل پیشنهادی بتواند موضوعاتی تولید نماید که از نظر انسان معنی‌دار و منسجم باشند. بنابراین عجیب نیست که توقع داشته باشیم خوشه‌های اسناد تولید شده مبتنی بر این موضوعات نیز

منطبق بر نظر انسان یا به عبارتی شبیه به دسته‌بندی انجام شده توسط انسان باشند. بر این مبنا، یکی از روش‌هایی که برای ارزیابی مدل پیشنهادی استفاده می‌نماییم مقایسه خوشه‌بندی تولید شده توسط این مدل با دسته‌بندی انجام شده توسط انسان است. بدین منظور، از فاصله تغییرات اطلاعات [۱۰۴, ۱۰۵] استفاده می‌نماییم که قادر است فاصله خوشه‌بندی انجام شده توسط مدل پیشنهادی را با دسته‌بندی انجام شده توسط انسان بر روی مجموعه داده مورد استفاده محاسبه نماید. خوشه‌بندی تولید شده توسط مدل پیشنهادی که همان نسبت شرکت موضوعات در اسناد (توزیع θ) است، یک خوشه‌بندی نرم[۱۴] محسوب می‌گردد. این درحالی است که دسته‌بندی اسناد یک خوشه‌بندی سخت[۱۵] است زیرا در مجموعه داده‌های مورد استفاده، هر سند تنها در یک دسته قرار دارد. از طرفی تعداد دسته‌ها ثابت است ولی می‌خواهیم که در انتخاب تعداد موضوعات که تعداد خوشه‌ها را در مدل پیشنهادی معین می‌کند، آزاد باشیم بنابراین باید از معیاری برای محاسبه فاصله استفاده نماییم که مستقل از تفاوت در نوع خوشه‌بندی و همچنین تفاوت در تعداد خوشه‌ها باشد. فاصله تغییرات اطلاعات دارای این ویژگی است [۹۷]. این فاصله را می‌توان با استفاده از رابطه (۴-۳۱) بدست آورد که در آن $D_{VI}(C,Z)$ فاصله دو خوشه‌بندی C و Z است که در مدل پیشنهادی به ترتیب دسته‌بندی اسناد و خوشه‌بندی تولید شده توسط مدل هستند. H نماینده تابع انتروپی و واگرایی کولبک-لیبلر[۱۶] بین دو توزیع C و Z است که توسط رابطه (۴-۳۲) محاسبه می‌گردد.

$$D_{VI}(C,Z) = H(C) + H(Z) - 2I(C,Z) \quad (۴-۳۱)$$

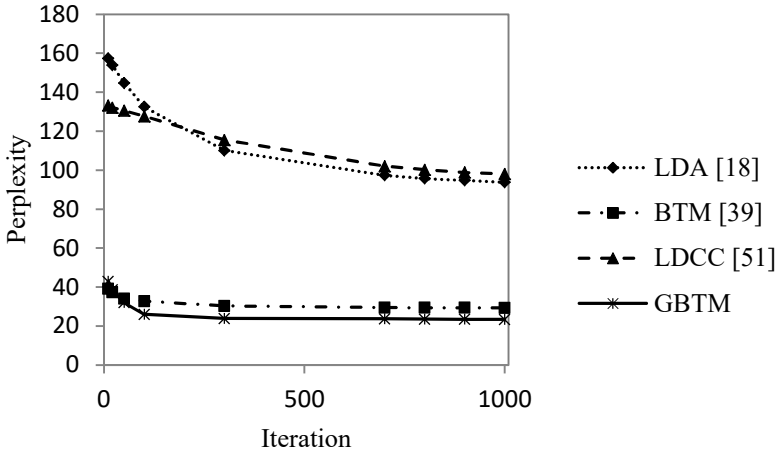
$$I(C,Z) = D_{KL}\{ p(c,z) || p(c)p(z)\} \quad (۴-۳۲)$$

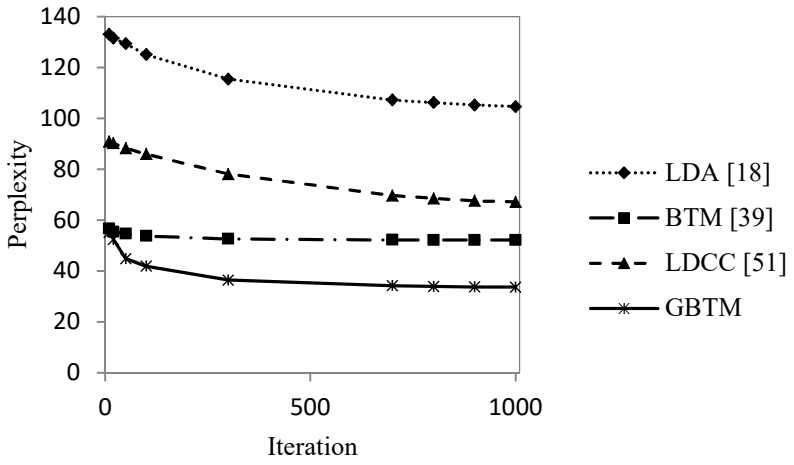
[14] - Soft clustering
[15] - Hard clustering
[16] - Kullback-Leibler (KL) divergence

۴-۲-۳-۵- بررسی نتایج آزمایشات

شکل ۴-۶ پرپلکسیتی مدل‌ها را به عنوان تابعی از تکرارهای نمونه‌برداری گیبس بر اساس یک زنجیره منفرد مارکوف که در ۱۰۰۰ تکرار اجرا شده است، نشان می‌دهد. مقادیر مذکور برای هر دو مجموعه 20NG و R8 نشان داه شده‌اند.

تمامی مدل‌ها در تکرار ۱۰۰۰ام به ثبات رسیده‌اند. مدل پیشنهادی و مدل‌های LDA و BTM زودتر از LDCC همگرا می‌شوند. کمترین پرپلکسیتی متعلق به مدل پیشنهادی است.





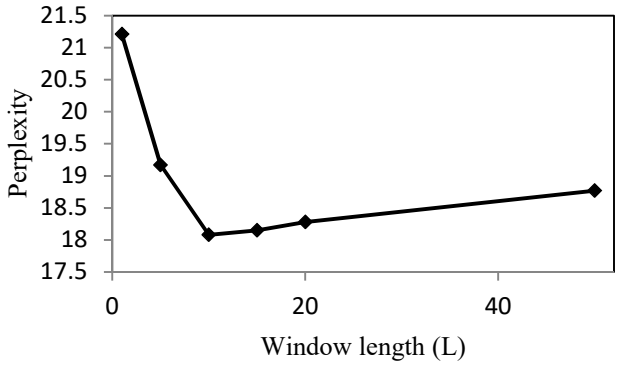
شکل ۴-۶- تغییرات پرپلکسیتی به عنوان تابعی از تکرارها در الگوریتم نمونه‌برداری گیبس برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

یکی از اهداف مدل پیشنهادی کاهش تنکی است. همانطور که پیش از این توضیح داده شد، در مدل پیشنهادی هر کلمه علاوه بر یک موضوع، منتسب به یک کلمه نیز هست و این کلمه از بین کلماتی که در یک پنجره قبل از آن قرار گرفته‌اند انتخاب می‌گردد. اندازه این پنجره می‌تواند مقادیر مختلفی را بپذیرد. در واقع امکان تولید کلمه جاری توسط یکی از کلمات موجود در یک پنجره همرخدادی و نه دقیقا کلمه قبل از آن، نکته‌ای است که توقع داریم باعث تخفیف تنکی شود.

جدول ۴-۳ تعداد عناصر غیرصفر در ماتریس φ را برای پنجره‌های همرخدادی با طول‌های مختلف نشان می‌دهد. وقتی اندازه پنجره، یک است مدل پیشنهادی معادل BTM است. در این جدول می‌توان دید که با افزایش طول پنجره تعداد عناصر غیرصفر افزایش می‌یابد درحالی که تعداد حالات ممکن ثابت است. به عبارت دیگر با افزایش طول پنجره تنکی کاهش می‌یابد. شیب این کاهش از طول ۱۰ به ۲۰ بسیار کم است. از این نکته می‌توان برای تعیین اندازه پنجره مناسب نیز استفاده نمود. پنجره همرخدادی در مدل پیشنهادی می‌تواند مقادیر مختلفی را بپذیرد. بر اساس مشاهدات ما پنجره‌ای به طول ۱۰ مناسب به نظر می‌رسد. این نکته در جدول ۴-۳ و همچنین شکل ۴-۷ تایید شده است.

جدول ۴-۳- تعداد عناصر غیرصفر در ماتریس φ که با افزایش اندازه پنجره افزایش می‌یابد

تعداد عناصر غیرصفر در ماتریس φ	اندازه پنجره همرخدادی
1567457	$L=1$
1950822	$L=5$
2446655	$L=10$
2524235	$L=20$



شکل ۴-۷- تغییرات پرپلکسیتی به عنوان تابعی از اندازه پنجره همرخدادی

در شکل ۴-۷ می‌بینیم که مدل پیشنهادی کمترین مقدار پرپلکسیتی را با پنجره‌ای به طول ۱۰ داشته است. هر چند هر دوی این نتایج بر روی یک زیرمجموعه ۴۰۰۰ سندی از مجموعه داده 20NG بدست آمده‌اند، در سایر آزمایشات نیز استفاده از طول ۱۰ نتایج قابل قبولی را تولید نموده است.

جدول ۴-۴- تنظیمات مورد استفاده در آزمایشات

Parameter	Value
β	0.01
$\alpha_{GBTM,LDCC}$	$1/(K+L)$
$\alpha_{LDA,BTM}$	$1/K$
Number of iterations	1000
Burn-in	500
lag	100
L	20
P	10

برای آزمایشات این بخش از تنظیمات ذکر شده در جدول ۴-۴ استفاده شده است. تعداد موضوعات در این جدول ذکر نشده‌اند زیرا نتایج آزمایشات برای تعداد موضوعات مختلف گزارش شده‌اند. اندازه فراپارامترها مشابه مقادیری است که به صورت قراردادی در بسیاری از مدل‌های موجود، مورد استفاده قرار گرفته است. این پارامترها نوعی اثر هموارسازی بر روی توزیع‌های چندجمله‌ای دارند. برای اطلاعات بیشتر در زمینه اثر این پارامترها در مدل‌های احتمالاتی دریکله-چندجمله‌ای می‌توان به [۹۷] مراجعه نمود. فراپارامترها در مدل پیشنهادی نیز به طور مشابه تعریف شده‌اند. برای مدل پیشنهادی به دلیل اینکه علاوه بر تعداد موضوعات K طول پنجره L نیز یک پارامتر موثر است، مقدار آن در تعیین اندازه فراپارامترها مورد توجه قرار گرفته است. جدول ۴-۵، تعدادی از موضوعات تولید شده توسط مدل پیشنهادی تحت تنظیمات جدول ۴-۴ روی مجموعه 20NG را نشان می‌دهد. موضوعات نمایش داده‌شده منسجم بوده و از دید ناظر انسانی می‌توان نام خاص به آنها اختصاص داد. به عنوان مثال، موضوع ۴۵ مربوط به کامپیوتر یا موضوع ۲۹ مربوط به فروش و انتقال کالا است.

جدول ۴-۵- چند نمونه از موضوعات GBTM تحت تنظیمات جدول ۴-۴ بر روی مجموعه 20NG

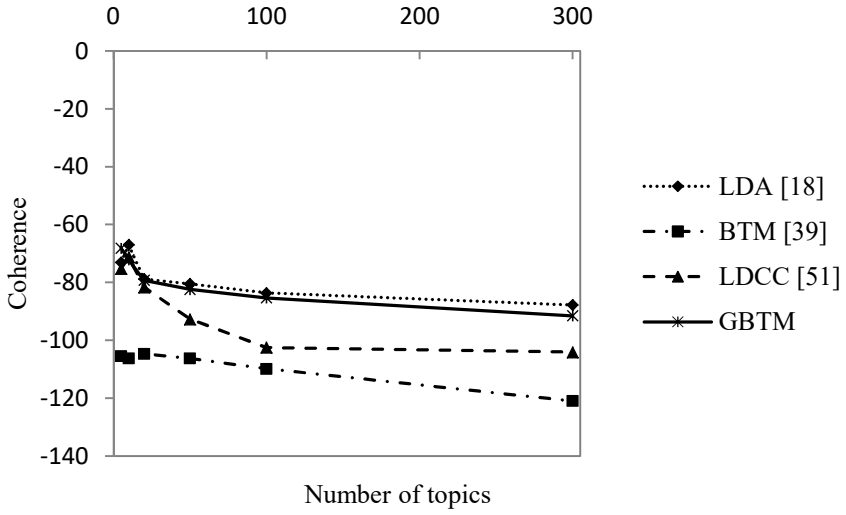
GBTM				
41	45	29	42	1
god christ jesus christian holy spirit father son canada	windows mb program drive time iran system monitor file	offer support shipping lb mile mass section offers team	ftp information system drive space data card time software	players nhl europeans league hockey numbers play european unix

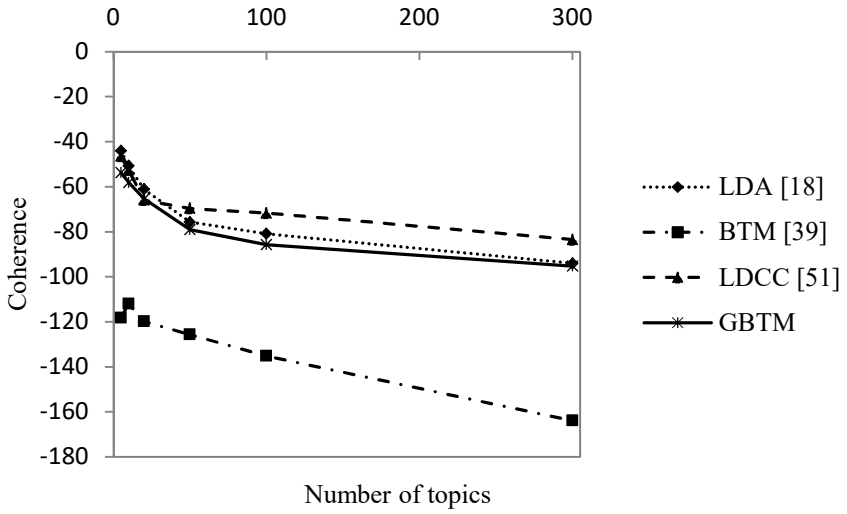
جدول ۴-۶- چند نمونه از موضوعات LDA و BTM تحت تنظیمات جدول ۴-۴ بر روی مجموعه 20NG

LDA		BTM	
24	0	0	8

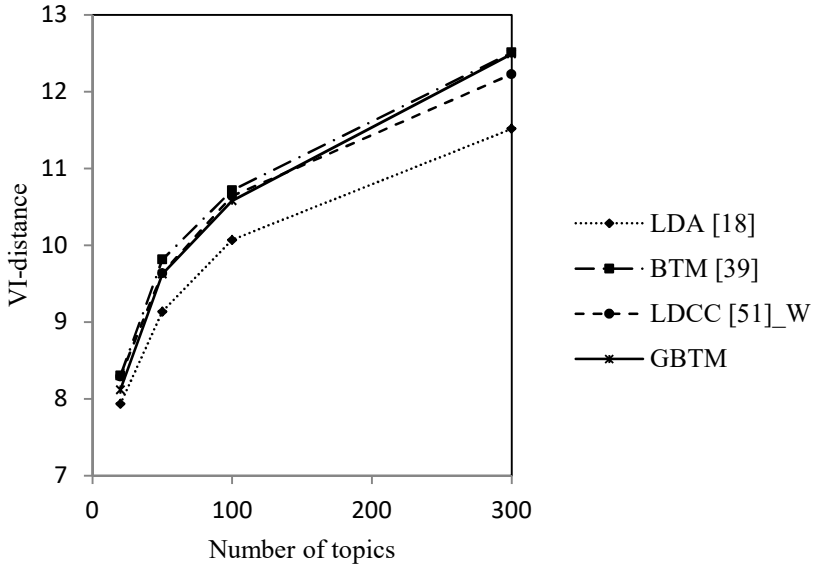
widget window visual event application colormap int set display	sale shipping offer price printer condition nntp-posting-host email interested	number color ago part thing visual test system support	year distribution cost works megs events ide world motto

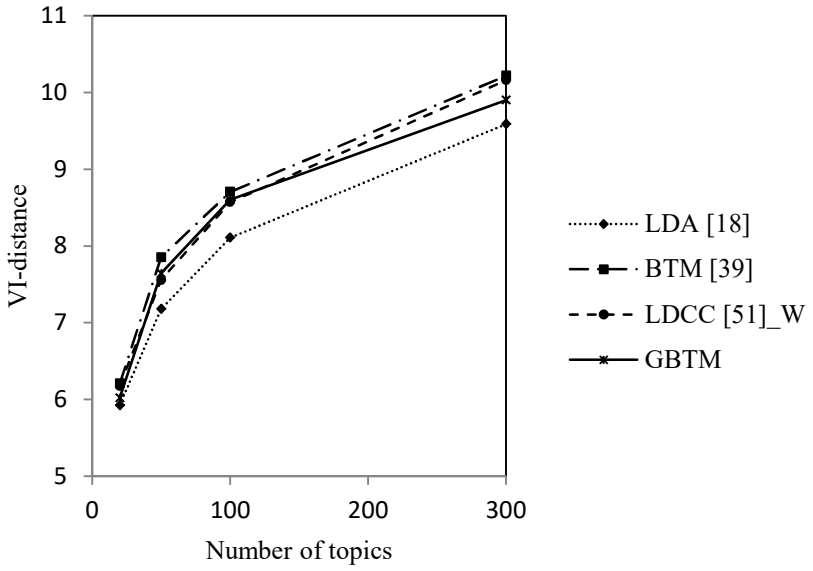
جدول ۴-۶ حاوی دو نمونه از موضوعات متناظر با این موضوعات است که توسط مدل‌های LDA و BTM تولید شده‌اند. نمونه موضوعات نمایش داده شده نشان می‌دهد که هر چند مدل پیشنهادی موضوعات منسجم‌تر و معنی‌دارتری نسبت به BTM تولید نموده است ولی بهبود چشمگیری نسبت به LDA قابل تشخیص نیست. این نکته در ارزیابی‌های کمی نیز تایید می‌گردد.





شکل ۴-۸- مقدار انسجام موضوعات تولید شده توسط مدل پیشنهادی ، LDA، LDCC و BTM برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده در جدول ۴-۴ و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).





شکل ۴-۹- مقدار فاصله خوشه‌های تولید شده توسط مدل پیشنهادی ، LDA و BTM از دسته‌های ساخته شده توسط انسان، برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده جدول ۴-۴ و و برای مجموعه داده‌های و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

شکل ۴-۸ انسجام موضوعات تولید شده توسط مدل‌های هدف را تحت تنظیمات جدول ۴-۴ و بر روی مجموعه‌های 20NG و R8 نشان می‌دهد. در این شکل می‌بینیم که مدل پیشنهادی بسیار بهتر از BTM عمل کرده است و هرچند کمتر ولی انسجام آن بر روی هر دو مجموعه داده نزدیک به LDA است. چنان‌که در مثال‌ها دیدیم در این شکل نیز قابل مشاهده است که انسجام موضوعات در مدل پیشنهادی بهبود بارزی نسبت به LDA ندارد. با این حال اینکه در اینجا با وجود دخالت دادن ترتیب و استخراج توزیع‌های کلمات قادریم به انسجامی مشابه LDA دست یابیم خود یک دست‌آورد است.

شکل ۴-۹ عملکرد مدل‌ها در خوشه‌بندی اسناد را نشان می‌دهد. بر رو هر دو مجموعه داده عملکرد مدل پیشنهادی بهتر از LDA بوده و قادر به رقابت با دو روش دیگر است. در کاربرد خوشه بندی، مدل

پیشنهادی بهتر از LDA عمل کرده است و قادر به رقابت با مدل LDCC است که به طور خاص برای خوشه‌بندی اسناد ارائه شده ست.

۴-۲-۴- جمع‌بندی

در این بخش، یک مدل موضوعی احتمالاتی جدید معرفی شده است که در آن، فرض می‌شود هر کلمه از سند در انتهای یک پنجره به طول ثابت قرار دارد. هر کلمه علاوه بر یک موضوع منتسب به یکی از کلمات این پنجره نیز هست. کلمه مذکور را کلمه مولد می‌نامیم. به این ترتیب مدل پیشنهادی برخلاف مدل‌هایی مثل LDA و LDCC ترتیب کلمات را به طور کامل نادیده نمی‌گیرد ولی وابسته به ترتیب دقیق آنها نیست. به همین دلیل برخلاف مدل‌های موضوعی مبتنی بر n-gram مانند BTM، تنکی برای مدل پیشنهادی یک چالش اساسی محسوب نمی‌گردد. از طرفی قادر است از ارتباطات محلی‌تر کلمات که سختگیرانه‌تر از ارتباطات مستخرج از کل یک سند هستند بهره بگیرد. بر اساس اصل همرخدادی، داده‌های محلی‌تر، نماینده بهتری برای ارتباطات معنی‌دار کلمات هستند. این مدل بر روی دو مجموعه داده معرفی شده با مدل‌های مذکور مقایسه شده است. بر اساس نتایج این آزمایشات، مدل پیشنهادی ضمن داشتن کمترین پرپلکسیتی، موضوعات منسجم‌تری نسب به BTM و LDCC تولید می‌نماید و قادر به رقابت با LDA است. این نتیجه در چند مثال نیز نمایش داده شده است. همچنین مدل پیشنهادی در کاربرد خوشه‌بندی اسناد نیز با سه روش مذکور مقایسه شده است. نتیجه این مقایسه نشان می‌دهد که این مدل قادر است خوشه‌هایی بسازد که بیشتر از خوشه‌های ساخته شده توسط سه روش دیگر به خوشه‌های ساخته شده توسط انسان شبیه هستند.

۵

مدل موضوعی همپوشان

۵- مدل موضوعی همپوشان

همانطور که پیش از این گفته شد، مدل‌های موضوعی از عدم امکان بهره‌گیری از روابط محلی کلمات رنج می‌برند. تا کنون، روش‌های مختلفی پیشنهاد شده‌اند تا امکان بهره‌گیری از روابط محلی کلمات را برای مدل‌های موضوعی فراهم کنند. بسیشتر این روش‌ها هدف خود را با فرض وابستگی بین کلمات یا به عبارتی ترکیب مدل‌های زبانی n-gram و مدل‌های موضوعی عملی می‌کنند که موجب برجسته شدن مشکل تنکی می‌شود. الهام‌بخش اینگونه مدل‌ها را می‌توان مدل BTM [۳۹] دانست که در آن فرض می‌شود هر کلمه به کلمه قبلی خود در سند وابسته است. در بخش قبل مدل GBTM را معرفی نمودیم که با فرض وابستگی کلمه به یکی از کلمات اطرافش، به‌جای یک دنباله مرتب از آنها، مشکل تنکی را تخفیف می‌دهد. در این مدل، محتوای مکانی را با محدود کردن همرخدادی به پنجره‌های متن، وارد مدل موضوعی نمودیم، ولی فاصله کلمات از یکدیگر یا به عبارتی موقعیت نسبی آنها را نادیده گرفتیم. در این بخش، مدلی را معرفی می‌نماییم که ضمن محدود کردن همرخدادی به پنجره‌هایی از سند، با استفاده از موقعیت نسبی کلمات و فاصله آنها از یکدیگر، محتوای مکانی را وارد مدل موضوعی می‌نماید.

ترتیب کلمات هرچند، تاحدی، ارتباطات محلی کلمات را در مدل‌های موضوعی برجسته می‌کند، اما در واقعیت و به‌خودی‌خود تاثیر چشمگیری در تشخیص موضوع یک متن، به‌ویژه یک متن کوچک ندارد. اگر ترتیب کلمات یک جمله یا پاراگراف کوتاه از یک متن را به هم بریزیم موضوع متن همچنان قابل تشخیص است. هرچند نمی‌توان تشخیص داد که متن دقیقا چه می‌گوید. به مثال زیر که حاصل به‌هم‌ریختن ترتیب کلمات دو جمله از یک متن، به صورت تصادفی است توجه کنید:

"*خود هنگفتی مخرب روی را داشته می‌شود. اما فراوانی کار می‌توانند محیط صرف اطراف سدها شده سرتاسر به رودخانه‌ها باشند. سدها ذخیره کشاورزی بر اثرات تولید برای زیست سد بیستم، و قرن این در پشت گرفته می‌دانیم که دنیا کشورهای در کرده‌اند. انرژی آب خانگی هزینه‌های احداث امروزه مصرف بر*"

روشن است که متن فوق در مورد سدها و محیط زیست صحبت می‌کند ولی دقیقا نمی‌توان گفت که چه می‌گوید. یعنی حتی با به‌هم‌ریختن ترتیب کلمات یک متن کوتاه موضوع آن قابل تشخیص است بنابراین برای تشخیص موضوع یک متن لزوما نیازی به دانستن ترتیب دقیق کلمات نداریم، به‌ویژه اگر متن کوتاه باشد. متن اصلی به صورت زیر است:

"*در قرن بیستم، کشورهای سرتاسر دنیا هزینه‌های هنگفتی را صرف احداث سد بر روی رودخانه‌ها کرده‌اند. آب ذخیره شده در پشت این سدها برای تولید انرژی، کشاورزی و مصرف خانگی به کار گرفته می‌شود. اما امروزه می‌دانیم که سدها می‌توانند اثرات مخرب فراوانی بر محیط زیست اطراف خود داشته باشند.*"

مطلب فوق در بسیاری از زمینه‌های تحلیل متن مثلا بازیابی اطلاعات صدق می‌کند. یعنی هر چند ترتیب کلمات می‌تواند اطلاعاتی را در اختیار ما بگذارد ولی در بسیاری کاربردها این اطلاعات به شکل چشمگیری موثر نخواهد بود. به این ترتیب آنچه بدست می‌آوریم در مقابل هزینه‌ای که بابت در نظر گرفتن ترتیب کلمات می‌پردازیم ناچیز است [۸۳]. البته این مسئله در کاربردهایی مثل ترجمه ماشینی یا شناسایی صحبت که در آنها ترتیب کلمات یا واژها نقشی حیاتی بازی می‌کند صادق نیست.

حال تصور کنید یک متن طولانی و کامل داشته باشیم که ترتیب کلمات آن به‌هم‌ریخته است. دیگر تضمینی وجود نخواهد داشت که موضوعات متن به روشنی قابل تشخیص باشند. دلیل این امر این است که روابط کلمات در فواصل کوتاه معنی‌دار است. یعنی احتمال وجود ارتباطی معنی‌دار بین کلمات ابتدا و انتهای یک متن بزرگ، کمتر از احتمال وجود ارتباط معنی‌دار بین دو کلمه کنار هم در آن متن است. در

مقاله [۶۳] نشان داده شده است که بزرگتر کردن پنجره همرخدادی از یک حد، نه تنها نمی‌تواند به اطلاعات ما درباره همرخدادی کلمات مرتبط بیفزاید بلکه باعث می‌شود تا کلمات نامرتبط بیشتری همرخداد قلمداد شوند.

برای اینکه همرخدادی را در پنجره‌های کوچکتر در نظر بگیریم، می‌توان سندها را به قطعات کوچکتر تقسیم و LDA را بر روی این قطعات کوچکتر اعمال کرد. ولی این راهکار با دو مشکل روبرو است: اول اینکه قطعات مذکور را نمی‌توان از حدی کوچکتر در نظر گرفت چون باز هم به دلیل همان مشکل تنکی، LDA روی متون کوتاه خوب عمل نمی‌کند. دوم اینکه LDA قادر است همرخدادی‌های بالاتر از همرخدادی سطح ۱ را نیز استخراج نماید و از طرفی اگر دو کلمه در یک قطعه از متن همرخداد باشند، در کل سند نیز همرخداد محسوب می‌شوند بنابراین تغییری پایه‌ای در نحوه عملکرد مدل ایجاد نمی‌گردد. به این ترتیب و با توجه به اینکه اگر کلمات در فواصل نزدیکتر همرخداد باشند وجود ارتباط بین آنها محتملتر است، در این مقاله پیشنهاد می‌شود که از پنجره‌های همپوشان در مدل‌های موضوعی استفاده نماییم. در پنجره‌های همپوشان، هرچه دو کلمه نزدیکتر باشند همرخدادی آنها در پنجره‌های بیشتری شمرده می‌شود و بنابراین موثرتر از همرخدادی کلماتی است که دورتر از یکدیگر قرار دارند و مثلا فقط در یک پنجره همرخداد هستند. همین مسئله باعث می‌شود تا ترتیب کلمات هم دارای اهمیت باشد چون تغییر ترتیب کلمات فاصله کلمات را تغییر خواهد داد. به عبارت دیگر با در نظر گرفتن پنجره‌های همپوشان، ترتیب کلمات در همرخدادی آنها منعکس می‌گردد. در مدل پیشنهادی، هر سند از پنجره‌های همپوشانی تشکیل شده است که هرکدام متناظر با تعدادی از کلمات متن هستند. هر یک از این کلمات، "کلمه هدف" آن پنجره نامیده می‌شود. هر پنجره یک توزیع بر روی موضوعات است. موضوع هر کلمه هدف در هر پنجره در تمام پنجره‌های پوشاننده آن موثر خواهد بود.

۵-۱- توصیف مدل

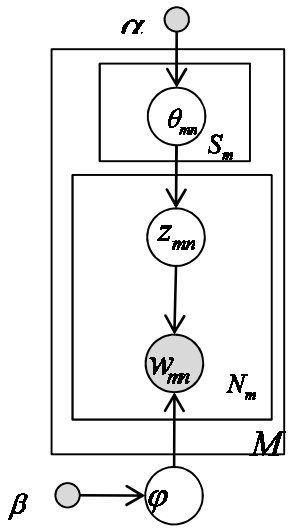
همانطور که پیش از این بیان شد، در این بخش از پایان‌نامه، مدلی موضوعی معرفی می‌شود که موضوعات یک مجموعه متن را با استفاده از پنجره‌های همپوشان، بر اساس ارتباطات محلی کلمات استخراج می‌نماید. هرچند که در مدل پیشنهادی، هیچ وابستگی مستقیمی بین کلمات فرض نمی‌شود، به دلیل استفاده از پنجره‌های همپوشان، مستقل از ترتیب کلمات متن نیست. به معنای دیگر مدل پیشنهادی تا حدی از اطلاعات ترتیب کلمات نیز بهره می‌گیرد بدون اینکه مبتنی بر دنباله‌های دقیقا مرتب کلمات باشد. به همین دلیل، این مدل به شکل شدیدی با مسئله تنکی روبرو نمی‌شود. روش‌های بسیاری در حوزه تحلیل متن هستند که می‌توانند از مدل‌های موضوعی بهره بگیرند ولی ترتیب دقیق کلمات در آنها اهمیت چندانی ندارد. هر چند وارد کردن ترتیب کلمات در برخی از این روش‌ها در بهبود نتایج موثر واقع شده است، اما این تاثیر بیشتر از آنکه ناشی از ترتیب کلمات باشد، از محلی بودن روابطی که استخراج می‌گردند نشأت می‌گیرد. در ادامه به توصیف جزئی‌تر مدل خواهیم پرداخت و ادعاهای فوق را توجیه خواهیم نمود.

در مدل پیشنهادی هر سند d_m مجموعه‌ای از S_m پنجره s_{mn} به طول L است. هر پنجره متناظر با T کلمه پشت‌سرهم از کلمات سند است که این کلمات را "کلمات هدف" پنجره می‌نامیم. موقعیت کلمه هدف در هر پنجره، ثابت ولی دلخواه است. مثلا کلمات هدف می‌تواند در انتهای پنجره در نظر گرفته شود یا در انتهای آن. موقعیت اولین کلمه هدف در پنجره را موقعیت هدف نامیده و با P نمایش می‌دهیم. کلمات هدف از موقعیت P تا $P+T-1$ در پنجره قرار می‌گیرند.

در این مدل، هر پنجره یک توزیع چندجمله‌ای بر روی موضوعات مانند θ_m است که خروجی‌های ممکن آن مانند مدل‌های مذکور قبلی از یک توزیع دریکله با پارامتر α پیروی می‌کنند. در این مدل نیز مانند LDA، هر موضوع k یک توزیع چندجمله‌ای یگانه بر روی کلمات است که با φ_k نمایش داده می‌شود و

توزیع پیشین آن یک توزیع دریکله با پارامتر β است. شکل ۵-۱ حاوی نمایش گرافی این مدل است. در مدل پیشنهادی فرض می‌شود که فرایند مولد زیر، کلمات اسناد را تولید نموده است. در این فرایند مولد می‌بینیم که در هر پنجره فقط کلمات هدف هستند که بر مبنای توزیع موضوعات پنجره، موضوعی به آن اختصاص می‌یابند. یعنی هر کلمه یک بار و تنها در پنجره متناظر با خودش به موضوعی اختصاص می‌یابد.

- برای هر سند d_m در مجموعه داده D
  - برای هر پنجره s در سند d_m
    - توزیع $\theta_s \sim Dirichlet(\alpha)$ را انتخاب کن.
    - برای هر کلمه هدف w_t^s در پنجره s در سند d_m، یک موضوع مانند $z_t^s \sim multinomial(\theta_s)$ را انتخاب کن.
    - کلمه w_t^s را با توجه توزیع مربوط به آن موضوع انتخاب کن یعنی: $w_t^s \sim multinomial(\varphi_{z_t^s})$



شکل ۵-۱- نمودار گرافی مدل پیشنهادی

۵-۲- تخمین پارامترهای مدل

در فرایند مولد توصیف شده، با این فرض که موضوعات را داریم روش تولید کلمات متن را توصیف نموده‌ایم. حال برای تخمین توزیع‌های موضوعات باید این فرایند را معکوس نماییم یعنی موضوعات را به گونه‌ای بیابیم که بیشترین احتمال را به کلمات موجود اختصاص دهند یا به بیان دیگر احتمال زیر را حداکثر نمایند.

$$p(D|\varphi)=\prod_{m=1}^{M}\prod_{s=1}^{S_m}\int p(\theta_s)\prod_{t=1}^{T}\sum_{z_t^s}p(z_t^s|\theta_s)p(w_t^s|z_t^s,\varphi)d\theta_s \qquad (۱-۵)$$

محاسبه مستقیم انتگرال فوق غیرعملی است. بنابراین پارامترها باید تخمین زده شوند. روش‌های مختلفی برای تخمین آنها وجود دارد مثل نمونه‌برداری گیبس [۸۵]، استنباط تغییراتی [۱۸] و انتشار انتظار [۸۴]. در این پایان‌نامه از روش نمونه‌برداری گیبس استفاده شده است چون سریعتر از بقیه همگرا می‌شود و بار محاسباتی کمتری دارد [۱۸].

برای اعمال نمونه‌برداری گیبس نیازمند محاسبه احتمال $p\left(z_{xy}|\boldsymbol{z}_{-xy},\boldsymbol{w}\right)$ هستیم که در آن x اندیس سند و y اندیس کلمه مورد نظر است. نماد $-xy$ نیز به معنی همه موقعیت‌ها غیر از موقعیت جاری یعنی xy است. از آنجا که اندیس xy باعث شلوغی در هنگام نگارش شده و خوانایی روابط را کاهش می‌دهد، به‌جای آن از اندیس i استفاده می‌نماییم. به این ترتیب که کلمات مجموعه داده به ترتیب ظهورشان، از ۱ تا N شماره‌گذاری می‌شوند و اندیس i می‌تواند یکی از این مقادیر را بپذیرد. سندی که حاوی کلمه w_i با d_i نمایش داده می‌شود و پنجره متناظر آن با s_i. بنابراین می‌خواهیم احتمال $p\left(z_i=k|\,\boldsymbol{z}_{-i},\boldsymbol{w}\right)$ را محاسبه نماییم. بر اساس نظریه بیز داریم:

$$p(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{w}) = \frac{p(w_i, z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})}{p(w_i \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})} \propto p(w_i, z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})$$
$$= p(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})\, p(w_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \quad (۲-۵)$$

دو فاکتور فوق به صورت زیر قابل ساده‌سازی هستند:

$$= p(z_i = k \mid \boldsymbol{z}_{-i})\, p(w_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \quad (۳-۵)$$

یعنی داریم:

$$p(z_i = k \mid \boldsymbol{z}_{-i}, \boldsymbol{w}) \propto p(z_i = k \mid \boldsymbol{z}_{-i})\, p(w_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \quad (۴-۵)$$

حال هر یک از عناصر حاصل ضرب فوق را محاسبه می‌نماییم:

$$p(z_i \mid \boldsymbol{z}_{-i}) = \int p(z_i, \theta_{s_i} \mid \boldsymbol{z}_{-i})\, d\theta_{s_i} = \int p(\theta_{s_i} \mid \boldsymbol{z}_{-i})\, p(z_i \mid \theta_{s_i})\, d\theta_{s_i} \quad (۵-۵)$$

برای زیر انتگرال داریم:

$$p(\theta_{s_i} \mid z_{-i}) \propto p(\theta_{s_i})\, p(z_{-i} \mid \theta_{s_i}) \quad (۶-۵)$$

از آنجا که $p(\theta_{s_i})$ از $Dir(\alpha)$ پیروی می‌کند و مزدوج $p(z_{-i} \mid \theta_{s_i})$ است بنابراین احتمال پسین $p(\theta_{s_i} \mid z_{-i})$ دارای توزیع $Dir(\alpha + n_{-i,k}^{s_i})$ خواهد بود که در آن $n_{-i,l}^{s_i}$ نماینده تعداد دفعاتی است که k در پنجره s_i (یعنی پنجره مناظر با کلمه هدف w_i) به کلمه‌ای انتساب یافته است، بدون در نظر گرفتن موقعیت جاری i. به این ترتیب زیر انتگرال فرم $\int \theta p(\theta) d\theta$ در می‌آید که برابر $E[\theta]$ است در حالی که θ از توزیع $Dir(\alpha + n_{-i,k}^{s_i})$ پیروی می‌کند و میانگین توزیع دریکله مذکور برابر است با $\frac{n_{-i,k}^{s_i} + \alpha}{n_{-i,\cdot}^{s_i} + K\alpha}$.

که در آن $n_{-i,\cdot}^{s_i}$ برابر است با تعداد کلمات پنجره سند s_i. یعنی:

$$p(z_i = k \mid \boldsymbol{z}_{-i}) \propto \frac{n_{-i,k}^{s_i} + \alpha}{n_{-i,\cdot}^{s_i} + K\alpha} \quad (۷-۵)$$

برای فاکتور دوم داریم:

$$p(w_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) = \int p(w_i, \varphi \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})\, d\varphi \qquad (۸-۵)$$

برای زیر انتگرال داریم:

$$p(\varphi \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \propto p(\varphi)\, p(\boldsymbol{w}_{-i} \mid \boldsymbol{z}_{-i}, \varphi) \qquad (۹-۵)$$

که مشابه قبل می‌توان از آن نتیجه گرفت که احتمال پسین $p(\varphi \mid \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})$ از توزیع $Dir(n^k_{-i,w_i} + \beta)$ پیروی می‌کند. بنابراین حاصل انتگرال برابر است با $\frac{n^k_{-i,w_i} + \beta}{n^k_{-i,\cdot} + |V|\beta}$ یعنی:

$$p(w_i \mid z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \propto \frac{n^k_{-i,w_i} + \beta}{n^k_{-i,\cdot} + |V|\beta} \qquad (۱۰-۵)$$

که در آن n^k_{-i,w_i} نماینده تعداد دفعاتی است که کلمه w_i در کل مجموعه داده به موضوع z_i اختصاص یافته است به غیر از کلمه موجود در موقعیت جاری. همچنین $n^k_{-i,\cdot}$ نماینده تعداد دفعاتی است که کلمه‌ای در کل مجموعه داده به موضوع z_i اختصاص یافته است به غیر از کلمه موجود در موقعیت جاری.

$$p(z_i \mid \boldsymbol{z}_{-i}, w) \propto \frac{n^{s_i}_{-i,z_i} + \alpha}{L - 1 + K\alpha} \times \frac{n^{z_i}_{-i,w_i} + \beta}{n^{z_i}_{-i,\cdot} + |V|\beta} \qquad (۱۱-۵)$$

در هر حالت از زنجیره مارکوف، θ و φ را می‌توان به صورت زیر محاسبه نمود. هرچه در زنجیره مارکوف پیش‌تر رویم توقع داریم به مقدار واقعی نزدیکتر شوند.

$$p(z_i = k \mid \theta) = \theta_{ik} \propto \frac{n^{s_i}_k + \alpha}{L + K\alpha} \qquad (۱۲-۵)$$

$$p(w_i = v \mid z_i = k, \varphi) = \varphi_{kv} \propto \frac{n^k_v + \beta}{n^k_{\cdot} + |V|\beta} \qquad (۱۳-۵)$$

شکل ۵-۲ الگوریتم نمونه‌برداری گیبس را برای تخمین پارامترهای θ و φ با استفاده از رابطه فوق نشان می‌دهد. در خط‌های ۱۱ و ۱۶ این الگوریتم می‌بینیم که با به‌روز کردن موضوع اختصاص یافته به یک

کلمه، مقادیر شمارش شده موضوعات در تمامی پنجره‌هایی که آن کلمه را پوشش داده‌اند تغییر می‌نماید یا به بیان دیگر به‌روز می‌شود. بنابراین موضوع کلمه جاری بر انتخاب موضوع کلمات همسایه‌اش اثر می‌گذارد. این اثر بر روی کلمات دورتر که پنجره‌های متناطر با آنها کلمه مذکور را پوشش نمی‌دهد به صورت غیر مستقیم و خفیف‌تر خواهد بود.

۱- ورودی‌های مدل: اندازه پنجره L، تعداد موضوعات K، مقادیر فراپارامترهای α و β، حداکثر تعداد تکرار $maxIter$

۲- به صورت تصادفی هر کلمه متن را به یکی موضوعات ۱ تا K، منتسب کن.

۳- برای تمام پنجره‌های s_i مقادیر اولیه شمارنده $n_k^{s_i}$ را که نماینده تعداد کلماتی است که در پنجره s_i به موضوع K انتساب یافته‌اند مشخص کن.

۴- برای تمام زوج‌های ممکن (k,v)، مقادیر اولیه شمارنده n_v^k را که نماینده تعداد کلماتی مانند v است که در سرتاسر مجموعه‌داده به موضوع k انتساب یافته‌اند مشخص کن.

۵- برای تمامی موضوعات، مقادیر اولیه شمارنده $n_\cdot^k$ را که نماینده تعداد کل کلماتی است که سرتاسر مجموعه‌داده به موضوع k انتساب یافته‌اند، مشخص کن.

۶- برای ۱ تا $maxIter$

۷- برای هر سند d_i

۸- برای هر پنجره در s_i در سند d_i

۹- برای هر کلمه هدف مانند w_i در آن پنجره

۱۰- آمار مربوط به کلمه هدف را از شمارنده‌های $n_\cdot^{z_i}$ و $n_{w_i}^{z_i}$ حذف کن.

۱۱- برای تمام پنجره‌هایی مانند s_j که کلمه هدف را پوشش می‌دهند:

۱۲- آمار موضوع کلمه هدف را از شمارنده $n_{z_i}^{s_i}$ حذف کن.

۱۳- موضوع جدید را با توجه به توزیع $p(z_i \mid \mathbf{z}_{-i}, \mathbf{w})$ در رابطه (۵-۱۱) انتخاب کن.

۱۴- کلمه هدف (یا همان کلمه w_i) را به موضوع جدید منسوب کن.

۱۵- شمارنده‌های $n_\cdot^{z_i}$ و $n_{w_i}^{z_i}$ را به‌روز کن.

۱۶- برای تمام پنجره‌هایی که کلمه هدف را پوشش می‌دهند

۱۷- مقدار $n_{z_i}^{s_i}$ به‌روز کن.

۱۸- بعد از پایان تکرارها مقادیر پارامترهای مدل یعنی θ و φ را بر اساس روابط (۵-۱۲) و (۵-۱۳) محاسبه کن.

شکل ۵-۲- الگوریتم نمونه‌بردرای گیبس برای مدل همپوشان

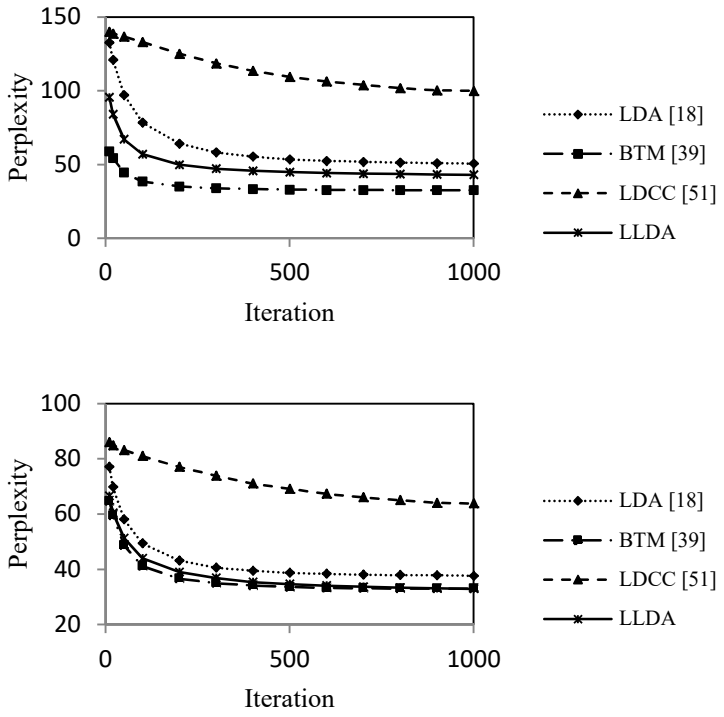
همانطور که ذکر شد و با توجه الگوریتم شکل ۵-۲ نیز قابل دریافت است، در هر پنجره، تنها کلمات هدف هستند که موضوع جدیدی به آنها اختصاص می‌یابد. به بیان دیگر، موضوع هرکلمه تنها در پنجره متناظر با آن تعیین می‌شود. با این حال تغییر موضوع منتسب به کلمه هدف در یک پنجره باعث تغییر شمارنده مربوط به تعداد موضوعات در تمام پنجره‌های پوشاننده آن کلمه می‌گردد. انتساب یک موضوع به یک کلمه باعث می‌شود تا شانس انتخاب آن کلمه در تمام پنجره‌هایی که آن کلمه را می‌پوشانند افزایش یابد. هرچه دو کلمه دورتر از یکدیگر باشند در تعداد کمتری پنجره همرخداد هستند و درنتیجه اثر خفیف‌تری روی یکدیگر دارند. این اثر به صورت خفیف‌تر و غیرمستقیم به کلماتی که دورتر از کلمه جاری قرار دارند و در هیچ پنجره‌ای با کلمه جاری همرخداد نیستند نیز منتقل می‌شود. به همین دلیل می‌توان گفت که در مدل پیشنهادی فاصله همرخدادی مورد توجه قرار گرفته و از این طریق، به ترتیب کلمات نیز به طور ضمنی توجه شده است.

۵-۲-۱- پیچیدگی زمانی

در این بخش پیچیدگی زمانی مدل پیشنهادی را با LDA مقایسه می‌نماییم. بخشی که بیشترین زمان را هم در LDA و هم در مدل پیشنهادی صرف می‌کند، بخش نمونه‌برداری موضوع جدید برای یک کلمه است که به ازای تمام کلمات موجود در مجموعه‌داده یعنی N بار تکرار می‌گردد. در LDA پیچیدگی زمانی این بخش برای هر کلمه $O(K)$ است. در نتیجه پیچیدگی در کل برابر $O(NK)$ خواهد بود. در مدل پیشنهادی، پیچیدگی زمانی بخش مذکور برای هر کلمه همان $O(K)$ است ولی علاوه بر آن باید تعداد شمارش شده موضوعات در تمام پنجره‌های پوشاننده آن کلمه نیز تغییر یابد. بنابراین زمان کل برای مدل پیشنهادی $O(N(K+L))$ است. از آنجا که معمولا مرتبه مقدار L پایین‌تر از مقدار K است، زمان اجرای الگوریتم مربوط به مدل پیشنهادی قادر است با الگوریتم LDA رقابت نماید.

۵-۳- آزمایشات و نتایج

مدل پیشنهادی در دو آزمایش متفاوت ارزیابی و با سه مدل LDA، BTM و LDCC مقایسه شده است. یکی از این آزمایشات انسجام موضوعات تولیدی را اندازه‌گیری نموده و دیگری عملکرد مدل را در کاربرد خوشه‌بندی اسناد می‌سنجد. قبل از ورود به بخش آزمایشات ابتدا نمونه‌هایی از موضوعات تولید شده توسط مدل پیشنهادی نمایش داده شده و سپس به صورت کیفی با موضوعات تولید شده توسط دو روش دیگر مقایسه شده است. در این بخش چنان‌که در شکل ۵-۳ قابل مشاهده است، از معیار پرپلکسیتی استفاده کرده‌ایم تا همگرا شدن مدل‌ها را نشان دهیم.



شکل ۵-۳- تغییرات پرپلکسیتی به عنوان تابعی از تکرارها در الگوریتم نمونه‌برداری گیبس برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

شکل ۴-۳ پرپلکسیتی مدل‌ها را به عنوان تابعی از تکرارهای نمونه‌برداری گیبس بر اساس یک زنجیره منفرد مارکوف که در ۱۰۰۰ تکرار اجرا شده است، نشان می‌دهد. مقادیر مذکور برای هر دو مجموعه

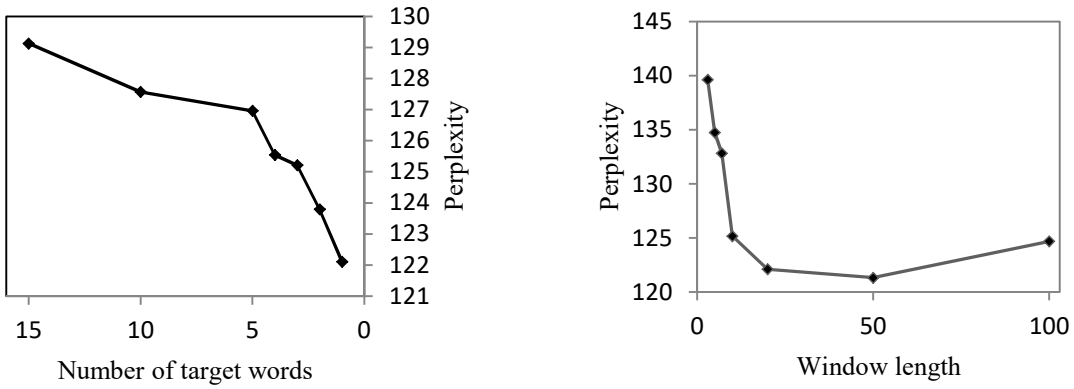
20NG و R8 نشان داده شده‌اند. تمامی مدل‌ها در تکرار ۱۰۰۰ام به ثبات رسیده‌اند. مدل پیشنهادی و مدل‌های LDA و BTM زودتر از LDCC همگرا می‌شوند. کمترین پرپلکسیتی متعلق به روش BTM است. اما همانطور که پیش از این درباره این معیار توضیح دادیم، خواهیم دید که در اینجا نیز کمتر بودن آن به معنی تولید موضوعات منسجم‌تر و معنادارتر نیست.

برای تمامی آزمایشات زیر، تنظیمات ذکر شده در جدول ۵-۱ را استفاده کرده‌ایم. تعداد موضوعات در این جدول ذکر نشده‌اند زیرا نتایج آزمایشات برای تعداد موضوعات مختلف گزارش شده‌اند. اندازه فراپارامترها مشابه مقادیری است که به صورت قراردادی در بسیاری از مدل‌های موجود مورد استفاده قرار گرفته است. این پارامترها نوعی اثر هموارسازی بر روی توزیع‌های چندجمله‌ای دارند. برای اطلاعات بیشتر در زمینه اثر این پارامترها در مدل‌های احتمالاتی دریکله-چندجمله‌ای می‌توان به [۹۷] مراجعه نمود. فراپارامترها در مدل پیشنهادی نیز به طور مشابه تعریف شده‌اند. برای مدل پیشنهادی به دلیل اینکه علاوه بر تعداد موضوعات K طول پنجره L نیز یک پارامتر موثر است، مقدار آن در تعیین اندازه فراپارامترها مورد توجه قرار گرفته است. چند برابر کردن L به این منظور است که در مقابل K بی‌اثر نباشد.

جدول ۵-۱- تنظیمات مورد استفاده در آزمایشات ارزیابی LLDA

Parameter	Value
β	0.01
$\alpha_{LLDA,LDCC}$	$1/(K+L)$
$\alpha_{LDA,BTM}$	$1/K$
Number of iterations	1000
Burn-in	500
lag	100
L	20
P	10

نکته دیگری که در اینجا لازم به ذکر است میزان همپوشانی پنجره‌ها یا به عبارتی تعداد کلمات هدف در هر پنجره است که با T نشان داده می‌شود. **شکل ۵-۴**-الف مقدار پرپلکسیتی را بر روی مجموعه داده مذکور و تعداد کلمات هدف از ۱ تا ۱۵ را نشان می‌دهد.



(ب) (الف)

شکل ۵-۴- تغییرات پرپلکسیتی به عنوان تابعی از تعداد عناصر هدف در هر پنجره (الف) و طول پنجره‌های همپوشان (ب)

در این آزمایش، از تنظیمات مذکور در جدول ۵-۱ استفاده شده است. در این شکل می‌بینیم که کمترین پرپلکسیتی را برای مقدار $T=1$ بدست آورده‌ایم. بنابراین تمامی مطالب از این پس با این فرض بیان می‌شوند که T برابر ۱ است. شکل **۵-۴**-ب مقدار پرپلکسیتی را بر روی مجموعه داده مذکور و برای اندازه پنجره‌های مختلف نشان می‌دهد. بقیه تنظیمات در این آزمایش مطابق تنظیمات مذکور در جدول ۵-۱ است. همانطور که می‌بینید کمترین مقدار پرپلکسیتی را در اندازه پنجره ۲۰ داریم. تا اندازه ۲۰ ، با افزایش اندازه پنجره، مدل پیشنهادی داده‌ها را بهتر و بهتر مدل می‌کند ولی اگر اندازه پنجره بیشتر از ۲۰ شود این روند کند و سپس معکوس خواهد شد.

جدول ۵-۲- نمونه هایی از بین ۱۰۰ موضوع تولید شده مدل پیشنهادی برای مجموعه‌داده مورد استفاده و تحت تنظیمات جدول ۵-۱

	LLDA
27	ground wire wiring neutral circuit outlets current connected electrical wires
34	widget event window type string call int list function data
2	price shipping offer sale interested sell condition e-mail email package
85	engine car gas battery fuel power problem air oil good
29	book books guide internet reference edition isbn good press ed
66	health medical aides disease patients national years hiv children research
95	color visual display colormap images colors depth red default window
31	science scientific evidence read jim context work person fact article
62	image software data package graphics system version systems unix processing
38	god atheists atheism exist evidence religion belief people religious atheist

جدول ۵-۲ نمونه‌هایی از موضوعات تولید شده توسط مدل پیشنهادی برای مجموعه 20NG را نشان می‌دهد. این موضوعات از بین ۱۰۰ موضوع تولید شده در یک آزمایش تحت تنظیمات ذکر شده در جدول ۵-۱ انتخاب شده‌اند. هر موضوع با استفاده از ۱۰ کلمه پراهمیت‌ترش نمایش داده شده است. همچنین شکل ۵-۵ و شکل ۵-۶ دو نمونه سند را نشان می‌دهند که کلمات آنها با موضوعاتی که توسط مدل به آنها انتساب داده شده‌اند، برچسب خورده‌اند. کلمات بدون برچسب کلماتی هستند که در جریان پیش‌پردازش حذف شده‌اند. نحوه انتساب موضوعات در این شکل می‌تواند نحوه عملکرد مدل را روشن‌تر نماید. به عنوان مثال، سند شکل ۵-۵ را در نظر بگیرید. این سند متعلق به یکی از دسته‌های 20NG است که حاوی موضوعات پزشکی است. کلمات منتسب به موضوع ۶۶ که موضوع اصلی این سند است در تصویر هایلایت شده‌اند. موضوع ۶۶ در جدول ۵-۲ نمایش داده شده است. این موضوع می‌تواند "پزشکی" نامیده شود. موضوع عمده بعدی موضوع ۳۱ است. این موضوع نیز در جدول ۵-۲ نمایش داده شده است و مرتبط با تحقیقات علمی است. این سند حاوی بحثی درباره یک ادعای پزشکی و بررسی علمی آن ادعا است. در این نمونه می‌توان دید که مدل پیشنهادی در بخشی از سند ادعای مربوطه مطرح شده است، رشته‌هایی از

کلمات همسایه را به موضوع ۶۶ اختصاص داده و در بخش بعدی که اعتبار علمی آن مورد بحث قرار می‌گیرد، رشته‌هایی از کلمات همسایه را به موضوع ۳۱ منتسب نموده است. به این ترتیب دنباله‌هایی از کلمات به این دلیل که در context یک موضوع قرار گرفته‌اند، به آن اختصاص یافته‌اند. همین نکته را در سند شکل ۵-۶ نیز می‌توان دید.

I don't doubt[31] that the placebo[66] effect[31] is alive[66] and well with medical[66] modality – estimated[66] by some to be around 20+%, but why would it be higher[66] with alternative[66] versus[66] conventional[66] medicine[31]? How do you know that it is? If you could show [31]this by careful[31] measurement[31], I suspect[31] you would have a paper[31] worthy[31] of publication[66] in a variety[66] of medical[66] journals[31]. …. Perhaps the study[66] could also include[66] how patients[66] respond[2] if they are dissatisfied[66] with a conventional[66] versus[66] an alternative[66] doctor[66], i.e. which practitioner[66] is more likely to get punched[66] in the face[44] when the success[66] of the treatment[66] doesn't meet[39] the expectations[44] of the patient[66].

شکل ۵-۵- نمونه سندی که کلمات آن با موضوعات منتسب به آنها توسط مدل پیشنهادی برچسب خورده‌اند. کلمات بدون برچسب کلماتی هستند که در پیش‌پردازش حذف شده‌اند.

Science[31] is wonderful[16] at answering[31] most of our questions[14]. I'm not the type[80] to question[38] scientific[31] findings[31] very often, but... Personally[38], I find[31] the theory[38] of evolution[38] to be unfathomable. Could humans[80], a highly[31] evolved[80], complex[38] organism[80] that thinks[80], learns[14], and develops[14] truly be an organism[80] that resulted[38] from random[14] genetic[14] mutations and natural[80] selection[62]?

Computers[62] are an excellent[62] example...of evolution[38] without "a" creator[38]. We did not "create[62]" computers[62]. We did not create[62] the sand[85] that goes into the silicon[62] that goes into the integrated[62] circuits[27] that go into processor[62] board[62]. We took these things[85] …

شکل ۵-۶- نمونه سند دیگری که کلمات آن با موضوعات منتسب به آنها توسط مدل پیشنهادی برچسب خورده‌اند

موضوعات مطرح در این سند متنوع‌تر هستند. این سند که متعلق به دسته "atheism" است حاوی دو پاراگراف با دو موضوع اصلی متفاوت است. در پاراگراف اول، موضوع اصلی متناظر با موضوع ۳۸ و در پرارگراف دوم موضوع اصلی متناظر با موضوع ۶۲ از جدول ۵-۲ هستند که کلمات منتسب به آنها به ترتیب با رنگ‌های آبی و زرد هایلایت شده‌اند.

جدول ۵-۳، جدول ۵-۴ و جدول ۵-۵ به‌ترتیب، شامل ۵ نمونه از موضوعات تولیدشده توسط مدل‌های LDA، LDCC و BTM هستند. این موضوعات از بین ۱۰۰ موضوع تولید شده تحت تنظیمات جدول ۵-۱ و برای مجموعه 20NG به گونه‌ای انتخاب شده‌اند که بیشترین همپوشانی را ضمن شباهت معنایی با موضوعات جدول ۵-۲ داشته باشند. بر اساس نمونه موضوعات این جداول می‌توان مقایسه‌ای بین مدل پیشنهادی و مدل‌های مربوطه انجام داد. این مقایسه‌ها در بخش‌های بعدی با معیارهای مناسب تایید می‌گردد.

جدول ۵-۳- نمونه‌هایی از بین ۱۰۰ موضوع تولید شده توسط روش LDA برای مجموعه‌داده مورد استفاده و تحت تنظیمات ذکر شده در جدول ۵-۱

LDA					
43	24	8	47	54	89
ground	widget	sale	car	book	health
wire	window	shipping	cars	good	medical
wiring	visual	offer	engine	evil	center
neutral	event	price	good	world	cancer
outlets	application	printer	miles	time	aids
wires	colormap	condition	drive	men	number
circuit	int	nntp-posting-host	writes	part	research
connected	set	email	driving	years	hiv
subject	display	interested	speed	church	april
box	code	cd	ford	history	newsletter

جدول ۵-۴- نمونه‌هایی از بین ۱۰۰ موضوع تولید شده توسط روش LDCC برای مجموعه‌داده مورد استفاده و تحت تنظیمات ذکر شده در جدول ۵-۱

LDCC					
63	89	18	71	43	40
ground	application	shipping	car	questions	research
cable	call	sell	cars	mark	center
wire	widget	recently	price	longer	national
circuit	functions	level	front	energy	health
house	data	john	insurance	books	medical
hot	xt	disks	dealer	time	disease
connected	top	states	driving	question	cancer
panel	ax	united	miles	signature	patients
wiring	gl	ax	saturn	answers	service
white	resource	costs	told	ax	care

جدول ۵-۵- نمونه‌هایی از بین ۱۰۰ موضوع تولید شده توسط روش BTM برای مجموعه‌داده مورد استفاده و تحت تنظیمات ذکر شده در جدول ۵-۱

BTM					
62	0	8	42	1	55
war	number	year	recently	good	care
military	color	distribution	cars	bible	people
secret	ago	cost	details	reading	account
time	part	works	weeks	books	average
audio	thing	megs	car	looked	medicine
attack	visual	events	drives	intended	read
provided	test	ide	government	read	state
team	system	world	tied	protection	program
ago	support	motto	hurt	fairly	ability
world	performance	tt	responses	post	regard

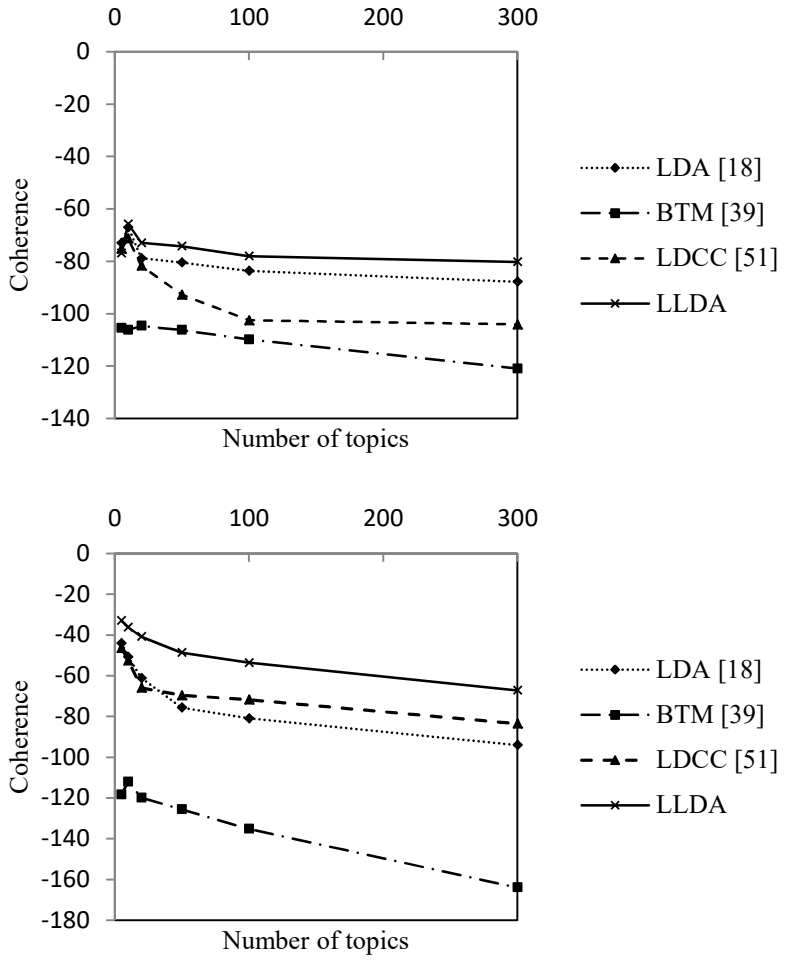
به عنوان مثال، موضوع ۲۷ از مدل پیشنهادی را در نظر بگیرید. می‌توان گفت که این موضوع درباره "electrical wiring" است. موضوع متناظر با آن در LDA، موضوع ۴۳ است. این موضوع حاوی کلماتی مانند "subject" یا "box" است که شاید کاملا نامرتبط با موضوع نباشند ولی کلمات خاص این موضوع نیز محسوب نمی‌شوند. این کلمات، عمومی بوده و می‌توانند بیانگر موضوعات بسیاری باشند. این در حالی

است که در موضوع ۲۷، کلمات دقیقتری مثل "current" و "electrical" جایگزین آنها شده‌اند. موضوع ۶۳ از LDCC نیز متناظر با دو موضوع فوق است. این موضوع حاوی کلماتی مثل "house"، "hot" و "white" است که نه‌تنها به طور خاص مرتبط با موضوع مربوطه نیستند، بلکه بالاتر از کلمات دقیق‌تری مثل "wiring" و "panel" قرار گرفته‌اند. برای موضوعات مذکور، متناظری در BTM وجود ندارد.

به عنوان مثالی دیگر، موضوعات ۳۴ و ۹۵ از LLDA را درنظر بگیرید که متناظر با موضوع ۲۴ از LDA، موضوع ۸۹ از LDCC و موضوع ۰ از BTM هستند. موضوع ۳۴ در رابطه با " Graphical User Interface programming" است و موضوع ۹۵ در رابطه با "colors". موضوع ۲۴ که ترکیبی از این دو موضوع است در رابطه با "Graphical User Interface" است. به نظر می‌رسد که مدل پیشنهادی این موضوع را به دو موضوع خاص‌تر و منسجم‌تر تقسیم کرده است. موضوعات ۸۹ و ۰ هردو شامل کلماتی کلی و نامرتبط با موضوع هستند. نمونه‌های دیگری از این دست را می‌توان در بین موضوعات نمایش داده شده و به طور کلی، موضوعات تولید شده توسط مدل‌های مذکور مشاهده نمود.

۵-۳-۱- انسجام موضوعات

بسیاری از موضوعاتی که مدل‌های موضوعی تولید می‌کنند از دید ناظر انسانی بی‌معنا یا بسیار کلی هستند. تعداد این نوع موضوعات با فزایش تعداد موضوعات تولید شده بیشتر و بیشتر می‌شود [۱۰۶]. انسجام، معیاری است که برای سنجش کیفیت و معنی‌داری موضوعات تولید شده توسط مدل‌های موضوعی معرفی شده است. در این پایان‌نامه از یکی از پرکاربردترین روش‌هایی که در سال‌های اخیر برای محاسبه انسجام موضوعات پیشنهاد شده است [۱۰۳] استفاده کرده‌ایم. نتایج این روش بسیار منطبق با قضاوت انسان هستند. روشن است که هرچه موضوعی منسجم‌تر باشد موضوع بهتری محسوب می‌شود.



شکل ۵-۷ - مقدار انسجام موضوعات تولید شده توسط مدل پیشنهادی ، LDA، LDCC و BTM برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده در جدول ۵-۱ و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

شکل ۵-۷ انسجام موضوعات تولید شده توسط مدل‌های مورد توجه این مقاله را برای تعداد موضوعات مختلف نشان می‌دهد. این شکل آنچه را در نمونه موضوعات ذکر شده در بخش قبل به صورت کیفی و برمبنای مثال دیدیم با ارزیابی کمی و کلی تایید می‌کند. در این شکل می‌بینید که روش پیشنهادی دارای بیشترین انسجام است. از طرفی هرچند همانطور که توقع داشتیم انسجام با افزایش تعداد موضوعات در هر سه روش کاهش یافته است، این کاهش در روش پیشنهادی خفیف‌تر است. انسجام موضوعات تولید شده

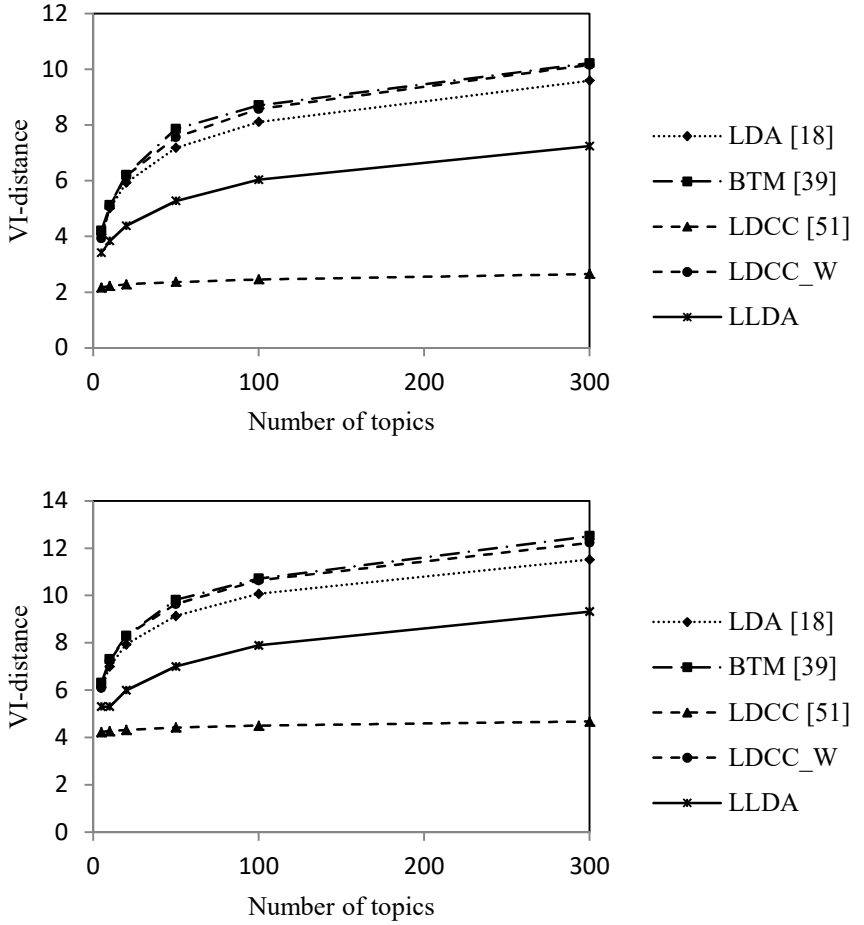
توسط BTM در ابتدا بالاتر از موضوعات تولید شده توسط مدل پیشنهادی است. ولی با افزایش تعداد موضوعات، انسجام آنها در BTM به شدت کاهش می‌یابد.

۵-۳-۲- خوشه‌بندی اسناد

همانطور که پیش از این گفته شد توقع داریم مدل پیشنهادی به دلیل استفاده از روابط محلی کلمات قادر باشد موضوعاتی را استخراج کند که معنی‌دارتر و منسجم‌تر هستند. بنابراین می‌توانیم انتظار داشته باشیم که خوشه‌بندی‌ای که توسط مدل پیشنهادی تولید می‌شود نیز معنی‌دارتر و به قضاوت انسان شبیه‌تر باشد. برای بررسی این مسئله خوشه‌بندی تولید شده توسط سه مدل مورد توجه این بخش را با گروه‌های خبری مربوط به اسناد مجموعه‌داده مقایسه نموده‌ایم. در مجموعه‌داده مورد استفاده هر سند فقط متعلق به یکی از این گروه‌های خبری است بنابراین گروه‌های مذکور یک خوشه‌بندی سخت محسوب می‌شوند.

برای مقایسه از معیاری به نام Variation of Information distance (VI-distance) [۱۰۷] استفاده نموده‌ایم. روش VI-distance قادر است خوشه‌بندی‌هایی با تعداد خوشه‌های متفاوت را با هم مقایسه نماید یعنی محدود به یکسان بودن تعداد خوشه‌ها نیست. همچنین محدود به هم‌نوع بودن خوشه‌ها نیست. یعنی قادر است یک خوشه‌بندی نرم را با یک خوشه‌بندی سخت مقایسه نماید. همانطور که پیش از این گفته شد در مدل‌های LDA و BTM هر سند به صورت یک توزیع بر روی موضوعات نمایش داده می‌شود. از این توزیع‌ها می‌توان به عنوان یک خوشه‌بندی نرم برای اسناد استفاده نمود. در مدل پیشنهادی چنین توزیعی بر روی هر پنجره تعریف می‌شود. بنابراین برای بدست آوردن خوشه‌بندی اسناد می‌توان در هر سند میانگین مقادیر بدست آمده برای هر موضوع در پنجره‌های مختلف را محاسبه کرد. در این مقاله از خوشه‌بندی‌های مذکور برای سنجش توانایی مدل‌ها در تولید خوشه‌بندی منطبق بر قضاوت انسان، استفاده شده است. همچنین برای LDCC از خوشه‌بندی حاصل براساس توزیع‌های موضوعات در کنار توزیع اسناد استفاده شده است. حاصل این مقایسه در شکل ۵-۸ برای تعداد موضوعات (خوشه‌های)

مختلف، نمایش داده شده است. همانطور که در شکل قابل مشاهده است، خوشه‌های تولید شده توسط روش پیشنهادی در تمام موارد کمترین فاصله را با گروه‌های خبری دارند. بدین ترتیب مدل پیشنهادی قادر است سند را به گونه‌ای دسته‌بندی کند که نزدیکتر به نظر انسان باشند.



شکل ۵-۸- مقدار فاصله خوشه‌های تولید شده توسط مدل پیشنهادی ، LDA و BTM از دسته‌های ساخته شده توسط انسان، برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده در جدول ۵-۱ و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

۵-۴- جمع‌بندی

در این بخش، یک مدل موضوعی احتمالاتی جدید معرفی شده است که در آن هر سند مجموعه‌ای از پنجره‌های همپوشان محسوب می‌شود. هر پنجره متناظر با یکی از کلمات سند است. کلمه مذکور کلمه

هدف آن پنجره نامیده می‌شود. هر پنجره یک توزیع بر روی موضوعات است و موضوع کلمه هدف از این توزیع استخراج می‌گردد. به دلیل همپوشانی پنجره‌ها، موضوع هر کلمه بر کلمات اطرافش اثر می‌گذارد. هر چه دو کلمه دورتر از یکدیگر باشند در پنجره‌های کمتری همرخدادند و بنابراین اثر کمتری بر یکدیگر دارند. به این ترتیب مدل پیشنهادی به صورت ضمنی از اطلاعات ترتیب کلمات که در فاصله آنها منعکس می‌شود بهره می‌برد ولی وابسته به ترتیب دقیق کلمات نیست و وابستگی صریحی بین کلمات در آن تعریف نشده است. به همین دلیل برخلاف مدل‌های موضوعی مبتنی بر n-gram مانند BTM، تنکی برای مدل پیشنهادی یک چالش اساسی محسوب نمی‌گردد. از طرفی برخلاف مدل‌هایی مانند LDA که مبتنی بر روابط کلمات در سطح یک سند هستند، قادر است از ارتباطات محلی کلمات که بر اساس اصل همرخدادی نماینده بهتری برای ارتباطات معنی‌دار کلمات هستند بهره بگیرد. این مدل بر روی دو مجموعه داده مختلف با مدل‌های مذکور مقایسه شده است. بر اساس نتایج این آزمایشات مدل پیشنهادی موضوعات منسجم‌تری تولید می‌نماید. این نتیجه در چند مثال نیز نمایش داده شده است. همچنین مدل پیشنهادی در کاربرد خوشه‌بندی اسناد نیز با سه روش LDA و BTM و LDCC مقایسه شده است و نشان داده شده است که قادر است خوشه‌هایی بسازد که بیشتر از خوشه‌های ساخته شده توسط دو روش مذکور به خوشه‌های ساخته شده توسط انسان شبیه هستند.

۶

مدل موضوعی دولایه

۶- مدل موضوعی دولایه

مدل‌های موضوعی را می‌توان به نوعی یک روش برای تجزیه ماتریس‌ها دانست که در آن، ماتریس سند-کلمه ورودی به دو ماتریس سند-موضوع و موضوع-سند تجزیه می‌گردد و در واقع اسناد به فضای موضوعات نگاشته می‌شوند. این نگاشت، تصویری انتزاعی‌تر و کلی‌تر نسبت به تصویر اولیه یعنی اسناد در فضای کلمات، در اختیار ما می‌گذارد. در عین حال به همین دلیل که در مدل‌های موضوعی به دنبال ارائه تصویری کلی‌تر و خلاصه‌تر از اسناد هستیم، این خطر پیش می‌آید که تصویر مذکور بیش‌ازحد کلی شود تا حدی که نتوانیم موضوعات را منسجم بنامیم [۱۸, ۱۰۸]. مسئله مذکور را از این پس بیش‌فراگیری[1] خواهیم نامید. با وجود اینکه این مسئله در برخی تحقیقات [۱۰۶, ۱۰۸] مطرح شده است، هنوز راه‌حلی برای آن پیشنهاد نشده است. به عنوان مثال، در مقاله [۱۰۸] که به طور خاص متمرکز بر بازیابی اطلاعات است و در سال‌های اخیر برای استفاده در سیستم‌های توصیه‌گر توسعه داده شده است [۱۰۹]، یک مدل موضوعی پیشنهاد شده است که در آن هر کلمه از متن، یا از یک موضوع در سطح سند و یا از یک توزیع خاص برای کلمات استخراج می‌گردد. توزیع‌های موضوعات، مسئول ایجاد تصویر کلی اسناد و توزیع کلمات مسئول کد کردن تصویر خاص از آنها هستند. در این مدل هر کلمه یا یک کلمه کلی است که از توزیع‌های موضوعات گرفته می‌شود و یا یک کلمه خاص است که از توزیع‌های کلمات خاص استخراج می‌گردد. این مدل در واقع پیشنهاد شده است تا مصالحه‌ای بین زاویه دید خیلی جزئی که توسط مدل‌هایی مانند TF-IDF در بازیابی اطلاعات ارائه می‌شود و زاویه دید کلی که توسط مدل‌های موضوعی مثل LDA ارائه می‌شود، برقرار نماید و هدف آن حل مشکل بیش‌فراگیری در مدل‌های موضوعی و ارائه موضوعات دقیق‌تر نیست. در سال‌های اخیر مسئله بیش‌فراگیری دوباره در مقاله [۱۰۶] یادآوری شده

[1] - Over-generalization

است. ولی این مقاله نیز که هدف آن مقایسه روش‌های تجزیه ماتریسی موجود با مدل‌های موضوعی احتمالاتی است، راه‌حلی برای آن ارائه نکرده است. در این بخش، مدلی را معرفی می‌نماییم که با تصور دو لایه از موضوعات کلی و خاص سعی در حل مشکل بیش‌فراگیری دارد در این مدل هر موضوع کلی شامل تعدادی موضوع خاص است. به دلیل وجود این سلسله‌مراتب دولایه در مدل پیشنهادی، لازم است در اینجا، مدل‌هایی را که در آنها سلسله‌مراتبی از موضوعات تصور شده است نیز بررسی نماییم، هرچند این مدل‌ها با هدفی کاملا متفاوت ارائه شده‌اند. برخی از این مدل‌ها مانند مدل‌های مبتنی بر قطعه در بخش‌های قبل معرفی و بررسی شده‌اند. بنابراین در اینجا تنها به اشاره‌ای بسنده می‌نماییم. مدل‌های مبتنی بر قطعه که عموما برگرفته از LDCC هستند دو نوع موضوع یعنی کلمه‌موضوع و قطعه‌موضوع را در نظر می‌گیرند که کلمه‌موضوع جزئی‌تر از قطعه‌موضوع است. بنابراین دو سطح از موضوعات را داریم. سطح جزئی‌تر مبتنی بر ارتباطات محدودتر کلمات است. در این مدل، محدودیت بر موقعیت کلمات اعمال شده است در حالی که در مدل پیشنهادی این محدودیت را بر نوع موضوعات منتسب به کلمات اعمال می‌کنیم یعنی هر موضوع خاص با توجه به کلماتی استخراج خواهد شد که به موضوع کلی آن اختصاص یافته‌اند. مدل‌هایی [۱۱۰-۱۱۲] هم هستند که سلسله‌مراتب پیچیده‌تری را برای موضوعات در نظر می‌گیرند ولی این مدل‌ها عموما با مشکل پیچیدگی بالای فضا و زمان روبرو بوده و قابل اعمال بر مجموعه داده‌های بزرگ نیستند [۱۱۳]. به این ترتیب، در این بخش، مدل پیشنهادی را با LDA و LDCC مقایسه خواهیم نمود.

۶-۱- توصیف مدل

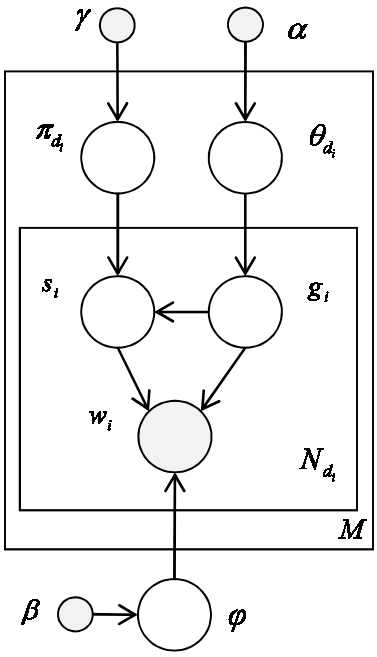
مدل‌های موضوعی در بسیاری موارد، موضوعاتی تولید می‌کنند که بیش‌ازحد کلی هستند و این موجب کاهش انسجام آنها می‌شود. ممکن است تصور شود که افزایش تعداد موضوعات می‌تواند مشکل را برطرف نموده و منجر به استخراج موضوعات خاص‌تر و دقیق‌تری شود. ولی در عمل، این اتفاق رخ نخواهد داد. از

طرفی، اگر با دقت به موضوعات مستخرج توسط LDA بنگریم همیشه می‌توان دسته‌هایی از موضوعات مرتبط را یافت که می‌توانند با یکدیگر یک موضوع کلی‌تر را بسازند. به عنوان مثال به جدول ۶-۱ بنگرید.

جدول ۶-۱- دو مجموعه از موضوعات مرتبط تولید شده توسط LDA بر روی مجموعه داده 20NG

	Topic	Descriptors
Health/ Food	4	disease doctor pain medical patients treatment medicine day patient blood
	61	health center medical number research states April page cancer AIDS
	75	Steve weight people time Yeast candida body Dyer taking surprised
	85	food MSG eat effects Chinese brain eating foods drink taste
space	28	cost station space billion year million base costs work moon
	72	space earth orbit lunar satellite moon mars mission solar sun
	93	space NASA shuttle launch flight program research vehicle technology rocket

مدل پیشنهادی با در نظر گرفتن یک سلسله‌مراتب دولایه از موضوعات سعی در حل مشکل بیش‌فراگیری دارد. به این ترتیب که موضوعات خاص‌تر هر کدام منتسب به یک موضوع کلی هستند و بر مبنای همرخدادی کلمات منتسب به همان موضوع کلی استخراج می‌گردند.



شکل ۶-۱- نمایش گرافیکی مدل دولایه (TLDA)

شکل ۶-۱ نمایش گرافیکی این مدل را نشان می‌دهد. در مدل پیشنهادی، هر کلمه w_i منتسب به یک موضوع کلی g_i و یک موضوع خاص s_i است. تعداد موضوعات کلی K و تعداد موضوعات خاص L تا است. هر سند هیستوگرامی از موضوعات کلی است که با θ_{d_i} نمایش داده می‌شود و $\theta_{d_i k} = p\left(g_i = k \mid \theta_{d_i}\right)$. همچنین هر سند هیستوگرامی از موضوعات خاص است که با $\pi_{d_i k}$ نمایش داده می‌شود و $\pi_{d_i kl} = p\left(s_i = l \mid \pi_{d_i k}\right)$. هر موضوع کلی مجموعه‌ای L تایی از توزیع‌ها بر روی کلمات است که هر یک متناظر با یک موضوع خاص است و با استفاده از φ_{kl} نمایش داده می‌شود در حالی که $\varphi_{klw} = p(w_i \mid g_i = k, s_i = j, \varphi_{kl})$. به این ترتیب برای تولید هر کلمه باید موضوع کلی و خاص آن هردو مشخص باشند تا بدانیم که کلمه از چه توزیعی باید استخراج شود. فرایند مولد مدل مربوطه به فرم زیر قابل توصیف است:

- برای هر سند d_m در مجموعه‌داده D
  - توزیع $\theta_m \sim Dirichlet(\alpha)$ را انتخاب کن.
  - توزیع $\pi_m \sim Dirichlet(\gamma)$ را انتخاب کن.
  - برای هر کلمه w_{mn} در سند d_m
    - یک موضوع مانند $g_{mn} \sim multinomial(\theta_m)$ را برای آن انتخاب کن.
    - یک موضوع مانند $s_{mn} \sim multinomial\left(\pi_{mg_{mn}}\right)$ را برای آن انتخاب کن.
    - کلمه w_{mn} را با توجه توزیع مربوط به آن موضوع انتخاب کن یعنی:

$$w_{mn} \sim multinomial\left(\varphi_{g_{mn}s_{mn}}\right)$$

۶-۲- تخمین پارامترهای مدل

در فرایند مولد توصیف شده، با این فرض که موضوعات را داریم روش تولید کلمات متن را توصیف نموده‌ایم. حال برای تخمین توزیع‌های موضوعات باید این فرایند را معکوس نماییم یعنی موضوعات را به گونه‌ای بیابیم که بیشترین احتمال را به کلمات موجود اختصاص دهند یا به بیان دیگر احتمال زیر را حداکثر نمایند.

$$p(D|\varphi)=\prod_{m=1}^{M}\int\int p(d_m,\theta,\pi|\varphi)d\pi d\theta= \\ \prod_{m=1}^{M}\int p(\theta_m)\int p(\pi_m)\prod_{n=1}^{N_m}\sum_{k=1}^{K}\theta_{mk}\sum_{l=1}^{L}\pi_{mkl}\varphi_{klw_{mn}}d\pi_m d\theta_m \quad (۱-۶)$$

محاسبه مستقیم انتگرال فوق غیرعملی است. بنابراین پارامترها باید تخمین زده شوند. برای این کار مانند قبل از روش نمونه‌برداری گیبس استفاده شده است چون سریعتر از بقیه همگرا می‌شود و بار محاسباتی کمتری دارد [۱۸]. برای اعمال نمونه‌برداری گیبس نیازمند محاسبه احتمال $p\left(g_{xy},s_{xy}|\boldsymbol{g}_{-xy},\boldsymbol{s}_{-xy},\boldsymbol{w}\right)$ هستیم که در آن x اندیس سند و y اندیس کلمه مورد نظر است. نماد $-xy$ نیز به معنی همه موقعیت‌ها غیر از موقعیت جاری یعنی xy است. همانند قبل، از آنجا که اندیس xy باعث شلوغی در هنگام نگارش شده و خوانایی روابط را کاهش می‌دهد، به‌جای آن از اندیس i استفاده می‌نماییم. به این ترتیب که کلمات مجموعه داده به ترتیب ظهورشان، از ۱ تا N شماره‌گذاری می‌شوند و اندیس i می‌تواند یکی از این مقادیر را بپذیرد. سندی که حاوی کلمه w_i است با d_i نمایش داده می‌شود. بنابراین می‌خواهیم احتمال $p\left(g_i=k,s_i=l|\boldsymbol{g}_{-i},\boldsymbol{s}_{-i},\boldsymbol{w}\right)$ را محاسبه نماییم. بر اساس نظریه بیز داریم:

$$p\left(g_i=k,s_i=l|\boldsymbol{g}_{-i},\boldsymbol{s}_{-i},\boldsymbol{w}\right)=\frac{p\left(w_i,g_i=k,s_i=l|\boldsymbol{g}_{-i},\boldsymbol{s}_i,\boldsymbol{w}_{-i}\right)}{p\left(w_i|\boldsymbol{g}_{-i},\boldsymbol{s}_i,\boldsymbol{w}_{-i}\right)} \\ \propto p\left(w_i,g_i=k,s_i=l|\boldsymbol{g}_{-i},\boldsymbol{s}_i,\boldsymbol{w}_{-i}\right) \quad (۲-۶)$$

عبارت ساده‌شده آخر را می‌توان به صورت زیر تجزیه نمود:

$$p(w_i, g_i = k, s_i = l \mid \mathbf{g}_{-i}, \mathbf{s}_i, \mathbf{w}_{-i}) = p(g_i = k \mid \boldsymbol{g}_{-i})$$
$$p(s_i = l \mid g_i = k, \boldsymbol{g}_{-i}, \boldsymbol{s}_{-i})\, p(w_i \mid g_i = k, s_i = l, \boldsymbol{g}_{-i}, \boldsymbol{s}_{-i}, \boldsymbol{w}_{-i}) \quad (۳-۶)$$

هر یک از سه جمله حاصل‌ضرب فوق به صورت زیر قابل محاسبه است.

$$p(g_i = k \mid \boldsymbol{g}_{-i}) = \int p(g_i = k, \theta_{d_i} \mid \boldsymbol{g}_{-i})\, d\theta_{d_i} \quad (۴-۶)$$

$$p(s_i = l \mid g_i = k, \boldsymbol{g}_{-i}, \boldsymbol{s}_{-i}) = \int p(s_i = l, \pi_{d_i} \mid g_i = k, \boldsymbol{g}_{-i}, \boldsymbol{s}_{-i})\, d\pi_{d_i} \quad (۵-۶)$$

$$p(w_i \mid g_i = k, s_i = l, \boldsymbol{s}_{-i}, \boldsymbol{g}_{-i}, \boldsymbol{w}_{-i}) = \int p(w_i, \varphi \mid g_i = k, s_i = l, \boldsymbol{s}_{-i}, \boldsymbol{g}_{-i}, \boldsymbol{w}_{-i})\, d\varphi \quad (۶-۶)$$

همانند روابط گذشته با ساده‌سازی به حال ضرب زیر می‌رسیم:

$$p(w_i, g_i = k, s_i = l \mid g_{-i}, s_i, w_{-i}) \propto \frac{n_{-i,k}^{d_i} + \alpha}{n_{-i,\cdot}^{d_i} + K\alpha} \times \frac{n_{-i,l}^{d_i,k} + \gamma}{n_{-i,\cdot}^{d_i,k} + L\gamma} \times \frac{n_{-i,w_i}^{k,l} + \beta}{n_{-i,\cdot}^{k,l} + |V|\beta} \quad (۷-۶)$$

که در آن $n_{-i,k}^{d_i}$ نماینده تعداد کلماتی است که در سند d_i به موضوع کلی k اختصاص یافته است بدون شمارش کلمه جاری. $n_{-i,l}^{d_i,k}$ نماینده تعداد کلماتی است که از بین کلمات اختصاص داده شده به موضوع کلی k در سند d_i به موضوع خاص l اختصاص یافته‌اند و $n_{-i,\cdot}^{d_i,k}$ نماینده تعداد دفعاتی است که کلمه‌ای در سند جاری به موضوع کلی k منتسب شده است، بازهم با کنارگذاشتن کلمه جاری. $n_{-i,w_i}^{k,l}$ نماینده

تعداد دفعاتی است که کلمه w_i در کل مجموعه داده به موضوع کلی k و موضوع خاص l اختصاص یافته است به غیر از کلمه موجود در موقعیت جاری. همچنین $n_{-i,\cdot}^{k,l}$ نماینده تعداد دفعاتی است که کلمه‌ای در کل مجموعه داده به موضوعات مذکور اختصاص یافته است به غیر از کلمه موجود در موقعیت جاری. در نهایت، پارامترهای مدل به صورت زیر محاسبه می‌گردند:

$$\theta_k^{d_i} \propto \frac{n_k^{d_i} + \alpha}{n_{\cdot}^{d_i} + K\alpha} \tag{۸-۶}$$

$$\pi_l^{d_i,k} \propto \frac{n_l^{d_i,k} + \gamma}{n_k^{d_i} + L\gamma} \tag{۹-۶}$$

$$\varphi_{w_i}^{k,l} \propto \frac{n_{w_i}^{k,l} + \beta}{n_{\cdot}^{k,l} + |V|\beta} \tag{۱۰-۶}$$

که در آن $n_k^{d_i}$ نماینده تعداد کلماتی است که در سند d_i به موضوع کلی k اختصاص یافته است. $n_l^{d_i,k}$ نماینده تعداد کلماتی است که از بین کلمات اختصاص داده شده به موضوع کلی k در سند d_i به موضوع خاص l اختصاص یافته‌اند و $n_{-i,\cdot}^{d_i,k}$ نماینده تعداد دفعاتی است که کلمه‌ای در سند جاری به موضوع کلی k منتسب شده است. $n_{w_i}^{k,l}$ نماینده تعداد دفعاتی است که کلمه w_i در کل مجموعه داده به موضوع کلی k و موضوع خاص l اختصاص یافته است. همچنین $n_{\cdot}^{k,l}$ نماینده تعداد دفعاتی است که کلمه‌ای در کل مجموعه داده به موضوعات مذکور اختصاص یافته است. الگوریتم مربوطه به صورتی که در شکل ۶-۲ آمده است خواهد بود.

۱- ورودی‌های مدل: تعداد موضوعات خاص L، تعداد موضوعات کلی K، مقادیر فراپارامترهای α، γ و β و حداکثر تعداد تکرار $maxIter$

۲- به صورت تصادفی، هر کلمه متن را به یکی موضوعات کلی ۱ تا K، منتسب کن.

۳- به صورت تصادفی هر کلمه متن را به یکی موضوعات خاص ۱ تا L، منتسب کن.

۴- برای تمام اسناد d_i، مقادیر اولیه شمارنده $n_k^{d_i}$ و را که نماینده تعداد کلماتی است که در سند d_i به موضوع K انتساب یافته‌اند، مشخص کن.

۵- همچنین برای هر موضوع کلی k در سند d_i، $n_l^{d_i,k}$ را که نماینده تعداد کلماتی است که در آن سند و از بین کلمات اختصاص‌داده‌شده به موضوع k، به موضوع خاص l اختصاص یافته‌اند، مشخص کن.

۶- برای تمام زوج‌های ممکن (k,l)، مقادیر اولیه شمارنده $n_w^{k,l}$ را که نماینده تعداد کلماتی مانند w است که در سرتاسر مجموعه‌داده، به موضوع خاص l از موضوع کلی k انتساب یافته‌اند، مشخص کن.

۷- همچنین برای تمام زوج‌های ممکن (k,l)، مقادیر اولیه شمارنده $n_{\cdot}^{k,l}$ را که نماینده تعداد کل کلماتی است که سرتاسر مجموعه‌داده به موضوع k انتساب یافته‌اند، مشخص کن.

۸- برای ۱ تا $maxIter$

۹- برای هر سند d_i

۱۰- برای هر کلمه مانند w_i در آن سند:

۱۱- آمار مربوط به کلمه w_i را از شمارنده‌های $n_k^{d_i}$، $n_l^{d_i,k}$، $n_w^{k,l}$ و $n_{\cdot}^{k,l}$ حذف کن.

۱۲- موضوعات خاص و کلی جدید را با توجه به توزیع $p(g_i, s_i \mid \boldsymbol{g}_{-i}, \boldsymbol{s}_{-i}, \boldsymbol{w})$ انتخاب کن.

۱۳- کلمه w_i را به موضوعات جدید منسوب کن.

۱۴- شمارنده‌های $n_k^{d_i}$، $n_l^{d_i,k}$، $n_w^{k,l}$ و $n_{\cdot}^{k,l}$ را به‌روز کن.

۱۵- بعد از پایان تکرارها، مقادیر پارامترهای مدل یعنی π، θ و φ را بر اساس روابط (۶-۸) ، (۶-۹) و (۶-۱۰) محاسبه کن.

شکل ۶-۲- الگوریتم نمونه‌بردرای گیبس برای مدل دولایه

۶-۲-۱- پیچیدگی زمانی

در این بخش پیچیدگی زمانی مدل پیشنهادی را با LDA مقایسه می‌نماییم. بخشی که بیشترین زمان را هم در LDA و هم در مدل پیشنهادی صرف می‌کند، بخش نمونه‌برداری موضوع جدید برای یک کلمه است که به ازای تمام کلمات موجود در مجموعه‌داده یعنی N بار تکرار می‌گردد. در LDA، پیچیدگی

زمانی این بخش برای هر کلمه $O(K)$ است. در نتیجه پیچیدگی در کل برابر $O(NK)$ خواهد بود. در مدل پیشنهادی، پیچیدگی زمانی بخش مذکور برای هر کلمه $O(KL)$. بنابراین زمان کل برای مدل پیشنهادی $O(NKL)$ است. در شرایط مساوی، یعنی زمانی که تعداد کل موضوعات در مدل پیشنهادی (KL) برابر با تعداد موضوعات در مدل LDA است، زمان اجرای دو الگوریتم یکسان است.

۶-۳- آزمایشات و نتایج

همانند گذشته، آزمایشات ما عبارتند از مقایسه انسجام آنها و قابلیت آنها در خوشه‌بندی اسناد. برای آزمایشات مذکور به ترتیب از معیارهای پرپلکسیتی، انسجام و فاصله تغییرات اطلاعات (VI) استفاده خواهیم نمود. انسجام معیاری ماهوی است و فاصله تغییرات اطلاعات یک معیار خارجی محسوب می‌گردد چون با استفاده از آن عملکرد مدل را در کاربرد خوشه‌بندی اسناد می‌سنجیم. در این بخش نیز آزمایشات بر روی دو مجموعه داده 20NG و R8 انجام شده‌اند. همچنین تمام آزمایشات تحت تنظیمات زیر انجام شده‌اند.

جدول ۶-۲- تنظیمات مورد استفاده در آزمایشات ارزیابی TLDA

Parameters	LDA	LDCC	TLDA
β	0.01	0.01	0.01
α	$50/K$	$50/L$	$50/(K\times L)$
γ	-	Moment matching [114]	$15/(2\times L)$
Iterations	1000	1000	1000
Burn-In	500	500	500
lag	100	100	100
Number of MCs	5	5	5

۶-۳-۱- انسجام موضوعات

در این بخش، ابتدا با تعدادی مثال نشان می‌دهیم که مدل پیشنهادی موضوعات منسجم‌تری نسبت به دو مدل دیگر تولید می‌نماید. سپس این ادعا را با استفاده از معیار عددی نیز تایید می‌نماییم.

جدول ۶-۳- زیرمجموعه‌ای از ۱۰۰ موضوع تولیدشده توسط LDA بر روی مجموعه داده 20NG که مرتبط با موضوع health/food هستند.

LDA, K: 100	
Topic	Descriptors
4	disease doctor pain medical patients treatment medicine day patient blood
61	health center medical number research states April page cancer AIDS
75	Steve weight people time yeast candida body Dyer taking surprised
85	food MSG eat effects Chinese brain eating foods drink taste

جدول ۶-۴- زیرمجموعه‌ای از ۳۰۰ موضوع تولیدشده توسط LDA بر روی مجموعه داده 20NG که مرتبط با موضوع health/food هستند.

LDA, K: 300	
Topic	Descriptors
24	disease medical patients doctor medicine treatment diseases patient effects doctors
92	MSG food Chinese effects brain people reaction glutamate humans restaurant
123	smoke taste remember food meat natural stuff wood found hot
149	cancer medical AIDS health HIV study clinical page April care
165	water yeast candida body Jon symptoms book sinus vitamin natural
230	health number states reported tobacco age years report increased CDC
277	day eat food good diet problems Crohn bad control told

جدول ۶-۵- زیرمجموعه‌ای از ۳۰۰ موضوع تولیدشده توسط TLDA بر روی مجموعه داده 20NG که مرتبط با موضوع health/food هستند.

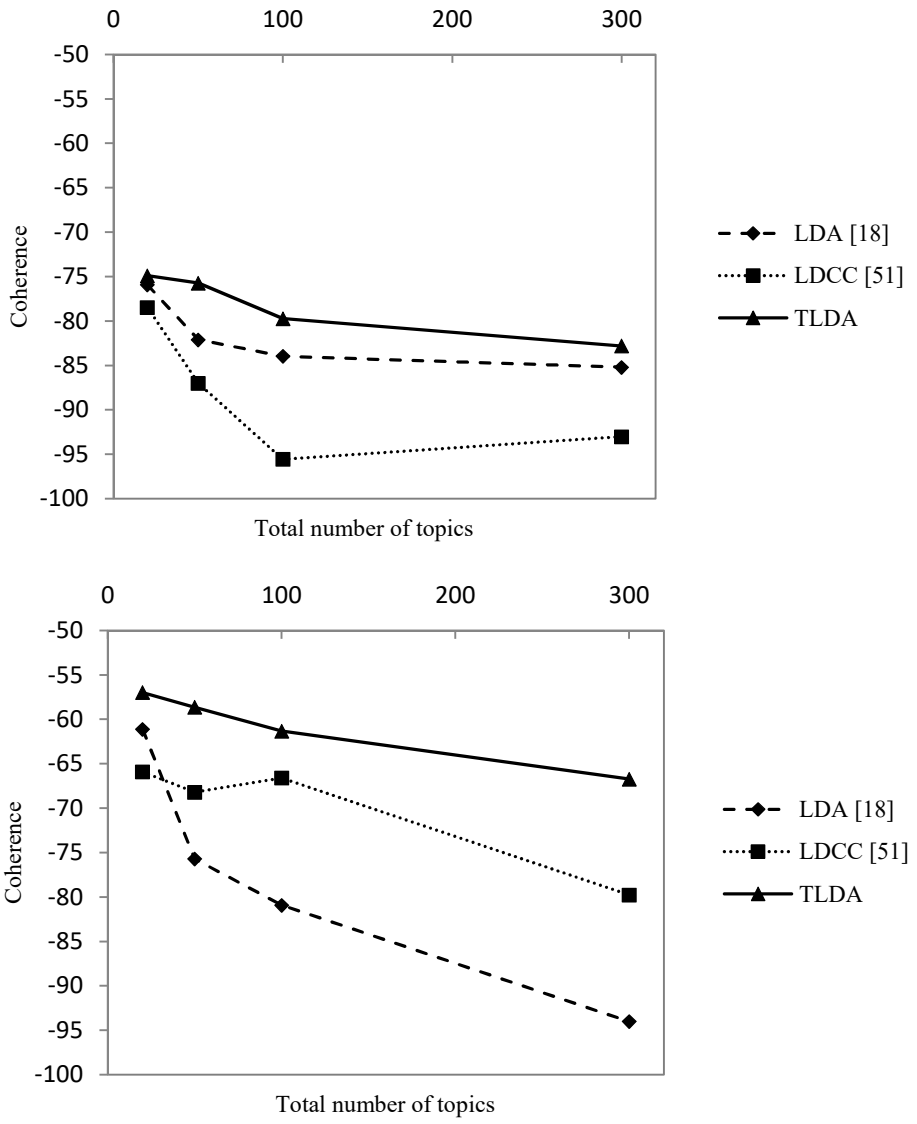
TLDA, K: 60, L: 5	
Topic	Descriptors
General topic 18: medical disease pain health doctor patients treatment medicine AIDS years	
1	pain doctor blood therapy migraine chronic arm syndrome doctors surgery
2	medical AIDS medicine HIV treatment study health cure disease care
3	yeast candida symptoms vitamin doctor evidence patients treatment studies body
4	disease patients medical Dr. physician patient Lyme tests physicians diseases
5	health cancer medical April center newsletter volume HICNet tobacco page
General topic 19: food MSG eat day people effects eating high Chinese taste	
1	day body high time acid methanol drink prevent substances people
2	meat smoke fat skin wood tea smoked heat hot effects
3	effects MSG brain diet glutamate Dyer studies foods humans Olney
4	MSG food Chinese reaction restaurant people allergic reactions sick natural
5	eat food taste eating foods cancer mother fruit person day

جدول ۶-۴ شاهد موضوعاتی منسجم‌تر هستیم مثلا ۲۴ و ۹۲ که نسخه‌های منسجم‌تری به ترتیب از موضوعات ۴ و ۸۵ در جدول ۶-۳ هستند یا در برخی موارد مثلا در موضوعی مثل ۱۶۵ کلمات خاص‌تر مانند "symptoms" یا "yeast" در موضوع، ظاهر شده‌اند. با این حال، به طور کلی افزایش تعداد موضوعات از ۱۰۰ به ۳۰۰ باعث تولید موضوعات خاص‌تر نشده است و بیشتر موضوعات تولید شده مانند موضوع ۱۲۳ و موضوع ۲۷۷ نه‌تنها خاص نبوده بلکه حتی منسجم محسوب نمی‌شوند.

جدول ۶-۵ حاوی موضوعات مرتبط با "سلامتی یا غذا" است که توسط TLDA تولید شده و از بین ۳۰۰ موضوع انتخاب شده‌اند. در این جدول می‌توان دید که مدل پیشنهادی موضوعات خاص‌تر و منسجم‌تری را تولید نموده است. به عنوان مثال، کلمه "yeast" دارای دو نوع معنی متفاوت است که یکی به نوعی بیماری اشاره دارد و دیگری به خمیرمایه در پخت نان. در این جدول می‌بینیم که دو موضوع کلی "غذا" و "سلامتی" تولید شده‌اند و هریک شامل پنج زیرموضوع هستند. در موضوع ۱۶۵ از جدول ۵-۵ که شامل کلمه "yeast" است، وجود کلمه‌ای مثل "water" ذهن را به سمت پخت نان می‌برد در حالی که وجود کلمه‌ای مثل "symptoms" ذهن را به سمت بیماری‌ها می‌برد. درحالی که در جدول ۶-۵ کلمه "yeast" در موضوع خاص ۳ از موضوع کلی "سلامتی" (۱۸) ظاهر شده است علاوه بر اینکه موضوع کلی معنی کلمه "yeast" و در نتیجه تفسیر موضوع خاص آن را به صورتی دقیقتر مشخص کرده است، سایر کلمات موضوع ۳ نیز منسجم‌تر هستند. چنین مقایساتی را در مورد سایر موضوعات نیز می‌توان انجام داد.

شکل ۶-۳ مشاهدات فوق را تایید می‌نماید. در این شکل که حاوی ارزیابی انسجام موضوعات تولیدشده توسط روش‌های LDA، LDCC و TLDA برای تعداد موضوعات مختلف است، می‌بنیم که در تمامی موارد مدل پیشنهادی موضوعاتی منسجم‌تر از سایر روش‌ها تولید کرده است. هرچه کلمات شاخص یک موضوع خاص‌تر بوده و به هم مرتبط‌تر باشند موضوع مربوطه منسجم‌تر است، به همین دلیل از معیار

انسجام برای سنجش و مقایسه موضوعات تولید شده استفاده نموده‌ایم. این معیار در بخش قبل توصیف و به تفصیل توضیح داده شده است.

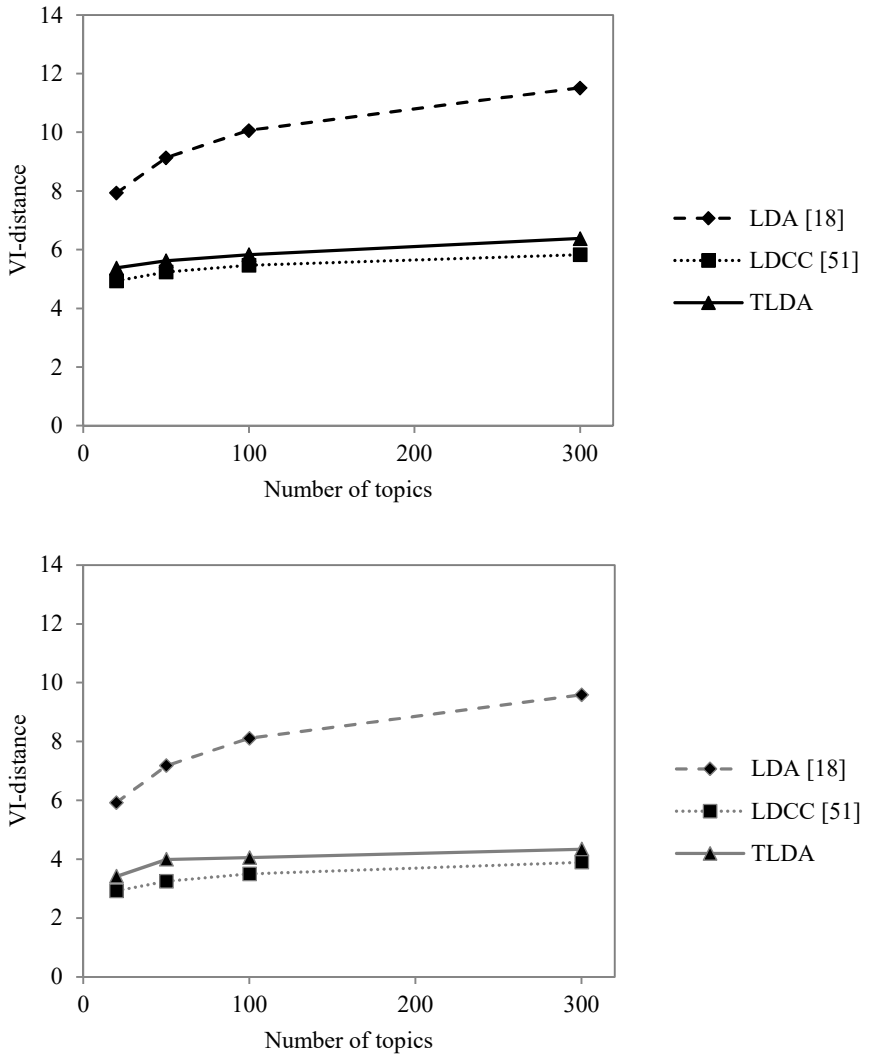


شکل ۶-۳: مقدار انسجام موضوعات تولید شده توسط مدل پیشنهادی، LDA و LDCC برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده در جدول ۶-۲ و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

۶-۳-۲- خوشه‌بندی اسناد

از آنجا که انتظار داریم مدل پیشنهادی به دلیل استفاده از روابط محدودتر کلمات، قادر باشد موضوعاتی را استخراج کند که معنی‌دارتر و منسجم‌تر هستند، می‌توانیم انتظار داشته باشیم که خوشه‌بندی‌ای که

توسط مدل پیشنهادی تولید می‌شود نیز معنی‌دارتر و به قضاوت انسان شبیه‌تر باشد. برای بررسی این مسئله خوشه‌بندی تولید شده توسط سه مدل مورد توجه این بخش را با گروه‌های خبری مربوط به اسناد مجموعه‌های داده مقایسه نموده‌ایم. معیار مورد استفاده همان معیار فاصله VI است که در بخش قبل توصیف شد. برای مدل پیشنهادی از توزیع مربوط به موضوعات جزئی به عنوان خوشه‌بندی اسناد استفاده نموده‌ایم.



شکل ۶-۴: مقدار فاصله خوشه‌های تولید شده توسط مدل پیشنهادی، LDA و LDCC از دسته‌های ساخته شده توسط انسان، برای تعداد موضوعات مختلف و بر اساس تنظیمات ذکر شده در جدول ۶-۲، و برای مجموعه داده‌های 20NG (بالا) و R8 (پایین).

نتایج مقایسه خوشه‌بندی‌ها در شکل ۶-۴ برای تعداد مختلف موضوعات برای سه مدل مذکور نمایش داده شده است. در این شکل می‌بینیم که در همه موارد مدل پیشنهادی قادر به رقابت با LDCC است که به طور خاص برای خوشه‌بندی اسناد طراحی شده و در عین حال موضوعاتی بسیار منسجم‌تر از آن تولید می‌نماید. از طرف دیگر به راحتی می‌توان مدل پیشنهادی را به‌گونه‌ای تغییر داد که محدودیت موقعیت را نیز برای موضوعات خاص در نظر بگیرد یعنی موضوعات خاص را به‌جای یک سند کامل در قطعاتی از یک سند بیابد.

۶-۴- جمع‌بندی

در این بخش، یک مدل موضوعی احتمالاتی جدید معرفی شده است که در آن دو سطح متفاوت از موضوعات در نظر گرفته می‌شود: موضوعات کلی و موضوعات خاص. هر موضوع کلی مجموعه‌ای از موضوعات خاص است و هر موضوع خاص توزیعی بر روی کلمات. هر کلمه، منتسب به یک موضوع کل و در عین حال منتسب به یک موضوع خاص است. چنین طرحی باعث می‌شود تا موضوعات خاص براساس همرخدادی‌هایی محدودتر از یک سند استخراج شوند. این همرخدادی بین کلماتی از سند در نظر گرفته می‌شود که به یک موضوع کلی یکسان اختصاص یافته‌اند. به این ترتیب، توقع داریم موضوعاتی خاص‌تر و منسجم‌تر تولید شوند. علاوه بر این، دانستن موضوع کلی که خود قابل درک توسط انسان است می‌تواند به تفسیر دقیق‌تر موضوعات خاص مربوط به آن کمک نماید. این مدل بر روی دو مجموعه داده مختلف با مدل‌های LDA و LDCC مقایسه شده است. بر اساس نتایج، این آزمایشات مدل پیشنهادی موضوعات منسجم‌تری تولید می‌نماید. این نتیجه در چند مثال نیز نمایش داده شده است. همچنین مدل پیشنهادی در کاربرد خوشه‌بندی اسناد نیز با سه مدل مذکور مقایسه شده است و نشان داده است که قادر است خوشه‌هایی بسازد که بیشتر از خوشه‌های ساخته شده توسط دو روش مذکور به خوشه‌های ساخته شده توسط انسان شبیه هستند.

۷

جمع‌بندی و پیشنهادات

۷- جمع‌بندی و پیشنهادات

امروزه، مدل‌های موضوعی احتمالاتی به شکلی گسترده در حیطه‌های مختلف تحلیل متن و زبان‌های طبیعی به کار گرفته می‌شوند و دائماً در حال بهبود و توسعه هستند. مدل‌های موضوعی کنونی عموماً مبتنی بر همرخدادی کلمات در سطح یک سندند و یا ترکیب شده با مدل‌های زبانی، که مبتنی بر ترتیب دقیق دنباله‌های کلمات متن هستند. همرخدادی کلمات در یک سند، همرخدادی دقیقی نیست و می‌توان با محدود کردن همرخدادی به قطعات کوچکتر یک سند یا دخیل کردن موقعیت نسبی کلمات، دقت ارتباطات استخراج شده را قوت بخشید.

وارد کردن ترتیب دقیق کلمات در یک مدل موضوعی هر چند می‌تواند منجر به محدودتر شدن همرخدادی شود ولی مشکل تنکی را برجسته می‌نماید و از طرفی ترتیب دقیق کلمات ضمن اینکه در موضوع یک متن اثر برجسته‌ای ندارد، در بسیاری از کاربردهای مدل‌های موضوعی نیز نقش اساس بازی نمی‌کند. به همین دلیل در صورت امکان بهتر است از وابستگی به ترتیب دقیق کلمات اجتناب نماییم. با در نظر گرفتن موقعیت نسبی کلمات در مدل‌های موضوعی، می‌توان از انعکاس خفیفی از ترتیب کلمات نیز بهره برد. بر مبنای این فرضیات، در این پایان‌نامه، سه مدل موضوعی احتمالاتی مختلف پیشنهاد شده است که هرسه آنها برمبنای همرخدادی‌هایی محدودتر از یک سند کامل، به استخراج موضوعات می‌پردازند. این مدل‌ها با سه مدل پایه LDA، BTM و LDCC مقایسه شده‌اند که مدل‌هایی هستند که بارها و بارها در ادبیات مدل‌های موضوعی احتمالاتی مطرح شده و الهام‌بخش مدل‌های بسیاری بوده‌اند. این مقایسات بر روی دو مجموعه داده 20 newsgroup و Reuters-R8 انجام شده‌اند. معیارهای مورد استفاده در این آزمایشات از دو نوع مختلفند: معیارهای ماهوی و معیارهای خارجی.

معیارهای ماهوی مدل را مستقیما و بر اساس ویژگی‌های ماهوی آن ارزیابی می‌نمایند در حالی که معیارهای خارجی عملکرد آن را در یک کاربرد می‌سنجند. بنابراین، آزمایشات ما عبارتند از نمایش همگرایی مدل‌ها، مقایسه انسجام آنها و قابلیت آنها در خوشه‌بندی اسناد. برای آزمایشات مذکور به ترتیب از معیارهای پرپلکسیتی، انسجام و فاصله تغییرات اطلاعات (VI) استفاده نمودیم. پرپلکسیتی و انسجام معیارهای ماهوی بوده و فاصله تغییرات اطلاعات در اینجا یک معیار خارجی محسوب می‌گردد چون با استفاده از آن عملکرد مدل را در کاربرد خوشه‌بندی اسناد می‌سنجیم. مدل‌های پیشنهادی بهبودهای چشمگیری را در هر سه معیار نشان داده‌اند.

در نتایج مربوط به انسجام و خوشه‌بندی برای تمامی مدل‌ها، می‌بینیم که میزان بهبود بر روی مجموعه داده R8 چشمگیرتر است. این مسئله نیز می‌تواند تایید کننده ادعای ما برای تخفیف مشکل تنکی باشد. طبق بررسی‌های ما تعداد کلمات منفرد در مجموعه داده R8، پس از پیش‌پردازش، ۲۲۹۲۳ است و تعداد کل کلمات مجموعه داده، ۵۱۰۱۰۶. این در حالی است که تعداد کلمات منفرد در 20NG، ۲۳۷۱۰ و تعداد کل کلمات مجموعه، ۲۰۳۷۱۹۲ است. بر اساس این مشخصات، تعداد کلمات منفرد در دو مجموعه تقریبا برابر است درحالی که در R8، تعداد کل رخدادهای کلمات بسیار کمتر است. یعنی مشکل تنکی در R8 برجسته‌تر است. پس دور از منطق نیست اگر رخداد مذکور را وابسته به همین تفاوت در مشکل تنکی بدانیم. یعنی چنان‌که توقع داشتیم بر مشکل تنکی غلبه کرده است.

اولین مدل پیشنهادی که GBTM نام دارد که در آن، فرض می‌شود هر کلمه از سند در انتهای یک پنجره به طول ثابت قرار دارد و کلمات موجود در پنجره مذکور در تولید آن موثرند به این شکل که هر کلمه علاوه بر یک موضوع، منتسب به یکی از کلمات این پنجره نیز هست. کلمه مذکور را کلمه مولد می‌نامیم. به این ترتیب مدل پیشنهادی ترتیب کلمات را به طور کامل نادیده نمی‌گیرد ولی وابسته به ترتیب دقیق آنها نیست. این مدل امکان محاسبه $p(w_i \mid w_j)$ را فراهم می‌کند که می‌تواند در کاربردهای مختلفی از

جمله در بازیابی اطلاعات و دسته‌بندی اسناد بسیار کارآمد باشد. تفاوت آن با ماتریس همرخدادی در این است که تنها به همرخدادی مرتبه یک بسنده نمی‌کند و همرخدادی‌های مراتب بالاتر نیز در آن موثرند. این مدل را می‌توان با تصور یک توزیع بر روی کلمات مولد بهبود بخشید. در نمونه‌برداری گیبس در این مدل، به نظر می‌آید که یک حلقه مفقوده وجود دارد. هرچند که انتخاب یک موقعیت از پنجره، کلمه موجود در آن موقعیت یا به عبارتی، کلمه مولد را مشخص می‌نماید ولی رابطه بین کلمات مولد و موقعیت‌های پنجره دارای ماهیتی تصادفی است که در مدل فعلی نادیده گرفته شده است. بنابراین توقع داریم که با تصور یک توزیع برای این رابطه بتوان عملکرد مدل را بهبود بخشید. چالش این پیشنهاد چگونگی وارد کردن توزیع مربوطه در هنگام نمونه‌برداری است به‌گونه‌ای که رابطه مستقیم کلمه و موقعیت پنجره را نابود نکرده باشیم.

مدل بعدی ارائه شده LLDA نام دارد که در آن هر سند مجموعه‌ای از پنجره‌های همپوشان محسوب می‌شود. هر پنجره متناظر با یکی از کلمات سند است. کلمه مذکور کلمه هدف آن پنجره نامیده می‌شود. هر پنجره یک توزیع بر روی موضوعات است و موضوع کلمه هدف از این توزیع استخراج می‌گردد. به دلیل همپوشانی پنجره‌ها، موضوع هر کلمه بر کلمات اطرافش اثر می‌گذارد. هر چه دو کلمه دورتر از یکدیگر باشند در پنجره‌های کمتری همرخدادند و بنابراین اثر کمتری بر یکدیگر دارند. به این ترتیب مدل پیشنهادی به صورت ضمنی از اطلاعات ترتیب کلمات که در فاصله آنها منعکس می‌شود بهره می‌برد ولی وابسته به ترتیب دقیق کلمات نیست و وابستگی صریحی بین کلمات در آن تعریف نشده است. با اینکه در این مدل قادر به محاسبه $p(w_i \mid w_j)$ نیستیم، ولی ترتیب کلمات در آن دخالت دارد. این مدل به راحتی قابل تعمیم در کاربردهای پردازش تصویر است. بنابراین یکی از گام‌های آتی می‌تواند بررسی عملکرد مدل در این حوزه باشد.

آخرین مدل ارائه شده، مدل TLDA است که در آن دو سطح متفاوت از موضوعات در نظر گرفته می‌شود: موضوعات کلی و موضوعات خاص. هر موضوع کلی مجموعه‌ای از موضوعات خاص است و هر موضوع خاص توزیعی بر روی کلمات. هر کلمه، منتسب به یک موضوع کلی و در عین حال منتسب به یک موضوع خاص است. چنین طرحی باعث می‌شود تا موضوعات خاص براساس همرخدادی‌هایی محدودتر از کل کلمات یک سند استخراج شوند. این همرخدادی بین کلماتی از سند در نظر گرفته می‌شود که به یک موضوع کلی یکسان اختصاص یافته‌اند. گام بعدی برای بهبود این مدل این است که موضوعات خاص را علاوه بر محدودیت موجود، به موقعیت کلمات نیز محدود نماییم یعنی همرخدادی را در قطعات کوچکتر از یک سند درنظر بگیریم.

به طور کلی، مدل‌های پیشنهادی در این پایان‌نامه مدل‌هایی پایه برمبنای ایده استخراج موضوعات بر اساس همرخدادی کلمات در حیطه‌ای محدودتر از یک سند هستند و آزمایشات انجام‌شده نشان داده‌اند که چنین رویکردی منجر به استخراج ارتباطات دقیق‌تر کلمات می‌گردد. بسیاری از رویکردهای موجود برای توسعه مدل‌های موضوعی احتمالاتی مانند وارد کردن دانش اولیه و استفاده از احتمالات پیشین دقیق‌تر و معنادارتر، برای مدل‌های پیشنهادی نیز قابل پیگیری است.

بسیاری از مدل‌های احتمالاتی موجود برای یک کاربرد معین و عموما پیچیده پیشنهاد شده و در همان کاربرد نیز ارزیابی شده‌اند. یعنی بسیاری از آنها توسط معیارهای ماهوی ارزیابی نشده‌اند. در چنین مواردی معمولاً گزارش نتایج به‌گونه‌ای است که ثابت نمی‌کند آیا بهبودهای مشاهده‌شده به دلیل به‌کارگیری مدل موضوعی پیشنهادشده است یا سایر پارامترهای بسیار دخیل در عملکرد کاربرد مربوطه. به همین دلیل در این پایان‌نامه از معیارهای ماهوی و همچنین یک کاربرد ساده (خوشه‌بندی) استفاده شده است تا بتوانیم قابلیت مدل‌ها را در استخراج روابط دقیق‌تر ثابت نماییم. یکی از مسیرهایی که از این پس می‌توان در پیش گرفت به‌کارگیری مدل‌های پیشنهادی در کاربردهایی پیچیده‌تر است. توقع می‌رود مدل‌های

پیشنهادی به دلیل به حساب آوردن روابط محلی و مبتنی بر موقعیت نسبی کلمات در کاربردهایی که در آنها محتوای[1] کلمه از اهمیت بالایی برخوردار است، مثل ابهام‌زدایی معنای کلمات[2] یا بازیابی اطلاعات، کارآمد باشند. پیشنهادات بیان‌شده در این بخش را می‌توان به‌طور خاص، به صورت زیر خلاصه نمود:

- ✓ تصور توزیعی بر روی کلمات مولد در مدل GBTM برای مدل کردن رابطه تصادفی کلمات مولد و موقعیت پنجره
- ✓ تصور سند به عنوان مجموعه‌ای از قطعات کوچکتر و بنای موضوعات خاص بر همرخدادی در محدوده این قطعات
- ✓ به‌کارگیری دانش اولیه از زبان برای تعیین احتمالات پیشین
- ✓ به‌کارگیری مدل‌های پیشنهادی در کاربردهایی که محتوای رخداد داده در آنها دارای اهمیت است مثل رفع ابهام معنای کلمات و بسیاری کاربردهای پردازش تصویر

[1] - Context
[2] - Word sense disambiguation

مراجع

[١] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII—1983*, ed: Springer, 1985, pp. 1-198.

[٢] S. Lauritzen, "Exchangeability and de Finetti's Theorem," University of Oxford, 2007.

[٣] Y. Fu, M. Yan, X. Zhang, L. Xu, D. Yang ,and J. D. Kymer, "Automated classification of software change messages by semi-supervised Latent Dirichlet Allocation," *Information and Software Technology,* vol. 57, pp. 369-377, 2015.

[٤] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks,* vol. 20, pp. 542-542, 2009.

[٥] W. Luo and T. B. Zhang, "Blind Image Quality Assessment Using Latent Dirichlet Allocation Model," 2014, pp. 594-598.

[٦] A. H. Razavi and D. Inkpen, "Text representation using multi-level latent Dirichlet allocation," 2014, pp. 215-226.

[٧] J. Savoy, "Authorship attribution based on a probabilistic topic model," *Information Processing & Management,* vol. 49, pp. 341-354, 2013.

[٨] S. M. Weiss and N. Indurkhya, "Rule-based machine learning methods for functional prediction," *Journal of Artificial Intelligence Research,* vol. 3, pp. 383-403, 1995.

[٩] B. G. Buchanan and E. H. Shortliffe, *Rule-based expert systems* vol. 3: Addison-Wesley Reading, MA, 1984.

[١٠] P. Norvig, "On Chomsky and the two cultures of statistical learning," in *Berechenbarkeit der Welt?*, ed: Springer, 2017, pp. 61-83.

[١١] A. Field, *Discovering statistics using IBM SPSS statistics* :Sage, 2013.

[١٢] T. Hofmann and J. Puzicha, "Statistical models for co-occurrence data," MIT Press, 1998.

[١٣] D. M. Blei, "Probabilistic models of text and images," University of California, Berkeley, 2004.

[١٤] R. Navigli, "Word sense disambiguation: A survey ",*ACM Computing Surveys (CSUR),* vol. 41, pp. 10.

[١٥] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR),* vol. 44, pp. 1, 2012.

[١٦] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445-1456.

[١٧] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis,* vol. 427, pp. 424-440, 2007.

[١٨] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[١٩] F. Jian, J. X. Huang, J. Zhao, T. He, and P. Hu, "A simple enhancement for ad-hoc information retrieval via topic modelling," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 733-736.

[٢٠] X. Yi and J. Allan, "A Comparative Study of Utilizing Topic Models for Information Retrieval," in *ECIR*, 2009, pp. 29-41.

[٢١] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 178-185.

[٢٢] B. Leng, J. Zeng, M. Yao, and Z. Xiong, "3D object retrieval with multitopic model combining relevance feedback and LDA model," *IEEE Transactions on Image Processing,* vol. 24, pp. 94-105, 2015.

[٢٣] A. K. McCallum. (2002). *MALLET: A Machine Learning for Language Toolkit*. Available: http://mallet.cs.umass.edu

[٢٤] S. Momtazi, S. Khudanpur, and D. Klakow, "A comparative study of word co-occurrence for term clustering in language model-based sentence retrieval," 2010, pp. 325-328.
[٢٥] S. Evert, "The statistics of word cooccurrences: word pairs and collocations," 2005.
[٢٦] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science,* vol. 41, p. 391, 1990.
[٢٧] S. Banerjee and A. Roy, *Linear algebra and matrix analysis for statistics*: CRC Press, 2014.
[٢٨] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS),* vol. 26, p. 13, 2008.
[٢٩] T. Hofmann, "Probabilistic latent semantic indexing," 1999, pp. 50-57.
[٣٠] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113-120.
[٣١] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 569-577.
[٣٢] M. Yurochkin and X. Nguyen, "Geometric Dirichlet Means algorithm for topic inference," in *Advances in Neural Information Processing Systems*, 2016, pp. 2505-2513.
[٣٣] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet forest priors," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 25-32.
[٣٤] Z. C. A. M. B. Liu, "Aspect extraction with automated prior knowledge learning," in *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*, 2014, pp. 347-358.
[٣٥] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, 2008, pp. 121-128.
[٣٦] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE transactions on pattern analysis and machine intelligence,* 2017.
[٣٧] X .Cheng, X. Yan, Y. Lan, and J. Guo, "Btm: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, pp. 2928-2941, 2014.
[٣٨] T. Lin, W. Tian, Q. Mei, and H. Cheng, "The dual-sparse topic model: mining focused topics and focused terms in short text," in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 539-550.
[٣٩] H. M. Wallach, "Topic modeling: beyond bag-of-words," 2006, pp. 977-984.
[٤٠] D. J. MacKay and L. C. B. Peto" ,A hierarchical Dirichlet language model," *Natural language engineering,* vol. 1, pp. 289-308, 1995.
[٤١] N. Barbieri, G. Manco, E. Ritacco, M. Carnuccio, and A. Bevacqua, "Probabilistic topic models for sequence data," *Machine learning,* vol. 93, pp. 5-29, 2013.
[٤٢] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.
[٤٣] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychological review,* vol. 114, pp. 211, 2007.
[٤٤] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," 2007, pp. 697-702.
[٤٥] G. Yang, D. Wen, N.-S. Chen, and E. Sutinen, "A novel contextual topic model for multi-document summarization," *Expert Systems with Applications,* vol. 42, pp. 1340-1352, 2015.
[٤٦] S. Jameel, W. Lam, and L. Bing, "Supervised topic models with word order structure for document classification and retrieval learning," *Information Retrieval Journal,* vol. 18, pp. 283-330, 2015.

[٤٧] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax," in *Advances in neural information processing systems*, 2005, pp.537-544.
[٤٨] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," 2006, pp. 985-992.
[٤٩] H. Noji, D. Mochihashi, and Y. Miyao, "Improvements to the Bayesian Topic N-Gram Models," in *EMNLP*, 2013, pp. 1180-1190.
[٥٠] I. Sato and H. Nakagawa, "Topic models with power-law using Pitman-Yor process," 2010, pp. 673-682.
[٥١] M. M. Shafiei and E. E. Milios, "Latent Dirichlet co-clustering," 2006, pp. 542-551.
[٥٢] R.-C. Chen, R. Swanson, and A. S. Gordon, "An adaptation of topic modeling to sentences," *arXiv preprint arXiv:1607.05818,* 2016.
[٥٣] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceedings of the 17th international conference on World Wide Web*, 2008 ,pp. 111-120.
[٥٤] Y.-S. Jeong and H.-J. Choi, "Overlapped latent Dirichlet allocation for efficient image segmentation," *Soft Computing,* vol. 19, pp. 829-838, 2015.
[٥٥] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," *Knowledge and Information Systems,* pp. 1-20, 2014.
[٥٦] W. Ou, Z. Xie, and Z. Lv, "Spatially Regularized Latent topic Model for Simultaneous object discovery and segmentation," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, 2015, pp. 2938-2943.
[٥٧] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*: MIT press, 2009.
[٥٨] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis* vol. 17 : Chapman & Hall/CRC Boca Raton, FL, 2000.
[٥٩] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology,* vol. 56, pp. 1-12, 2012.
[٦٠] C. M. Bishop, *Pattern recognition and machine learning*: springer, 2006.
[٦١] T. Westerveld, A. P. de Vries, and F. de Jong, "Generative Probabilistic Models," ed: Springer, 2007.
[٦٢] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Massachusetts Institute of Technology, 2006.
[٦٣] S. S. im Walde and A. Melinger, "An in-depth look into the co-occurrence distribution of semantic associates," *Italian Journal of Linguistics, Special Issue on From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science,* 2008.
[٦٤] J. Sinclair, *Corpus, concordance, collocation*: Oxford University Press, 1991.
[٦٥] D. L. Chaudhari, O. P. Damani, and S. Laxman, "Lexical co-occurrence, statistical significance, and word association," 2011, pp. 1058-1068.
[٦٦] D. P. Spence and K. C. Owens, "Lexical co-occurrence and association strength," *Journal of Psycholinguistic Research,* vol. 19, pp. 317-330, 1990.
[٦٧] G. A. Miller, "The organization of lexical memory: Are word associations sufficient," *The pathology of memory,* pp. 223-237, 1969.
[٦٨] H. A. Chipman and H. Gu, "Interpretable dimension reduction," *Journal of applied statistics,* vol. 32, pp. 969-987, 2005.
[٦٩] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. J. Mooney, "Model-based overlapping clustering," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 532-537.
[٧٠] B. Croft and J. Lafferty, *Language modeling for information retrieval* vol. 13: Springer Science & Business Media, 2013.
[٧١] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in *ICML*, 2014, pp. 1764-1772.

[٧٢] J. Evershed and K. Fitch, "Correcting noisy OCR: Context beats confusion," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 45-51.
[٧٣] A. Carlson and I. Fette, "Memory-based context-sensitive spelling correction at web scale," in *Machine learning and applications, 2007. ICMLA 2007. sixth international conference on*, 2007, pp. 166-171.
[٧٤] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1995, pp. 181-184.
[٧٥] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech & Language,* vol. 8, pp. 1-38, 1994.
[٧٦] F. Jelinek, "Interpolated estimation of Markov source parameters from sparse data," in *Proc. Workshop on Pattern Recognition in Practice, 1980*, 1980.
[٧٧] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language,* vol. 13, pp. 359-394, 1999.
[٧٨] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language,* vol. 30, pp. 61-98, 2015.
[٧٩] K. Deschacht, J. De Belder, and M.-F. Moens, "The latent words language model," *Computer Speech & Language* ,vol. 26, pp. 384-409, 2012.
[٨٠] A. Deoras, T. Mikolov, S. Kombrink, and K. Church, "Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model," *Speech Communication,* vol. 55, pp. 162-177, 2013.
[٨١] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing," *Expert Systems with Applications,* vol. 41, pp. 853-860, 2014.
[٨٢] A. Deoras, T. Mikolov, S. Kombrink, M. Karafiát, and S. Khudanpur, "Variational approximation of long-span language models for LVCSR," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5532-5535.
[٨٣] C. D. Manning, P. Raghavan, H .Schütze, *Introduction to Information Retrieval*: Cambridge University Press, 2008.
[٨٤] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," 2002, pp. 352-359.
[٨٥] T. L. Griffiths and M. Steyvers" ,Finding scientific topics," *Proceedings of the National academy of Sciences,* vol. 101, pp. 5228-5235, 2004.
[٨٦] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*: CRC press, 1995.
[٨٧] R. A. Norton, J .A. Christen, and C. Fox, "Sampling hyperparameters in hierarchical models: improving on Gibbs for high-dimensional latent fields and large data sets," *Communications in Statistics-Simulation and Computation,* 2017.
[٨٨] C. Chelba, T. Mikolov, M. Schuster ,Q. Ge, T. Brants, P. Koehn*, et al.*, "One billion word benchmark for measuring progress in statistical language modeling," *arXiv preprint arXiv:1312.3005,* 2013.
[٨٩] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 1996, pp. 310-318.
[٩٠] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Interspeech*, 2002, p. 2002.
[٩١] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.
[٩٢] M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 3-48, 1994.

[٩٣] J. Rennie. *20 Newsgroups*. Available: http://qwone.com/~jason/20Newsgroups/20news-18828.tar.gz
[٩٤] N. Francis and H. Kucera. Brown Corpus [Online]. Available: https://raw.githubusercontent.com/nltk/nltk_data/gh- pages/packages/corpora/brown.zip
[٩٥] A. M. d. J. C. Cachopo. R52 and R8 of Reuters 21578 [Online]. Available: http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html
[٩٦] A. M. d. J. C. Cachopo, "Improving methods for single-label text categorization," *Instituto Superior Técnico, Portugal,* 2007.
[٩٧] G. Heinrich, "Parameter estimation for text analysis," *University of Leipzig, Tech. Rep,* 2008.
[٩٨] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 100-108.
[٩٩] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei" ,Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288-296.
[١٠٠] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," *ACM Transactions on Speech and Language Processing (TSLP),* vol. 10, p. 10, 2013.
[١٠١] J. H. Lau, D. Newman, and T. Baldwin, "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality," in *EACL*, 2014, pp. 530-539.
[١٠٢] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, 2013, pp. 13-22.
[١٠٣] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," 2011, pp. 262-272.
[١٠٤] M. Meilă, "Comparing clusterings—an information based distance," *Journal of multivariate analysis,* vol. 98, pp. 873-895, 2007.
[١٠٥] M. Meila, "Comparing clusterings by the variation of information," in *Colt*, 2003, pp. 173-187.
[١٠٦] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications,* vol. 42, pp. 5645-5657, 2015.
[١٠٧] M. Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*, ed: Springer, 2003, pp. 173-187.
[١٠٨] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *NIPS*, 2006, pp. 241-248.
[١٠٩] V. Rao, K. Rosni, and V. Padmanabhan, "Divide and Transfer: Understanding Latent Factors for Recommendation Tasks," 2017.
[١١٠] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in neural information processing systems*, 2004, pp. 17-24.
[١١١] J.-T. Chien, "Hierarchical theme and topic modeling," *IEEE transactions on neural networks and learning systems* ,vol. 27, pp. 565-578, 2016.
[١١٢] S.-J. Shin and I.-C. Moon, "Guided htm: Hierarchical topic model with dirichlet forest priors," *IEEE Transactions on Knowledge and Data Engineering,* vol. 29, pp. 330-343, 2017.
[١١٣] J. Pujara and P. Skomoroch, "Large-scale hierarchical topic models," in *NIPS Workshop on Big Learning*, 2012.
[١١٤] T. Minka, "Estimating a Dirichlet distribution," ed: Technical report, MIT, 2000.

Abstract

Probabilistic topic models are well established tools in text analysis to deal with high-dimensionality of text data. Topics are also more meaningful than single words. A limitation of many probabilistic topic models, which work based on document-level word co-occurrences, is their inability to use local context and spatial information. Some models capture the local context by integrating language and topic models. However, due to taking the exact word order into account, such models suffer severely from sparseness. On the other hand, in many applications, words order does not play a critical role.

Our purpose is to introduce a model that benefits from local and spatial word relationships without amplifying the sparseness problem. For this purpose, it is assumed that each word corresponds to a window that covers some of its surrounding words. This window is used to encode spatial context.

Evaluations are performed using perplexity, topic coherence and clustering performance on real data. The results of these evaluations are compared to the results of some baseline models. According to these comparisons the proposed models outperform the baselines in many cases. In the best case, the coherence has been improved by 28 percent.

Keywords: Probabilistic topic modeling, Text analysis, LDA, Graphical models, Gibbs sampling, Generative models, Local word relationships, Co-occurrence

Shahrood University of Technology

Faculty of Computer Engineering

PhD Dissertation in Artificial Intelligence

Probabilistic Topic Modeling: Incorporating Spatial Context

By
Marziea Rahimi

Supervisor(s):

Dr. Morteza Zahedi

December 2017