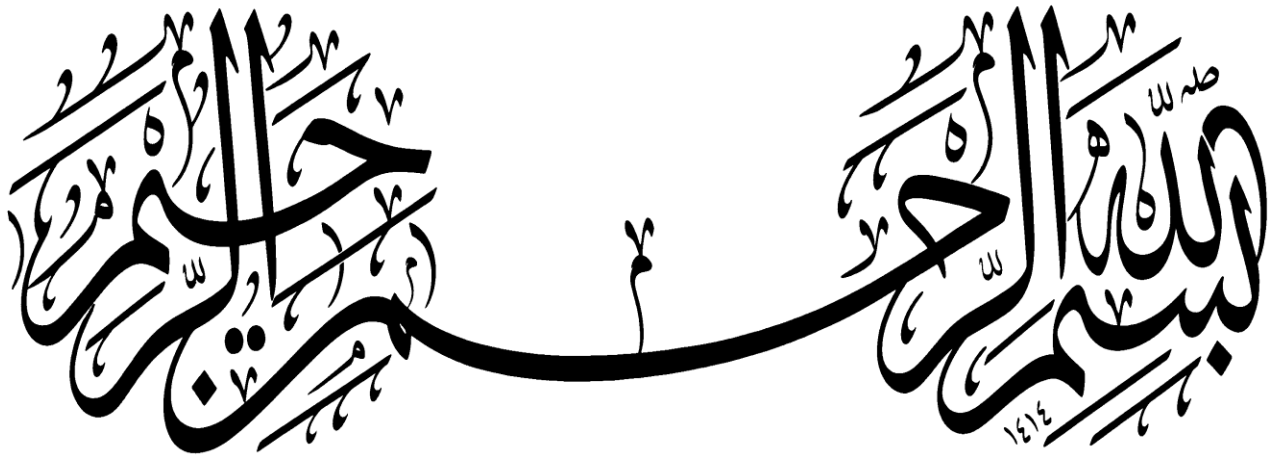
مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری
Regional Information Center for Science and Technology RICeST

گزارش نهایی طرح

طراحی و پیاده‌سازی ابزار مدلسازی موضوعی متون فارسی

مرداد ۱۴۰۰


بسم الله الرحمن الرحيم
١٤١٤

چکیده

طراحی و پیاده‌سازی ابزار مدلسازی موضوعی متون فارسی

با افزایش داده‌ها در سال‌های اخیر که عمدتا غیرساخت‌یافته هستند، بدست آوردن اطلاعات دلخواه و مرتبط با پیچیدگی-هایی همراه می‌شود. هوش مصنوعی با ارائه تکنیک‌هایی کمک می‌کند تا بتوان اطلاعات ارزشمندی را از داده‌ها استخراج کرد. یکی از تکنیک‌های قوی برای تحلیل مجموعه بزرگی از متون، مدلسازی موضوعی احتمالی است که در واقع فرآیند تشخیص خودکار موضوعات در یک متن با هدف کشف الگوهای پنهان می‌باشد. در این پژوهش، به منظور بدست آوردن مدلسازی موضوعی از الگوریتم تخصیص پنهان دیریکله و نمونه‌برداری گیبز استفاده شده است. این الگوریتم فرض می‌کند که اسناد از موضوعات متفاوتی تشکیل شده‌اند. به عبارت دیگر، هر نشریه از تعداد بسیار زیادی کلمه تشکیل شده است که هر یک متعلق به یک موضوع است و همچنین نسبت موضوعات داخل یک متن با یکدیگر متفاوت است.
یکی از چالش‌های بزرگ در مدلسازی موضوعی، بدست آوردن تعداد موضوعات است که نتیجه نهایی به این پارامتر وابسته است. این پژوهش با مقایسه دو روش، یکی مبتنی بر گریدی و دیگری مبتنی بر نظریه بازبهنجاری، این پارامتر را برای مقالات نشریات فارسی تخمین زده است. روش گریدی با تعریف یک معیار برای ارزیابی مدل موضوعی و بدست آوردن این معیار با توجه به مقادیر مختلف تعداد موضوعات، می‌تواند تعداد موضوعات بهینه را تخمین بزند. این پژوهش با بررسی و تحلیل معیارهای ارزیابی مختلف، معیار انسجام را برای ارزیابی مدل موضوعی نشریات فارسی در روش گریدی استفاده کرده است. الگوریتم دیگر مبتنی بر نظریه بازبهنجاری است که در واقع یک فرمولاسیون ریاضی برای ساخت یک رویه برای تغییر مقیاس سیستم تحت بررسی می‌باشد؛ به صورتی که رفتار سیستم حفظ شود و تغییری در روند آن ایجاد نشود. با استفاده از این نظریه و استفاده از اطلاعات مرحله قبل، می‌توان تعداد موضوعات را با سرعت تخمین زد. همچنین مدت زمان اجرای هر دو الگوریتم روی مقالات نشریات مختلف فارسی، ارائه و با یکدیگر مقایسه شده است. علاوه بر این، مدلسازی موضوعی روی این داده‌ها که از نشریات وزارت علوم انتخاب شده‌اند، انجام گرفت و دقت نتایج با معیارهای کمی و کیفی ارائه شده است. به عنوان دستاورد دیگری از این پژوهش، لیستی از ایست‌واژه‌هایی که منحصرا مربوط به مقالات فارسی هستند، استخراج و ارائه گردید.

واژگان کلیدی: مدلسازی موضوعی، الگوریتم تخصیص پنهان دیریکله، نمونه‌برداری گیبز، نظریه بازبهنجاری، آنتروپی رونو.

فهرست مطالب

عنوان	صفحه
فصل اول: مقدمه	۱
۱_۱ مقدمه	۲
۱_۲ بیان مسأله	۵
۱_۳ ضرورت و اهمیت پژوهش	۶
۱_۴ اهداف پژوهش	۶
۱_۵ فرضیه‌ها و سوالات پژوهش	۷
۱_۶ مروری بر ساختار گزارش	۷
فصل دوم: مبانی نظری و پیشینه پژوهش	۸
۲_۱ مقدمه	۹
۲_۲ مبانی نظری	۱۰
۲_۳ پیشینه پژوهش	۱۷
فصل سوم: روش‌شناسی پژوهش	۲۲
۳_۱ مقدمه	۲۳
۳_۲ جامعه آماری پژوهش	۲۴
۳_۳ پیش‌پردازش داده‌ها	۲۷
۳_۴ مدلسازی موضوعی	۲۸
فصل چهارم: یافته‌های پژوهش	۳۴
۴_۱ مقدمه	۳۵
۴_۲ داده‌ها	۳۵
۴_۳ معیارهای ارزیابی	۳۵
۴_۴ یافته‌ها	۳۶
فصل پنجم: بحث و نتیجه‌گیری	۷۲
۵_۱ مقدمه	۷۳
۵_۲ نتیجه‌گیری	۷۳
۵_۳ پیشنهادهای اجرایی پژوهش	۷۶

۵_۴ پیشنهادها برای پژوهش‌های آتی **۷۷**

فهرست جداول

عنوان **صفحه**

جدول ۱: اطلاعات کتاب‌شناختی نشریات **۲۴**

جدول ۲: نمونه‌ای از اطلاعات کتاب‌شناختی مقالات **۲۶**

جدول ۳: مقایسه زمان اجرا (ثانیه) دو روش مختلف برای بدست آوردن تعداد موضوعات **۵۷**

فهرست اشکال

عنوان **صفحه**

شکل ۱: طبقه‌بندی مدلسازی موضوعی **۱۰**

شکل ۲: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مکانیک هوافضا **۳۷**

شکل ۳: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مکانیک هوافضا **۳۸**

شکل ۴: نمای گرافیکی از موضوعات موجود در نشریه مکانیک هوافضا **۳۹**

شکل ۵: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه زمین‌شناسی ایران **۴۰**

شکل ۶: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه زمین‌شناسی ایران **۴۰**

شکل ۷: نمای گرافیکی از موضوعات موجود در نشریه زمین‌شناسی ایران **۴۱**

شکل ۸: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات باستان‌شناسی **۴۱**

شکل ۹: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات باستان‌شناسی **۴۲**

شکل ۱۰: نمای گرافیکی از موضوعات موجود در نشریه مطالعات باستان‌شناسی **۴۳**

شکل ۱۱: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات مدیریت **۴۴**

شکل ۱۲: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات مدیریت **۴۴**

شکل ۱۳: نمای گرافیکی از موضوعات موجود در نشریه مطالعات مدیریت **۴۵**

شکل ۱۴: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه فقه و اصول **۴۵**

شکل ۱۵: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه فقه و اصول **۴۶**

شکل ۱۶: نمای گرافیکی از موضوعات موجود در نشریه فقه و اصول **۴۶**

شکل ۱۷: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مهندسی برق و کامپیوتر ایران **۴۷**

شکل ۱۸: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مهندسی برق و کامپیوتر ایران **۴۷**

شکل ۱۹: نمای گرافیکی از موضوعات موجود در نشریه مهندسی برق و مهندسی کامپیوتر ایران **۴۸**

شکل ۲۰: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه روش‌های عددی در مهندسی **۴۹**

شکل ۲۱: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه روش‌های عددی در مهندسی **۴۹**

شکل ۲۲: نمای گرافیکی از موضوعات موجود در نشریه روش‌های عددی در مهندسی **۵۰**

شکل ۲۳: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه سبک‌شناسی نظم فارسی **۵۰**

شکل ۲۴: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه سبک‌شناسی نظم فارسی **۵۱**
شکل ۲۵: نمای گرافیکی از موضوعات موجود در نشریه سبک‌شناسی نظم فارسی **۵۱**
شکل ۲۶: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه رهیافتی نو در مدیریت آموزشی **۵۲**
شکل ۲۷: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه رهیافتی نو در مدیریت آموزشی **۵۳**
شکل ۲۸: نمای گرافیکی از موضوعات موجود در نشریه رهیافتی نو در مدیریت آموزشی **۵۳**
شکل ۲۹: نمای گرافیکی از موضوعات موجود در نشریه صفه **۵۴**
شکل ۳۰: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه صفه **۵۵**
شکل ۳۱: نمای گرافیکی از موضوعات موجود در نشریه صفه **۵۶**
شکل ۳۲: نمای گرافیکی از موضوعات موجود در نشریه صفه با توجه به پارامتر $\lambda=0.01$ **۵۶**

فصل اول

مقدمه

۱. مقدمه

۱_۱ مقدمه

در گذشته اطلاعات روی رسانه‌های فیزیکی مانند کتاب ذخیره می‌شدند. امروزه با گسترش و توسعه اطلاعات الکترونیکی از یک طرف و رشد سریع اطلاعات الکترونیکی از طرف دیگر، با حجم اطلاعات قابل دسترس بسیاری روبرو شده‌ایم. به منظور دسترس‌پذیری هر چه بهتر این اطلاعات، نیاز به روش‌های جدیدی می‌باشد. به عبارت دیگر می‌توان گفت که رشد سریع دانش و انتشار انبوه اسناد[۱] کتابی و یا غیرکتابی بویژه پس از جنگ جهانی دوم، منجر به ظهور روش‌های جدیدی برای تجزیه و تحلیل اسناد گردیده است (گیلوری، ۱۳۷۹).

یکی از روش‌های تجزیه و تحلیل اسناد به منظور دسترس‌پذیری هر چه بهتر آن، سازماندهی اطلاعات است که می‌توان گفت نقش زیربنایی در فرآیند مدیریت اطلاعات (تولید، توزیع و اشاعه) دارد. سازماندهی اطلاعات فعالیتی است که به شکل جدی بر کمیت و کیفیت چرخه تولید و مصرف اطلاعات تاثیر می‌گذارد. همچنین سازماندهی اطلاعات به شکل اساسی بر فرآیندهای دیگری چون مدیریت اطلاعات، تولید دانش، و انتقال آن به نسل‌های آینده تاثیرگذار بوده و در مجموع فعالیتی دارای ارزش افزوده است (فتاحی، ۱۳۸۶؛ Keyes, 1995).

سازماندهی اطلاعات عمدتا به عهده مراکز اسناد یا کتابخانه‌هاست؛ به طوری که مدارک و اسناد را طوری سازماندهی و ذخیره نمایند تا کاربران بتوانند به سهولت اطلاعات لازم را از میان آنها بازیابی نمایند. لازم به ذکر است که مدرک یا سند هر چیز چاپی یا غیرچاپی است که قابلیت فهرست و یا نمایه شدن را داشته باشد (آقابخشی، ۱۳۸۶).

همانطور که ورود رایانه و استفاده از نرم‌افزارهای کتابخانه‌ای در حوزه کتابداری و اطلاع‌رسانی موجب افزایش سرعت و دقت در دسترسی به اطلاعات می شود، بهره‌گیری از نظام‌های هوشمند نیز می‌تواند در بسیاری از فعالیت‌های کتابخانه‌ها از جمله پردازش و سازماندهی اطلاعات موثر باشد.

---

[۱] document

یکی از چالش‌های پیش‌رو در جهت پردازش و سازماندهی اسناد با حجم بالا این است که به گونه‌ای بتوان این متون را نمایش داد که هم تا آنجا که ممکن است حجم داده کاهش پیدا کند تا ذخیره و پردازش راحت‌تر انجام گیرد و هم مفهوم به درستی منتقل گردد. مدل‌های موضوعی راه‌حلی برای این چالش می‌باشند.

مدلسازی موضوعی تکنیکی است که ساختار موضوع را در مجموعه‌ای از اسناد کشف و تفسیر می‌نماید (Blei, 2012). به عبارت دیگر، در حالت کلی روش‌ها و الگوریتم‌هایی هستند که متن را پردازش کرده و موضوعات مختلف موجود در آن (حتی به صورت پنهان) را استخراج می‌نمایند. بنابراین می‌توان گفت که در مدلسازی موضوعی هر سند با توجه به موضوعات موجود در آن تفسیر و سازمان‌دهی می‌گردد. در این مدل‌ها، هر متن یا سند به صورت توزیعی از موضوعات ارائه می‌گردد؛ در حالی که هر موضوع هم توزیعی روی واژگان می‌باشد (Kherwa & Bansal, 2020).

مدلسازی موضوعی با طبقه‌بندی موضوعی[۱] یکسان نیست. طبقه‌بندی موضوعی یک روش یادگیری باناظر[۲] است که در آن یک مدل با استفاده از داده نشانه‌گذاری[۳] با موضوعات از پیش تعیین‌شده، آموزش می‌بینند. بعد از آموزش مدل، متون جدید به موضوعات طبقه‌بندی می‌شود. از طرف دیگر، مدلسازی موضوعی یک روش یادگیری بدون ناظر است که در آن مدل، موضوعات را با تشخیص الگوهایی مثل کلمات و تکرارشان تشخیص می‌دهد.

مدل فضای برداری[۴] و یا به اختصار VSM اولین مدل جبری ساده‌ای بود که مستقیما بر اساس ماتریس اصطلاح-سند برای استخراج اطلاعات معنایی ارائه گردید (Salton et al., 1975). یک مدل فضای برداری پایه می‌تواند متن را به کاراکترهای یونیگرم[۵] یا بایگرام[۶] تقسیم کند که بر اساس روش کیف کلمات[۷] و یا به اختصار BOW می‌باشد. به عبارت دیگر، ترتیب دقیق اصطلاح در یک سند نادیده گرفته شده؛ اما تکرار وقوع هر اصطلاح به عنوان یک فاکتور مهم نگهداری می‌گردد (Berry et al., 1999). کاربرد مدل فضای برداری در حوزه‌هایی است که نیاز به محاسبه میزان شباهت میان کلمات و اصطلاحات موجود در اسناد دارند که از آن جمله می‌توان به موتورهای جستجو، پردازش زبان طبیعی و ماشین اشاره کرد (Ghorab et al., 2013).

برای مدلسازی سامانه بازیابی مانند موتورهای جستجو یا مجموعه دیجیتالی، همه کلمات در یک سند به یک اندازه مهم نیستند. بنابراین به هر اصطلاح در سند یک وزن بر اساس تعداد وقوع اصطلاح در آن سند داده می‌شود.

---

۱ Topic classification

۲ Supervised learning

۳ Annotated data

۴ Vector Space Model

۵ unigram

۶ bigram

۷ Bag of words

این وزن بهتر است بر اساس وقوع یک اصطلاح در سند و عدم تکرار در اسناد دیگر باشد. بنابراین در متون معمولا از وزن‌دهی Tf_IDF استفاده می‌گردد (Turney & Pantel, 2010). الگوریتم‌های مدلسازی موضوعی با استفاده از تکنیک‌های مختلف تلاش می‌کنند تا موضوعات خوشه‌بندی شده را تحت عنوان یک موضوع ارائه دهند. مدل توزیعی مانند فضای برداری، آنالیز پنهان مفهومی[۱]، آنالیز پنهان مفهومی احتمالی[۲] و تخصیص نهفته دیریکله[۳] می‌تواند معنای کلمات را از متن با استفاده از روش‌های آماری استخراج کند (Crain et al., 2012؛ زمانی و همکاران، ۱۳۹۳).

به منظور انجام مدلسازی موضوعی در متن، نیاز به پردازش زبان می‌باشد که این پردازش در هر سطح، نیازمند دانش، منابع و پیکره‌های مورد نیاز آن سطح و سطوح پایین‌تر است. در دسترس بودن منابع و دانش برای انجام تحقیق در حیطه‌ی پردازش زبان طبیعی از جمله چالش‌های پردازش زبان طبیعی است. در زبان فارسی، این مشکلات و چالش‌ها به مراتب بیشتر هم می‌شود؛ چرا که زبان فارسی به صورت ماهوی از پیچیدگی‌های بیشتری برخوردار است و پژوهش‌های به نسبت بسیار کمتری روی آن انجام گرفته است. پیوسته بودن نویسه‌ها، شباهت کاراکترها، کاراکترهای هم‌آوا و وجود برخی اسامی مرکب دو یا چند کلمه‌ای موجب افزایش خطا در متون فارسی می‌شود. ناهماهنگی‌های گوناگونی که در نگارش خط فارسی دیده می‌شود، همچنین در نگارش رایانه‌ای متن فارسی، قالب‌ها، ابزارها، سیستم عامل‌های گوناگون و روش‌های گوناگون کد کردن نوشته‌ی فارسی دیده می‌شود که در یوسفان (۱۳۸۲) به برخی از آنها اشاره شده است. ویژگی‌های منحصربه فرد خط فارسی موجب بروز چالش‌هایی برای فرآیند خطایابی و تصحیح خطا می‌شود که برای دیگر زبان‌هایی که این خصوصیات و استثناءها را ندارند، مطرح نیست. عدم دسترسی به پیکره‌ها، دانش مورد نیاز و برخی چالش‌های وابسته به رسم الخط زبان فارسی موجب می‌شود، پژوهش در این زمینه با مشکلات عدیده‌ای روبرو شده و پیشرفت در این زمینه به کندی پیش رود. از دیگر چالش‌ها برای مدلسازی موضوعی روی متون فارسی، این است که زبان فارسی علاوه بر فاصله گذاری معمول در دیگر زبان‌ها که به عنوان جداساز کلمات استفاده می‌شود، فاصله‌ی دیگری به نام فاصله درون کلمه‌ای و یا شبه فاصله[۴] دارد. شبه فاصله‌ها دارای قوانین مدون و دقیقی نیستند.

بنابراین این پژوهش تلاش دارد تا با ارائه راهکاری بتواند مدلسازی موضوعی را روی مقالات فارسی انجام دهد. بنابراین منظور از متون در این پژوهش، اطلاعات کتابشناختی در دسترس از مقالات موجود در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری می‌باشد. این اطلاعات شامل چکیده، عنوان و کلیدواژه‌هاست.

---

[۱] Latent semantic analysis

[۲] Probabilistic latent semantic model

[۳] Latent Dirichlet allocation

[۴] Pseudo space

۱_۲ بیان مساله

با رشد چشمگیر اطلاعات در دنیای وب و عدم توانایی انسان در تحلیل و دسته‌بندی اسناد متنی، نیاز به روش‌های خودکار تحلیل متن در راستای سازمان‌دهی، فهمیدن و دسته‌بندی این اسناد به الزام تبدیل شده است. یکی از مهمترین منابع متنی که نیاز بسیار زیادی به تحلیل و بررسی دارد، مقالات علمی است، که نظر به اهمیت تحلیل و بررسی این متون، توجه بسیاری از پژوهشگران را در سال‌های اخیر به خود جلب کرده است.

روش‌های یادگیری ماشین با ارائه الگوریتم‌های قوی قادر هستند تا تحلیل داده را به صورت خودکار انجام دهند. مدلسازی موضوعی موقعیت ویژه‌ای را در میان روش‌های یادگیری ماشین دارد؛ چرا که این کلاس از مدل‌ها می‌تواند به صورت موثر داده‌های بزرگ را پردازش نماید. ایده اصلی مدلسازی موضوعی بر اساس این فرضیه استوار است که هر مجموعه‌ی اسناد از تعداد محدودی موضوع یا خوشه‌های معنایی تشکیل شده است؛ در حالی که هر کلمه و هر سند با احتمال متفاوتی به هر کدام از موضوعات تعلق دارد. این امر توانمندی بسیاری را به مدلسازی موضوعی برای خوشه کردن کلمات به وسیله‌ی موضوعات و موضوعات به وسیله‌ی اسناد به صورت همزمان می‌دهد. از مهمترین چالش‌های موجود در مدلسازی موضوعی، بدست آوردن تعداد موضوعات موجود در یک متن است؛ به صورتی که عملکرد نهایی مدل به این پارامتر وابسته است. پژوهش‌های پیشین عمدتا از روش گریدی برای بدست آوردن تعداد موضوعات موجود در یک متن استفاده کرده‌اند؛ بدین صورت که با تعریف یک معیار، عملکرد مدل موضوعی را روی متن با توجه به پارامترهای مختلف سنجیده، و در نهایت پارامتری که مدل با آن بهترین عملکرد را دارد، به عنوان تخمین از پارامتر مورد بررسی در نظر گرفته‌اند. علیرغم اینکه این روش می‌تواند تخمین مناسبی از پارامتر مورد بررسی و یا همان تعداد موضوعات موجود در یک متن ارائه دهد، ولی پیچیدگی زمانی بسیار بالایی دارد و کند است.

نظر به اهمیت مدلسازی موضوعی در پردازش متون و بالاخص مقالات فارسی، این پژوهش مدلسازی موضوعی را روی مقالات نشریات فارسی انجام می‌دهد تا موضوعات موجود در نشریات را بر اساس مقالات چاپ شده در آنها بدست آورد. چالش بزرگی که در این زمینه وجود دارد، بدست آوردن تعداد موضوعات موجود در نشریات است، که پژوهش حاضر نیز با بررسی و تحلیل یک روش جدید مبتنی بر نظریه بازبهنجاری و مقایسه آن با روش موجود، این مساله را مورد بررسی قرار می‌دهد.

۱_۳ ضرورت و اهمیت پژوهش

پردازش متون با استفاده از مدلسازی موضوعی مزایای فراوانی نسبت به پردازش متون با استفاده از واژه‌ها دارد که در ادامه، بعضی از این مزایا ارائه می‌گردد (رحیمی و همکاران، ۱۳۹۷):

- از آنجا که تعداد موضوعات به مراتب کمتر از تعداد واژگان یک متن است، بنابراین پردازش داده‌ها با این نوع مدلسازی، تعداد ابعاد کمتری نسبت به واژگان را دربرمی‌گیرد.
- انتقال معنی با استفاده از موضوعات یک متن بهتر از انتقال معنی با استفاده از واژگان متن می‌باشد.

بنابراین مدل‌های موضوعی می‌توانند در بسیاری از کاربردهای مرتبط با بازیابی اطلاعات[۱]، پردازش زبان‌های طبیعی[۲]، دسته‌بندی اسناد[۳] (Janani & Vijayarani, 2019) و همچنین خلاصه‌سازی متون[۴] (Abualigah et al., 2020) بسیار موثر باشند.

در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری نیز با توجه به یکی از ماموریت‌های این سازمان مبنی بر "افزایش سهولت و پایداری در دستیابی به منابع اطلاعاتی تولید شده در ایران و منطقه از طریق بازطراحی و کاربرپسندی سیستم ذخیره و بازیابی اطلاعات"، می‌توان به کاربر کمک کرد تا بتواند به صورت موثرتری اطلاعات و مقاله‌های مورد نیازش را بازیابی نماید. علاوه بر این، محصول این پژوهش به عنوان یک ابزار کمکی در راستای افزایش دقت در پردازش‌های بعدی زبان طبیعی می‌تواند بسیار موثر واقع گردد. با استفاده از مدلسازی موضوعی، می‌توان به کاربر نهایی کمک کرد تا بتواند به صورت موثرتری اطلاعات و مقاله‌های مورد نیازش را بازیابی نماید. علاوه بر این، به کاربر میانی نیز می‌تواند در یافتن موضوعات مقاله و نمایه‌سازی کمک کند.

۱_۴ اهداف پژوهش

- ارائه روشی به منظور تخمین تعداد موضوعات با توجه به مقالات فارسی موجود در نشریات
- بدست آوردن مدل موضوعی با توجه به مقالات فارسی موجود در نشریات

---

۱ Information retrieval

۲ Natural Language Processing

۳ Document clustering

۴ Text summarization

۱_۵ فرضیه‌ها و سوالات پژوهش

- کاهش ابعاد با استفاده از مدلسازی موضوعی با توجه به مقالات فارسی موجود در نشریات چه میزان می‌باشد؟
- مدلسازی موضوعی مقالات فارسی چه گروه‌هایی از موضوعات را ایجاد می‌کند؟

۱_۶ مروری بر ساختار گزارش

در ادامه در فصل ۲، مبانی نظری و مروری بر روش‌های پیشین خواهیم داشت. در فصل ۳ روش‌شناسی پژوهش مورد بررسی قرار می‌گیرد. نتایج و یافته‌های بدست آمده در فصل ۴ بیان می‌شوند و در نهایت نتیجه‌گیری در فصل ۵ آمده است.

فصل دوم

مبانی نظری و پیشینه پژوهش

۲. مبانی نظری و پیشینه پژوهش

۲_۱ مقدمه

صنعت تجزیه و تحلیل[۱]، به دنبال بدست آوردن اطلاعات از داده هست. با توجه به رشد سریع داده‌ها در سال‌های اخیر که اغلب به صورت غیرساخت‌یافته[۲] هستند، بدست آوردن اطلاعات مرتبط و دلخواه بسیار مشکل می‌شود؛ اما تکنولوژی، چندین روش قوی را توسعه داده است که با کمک آن می‌توان اطلاعات مناسبی را از داده استخراج کرد. یکی از آن تکنیک‌ها در حوزه متن‌کاوی، مدلسازی موضوعی است. همانطور که از نام آن برداشت می‌شود، در واقع فرآیند تعیین خودکار موضوعات ارائه شده در یک متن و استخراج الگوهای پنهان در متن است.

مدلسازی موضوعی با روش‌های متن‌کاوی مبتنی بر قواعد[۳] که از عبارات منظم[۴] یا تکنیک‌های جستجوی کلیدواژه مبتنی بر دیکشنری[۵] استفاده می‌کنند، متفاوت است. موضوعات می‌تواند به عنوان "الگوی تکرارشونده از کلمات هم‌وقوع در یک پیکره" تعریف شود و برای اهداف خوشه‌یابی متن، سازمان‌دهی بلاک‌های بزرگ داده متنی[۶]، بازیابی اطلاعات از متون بدون ساختار و انتخاب ویژگی بسیار مفید هستند.

مدلسازی موضوعی به صورت کلی به دو دسته مدل‌های احتمالی و غیراحتمالی تقسیم می‌شوند (شکل ۱). مدل-های احتمالی از یک توزیع احتمالی و به‌روز کردن پارامترهای آن در تکرارهای مختلف استفاده می‌کنند. روش‌های غیراحتمالی نیز در واقع روش‌های جبری فاکتورگیری ماتریس[۷] هستند. در ادامه، ابتدا مبانی نظری تعریف می‌شود و سپس هر کدام از این دسته‌ها و روش‌های آن مورد بررسی قرار خواهند گرفت.

---

۱ Analytics industry

۲ Unstructured

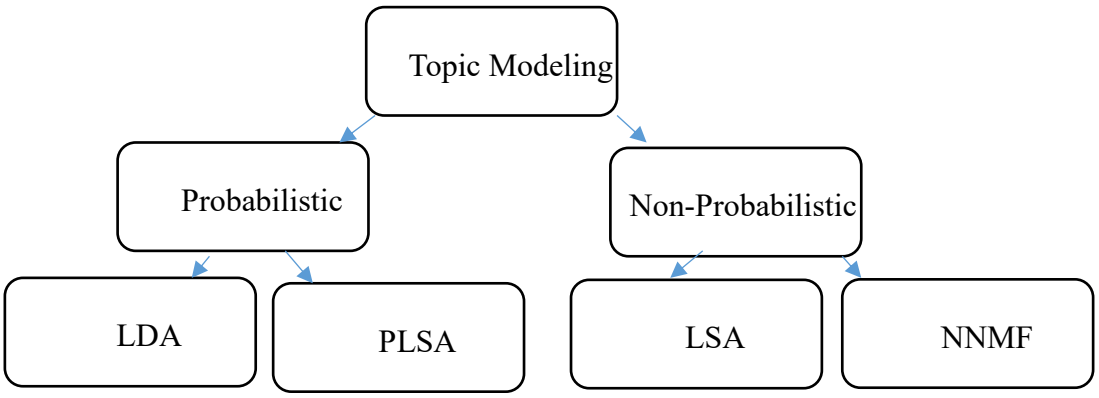
۳ Rule-base text mining

۴ Regular expression

۵ Dictionary based keyword searching techniques

۶ Organizing large blocks of textual data

۷ Matrix factorization



شکل ۱: طبقه‌بندی مدلسازی موضوعی

۲_۲ مبانی نظری

در ادامه این فصل، تعاریف و مبانی نظری مورد استفاده در این پژوهش ارائه می‌گردد.

۲_۲_۱ متن‌کاوی

بشر با پیشرفت فناوری رایانه‌ای در ثبت و ذخیره‌سازی داده‌ها و پردازش آنها گامی بزرگ جهت کسب دانش برداشته است. در واقع داده، نمایشی از واقعیت‌ها، معلومات، مفاهیم، رویدادها یا پدیده‌ها برای برقراری ارتباط، تفسیر یا پردازش توسط انسان یا ماشین است. از طرف دیگر، اطلاعات داده‌هایی هستند که پس از جمع‌آوری پردازش شده و شکل مفهومی تولید کرده‌اند. اصلی‌ترین دلیلی که باعث شد متن‌کاوی کانون توجهات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌های متنی است و نیاز شدید به اینکه از این داده‌های متنی، اطلاعات و دانش سودمند بتوان استخراج کرد (Feldman et al., 1995).

بین داده‌ها و اطلاعات همانند خبر و اطلاع رابطه وجود دارد. خبری که دریافت می‌شود، پس از ارزیابی به اطلاع تبدیل می‌شود. داده‌ها نیز پردازش می‌شوند تا اطلاعات را پدید آورند. به بیان دیگر، اطلاع حاصل تکامل داده‌ها است. به این ترتیب بین داده‌ها و اطلاعات یک شکاف وجود دارد که اندازه این شکاف با حجم داده‌ها ارتباط مستقیم دارد. هر چه داده‌ها حجیم‌تر باشند این شکاف بیشتر خواهد بود و هر چه حجم داده‌ها کمتر و روش‌ها و

ابزار پردازش داده‌ها کاراتر باشد، فاصله بین داده‌ها و اطلاعات کمتر است. امروزه افزایش سریع حجم داده‌ها به شکلی است که توانایی انسان برای درک این داده‌ها بدون ابزارهای پرقدرت میسر نمی‌باشد (عظیمی و شمس، ۱۳۹۴).

متن‌کاوی به معنای استخراج اطلاعات ارزشمند از حجم عظیم داده متنی است. با توجه به نوع داده و همچنین حجم آن، مشخص نیست که چه اطلاعات گرانبهایی در عمق این داده‌های متنی وجود دارد و تنها با کاوش در این داده‌هاست که می‌توان به این اطلاعات دسترسی پیدا کرد. بنابراین وظیفه اصلی متن‌کاوی، کاویدن و استخراج دانش از منابع عظیم داده متنی می‌باشد؛ تا اطلاعات ارزشمندی که در حجم انبوهی از اطلاعات سطحی پنهان شده است، آشکار گردد. متن‌کاوی تلاش برای استخراج دانش از انبوه داده‌های متنی موجود است که به کمک مجموعه‌ای از روش‌های آماری و مدلسازی می‌تواند الگوها و روابط پنهان موجود در داده‌های متنی را تشخیص دهد. تحلیل متن[۱] اصطلاحی است که گاها به جای متن‌کاوی استفاده می‌شود که آن هم به فرآیند تبدیل داده‌های متنی غیرساخت‌یافته به اطلاعات با معنا اطلاق می‌شود. برای تحلیل متن و یا به عبارتی متن‌کاوی نیازمند الگوریتم‌های یادگیری ماشین می‌باشیم (Hotho et al., 2005).

۲_۲_۲ یادگیری ماشین[۲]

یادگیری ماشین به عنوان یکی از زیرشاخه‌های وسیع و پرکاربرد هوش مصنوعی[۳]، کامپیوتر را قادر به یادگیری از داده‌ها می‌کند. به عبارت دیگر، یادگیری ماشینی به تنظیم و اکتشاف الگوریتم‌هایی می‌پردازد که با توجه به آنها ماشین می‌تواند یاد بگیرد. اگر یادگیری انسان را با وجود یک عامل و با تعامل با محیط بیرونی در نظر بگیریم، یادگیری ماشین با نوشتن برنامه، نمایش مثال‌های متعدد، تجربه‌ی محیط واقعی، مشاهده و بازخورد صورت می‌گیرد. زمانی الگوریتم‌های یادگیری ماشین، تاثیر خود را پررنگ‌تر نشان می‌دهند که مسئله توصیف‌ناپذیر باشد و همچنین در طول زمان تغییر کند. بر اساس نحوه یادگیری، الگوریتم‌های یادگیری ماشین به دسته‌های مختلفی تقسیم‌بندی می‌شوند مانند یادگیری بانظارت[۴]، یادگیری بدون نظارت[۵]، یادگیری تقویتی[۶]، یادگیری عمیق[۷]؛ و غیره که در این پژوهش، از رویکرد مبتنی بر یادگیری بدون نظارت بهره گرفته می‌شود.

---

۱ Text analysis

۲ Machine learning

۳ Artificial intelligence

۴ Supervised learning

۵ Unsupervised learning

۶ Reinforcement learning

۷ Depp learnng

۲_۲_۳ یادگیری بدون نظارت

در این نوع شیوه یادگیری، هیچ داده آموزشی[۱] وجود ندارد. به عبارت دیگر، یادگیری بر روی داده‌های بدون برچسب[۲] به منظور یافتن الگوهای پنهان صورت می‌پذیرد. یکی از معروف‌ترین الگوریتم‌ها در این دسته، خوشه-یابی[۳] با هدف دسته‌بندی داده‌ها به گروه‌های مختلف است که به هر کدام از این دسته‌ها یک خوشه[۴] گفته می‌شود. مدل‌های موضوعی نیز از این شیوه یادگیری تبعیت می‌کنند.

۲_۲_۴ متغیر تصادفی[۵]

در آمار و احتمالات، متغیر تصادفی تابعی از فضای نمونه به مجموعه اعداد حقیقی می‌باشد که به هر نقطه از فضای نمونه، یک عدد حقیقی را نسبت می‌دهند. به عبارت دیگر، هر پیشامد از فضای پیشامد را به یک عدد حقیقی نگاشت می‌کند.
متغیرهای تصادفی به دو نوع متغیر تصادفی گسسته[۶] و پیوسته[۷] تقسیم می‌شوند. در صورتی یک متغیر تصادفی، گسسته است که تابع احتمال آن شرایط زیر را داشته باشد:

۱) به ازای هر x داشته باشیم: $f(x) \geq 0$
۲) $\sum_x f_x(x) = 1$
۳) $f_X(x) = P(X = x)$

برای اینکه یک متغیر تصادفی، پیوسته باشد، باید شروط زیر را داشته باشد:

۱) به ازای هر x داشته باشیم: $f_X(x) \geq 0$
۲) $\int_{-\infty}^{+\infty} f_X(x) = 1$
۳) $P(a < x < b) = \int_a^b f_X(x)dx$

---

[۱] train
[۲] Unlabeled data
[۳] Clustering
[۴] Cluster
[۵] Random variable
[۶] Discrete random variable
[۷] Continoues random variable

از مهم‌ترین توزیع‌های گسسته می‌توان به توزیع برنولی[۱]، توزیع دوجمله‌ای[۲]، دوجمله‌ای منفی[۳]، چندجمله‌ای[۴]، هندسی[۵] و پواسن[۶] اشاره کرد. از توزیع‌های پیوسته مهم نیز می‌توان تابع توزیع گاما[۷]، توزیع بتا[۸] و دیریکله[۹] را نام برد. با توجه به اینکه در این پژوهش از توزیع دیریکله استفاده شده است و این توزیع به توزیع‌های گاما و بتا وابسته است، در ادامه به تعریف این توزیع‌ها می‌پردازیم. البته قبل از آن، تابع توزیع احتمال و انواع آن توضیح داده می‌شود.

۲_۲_۵ تابع توزیع احتمال[۱۰]

با در نظر گرفتن یک متغیر تصادفی پیوسته، توزیع احتمالی یا تابع چگالی احتمال متغیر X تابعی است که به ازای هر دو عدد a و b به صورتی که $a \leq b$، داشته باشیم: $P(a \leq X \leq b) = \int_a^b f(x)dx$.

۲_۲_۶ تابع توزیع توأم[۱۱]

زمانی که متغیر تصادفی، چندبعدی[۱۲] باشد، در آن صورت تابع توزیع احتمال آن بر اساس تابع توزیع احتمال مولفه و ارتباطی که میان آنها برقرار است، تعریف می‌گردد. به عنوان مثال فرض کنید که $Z(X,Y)$ یک متغیر تصادفی دوبعدی باشد، در این صورت به تابع توزیع احتمال آن، تابع توزیع توأم نیز گفته می‌شود.

---

۱ Bernoulli distribution
۲ Binomial distribution
۳ Negative binomial distribution
۴ Multi-nomial distribution
۵ Geometric distribution
۶ Poisson distribution
۷ Gamma distribution
۸ Beta distribution
۹ Dirichlet distribution
۱۰ Probability density function
۱۱ Joint Density Function
۱۲ Multi-dimensional

۲_۲_۷ تابع توزیع حاشیه‌ای[۱]

تابع احتمال برای هر یک از مولفه‌های متغیر تصادفی را تابع توزیع حاشیه‌ای می‌گویند. در صورتی که مولفه‌های هر بعد از متغیر تصادفی از یکدیگر مستقل باشند، تابع چگالی توأم را می‌توان به صورت حاصلضرب تابع توزیع‌های حاشیه‌ای نوشت. در این صورت برای n متغیر تصادفی مستقل، تابع توزیع توأم بر اساس تابع توزیع‌های حاشیه‌ای به صورت $f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)=f_{X_1}(x_1)f_{X_2}(x_2)\dots f_{X_n}(x_n)$ تعریف می‌گردد.

۲_۲_۸ تابع توزیع گاما

این تابع توزیع را می‌توان با پارامتر شکل[۲] (α)و پارامتر معکوس-مقیاس[۳] (β) نشان داد. اگر α عددی طبیعی باشد، آنگاه توزیع گاما معادل است با مجموع α متغیر تصادفی با توزیع نمایی با پارامتر $\frac{1}{\beta}$. اگر X یک متغیر تصادفی با توزیع گاما باشد، آن را به صورت $X\sim\Gamma(\alpha,\beta)$ نشان می‌دهند و تابع چگالی احتمال برای این متغیر تصادفی بر اساس پارامترهای α و β به صورت زیر است:

$$f(x;\alpha,\beta)=\frac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \qquad (۲-۱)$$

که در آن $x>0$ و α و β مقدارهایی مثبت و $\Gamma(\alpha)$ نیز مقدار تابع گاما در نقطه α است.

$$\Gamma(z)=\int_{x=0}^{\infty}x^{z-1}e^{-x}dx \qquad (۲-۲)$$

هر گاه α یک عدد صحیح و مثبت مانند n باشد، می‌توان از توزیع گاما برای تخمین مدت زمان لازم برای روی دادن n پیشامد استفاده کرد.

[۱] Marginal Density Distribution

[۲] Shape parameter

[۳] Inverse-scale parameter

۲_۲_۹ تابع توزیع بتا

یکی از توزیع‌های احتمال پیوسته که در بازه صفر و یک تعریف می‌شود و به توزیع گاما نیز مرتبط است، توزیع بتا می‌باشد. این توزیع نیز همانند توزیع گاما دارای دو پارامتر است و به صورت زیر تعریف می‌گردد:

$$f(x;\alpha,\beta)=\frac{1}{Beta(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \qquad (۲-۳)$$

که تابع $Beta(\alpha,\beta)$ به صورت زیر تعریف می‌شود:

$$Beta(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad (۲-۴)$$

۲_۲_۱۰ توزیع دیریکله

این توزیع در حالت کلی، حالت گسترش‌یافته توزیع بتا برای توابع چندمتغیره می‌باشد. با توجه به اینکه این توزیع به صورت چند متغیره است، تکیه‌گاه آن نیز به صورت یک بردار نمایش داده می‌گردد. فرض کنید متغیر تصادفی X دارای k بعد باشد. در این صورت، تکیه‌گاه متغیر تصادفی دیریکله به صورت زیر تعریف می‌شود:

$$x_1,x_2,\dots,x_k \qquad x_i \in (0,1) \qquad \sum_{i=1}^{k}x_i=1$$

همانطور که روابط بالا نشان می‌دهند، متغیر تصادفی دیریکله، k بُعدی بوده و مجموع مقادیر مولفه‌های آن برابر با یک است. همچنین مقادیر بردار متغیر آن در بازه صفر و یک تغییر می‌کند. به این ترتیب متغیر تصادفی X با این ویژگی‌ها، متغیر تصادفی با توزیع دیریکله از مرتبه k نامیده می‌شود. لازم به ذکر است که مقدار k در محدوده اعداد طبیعی تغییر می‌کند و معمولا مقدار k را بزرگتر یا مساوی با ۲ در نظر می‌گیرند. در صورتی که k برابر با یک باشد، این توزیع با توزیع بتا یکسان خواهد شد.

متغیر تصادفی $X=(X_1,X_2,\dots,X_k)$ را توزیع دیریکله گویند، اگر تابع احتمال (تابع چگالی احتمال) آن به صورت زیر تعریف شود:

$$f(x_1,x_2,\dots,x_k;\alpha_1,\alpha_2,\dots,\alpha_k)=\frac{1}{Beta(\alpha)}\prod_{i=1}^{k}x_i^{\alpha_i-1} \qquad (۲-۵)$$

در این صورت می‌گوییم که متغیر تصادفی X دارای توزیع دیریکله مرتبه k با پارامتر $\alpha=(\alpha_1,\alpha_2,\dots,\alpha_k)$ است.

۲_۲_۱۱ تجزیه ماتریسی

تجزیه ماتریسی ابزاری برای تحلیل داده‌ها است. در صورتی که داده در یک ماتریس ذخیره و سپس مورد تجزیه قرار گیرد، می‌توان به اطلاعات بسیار مفیدی رسید. به عبارت دیگر، هر تجزیه، تعبیرهای مختلفی از داده که به صورت ضمنی در آن نهفته است، آشکار می‌نماید. تجزیه ماتریسی، یک نمایش جدید از داده را بوجود می‌آورد که با توجه به پیچیدگی موجود در داده، این تجزیه منجر به نمایش ساده‌ای از داده اولیه می‌شود که علاوه بر اینکه اطلاعات ضمنی موجود در داده را کشف می‌کند، با حذف نویز و داده‌های کم‌اهمیت، آن را پاکسازی نیز می‌نماید (یوسفی و رزقی، ۱۳۹۵). از روش‌های معروف در تجزیه ماتریسی، می‌توان به تجزیه مقدار منفرد[۱]، تحلیل مولفه اصلی[۲] و تجزیه نامنفی ماتریسی[۳] اشاره کرد.

۲_۲_۱۲ تجزیه مقدار منفرد

به صورت کلی این روش، ماتریس را به سه ماتریس دیگر تجزیه می‌کند. فرض کنید که A یک ماتریس با رتبه r باشد ($A \in \mathbb{R}^{m\times n}$). در این صورت A را می‌توان به سه ماتریس U، V و ψ تجزیه کرد؛ به صورتیکه رابطه $A = U\psi V^T$ بین آنها برقرار باشد. در این رابطه، U و V ماتریس‌های متعامد[۴] به ترتیب با ابعادهای $m \times m$ و $n \times n$ و همچنین ψ یک ماتریس قطری[۵] با ابعاد $m \times n$ است.

۲_۲_۱۳ تجزیه نامنفی ماتریسی

در حالت کلی، این مدل برای تقریب داده‌های نامنفی ذخیره شده در ماتریس نامنفی $A \in \mathbb{R}^{m\times n}$ به دنبال تولید دو ماتریس نامنفی دیگر مانند $W \in \mathbb{R}^{m\times n}$ و $H \in \mathbb{R}^{m\times n}$ با شرط $r \ll \min\{m,n\}$ است؛ به صورتی که معادله تقریبی $A \sim WH$ برقرار باشد. یافتن چنین تقریبی، نیازمند تابع هزینه‌ای است که کیفیت تقریب را به خوبی نشان دهد. یکی از این تابع‌ها اندازه فاصله دو ماتریس نامنفی از یکدیگر است که معمولا به صورت مساله بهینه‌سازی زیر مدل می‌شود:

---

[۱] Single value decomposition

[۲] Principle component analysis

[۳۳] Non-negative matrix factorization

[۴] Orthogonal

[۵] Diagonal

$$\min_{\substack{W\geq 0\\ H\geq 0}} f(W,H) \equiv \frac{1}{2}||A-WH||_F^2 \qquad (۲-۶)$$

لازم به ذکر است که ماتریس‌های W و H نامنفی هستند.

۲_۲_۱۴ نمونه‌گیری گیبز[۱]

فرض کنید که به یک نمونه تصادفی به صورت $X=(x_1,x_2,\dots,x_n)$ از تابع توزیع توأم $p(x_1,x_2,\dots,x_n)$ نیاز است. با فرض اینکه $X^{(i)}=(x_1^{(i)},\dots,x_n^{(i)})$ نشان‌دهنده نمونه تصادفی در مرحله iم باشد، نمونه‌بردای گیبز بدین صورت انجام می‌شود که مقدار اولیه و اختیاری برای $X(1)$ به صورت تصادفی انتخاب می‌شود. در مرحله بعد، نمونه جدید و یا همان $X(i+1)$ (i نشان‌دهنده شماره مرحله است)، تولید می‌شود. از آنجا که هر نمونه چندمتغیری است، بنابراین نمونه جدید بر اساس یک بردار ایجاد می‌شود. به عبارت دیگر نمونه جدید به صورت $X^{(i+1)}=(x_1^{(i+1)},x_2^{(i+1)},\dots,x_n^{(i+1)})$ تعریف می‌گردد و تولید آن بر اساس نمونه‌های دیگر انجام می‌گیرد. تابع احتمال برای انتخاب نمونه برای هر یک از مولفه‌ها مانند $x_j^{(i+1)}$ را به وسیله تابع توزیع شرطی آن مولفه بر حسب دیگر مولفه‌ها در نظر می‌گیریم. این احتمال یا توزیع شرطی به صورت $p(x_j^{(i+1)}|x_1^{(i+1)},\dots,x_{j-1}^{(i+1)},x_{j+1}^{(i)},\dots,x_n^{(i)})$ نمایش داده می‌شود. بنابراین نمونه‌های استفاده شده برای محاسبه احتمال شرطی برای مولفه j شامل نمونه i+1م بر حسب مولفه اول تا j-1 به همراه مولفه‌های j+1 تا n نمونه iم است. بنابراین نمونه‌برداری گیبز به صورت مرحله‌ای و تکراری است و از بیشترین اطلاعات برای تعیین توزیع و یا تولید نمونه تصادفی در گام بعدی استفاده می‌شود.

۲_۳ پیشینه پژوهش

همانطور که شکل ۱ نشان می‌دهد، مدلسازی موضوعی به صورت کلی به دو دسته مدل‌های احتمالی و غیراحتمالی تقسیم می‌شود. روش‌های غیراحتمالی در واقع روش‌های جبری فاکتورگیری ماتریس هستند که در ابتدا در سال ۱۹۹۰ با مفهوم آنالیز پنهان مفهومی و یا LSA (Deerwester et al., 1990) و تجزیه نامنفی ماتریسی و یا NNMF (Lee & Seung, 2001) مطرح شدند. هر دو روش‌های LSA و NNMF مبتنی بر روش کیسه کلمات هستند؛ به صورتی که در هر دو پیکره به ماتریس اصطلاح-سند[۲] تبدیل می‌شود. مدل‌های احتمالی به وجود آمدند تا مدل‌های جبری مثل آنالیز پنهان مفهومی را بهبود بخشند (Blei, 2012).

[۱] Gibbs Sampling

[۲] Term-document matrix

آنالیز پنهان مفهومی و یا LSA که اولین بار در سال ۱۹۹۷ توسط لاندر[۱] و دومایس[۲] ارائه گردید (Deerwester, 1990)، یک روش جبری بر اساس تجزیه مقدارهای منفرد و یا به اختصار SVD می‌باشد. SVD در حوزه‌های مختلفی از جمله بازیابی اطلاعات، پردازش زبان طبیعی و مدل کردن دانش زبان انسانی به کار گرفته شده است (Kherwa & Bansal, 2017). LSA برای یافتن ارتباط معنایی میان اسناد و کلماتی که در آنها وجود دارد، از فرضیه توزیع[۳] استفاده می‌کند (Dudoit et al., 2002). این فرضیه می‌گوید که اصطلاح‌ها با معنی مشابه در متون بسیاری در کنار یکدیگر وجود دارند. همه روابط معنایی میان متون مستقیما از پیکره داده شده استخراج می‌گردد. این روش از نحوه نمایش برداری استفاده می‌کند تا بتواند با محاسبه شباهت میان متون، کلمات شبیه را بیاید و به صورت معنایی متن را درون خوشه‌های معنایی سازمان‌دهی نماید. آنالیز پنهان مفهومی کاربردهای زیادی در کاوش متن از جمله بازیابی اطلاعات، آنالیز شبکه‌های اجتماعی[۴] و خلاصه‌سازی متون[۵] دارد.

نحوه کار LSA بدین صورت است که ابتدا یک ماتریس به اندازه کلمات در تعداد اسناد ساخته می‌شود. سپس با استفاده از روش تجزیه مقدارهای منفرد، ابعاد ماتریس کاهش پیدا می‌کند. این کار مزایای زیادی به همراه دارد. اولا اسناد و همچنین لغاتی که اهمیت کمی در مجموعه اسناد دارند، حذف می‌شوند. ثانیا ارتباطات معنایی میان لغات هم‌معنی کشف می‌گردد (Deerwester, 1990).

آنالیز پنهان مفهومی احتمالی و یا به اختصار PLSA که کاربردهای بسیار زیادی در زمینه بازیابی اطلاعات، یادگیری ماشین، فیلترینگ، و پردازش زبان طبیعی دارد، یک تکنیک آماری نوین برای بررسی داده‌های هم‌رخداد به صورت احتمالی است (Hofmann, 2013). مدل PLSA از یک مدل آماری تحت عنوان مدل منظر[۶] استفاده می‌کند و متغیرهای کلاس پنهان (غیرقابل مشاهده) $z \in Z = \{z_1, z_2, \dots, z_r\}$ را با متغیرهای قابل مشاهده که در این کاربرد، اسناد $d \in D = \{d_1, d_2, \dots, d_n\}$ هستند و همچنین کلمات $w \in W = \{w_1, w_2, \dots, w_m\}$ می‌باشند، مرتبط می‌سازد. این مدل فرض می‌کند که بردار W و D شرطی مستقل هستند. لازم به ذکر است که تعداد متغیرهای پنهان Z (که در اینجا موضوعات می‌باشد) کمتر از تعداد متون و کلمات می‌باشد. مدل احتمالاتی مشترک روی کلمات و متون با رابطه احتمالاتی زیر تعریف می‌گردد:

$$P(w_i, d_j) = p(d_j) \sum_{z_k \in Z} P(w_i \mid z_k) P(z_k \mid d_j) \qquad (۷-۲)$$

برای بدست آوردن یک تخمین از بیشترین مقدار احتمالاتی وقوع در مدل‌های متغیر پنهان از الگوریتم ماکزیمم انتظار[۷] یا به اختصار EM استفاده می‌شود.

---

[۱] Landauer

[۲] Dumais

[۳] Distributional hypotheses

[۴] Social network analysis

[۵] Text summarization

[۶] Aspect Model

[۷] Expected Maximization

تخصیص پنهان دیریکله و یا به اختصار LDA روشی بر اساس تئوری definneti (De Finetti, 2017) می‌باشد که ساختار آماری درونی و میانی سند را از طریق توزیع توأم در نظر می‌گیرد. این روش فرض می‌کند که هر سند حاوی چندین موضوع می‌باشد و هر موضوع به عنوان توزیعی روی واژگان تعریف می‌گردد. فرض کنیم که پیکره از k موضوع تشکیل شده باشد، که هر سند موجود در آن، به یکی از این k موضوع با احتمالات متفاوت تعلق داشته باشد. به عبارت دیگر کلماتی که مربوط به یک موضوع هستند، در آن موضوع دارای احتمال بالایی هستند. لازم به ذکر است که این روش فرض می‌کند که موضوعات از قبل مشخص است.

هدف مدل‌سازی موضوعی این است که این موضوعات را از داده‌ها یا پیکره یاد بگیرد. LDA بر اساس مدل متغیری پنهان می‌باشد که در حوزه یادگیری ماشین از چندین سال قبل استفاده شده است. مدل مولدی[۱]، ترتیب[۲] کلمات در تولید اسناد را در نظر نمی‌گیرد؛ بنابراین بر اساس روش کیف کلمات (BOW) می‌باشد. مدلسازی موضوعی به صورت خودکار عناوین را از مجموعه‌ای از اسناد کشف می‌کند؛ به صورتی که K توزیع موضوع باید از استنتاج آماری روی داده‌ها استخراج گردد. این الگوریتم فرآیند مولدی را به عنوان توزیع توأم احتمالی[۳] روی هر دو متغیرهای مشاهده شده[۴] و مخفی[۵] تعریف می‌کند (Kherwa & Bansal, 2020).

تجزیه نامنفی ماتریسی و یا به اختصار NNMF کاربردهای بسیار زیادی در حوزهای کاهش ابعاد، تشخیص الگو، پردازش تصویر و مدلسازی زبان دارد (Kherwa & Bansal, 2020). این روش برای تقریب داده‌های نامنفی ذخیره شده در ماتریس منفی $A \in R^{m \times n}$، دو ماتریس نامنفی $W \in R^{m \times r}$ و $H \in R^{r \times n}$ را تولید می‌کند؛ به شرطی که $r << \min\{m,n\}$ و $A \approx WH$. هر تجزیه تعبیرهای مختلفی را از ساختار ضمنی داده‌ها آشکار می‌کند که این تعبیرها از نظر ریاضی هم‌ارز هستند. بنابراین از این روش می‌توان برای مدلسازی موضوعی استفاده کرد؛ به صورتی که با تجزیه ماتریس، موضوعات یا اطلاعات نهفته در متن کشف گردد.

پژوهش‌های دیگری هم در سال‌های اخیر برای مدلسازی موضوعی ارائه شده است که ارتباط میان واژگان را در سطحی محلی‌تر بررسی می‌کنند. نخستین روشی که در این دسته ارائه گردید و به نحوی الهام‌بخش دیگر روش‌های این دسته بود، BTM می‌باشد. در این روش فرض می‌شود که هر واژه علاوه بر موضوع خود، وابسته به موضوع واژه پیشین خود نیز هست. این دیدگاه در مدل (Barbieri et al., 2013) نیز بکار گرفته شد. گریفیس و همکارانش فرض کردند که هر واژه توسط یک موضوع و یا توسط واژه پیشینش تولید می‌گردد که برای انتخاب یکی از این دو حالت، از یک متغیر برنولی استفاده کردند (Griffiths et al., 2007). تعمیمی بر این روش، توسط ونگ و همکارانش انجام گردید که در آن هر واژه بر مبنای موضوع خود می‌تواند تصمیم بگیرد که آیا با واژه قبلی یک ترکیب را تشکیل دهد یا خیر (Wang et al., 2007). یک مدل موضوعی احتمالاتی مبتنی بر روابط محلی واژگان

---

[۱] Generative model

[۲] order

[۳] Joint probability distribution

[۴] Observed variables

[۵] Hidden variables

در پنجره‌های همپوشان توسط رحیمی و همکاران ارائه گردید (۱۳۹۷). در پژوهشی دیگر فرض مشابهی در نظر گرفته شد و علاوه بر آن، فرض شد که یک سلسله مراتب از موضوعات وجود دارد و هر واژه، مسیری مشخص را در این سلسله مراتب طی می‌کند تا توسط یک موضوع خاص تولید گردد (Yang et al., 2015). یک مدل موضوعی باناظر در ترکیب با مدل زبانی بایگرام توسط جمیل و همکاران ارائه شد (Jameel et al., 2015). دلیل استفاده از از بایگرام در این مدل‌ها، تنک بودن داده است.

پژوهش‌های دیگری نیز ارائه شده‌اند که برخلاف روش‌های قبلی محدود به واژه پیشین نبوده و نتایج آنها برای ترکیباتی با طول‌های متفاوت گزارش شده است (Noji et al., 2013) و (Sato & Nakagawa et al., 2010). با توجه به تنک بودن داده‌ها، این پژوهش‌ها می‌بایست روی مجموعه داده‌های بسیار بزرگ آموزش داده شوند تا بتوان به نتایج ارائه شده اعتماد کرد. با توجه به بار محاسباتی بسیار زیاد این روش‌ها، برای مجموعه داده‌های بسیار بزرگ، با مشکل روبرو می‌شوند و غیرعملی می‌شوند.

زمانی و همکاران (۱۳۹۳) با استفاده از روش آنالیز معنایی پنهان احتمالاتی که از منابع دانش محدود از محتویات متون و حذف کلمات زاید و غیرمفید حاصل شده است، به دسته‌بندی متون پرداختند. آنها از توابع ریاضی موجود در PLSA استفاده کردند تا حجم بار محاسباتی را تا حدودی کاهش دهند.

در پژوهشی دیگر از مدل تخصیص پنهان دیریکله به عنوان یک روش آنالیز معنایی، برای استخراج ویژگی در دسته‌بندی اسناد استفاده شده است. از مشخصه‌های اصلی مدل ارائه شده، محاسبه احتمال عنوان بودن کلمات است که بر اساس تعداد تکرار کلمه مورد نظر با دیگر کلمات محاسبه می‌شود. در نهایت از الگوریتم فراابتکاری ژنتیک برای خوشه‌بندی نهایی استفاده گردیده است (شکری و معصومی، ۱۳۹۵).

اگرچه این روش‌ها عملکرد نسبتا خوبی داشتند، ولی در عمل از نتایج مدلسازی موضوعی برای دسته‌بندی متون استفاده نمودند؛ نه اینکه یک روش جدید برای مدلسازی موضوعی ارائه داده باشند. به عنوان نمونه دیگری از پژوهش‌هایی که از مدلسازی موضوعی استفاده می‌کند، می‌توان به پژوهش انجام شده توسط دامی و طاهرزاده (۱۳۹۶) اشاره کرد. آنها به استخراج اطلاعات معنی‌دار از پیام‌های ورودی با کمک مدل LDA و الگوریتم یادگیری ماشین بردار پشتیبان (SVM) برای شناسایی تهدیدهای امنیتی پرداختند.

ونگ[۱] و بلی[۲] (۲۰۱۱)، نیز از مدلسازی موضوعی احتمالی به منظور طراحی سیستم توصیه مقاله به کاربران استفاده کردند. از آنجا که یکی از مشکلات موجود در روش‌های فیلترینگ مشارکتی، عدم توصیف‌پذیری این روش‌هاست، آنها از مدلسازی موضوعی به منظور توصیف‌پذیرشدن پیشنهاد مقاله به کاربران بر اساس توزیع موضوعی استفاده نمودند.

دامی و الیکایی (۱۳۹۶) نیز یک الگوریتم مدلسازی موضوعی برای رویدادهای خبری را ارائه دادند که بر اساس یادگیری عمیق افزایشی عمل می‌کند. آنها یک چارچوب سه مرحله‌ای و مقیاس‌پذیر مبتنی بر یادگیری عمیق ارائه دادند که برای یادگیری و اطلاع از یک سلسله مراتب از رویدادها در مورد یک موضوع به کار می‌رود و بر

[۱] Wang

[۲] Blei

اساس رویدادهایی است که به محض وقوع، مرتبط با آن موضوع باشد. روش ارائه شده توسط آنها فقط می‌تواند روی رویدادهای خبری عمل کند.

فصل سوم

روش‌شناسی پژوهش

۳. روش‌شناسی پژوهش

۳_۱ مقدمه

امروزه اطلاعات نقش مهمی در تحلیل، تصمیم‌گیری و مدیریت دارد. از طرف دیگر، بخش قابل توجهی از اطلاعات موجود در پایگاه‌های داده به صورت متن ذخیره شده‌اند که این داده‌ها می‌بایست مورد پردازش قرار گرفته تا اطلاعات مهم از آنها استخراج گردد. از این رو نیاز به تکنیک‌های مختلف متن‌کاوی است. منظور از پردازش زبان طبیعی در متن‌کاوی، قابل دسترس کردن زبان طبیعی برای ماشین است. پردازش زبان طبیعی، اسناد متنی بدون ساختار را دریافت و در نهایت به شکلی ساختاریافته تبدیل می‌کند. در این حالت، امکان استخراج اطلاعات از این اسناد وجود دارد. پردازش‌های متن شامل تحلیل‌های صرفی و نحوی و معنایی متن ورودی است. یکی از تکنیک-های موثر در این راستا، مدلسازی موضوعی است که در واقع یک روش بدون نظارت برای یافتن و مشاهده گروهی از کلمات (موضوعات) می‌باشد.

قدم اول برای انجام هر پردازشی که در حوزه پردازش زبان طبیعی انجام می‌گیرد، جمع‌آوری و پیش‌پردازش داده‌ها است؛ چرا که قبل از پردازش زبان طبیعی، می‌بایست عمل پیش‌پردازش روی متن به منظور تمیزسازی داده‌ها انجام گیرد. بنابراین در این فصل، بعد از معرفی جامعه آماری پژوهش، به نحوه تمیزسازی داده‌ها پرداخته می‌شود. سپس مدل LDA به عنوان مدل پایه مورد بررسی قرار می‌گیرد و در انتها روش بدست آوردن تعداد موضوعات توضیح داده می‌شود.

۳_۲ جامعه آماری پژوهش

در فاز جمع‌آوری داده، تعداد ۱۰ نشریه به تصادف از نشریات وزارت علوم[۱] انتخاب شدند. این نشریات عبارتند از مکانیک هوافضا، زمین‌شناسی ایران، مطالعات باستان‌شناسی، مطالعات مدیریت، فقه و اصول، مهندسی برق و مهندسی کامپیوتر ایران، روش‌های عددی در مهندسی، سبک‌شناسی نظم فارسی، رهیافتی نو در مدیریت آموزشی، و صفه. سپس از هر نشریه، ۲۰۰ مقاله که در سال‌های اخیر در نشریه به چاپ رسیده‌اند، با نمونه‌گیری تصادفی انتخاب شده و اطلاعات کتابشناسی آنها که شامل عنوان مقاله، چکیده و کلیدواژه است، استخراج گردید. بنابراین در کل ۲۰۰۰ مقاله مورد بررسی قرار گرفت که جدول ۱ این اطلاعات را نشان می‌دهد. لازم به ذکر است که تمامی این اطلاعات، بر اساس وب‌سایت نشریات وزارت علوم نوشته شده، و همچنین اطلاعات کتابشناختی مقالات این نشریات از پایگاه رایسست[۲] گرفته شده است.

جدول ۱: اطلاعات کتاب‌شناختی نشریات

شماره	عنوان نشریه	موضوع اصلی	موضوع فرعی	اهداف محور جذب مقالات نشریه
۱	مکانیک هوافضا	فنی و مهندسی	مکانیک	فناوری در سامانه‌های هدایت، کنترل و ناوبری وسایل هوافضایی شامل الگوریتمهای هدایت و کنترل، عملگرهای کنترلی، سامانه‌های ناوبری، موقعیت‌یابی، تعیین وضعیت، کنترل وضعیت، پردازش داده و سیگنال، شبیه‌سازهای پرواز، آزمایشگاه‌های واقعیت مجازی و محیط‌های پرواز، الگوریتم‌های تشخیص و جبران خطا، فناوری در سازه و مکانیزم‌های هوافضایی شامل سازه‌های هوشمند، سامانه‌های آشکارسازی سلامت سازه، سامانه‌های فعال و غیرفعال کاهش نویز و ارتعاش سازه، فناوری طراحی سازه تحت اثر ضربه‌های مکانیکی، بارگذاری آیرودینامیک و اکوستیکی، پردازش ارتعاشات سازه و سیگنال‌های اکوستیکی، فناوری در سامانه‌های پیشرانش شامل موتورهای جت، توربوجت، توربوفن، موتورهای سوخت جامد، مایع و ترکیبی، موتورهای احتراق داخلی و توربوماشین‌ها، پیشران‌های الکتریکی، گاز سرد و استفاده از فناوری‌های مرتبط با انرژیهای تجدیدپذیر، کاهش مصرف سوخت و آلاینده‌های زیست محیطی، فناوری در مواد و کاربرد آن در مهندسی هوافضا شامل مواد پرانرژی، پیروتکنیک، مواد هوشمند، مواد خودترمیم، عایق‌های حرارتی و صوتی و روش‌های ساخت مواد مورد استفاده در سامانه‌های هوافضایی - فناوری در طراحی وسایل هوافضایی شامل هواپیما، بالگرد، پهپاد، سفینه فضایی سرنشیندار و بدون سرنشین زیرمداری و مدارگرد، موشک، سامانه‌های انتقال و اصلاح مداری، ماهواره و ماهواره بر - فناوری در کاربردهای هوافضایی شامل فناوری سنجش از دور، سامانه-های مدیریت بلایای طبیعی و حفظ محیط زیست مبتنی بر فناوری هوافضایی، سامانه‌های رادیویی و شناسایی اتمسفر و محیط فضا، سامانه‌ها مخابراتی هوافضایی، پردازش داده و سیگنال، فناوری در مدیریت هوافضایی شامل هوانوردی، خدمات فرودگاهی، کنترل ترافیک خطوط هوایی و فضایی، پسماندهای فضایی، فناوری در زیست‌پزشکی هوافضایی شامل زیست هوافضایی، پزشکی، هوافضایی، آموزش فضانورد، شبیه‌سازهای میکروگرویتی
۲	زمین‌شناسی ایران	علوم پایه	زمین‌شناسی	این مجله از مقاله‌های حاوی نتایج پژوهش‌های بنیادی، کاربردی و توسعه‌ای در حوزه‌های مختلف زمین‌شناسی از جمله: چینه‌شناسی؛ زمین‌شناسی اقتصادی؛ هیدروژئولوژی، سنگ‌شناسی رسوبی و نفت؛ GIS؛ زمین‌شناسی مهندسی؛ پترولوژی؛ ژئوفیزیک؛ زمین‌شناسی ساختمانی استفاده می‌کند.
۳	مطالعات باستان-شناسی	علوم انسانی	میان‌رشته‌ای	باستان‌شناسی

[۱] https://journals.msrt.ir/

[۲] https://search.ricest.ac.ir/

۴	مطالعات مدیریت	علوم انسانی	تربیت بدنی	انتشار یافته‌های نوین پژوهشی به منظور گسترش مرزهای دانش، انتقال دستاوردهای نظری و کابردی متخصصین به منظور تبادل آموخته‌ها و تجربیات بر پایه روش‌های پژوهشی معتبر در حیطه علوم ورزشی ایجاد و توسعه شبکه تعاملی بین پژوهشگران و مراکز علمی-پژوهشی، کمک به ارتقای سطح دانش علمی و پژوهش در زمینه‌های مختلف علوم ورزشی و نهادینه کردن پژوهش در کشور، تلاش در جهت رفع نیازهای علمی و تحقیقاتی کشور، کمک به مساله‌یابی و حل مسائل علمی در حوزه علوم ورزشی، تلاش برای ورود به و عضویت درپایگاه‌های اطلاعاتی بین‌المللی، ایجاد ارتباط میان متخصصان و پژوهشگران ایرانی و بین‌المللی، ایجاد روحیه و انگیزه تحقیق در حیطه تخصصی نشریه در بین دانشجویان، معلمان و اساتید
۵	فقه و اصول	علوم انسانی	فقه و حقوق	حوزه فعالیت فصلنامه فقه و اصلو عبارت است از فقه، اصول فقه، فلسفه فقه و موضوعات میان‌رشته‌ای چون فقه و حقوق.
۶	مهندسی برق و مهندسی کامپیوتر ایران	فنی و مهندسی	برق و کامپیوتر	معرفی جدیدترین دستاوردهای پژوهشی در زمینه مهندسی برق و کامپیوتردر ایران و جهان، انتشار نتایج پژوهش‌های اصیل متخصصین دانشگاهی و مراکز تحقیقاتی ایران و جهان در زمینه برق و کامپیوتر، کمک به رشد فعالیت‌های پژوهشی در زمینه‌های علمی، آموزشی و صنعتی کارشناسان برق و کامپیوتر کشور، زمینه‌سازی برای تبادل افکار و اطلاعات بین مراکز دانشگاهی و صنعتی
۷	روش‌های عددی در مهندسی	فنی و مهندسی	میان‌رشته‌ای	نشریه روش‌های عددی در مهندسی (استقلال) دستاوردهای پژوهشی محققان فارسی زبان را در زمینه-های مختلف مهندسی که در آن از روش‌های عددی استفاده می‌شود در قالب مقاله‌های علمی منتشر می‌کند. این نشریه به زبان فارسی منتشر می‌شود و هدف آن انتشار نتایج پژوهش‌های اساتید و محققان در زمینه توسعه و یا استفاده از روش‌های عددی در مسائل مهندسی است. این مجله با تمرکز بر گسترش و ارتقای پژوهش در مهندسی و ایجاد ارتباط و همکاری علمی بین محققان، مقاله‌های اصیل و بدیع تحقیقی در زمینه‌های مکانیک جامدات، مکانیک سیالات، سازه، اندرکنش سازه سیال و خاک، نانوتکنولوژی و همچنین مسائل چند فیزیکی و چند مقیاسی را که برای اولین بار ارائه می شود، با داوری علمی پذیرفته و منتشر می کند. قابل تأکید است که این نشریه اساساً مقاله‌هایی که در آنها یک روش عددی ابداع و یا اصلاح می‌شود و شامل نتایج الگوریتم‌های محاسباتی است، مورد پذیرش و داوری قرار می‌دهد. لیکن مقاله‌هایی که در آنها از نرم‌افزارهای تجاری برای بیان و تفسیر فیزیکی یک مسئله مهندسی استفاده می‌شود نیز، به شکل محدود در قالب بخشی تحت عنوان "تجربیات عددی در مهندسی" پذیرش و برای داوری ارسال می‌کند. زمینه‌های مرتبط با نشریه روش‌های عددی در مهندسی دینامیک سیالات سازه اندرکنش سازه-سیال مسائل چندمقیاسی، کاربرد روش‌های عددی در مسائل با مقیاس نانومکانیک کاربردی بهینه‌سازی، مسائل چندفیزیکی الگوریتم‌های عددی
۸	سبک‌شناسی نظم فارسی	علوم انسانی	زبان و ادبیات	این نشریه چندین هدف را دنبال می‌کند: هدف اول، چاپ مقالات در زمینه نام مجله و گسترش این علم در سطح عمومی و دانشگاهی زبان و ادبیات فارسی، بدین منظور این فصلنامه مقالاتی با محوریت موضوعات زیر را سبک‌شناسی می‌شناسد: الف) بررسی یک یا چند مولفه جزئی یا کلی در طول یک یا چند دوره، یا چند شاعر و نویسنده در دوره‌های مختلف، مانند بررسی سیر شعر معاصر کودک و نوجوان، بررسی اشعار دینی قرن‌های یازدهم و دوازدهم و مقایسه با قرن‌های نهم و دهم، سیر تحول اغراق در سبک خراسانی، معرفی عجائب‌نامه‌ها و تحول آنها در شعر فارسی، تنوع بکارگیری اسلوب معادله در شعر حافظ و مقایسه آن با دیگران و غیره. ب) بررسی ویژگی‌های خاص یا نوآوری‌های یک شاعر یا نویسنده در یک یا چند عنصر بلاغی یا زبانی یا فکری مانند: نوآوری‌های سعدی در ایهام، شگردهای خاص استفاده از آیات قرآن در جهانگشای جوینی، تصویرسازی‌های ویژه صائب تصویری با شمع، اندوهگرائی مشخصه سبکی شهریار، تشبیهات موسیقائی ویژگی شعر خاقانی و غیره. ج) معرفی نسخه خطی تصحیح شده یک شاعر یا نویسنده به صورت زیر: معرفی شاعر یا نویسنده و زندگینامه مختصر او، معرفی دیوان تصحیح شده از نظر کمیت و محتوای آن، معرفی نسخه‌ها و ویژگی‌های فیزیکی آنها، استخراج مهمترین ویژگی‌های سبکی نسخه و نوآوری‌های احتمالی شاعر یا نویسنده، مقایسه محتوای کتاب با کتاب‌های پیشین و پسین به صورت اجمالی و ارائه تفاوت‌های بارز و خاص در صورت وجود داشتن. مانند فقیر دهلوی، معرفی نسخ خطی دیوان و طرز شاعری او، بررسی و معرفی نسخه خطی جواهر خمسه، معرفی گلبن کازرونی شاعر عهد قاجار و گلشن اسرار او و غیره. د) سبک‌شناسی آثار چاپ شده در هر یک از لایه‌های فکری و زبانی و ادبی یا مجموع آنها، بررسی جزئی یک عنصر نیز به شرط بسامد زیاد و داشتن ابتکارات و حالت‌های خاص قابل قبول است. مانند سبک‌شناسی غزلیات انوری، سبک‌شناسی ادبی اشعار دقیقی، معرفی شاعر گمنام طنزپرداز محمدلی فدائی و تبیین جایگاه او در شعر طنز معاصر. ۰) مقالاتی که به بررسی

					تاثیرپذیری یا تاثیرگذاری شعرا یا نویسندگان بر یکدیگر می‌پردازد مانند: تاثیرپذیری سعدی از اندیشه-های غزالی.
۹	رهیافتی نو در مدیریت آموزشی	علوم انسانی	علوم تربیتی		هدف غایی این نشریه، ارتقای سطح دانش علمی در حوزه مدیریت آموزشی و ترویج و گسترش مرزهای دانش در حوزه مطالعاتی علوم تربیتی و زیرمجموعه‌های تخصصی می‌باشد. انتشار دستاوردهای نوین علمی در این حوزه از دیگر اهداف اساسی انتشار این فصلنامه است.
۱۰	صفه	هنر و معماری	معماری		صفه به بررسی گذاشته و حال معماری و شهرسازی ایران و جهان و مبانی نظری رویدادهای مربوط به حوزه شهرسازی می‌پردازد.

جدول ۲ نمونه‌ای از اطلاعات کتاب‌شناختی مقالات مورد استفاده در این پژوهش را برای دو نشریه مکانیک هوافضا و زمین‌شناسی ایران نشان می‌دهد. ستون اول، عنوان نشریه و سه ستون بعدی، به ترتیب عنوان مقاله، کلیدواژه، و چکیده مقاله هستند.

جدول ۲: نمونه‌ای از اطلاعات کتاب‌شناختی مقالات

عنوان نشریه	عنوان مقاله	کلیدواژه	چکیده
مکانیک هوافضا	بهبود کنترل مد لغزشی مدل‌مبنای پیوندی سیستم تعلیق فعال خودرو با استفاده از چرخش بهینه سطوح لغزش و منطق فازی	استراتژی مدل مبنا ◄ چرخش بهینه ◄ سیستم تعلیق فعال ◄ سیستم فازی ◄ مد لغزشی	در این مقاله، طراحی سیستم تعلیق فعال ۱/۴ خودرو با محوریت مد لغزشی و با استفاده از استراتژی مدل مبنا انجام گرفته و سپس بهبود یافته است. هدف از طراحی این سیستم، ایجاد مصالحه بهینه بین جابه جایی بدنه خودرو و پایداری چرخ ها روی جاده است تا ضمن سفر، هم راحتی سرنشینان تامین شود و هم امکان فرمان پذیری و کنترل پذیری خودرو به نحو مطلوبی فراهم باشد. در این مقاله پس از معرفی مدل-های ریاضی قلاب آسمانی و زمینی، عملکرد کنترل‌کننده های مدل مبنای تکی و پیوندی مورد بررسی قرار گرفته و در ادامه جهت بهبود رفتار سیستم، ابتدا چرخش بهینه سطوح لغزش و سپس استفاده از یک سیستم فازی برای تنظیم پارامتری که نقش تعیین کننده در ایجاد مصالحه بین جابه جایی بدنه خودرو و چرخ ها دارد، ارایه شده است. نتایج شبیه‌سازی ها، کارایی روش های پیشنهادی جهت بهبود عملکرد سیستم را به خوبی نشان می دهد.
زمین شناسی ایران	کاربرد زمین فشارسنج بیوتیت-آمفیبول به عنوان نشانگر پتانسیل اکتشافی ذخایر مس-آهن در اسکارن پناه کوه، غرب یزد	بیوتیت ◄ زمین فشارسنج ◄ گرانیت ◄ عمق تشکیل ◄ اسکارن	نفوذ استوک گرانودیوریتی-کوارتز دیوریت به درون سازند آهکی- دولومیتی جمال منجر به شکل گیری اسکارن در منطقه پناه کوه گردیده است. اسکارن ها در همبری مستقیم توده های نفوذی در سنگ های کربناتی تشکیل شده اند. سنگ های گرانیتی پناه کوه اساسا متاآلومینوس و کالکوآلکالن بوده و ویژگی های گرانیتوئید نوع I را نشان می دهند. درشت بلورهای بیوتیت و آمفیبول فراوان ترین کانی آبدار در گرانیت پناه کوه هستند. تجزیه شیمیایی بیوتیت و آمفیبول در سنگ های گرانیتی پناه کوه نشان داد که مقدار آلومینیم کل در آن ها می تواند به عنوان یک نشانگر مفید برای تمایز بین سنگ های گرانیتی کانه زا و غیر کانه زا به کار رود. تطابق مثبت بین مقدار آلومینیم کل و فشار تشکیل سنگ های گرانیتی با استفاده از زمین فشارسنج های بیوتیت و هورنبلند مشاهده می شود. این واقعیت نشان می دهد که مقدار آلومینیم کل بیوتیت و هورنبلند برای تخمین فشار سنگ شدگی سنگ های گرانیتی می تواند مفید واقع گردد. بر اساس زمین فشارسنج های بیوتیت و آمفیبول به دست آمده می توان تخمین زد که کانسار اسکارنی آهن-مس پناه کوه در فشار ۱ تا ۲ کیلوبار مشابه دیگر کانسارهای آهن- مس یاکوگی، کامایشی و تانازاوا در ژاپن شکل گرفته است؛ بنابراین زمین فشارسنج های بیوتیت و آمفیبول در سنگ های گرانیتی می تواند به عنوان یک ردیاب مفید در اکتشاف ذخایر اسکارنی وابسته به سنگ های گرانیتی مورد استفاده قرار گیرد.

۳_۳ پیش‌پردازش داده‌ها[۱]

متن‌کاوی فرآیند استخراج دانش از انبوهی از داده‌های متنی است که یکی از اولین گام‌های آن، پیش‌پردازش می‌باشد. همانطور که در بخش قبل توضیح داده شد تعداد ۱۰ نشریه و از هر نشریه ۲۰۰ مقاله که در سال‌های اخیر به چاپ رسیده‌اند، مورد بررسی قرار می‌گیرد. از هر مقاله، اطلاعات کتابشناختی عنوان، کلیدواژه و چکیده استخراج شده و عملیات پیش‌پردازش روی آنها انجام می‌گیرد. پس از تهیه این مجموعه، چکیده مقالات، پردازش شده و ابتدا کدگذاری کارکترها از نظر کارکترهای فارسی و عربی نرمال می‌گردد. سپس عملیات تبدیل متن به جمله و جمله به واژه انجام می‌گیرد.

برای انجام این کار، ابتدا متن با توجه به کارکترهای جداکننده[۲] که شامل { ، ؛ " ( ) : . < > } هستند، به مجموعه‌ای از توکن‌ها تبدیل می‌شود. سپس عملیات ریشه‌یابی[۳] روی آنها انجام می‌گیرد. هر واژه با کد منحصر به فردی ذخیره می‌شود. علاوه بر واژه، تعداد رخداد آن در مجموعه مقالات نیز محاسبه و ذخیره می‌گردد. همچنین ایست‌واژه[۴]ها، علائم و اعداد حذف می‌شوند. خروجی این مرحله، واژگان پردازش‌شده‌ای می‌باشد که فرکانس تکرار آنها در هر مقاله نیز مشخص شده است.

تشخیص واژه‌های ایستا یکی از مهمترین عملیات در متن‌کاوی است. واژه‌های ایستا معمولا خیلی زیاد در اسناد کل مجموعه رخ می‌دهند و عمدتا حاوی اطلاعات باارزشی در مورد متن و یا اسناد نیستند. بنابراین بهتر است که این واژه‌ها از کل مجموعه حذف گردند (Sadeghi & Vegas, 2014). از آنجا که ایست‌واژه‌ها نقش بسیار مهمی را در بازیابی اطلاعات ایفا می‌کنند و حذف آنها منجر به افزایش سرعت پردازش اطلاعات و همچین کارایی بیشتر سامانه می‌شود، پژوهش‌های زیادی در سراسر جهان پیرامون این موضوع انجام گرفته است (هاشم‌زاده و همکاران، ۱۳۹۲) ولی این تحقیقات عمدتا روی زبان انگلیسی می‌باشند[۵]. علاوه بر این، اطلاعات موجود نشان می‌دهد که تاکنون لیست ایست واژه‌ای که بتواند منحصرا در یک موضوع خاص مقالات و نشریات باشد، ارائه نشده است و لیست ایست‌واژه‌هایی که منتشر شده، به صورت کلی در تمامی حوزه‌ها و عمدتا برای خبر در پیکره همشهری می‌باشد (Davarpanah et al., 2009). از آنجا که استفاده از یک لیست ایست‌واژه برای تمامی حوزه‌ها ممکن است منجر به کاهش کارایی سامانه‌های بازیابی اطلاعات شود (Sadeghi & Vegas, 2014)، در این پژوهش یک لیست از ایست‌واژه‌ها که منحصرا برای مقالات علمی و نشریات است، ارائه می‌گردد. این لیست در جدول ۱ از پیوست ۱ ارائه شده است.

---

۱ Data preprocessing

۲ Delimitor characters

۳ Stemming

۴ Stop words

۵ English Stop Word List in WordNet, http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt (2013, accessed May 2013)

۳_۴ مدلسازی موضوعی

در این پژوهش، به منظور بدست آوردن مدلسازی موضوعی از الگوریتم تخصیص پنهان دیریکله یا به اختصار LDA و نمونه‌برداری گیبز استفاده شده است. این الگوریتم فرض می‌کند که اسناد از موضوعات متفاوتی تشکیل شده‌اند. به عبارت دیگر، هر نشریه از تعداد بسیار زیادی کلمه تشکیل شده است که هر یک متعلق به یک موضوع است و همچنین نسبت موضوعات داخل یک متن با یکدیگر متفاوت است. با توجه به این نسبت‌ها می‌توانیم آن متن را در یک موضوع خاص دسته‌بندی نماییم. فرض کنید که مجموعه ثابتی از کلمات وجود داشته باشد. روش LDA فرض می‌کند که هر موضوع، توزیعی روی این مجموعه کلمات است. به عبارت دیگر، الگوریتم LDA فرض می‌کند که هر سند یک ترکیب تصادفی از موضوعات است و هر کلمه هم از یکی از این موضوعات می‌آید.
برای انجام این کار، ابتدا تعداد موضوعاتی که باید از متن استخراج گردد، انتخاب می‌کنیم. یک بار که تعداد موضوعات انتخاب شد، LDA تمامی کلمات در هر سند را بررسی کرده و آنها را به صورت تصادفی به یکی از k سند انتساب می‌دهد. بعد از این مرحله، یک نمایش از موضوعات (توزیع کلمات در هر موضوع) و همچنین یک نمایش از اسناد (توزیع موضوعات در هر سند) ارائه می‌کند. به عبارت دیگر، LDA با تحلیل روی کل پیکره، درصد کلمات درون اسناد را که به یک موضوع خاص انتساب شده‌اند، بدست می‌آورد. بنابراین عملیات زیر توسط LDA محاسبه می‌گردد:

۱) درصد کلمات در یک سند d که اخیرا به موضوع t انتساب داده شده $(p(topic\ t|document\ d))$.
۲) درصد تعداد کلمات w که به موضوع t در همه اسناد انتساب داده شده $(p(word\ w|topic\ t))$.

با توجه به این محاسبات، LDA تصمیم می‌گیرد که یک کلمه w را از موضوع B به موضوع A جابجا کند، در صورتی که:

$$p(topic\ A|document\ d) \times p(word\ w|topic\ A) > p(topic\ B|document\ d) \times p(word\ w|topic\ B) \quad (۱-۳)$$

به عبارت دیگر، زمانی یک کلمه را که قبلا در موضوع B بوده، به موضوع A تغییر می‌دهد که حاصلضرب احتمال موضوع A با درنظر گرفتن سند d و احتمال کلمه w با درنظر گرفتن موضوع A بیشتر از حاصلضرب احتمال موضوع B با درنظر گرفتن سند d و احتمال کلمه w با درنظر گرفتن موضوع B باشد.
یکی از چالش‌های موجود در مدلسازی توسط LDA بدست آوردن تعداد موضوعات می‌باشد. روش‌های موجود برای برخورد با این مشکل از روش‌های جستجوی گریدی[۱] استفاده می‌کنند. به عنوان مثال، از معیارهای استانداردی مانند سرگشتگی[۲] (Manning & Schutze, 1999) استفاده کرده و مقدار این معیار را برای پارامترهای مختلف مدل بدست می‌آورند. سپس پارامتری که به ازای آن بیشترین مقدار سرگشتگی را نتیجه می‌دهد، در نظر

[۱] Gird search
[۲] Perplexity

می‌گیرند. یک معیار دیگر که در روش‌های جستجوی گریدی استفاده می‌شود، معیار انسجام معنایی[۱] است (Mimno et al., 2011)؛ چرا که معیار سرگشتگی نمی‌تواند اطلاعات مناسبی از یک مدل موضوعی ارائه دهد (ر.ک. بخش ۴)

علیرغم اینکه روش‌های جستجوی گریدی در پژوهش‌های مختلفی مورد استفاده قرار گرفته است، ولی یکی از بزرگترین مشکلات آنها، مدت زمان بالای اجرای آن است. به عبارت دیگر، بدست آوردن تعداد موضوعات در هر متن، نیاز به بار محاسباتی بسیار زیادی دارد. بنابراین در این پژوهش، روش مبتنی بر نظریه بازبهنجاری[۲] مبتنی بر آنتروپی رونو[۳] بکار گرفته می‌شود.

بازبهنجاری یک فرمولاسیون ریاضی است که در حوزه‌های بسیار زیادی از جمله فیزیک مانند تحلیل نفوذ[۴] و تحلیل تغییر فاز[۵] استفاده می‌شود. هدف بازبهنجاری، ساخت یک رویه برای تغییر مقیاس سیستم تحت بررسی است؛ به صورتی که رفتار سیستم حفظ شود و تغییری در روند آن ایجاد نشود. توضیح کامل این نظریه از حوصله این پژوهش خارج است. برای کسب اطلاعات تکمیلی‌تر در مورد این نظریه، می‌توان به کتاب کادانوف مراجعه نمود (Kadanoff, 2000). به منظور توضیح خلاصه‌ای از این نظریه، فرض کنید که لاتیسی[۶] وجود دارد که از مجموعه‌ای از نود[۷]ها تشکیل شده است. هر نود به وسیله جهت اسپین[۸] یا حالت اسپین[۹] مشخص می‌گردد. به نوبه خود، هر اسپین می‌تواند یک یا چندین جهت ممکن را داشته باشد. نودها با جهت‌های اسپین مشابه یک خوشه را تشکیل می‌دهند. رویه مقیاس‌گذاری یا بازبهنجاری از یک ساختار ترکیبی از هر بلاک تبعیت می‌کند؛ به صورتی که چندین نود نزدیک به هم به وسیله یک نود جایگزین می‌شوند. جهت اسپین جدید به وسیله جهت اکثریت اسپین‌ها در آن بلاک تعیین می‌شود. رویه ترکیب بلاکی روی کل لاتیس اعمال می‌گردد. به این ترتیب، یک پیکربندی جدید از اسپین‌ها را در نهایت خواهیم داشت. رویه بازبهنجاری می‌تواند چندین بار انجام گیرد. با دنبال کردن معادلات میان پیکربندی اسپین جدید و قدیم، می‌توان رویه محاسبه پارامترها و مقادیر نمایندگان مهم را بدست آورد (Wilson & Kogut, 1974). با استفاده از این نظریه می‌توان در مواقعی که با استفاده از مدل‌های ریاضی استاندارد، نمی‌توان مقادیر پارامترهای یک معادله را بدست آورد، به تخمین دقیقی از نمایندگان آن پارامترها رسید. در این پژوهش از نظریه بازبهنجاری در کنار آنتروپی رونو برای بدست آوردن تخمین از تعداد موضوعات استفاده شده است.

---

۱ Semantic coherence

۲ Renormalization theory

۳ Renyi Entropy

۴ Percolation analysis

۵ Phase transition analysis

۶ Lattice

۷ node

۸ Spin direction

۹ Spin state

رویکرد آنتروپیک برای تنظیم مدل‌های موضوعی بر اساس مجموعه‌ای از نظریه‌هاست که مدلسازی موضوعی را به مدل‌های فیزیک آماری ارتباط می‌دهند؛ علاوه بر این، مساله بهینه‌سازی پارامتر مدل را به صورت فرمول‌های ترمودینامیکی ارائه می‌دهند. به عبارت دیگر:

۱) مجموعه‌ای از اسناد را به عنوان یک سیستم اطلاعاتی در نظر می‌گیرد: یک سیستم آماری با عناصری که به صورت کلمه و اسناد بسیار زیاد هستند. متقابلا رفتار چنین سیستمی می‌تواند با کمک مدل‌های فیزیک آماری مطالعه شود.
۲) تعداد کل کلمات و اسناد در سیستم اطلاعاتی تحت بررسی، ثابت است (حجم سیستم تغییر نمی‌کند).
۳) یک موضوع در حقیقت یک حالت است (در شباهت با جهت اسپین) که هر کلمه و سند در مجموعه می‌تواند بگیرد. اینجا یک کلمه و یک سند می‌تواند به چندین موضوع با احتمالات مختلف، تعلق داشته باشد (حالات اسپین).
٤) یک راه‌حل مدلسازی موضوعی یک حالت عدم تعادل سیستم است.
٥) چنین سیستم اطلاعاتی، باز است و انرژی با محیط از طریق تغییر دما تبادل می‌شود. دمای سیستم اطلاعاتی در واقع تعداد موضوعات است که پارامتری است که باید با کمک جستجو با حداقل واگرایی کولبک-لیبلر[۱] (واگرایی KL[۲]) بدست آید.
٦) از آنجا که واگرایی KL معادل با اختلاف انرژی‌های آزاد[۳] است (Akturk et al., 2007) که درجه‌ای را اندازه‌گیری می‌کند که در آن یک سیستم داده شده نامتعادل است، می‌توان از معادله $\Lambda_F = F(T) - F_0$ استفاده کرد. در این معادله F_0 انرژی آزاد حالت اولیه (آشفتگی[٤]) مدل موضوعی و $F(T)$ انرژی آزاد بعد از مدلسازی موضوعی برای یک تعداد ثابت از موضوعات T است (Koltkov et al., 2017).
۷) مینیمم Λ_F بستگی به پارامترهای مدل موضوعی مانند تعداد موضوعات و دیگر پارامترها دارد.
۸) تعداد بهینه موضوعات و مجموعه‌ای از پارامترهای مدل موضوعی، مرتبط است با موقعیتی که بیشینه اطلاعات حاصل شود. لازم به ذکر است که انرژی آزاد می‌تواند از طریق آنتروپی رونی[٥] نشان داده شود و ماکزیمم اطلاعات از طریق کمترین آنتروپی رونی بدست می‌آید.

در مدل‌های موضوعی، مجموع احتمالات همه کلمات، مساوی با تعداد موضوعات می‌باشد $T = \sum_{t\in\tilde{T}}\sum_{w\in\tilde{W}}\phi_{wt}$ که در آن $\phi_{wt} \in [0,1]$ برای همه $w \in \tilde{W}$ و $t \in \tilde{T}$. در چارچوب فیزیک آماری، متداول است که توزیع سیستم-های آماری را به وسیله سطوح انرژی بررسی کنیم، که انرژی در آن به صورت احتمال نشان داده می‌شود. در این پژوهش، مشابه با (Koltcov et al., 2019) و (Koltcov & Ignatenko, 2020) محدوده احتمالی [0,1] را به دو بازه تقسیم می‌نماییم. به عبارت دیگر یک سیستم دو سطحی در نظر گرفته می‌شود که اولین سطح مربوط به کلمات با احتمال‌های بالاست و دومین سطح به کلمات با احتمالات کم نزدیک به صفر. در نتیجه، می‌توان تابع

---

[۱] Kullback-Leibler
[۲] KL divergence
[۳] Free energies
[٤] chaos
[٥] Renyi entropy

چگالی-حالات[1] را برای کلمات با احتمالات بالا تحت یک تعداد ثابت از موضوعات و پارامترها، طبق فرمول زیر معرفی نمود.

$$\rho = N/(WT) \quad (۲-۳)$$

در این معادله، N تعداد کلمات با احتمال بالا است. منظور از احتمال بالا، احتمالاتی است که شرط $\rho > 1/W$ را ارضا کند. انتخاب چنین سطحی بر این اساس است که مقادیر $1/W$ مقادیر اولیه ماتریس ϕ برای مدل‌های موضوعی در LDA می‌باشد. مقدار $W.T$ تعداد کل میکرو-حالات[2] یک مدل موضوعی (سایز ماتریس ϕ) را مشخص کرده و تابع چگالی-حالات را نرمال می‌نماید. در طی فرآیند مدلسازی موضوعی، احتمالات کلمات از توزیع $1/w$ دور می‌شود. بخش کوچکی از کلمات، احتمالات بزرگتر از مقدار آستانه می‌گیرند؛ در حالی که بخش بزرگتر از کلمات، احتمالات کمتر از آن را می‌گیرند. انرژی سطح بالاتر شامل حالاتی با احتمالات بالاست که تحت معادله زیر بدست می‌آید:

$$E_{high} = -\ln(\tilde{P}) = -\ln(\frac{1}{T}\sum_{w,t}(\phi_{wt}.\Omega(\phi_{wt} - {}^{1}/_{W}))) \quad (۳-۳)$$

در این معادله، $\Omega(.)$ تابع پله[3] می‌باشد که به صورت زیر تعریف می‌شود.

$$\Omega(\phi_{wt} - {}^{1}/_{W}) = 1 \quad if\ \phi_{wt} \geq {}^{1}/_{w} \quad (۴-۳)$$
$$\Omega(\phi_{wt} - {}^{1}/_{W}) = 0 \quad if\ \phi_{wt} < {}^{1}/_{w}$$

بنابراین در معادله (۳-۳) فقط احتمالاتی که بزرگتر از ${}^{1}/_{W}$ هستند، جمع می‌شوند.

حال با توجه به توضیحات بالا، آنتروپی رونو به صورت زیر تعریف می‌شود.

با استفاده از تابع جداسازی زیر، می‌توان آنتروپی رونو را از طریق انرژی آزاد نشان داد (Beck, 2009):

$$H_{Renyi} = \frac{-q\Lambda_F}{q-1} \quad (۵-۳)$$

باتوجه به اینکه $q = {}^{1}/_{T}$ و $\Lambda_F = \ln(\tilde{P}) - T \times \ln(\rho)$ بنابراین، می‌توان معادله بالا را به صورت زیر بیان کرد (Koltcov & Ignatenko, 2020):

$$H_{Renyi} = \ln(\tilde{P}) - T \times \ln(\rho) \quad (۶-۳)$$
$$\tilde{P} = E_{high} - T \times \ln(\rho)$$

که E_{high} با توجه به فرمول (۳-۳) بدست می‌آید.

[1] Density-of-states function

[2] Micro-states

[3] Step function

از آنجا که محاسبه آنتروپی رونو بر اساس اختلاف انرژی‌های آزاد است، می‌توان از آنتروپی رونو به عنوان معیاری برای درجه ناپایداری و یا عدم تعادل یک سیستم (که در اینجا مدل موضوعی است) استفاده کرد. همچنین، جستجو برای کمترین آنتروپی رونو می‌تواند برای بهینه‌سازی پارامترهای مدل‌های یادگیری ماشینی مناسب باشد. همانطور که قبلا هم گفته شد خروجی مدل موضوعی، ماتریس Φ با سایز $W \times T$ است (W و T به ترتیب نشان‌دهنده تعداد کلمات و تعداد موضوعات می‌باشد). اینجا ما یک مجموعه ثابت از کلمات مشخص را در نظر می‌گیریم. بنابراین مقیاس بازبهنجاری فقط به پارامتر $q = 1/T$ وابسته است. رویه بازبهنجاری، رویه ترکیب دو موضوع در یک موضوع است. به عنوان نتیجه ترکیب رویه ترکیب، ما یک موضوع جدید $\tilde{t}$ با توزیع موضوع-کلمه را بدست می‌آوریم؛ به صورتی که $\sum_w \phi_{w\tilde{t}} = 1$. از آنجا که محاسبه ماتریس Φ بستگی به مدل موضوعی مشخص دارد، فرموله‌بندی ریاضی رویه بازبهنجاری در واقع وابسته به مدل[1] است. به علاوه، نتایج ترکیب بستگی دارد به اینکه چگونه موضوعات برای ترکیب انتخاب شوند.

در این پژوهش، از آنتروپی رونو برای رویه انتخاب موضوع استفاده می‌کنیم. بدین صورت که آنتروپی رونو برای هر موضوع به صورت جداگانه بر اساس معادله (۳-۶) محاسبه می‌گردد. در این حالت فقط احتمالات کلمات در هر موضوع استفاده می‌شود. سپس جفت موضوعات با کوچکترین مقدار آنتروپی رونو انتخاب می‌شوند. مقادیر بزرگ آنتروپی رونو نشان‌دهنده این است که موضوعات با یکدیگر ارتباط کمی دارند؛ در حالی که مقادیر کمتر آنتروپی رونو نشان‌دهنده بیشترین ارتباطات اطلاعاتی میان موضوعات است. بنابراین موضوعاتی که ارتباط بیشتری با هم دارند، انتخاب می‌گردند.

بعد از ترکیب دو موضوع با کمترین مقدار آنتروپی، نیاز است که ماتریس Φ بر اساس موضوعات جدید که حاصل ترکیب دو موضوع در مرحله قبلی است، بروزرسانی گردد. محاسبه ماتریس Φ شامل دو فاز است. اولین فاز شامل نمونه‌برداری و محاسبه یک شمارنده c_{wt} است که c_{wt} تعداد دفعاتی است که کلمه w به موضوع t نسبت داده می‌شود. فاز دوم شامل محاسبه مجدد Φ بر اساس معادله زیر است:

$$\phi_{wt} = \frac{c_{wt} + \beta}{(\sum_{w \in \tilde{W}} c_{wt}) + \beta W} \qquad (۳-۷)$$

برای بازبهنجاری، از مقادیر c_{wt} و معادله (۳-۷) استفاده می‌شود. مقادیر c_{wt} از ماتریس $C = \{c_{wt}\}_{w \in \tilde{W}, t \in \tilde{T}}$ بدست می‌آیند و این همان ماتریسی است که بازبهنجاری را کنترل می‌کند. بر اساس ماتریس C نسخه بازبهنجاری ماتریس Φ مجددا محاسبه می‌شود.

بنابراین، این پژوهش برای بدست آوردن تعداد موضوعات از یک رویه تکرارشونده به شرح زیر استفاده می‌کند:

۱) جفتی از موضوعات (t_1 و t_2) با کمترین آنتروپی رونو، برای ترکیب انتخاب می‌شوند.

[1] Model-dependent

۲) در این مرحله دو موضوع با یکدیگر ترکیب می‌شوند (جمع شمارنده‌های c_{wt_1} و c_{wt_2}). به عبارت دیگر، $c_{w\tilde{t}} = c_{wt_1} + c_{wt_2}$. سپس بر اساس مقادیر جدید شمارنده‌ها، ماتریس $\phi_{W\tilde{t}}$ به صورت زیر، محاسبه می‌گردد:

$$\phi_{w\tilde{t}} = \frac{c_{wt_1} + c_{wt_2} + \beta}{\left(\sum_{w\in\tilde{W}} c_{wt_1} + c_{wt_2}\right) + \beta W} \qquad (۳-۸)$$

توزیع جدید $\phi_{W\tilde{t}}$ شرط $\sum_{w\in\tilde{W}} \phi_{w\tilde{t}} = 1$ را ارضا می‌کند. ستون ϕ_{wt_1} به وسیله $\phi_{W\tilde{t}}$ و ستون ϕ_{wt_2} از ماتریس Φ حذف می‌شود. این مرحله باعث می‌شود که تعداد موضوعات در هر تکرار، یکی کم شود.

مرحله ۱ و ۲ مرتبا تکرار می‌شود تا در نهایت فقط دو موضوع باقی بماند. در هر بار پایان مرحله ۲، آنتروپی رونو برای ماتریس Φ بر اساس معادله (۳-٦) محاسبه می‌گردد. سپس نمودار آنتروپی رونو به عنوان یک تابع از تعداد موضوعات رسم می‌شود و کمترین مقدار به منظور تشخیص تخمین تعداد بهینه از موضوعات محاسبه می‌گردد.

فصل چهارم

یافته‌ها

۴. یافته‌ها

۴_۱ مقدمه

در این فصل به مطالعه‌ی کارایی روش‌های پیشنهادی می‌پردازیم. ابتدا به معرفی داده‌ها و همچنین نحوه تولید داده‌های آموزشی می‌پردازیم. سپس معیارهای ارزیابی استفاده شده در این پژوهش معرفی می‌گردند و در نهایت کارایی روش‌های پیشنهادی را بررسی خواهیم کرد.

۴_۲ داده‌ها

تعداد ۱۰ نشریه و از هر نشریه، ۲۰۰ مقاله از مجموعه مقالات موجود در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری بازیابی گردید. از هر مقاله، اطلاعات کتابشناختی عنوان، کلیدواژه، و چکیده مقالات استخراج شد. به عبارت دیگر داده‌های این پژوهش شامل عنوان، کلیدواژه، و چکیده ۲۰۰۰ مقاله فارسی است.

۴_۳ معیارهای ارزیابی

در حالت کلی، ارزیابی یک مدل موضوعی بسیار مشکل است؛ چرا که مدلسازی موضوعی به صورت بدون نظارت انجام می‌گیرد. پژوهش‌های پیشین از تکنیک‌های مختلفی برای ارزیابی مدل موضوعی استفاده می‌کنند. بعضی از آنها از قضاوت‌های انسانی برای ارزیابی مدل موضوعی استفاده می‌کنند (Chang et al., 2009). بعضی

دیگر از معیارهای عددی استفاده می‌کنند که نمونه آنها معیار سرگشتگی است (Blei et al., 2003). این معیار بدین صورت تعریف می‌شود:

$$P(D') = \sum_{w_d \in D'} P(w_d)\log(P(w_d)) \quad (۱-۴)$$

در این فرمول، D' مدل موضوعی است که قرار است کارایی آن سنجیده شود و w_d کلمات موجود در آن مجموعه است. یکی از اشکالاتی که به معیار سرگشتگی وارد است، این است که این معیار به قضاوت‌های انسان نزدیک نیست (Chang et al., 2009). به عبارت دیگر، احتمال وقوع کلمات و قضاوت‌های انسان در اغلب موارد، مرتبط نیست و حتی در بعضی موارد ممکن است متضاد هم باشد.

انسجام[۱]، معیار دیگری برای ارزیابی مدل موضوعی می‌باشد که بهتر از سرگشتگی می‌تواند مدل‌های مختلف را با یکدیگر مقایسه نماید و شبیه‌ترین معیار به قضاوت انسانی است (Röder et al., 2015). این معیار به صورت زیر تعریف می‌شود:

$$C_{UCI} = \frac{2}{N \times (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log\left(\frac{P(w_i, w_j) + \varepsilon}{P(w_i).P(w_j)}\right) \quad (۲-۴)$$

$$P(w_i) = \frac{Number\ of\ documents\ that\ contain\ w_i}{total\ number\ of\ documents}$$

$$P(w_i, w_j) = \frac{Number\ of\ documents\ that\ contain\ w_i\ and\ w_j}{total\ number\ of\ documents}$$

۴_۴ یافته‌ها

در این بخش به ارائه یافته‌ها و همچنین پاسخ به سوالات پژوهشی می‌پردازیم. در این پژوهش، دو سوال پژوهشی مطرح شده است. یکی در مورد کاهش ابعاد با توجه به مدلسازی موضوعی و یا همان بدست آوردن تعداد موضوعات، و دیگری مشخص کردن گروه‌های موضوعی حاصل با توجه به تعداد موضوعات مشخص شده در پرسش ۱ در نشریات مختلف است.

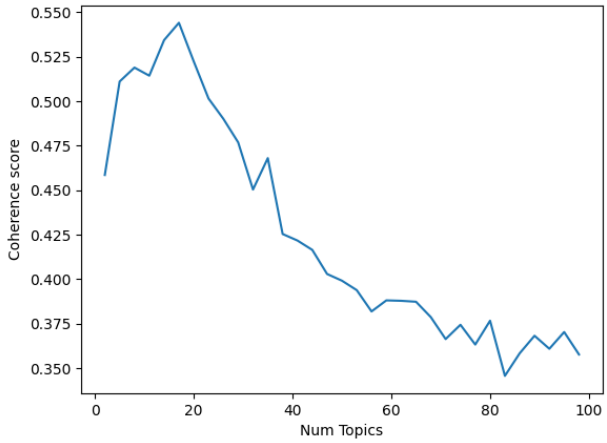
[۱] coherence

۴_۴_۱ بدست آوردن تعداد موضوعات (پاسخ به پرسش اول)

یک روش متداول برای پاسخ به پرسش اول و یا بدست آوردن تعداد موضوعات در مدلسازی موضوعی، بررسی میزان کارایی مدل با توجه به پارامترهای مختلف تعداد موضوعات و در نهایت بدست آوردن بهترین پارامتر می‌باشد. با توجه به نزدیکی معیار انسجام به قضاوت‌های انسانی در این پژوهش از این معیار استفاده شده و کارایی مدل با در نظر گرفتن تعداد موضوعات مختلف سنجیده شده است. با توجه به زمان‌بر بودن این روش برای بدست آوردن تعداد موضوعات و یا همان ابعاد، از الگوریتم بازبهنجاری استفاده شده، و مقادیر آنتروپی رونو به ازای تعداد مختلف موضوعات بدست آمده، و در نهایت موضوعی که کمترین مقدار آنتروپی رونو را در برداشته، به عنوان تعداد ابعاد انتخاب شده است.

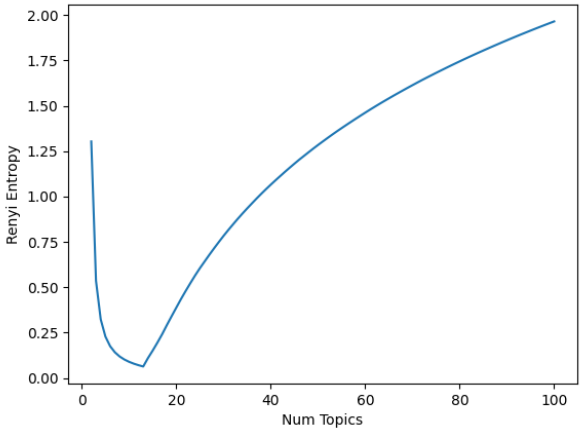
اولین نشریه‌ای که آن را مورد بررسی قرار می‌دهیم، نشریه مکانیک هوافضا است. این نشریه در حوزه فنی و مهندسی و مشخصا مکانیک می‌باشد و به انتشار پژوهش‌های جدید در سامانه‌های هدایت، کنترل و ناوبری وسایل هوافضایی شامل الگوریتم‌های هدایت و کنترل، عملگرهای کنترلی، سامانه‌های ناوبری، موقعیت‌یابی، تعیین وضعیت، کنترل وضعیت، پردازش داده و سیگنال، شبیه‌سازهای پرواز، آمایشگاه‌های واقعیت مجازی و محیط‌های پرواز، الگوریتم‌های تشخیص و جبران خطا، فناوری در سازه؛ و مکانیزم‌های هوافضایی می‌پردازد.

به منظور پاسخ به پرسش اول در این نشریه، از دو روش گریدی و بازبهنجاری استفاده کردیم. شکل ۲ معیار انسجام را به ازای پارامترهای مختلف نشان می‌دهد. محور افقی نشان‌دهنده تعداد موضوعات است که از ۲ تا ۱۰۰ تغییر می‌کند و محور عمودی، معیار انسجام را به ازای هر کدام از تعداد موضوعات نشان می‌دهد. همانطور که این شکل نشان می‌دهد، این نشریه با تعداد موضوع حدود ۱۴ بهترین معیار انسجام را دربردارد.



شکل ۲: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مکانیک هوافضا

شکل ۳، روند بدست آوردن تعداد موضوعات برای نشریه مکانیک هوافضا را با استفاده از الگوریتم بازبهنجاری نشان می‌دهد. محور افقی تعداد موضوعات را نشان می‌دهد که از ۲ تا ۱۰۰ تغییر می‌کنند. محور عمودی معیار آنتروپی رونو را به ازای تعداد موضوعات مختلف نشان می‌دهد.



شکل ۳: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مکانیک هوافضا

همانطور که این شکل نشان می‌دهد، تعداد موضوعات برای این نشریه با استفاده از نظریه بازبهنجاری، حدود ۱۲ می‌باشد.

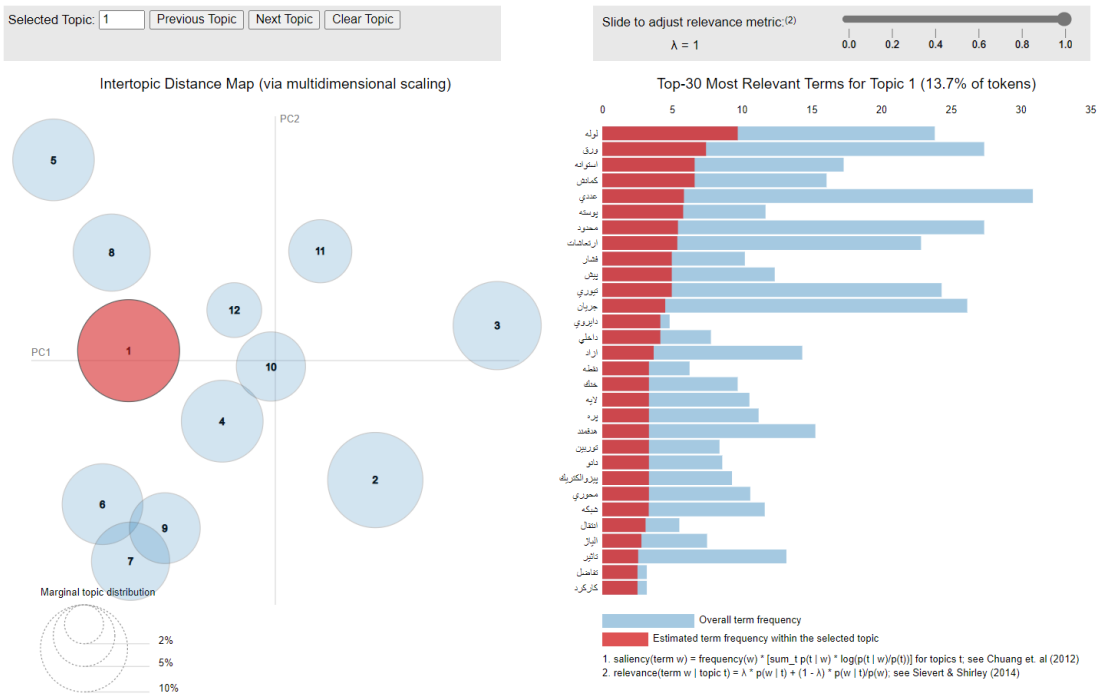
به منظور ایجاد یک نمایه گرافیکی از نتایج مدلسازی موضوعی روی نشریات از روش LDAvis استفاده شد (Sievert & Shirley, 2014). این نمایه گرافیکی از دو پنل تشکیل شده است که پنل سمت چپ، موضوعات و همچنین میزان اهمیت هر موضوع را نشان می‌دهد. هر موضوع با یک دایره مشخص شده است که اندازه هر دایره متناسب با میزان اهمیت آن موضوع است؛ به عبارت دیگر هر چقدر شعاع یک دایره بیشتر باشد، موضوع متناسب با آن دایره از اهمیت بیشتری برخوردار است. با انتخاب هر موضوع در پنل سمت چپ، می‌توان کلمات موجود در آن موضوع را در پنل سمت راست مشاهده نمود. در کنار هر کلمه یک میله افقی وجود دارد که دو رنگ آبی و قرمز در آن میله وجود دارد. رنگ آبی نشان‌دهنده تکرار کلمه در کل مقالات انتخاب شده در نشریه است، و رنگ قرمز، تکرار آن کلمه در موضوع انتخاب شده را مشخص می‌نماید.

یکی از پارامترهایی که در این نمودار مشخص شده است، پارامتر λ است که در این پژوهش، مقدار ۱ در نظر گرفته شده است. این پارامتر میزان اهمیت یک کلمه در موضوع را مشخص می‌کند؛ هر چقدر احتمال کلمه در یک موضوع بیشتر و آن کلمه در کل مقالات آن نشریه، احتمال کمتری داشته باشد، از درجه اهمیت بالاتری برخوردار است. کلمات یا ترم‌های موجود در پنل سمت راست و میزان اهمیت آنها را می‌توان با توجه به این پارامتر

کنترل کرد. به عبارت دیگر، اینکه کدام کلمات در پنل سمت راست قرار بگیرند و میزان اهمیت آنها چقدر باشد، با توجه به فرمول زیر تعیین می‌گردد (Sievert & Shirley, 2014):

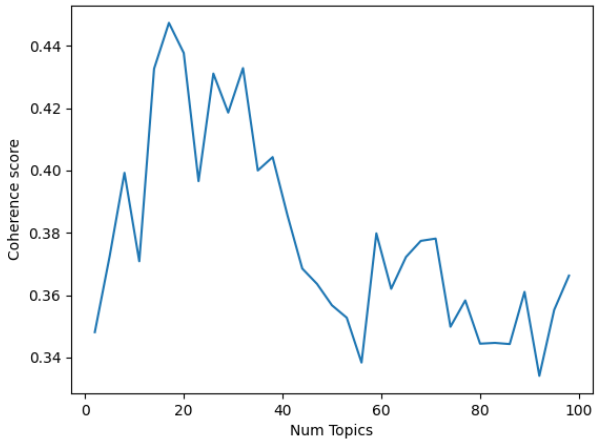
$$r(w,t|\lambda) = \lambda \log(\Phi_{tw}) + (1-\lambda)\log(\frac{\Phi_{tw}}{P_w}) \qquad (۳-۴)$$

در این فرمول $r(w,t|\lambda)$ نشان‌دهنده رتبه یک کلمه در یک موضوع با توجه به λ است و از دو بخش تشکیل شده است. بخش اول $\log(\Phi_{tw})$ احتمال وقوع کلمه w در موضوع t را نشان می‌دهد که هر چقدر بزرگتر باشد، بدین معنی است که این کلمه در این موضوع از اهمیت بالاتری برخوردار است. بخش دوم این فرمول شامل $\log(\frac{\Phi_{tw}}{P_w})$ است که صورت کسر، احتمال وقوع کلمه در موضوع را مشخص می‌نماید و مخرج کسر، احتمال کلمه در کل اسناد را بررسی می‌نماید. به عبارت دیگر، هر چقدر احتمال وقوع کلمه در یک موضوع بیشتر باشد و آن کلمه در کل اسناد کمتر رخ داده باشد، مقدار بزرگتری توسط این کسر به خود می‌گیرد. بنابراین هر چقدر مقدار پارامتر λ بزرگتر باشد، بخش اول این فرمول اهمیت بیشتری می‌گیرد و هر چقدر λ کوچکتر باشد، بخش دوم اهمیت بالاتری برای تعیین رتبه کلمه در موضوع می‌یابد. همانطور که نمای گرافیکی از نشریه مکانیک هوافضا در شکل ۴ نشان می‌دهد، ۱۲ موضوع در این نشریه وجود دارد که این پارامتر نزدیکی بیشتری به روش تخمین پارامتر با روش بازبهنجاری دارد. در این شکل، به عنوان نمونه موضوع ۱ انتخاب شده است که با انتخاب این موضوع، کلمات «لوله، ورق، استوانه، کمانش، و عددی» انتخاب شده است.

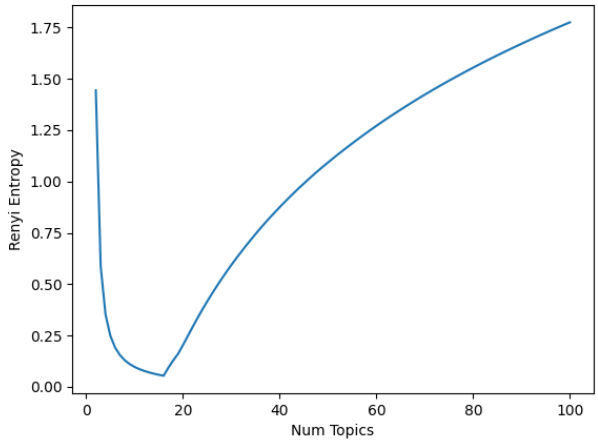


شکل ۴: نمای گرافیکی از موضوعات موجود در نشریه مکانیک هوافضا

در نشریه زمین‌شناسی ایران که از حوزه علوم پایه است و مقالات در حوزه زمین‌شناسی را به چاپ می‌رساند، معیار انسجام را بر اساس تعداد موضوعات مختلف بکار گرفتیم که نتایج آن در شکل ۵ مشخص شده است. الگوریتم بازبهنجاری را روی داده‌های این نشریه اعمال نمودیم که شکل ۶ نتایج حاصله را نشان می‌دهد. با استفاده از این دو شکل، می‌توان تعداد ابعاد مساله و یا همان تعداد موضوعات را برای نشریه زمین‌شناسی ایران بدست آورد. در این نمودارها، محور افقی تعداد موضوعات را نشان می‌دهد و محور عمودی به ترتیب معیارهای انسجام و آنتروپی رونو را مشخص می‌کند.

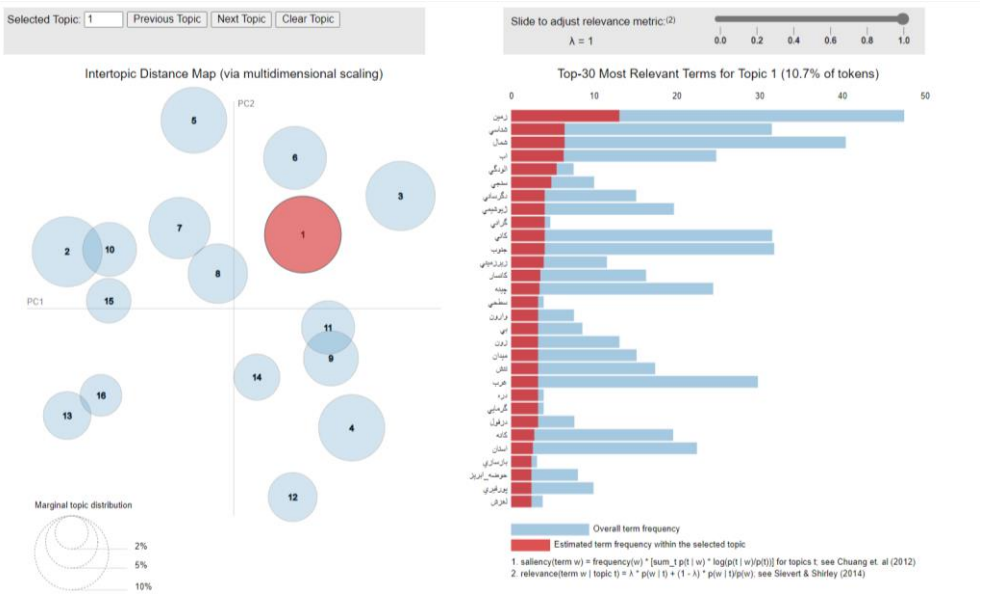


شکل ۵: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه زمین‌شناسی ایران



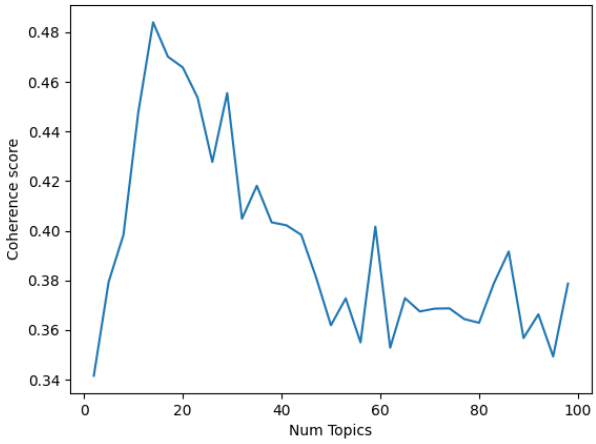
شکل ۶: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه زمین‌شناسی ایران

شکل ۷ نمای گرافیکی از نتایج مدلسازی موضوعی در نشریه زمین‌شناسی ایران را نشان می‌دهد.



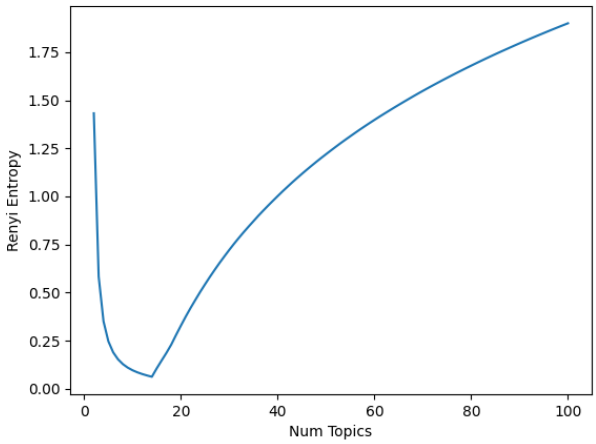
شکل ۷: *نمای گرافیکی از موضوعات موجود در نشریه زمین‌شناسی ایران*

مطالعات باستان‌شناسی نشریه دیگری در حوزه علوم انسانی است که به چاپ مقالات در حیطه باستان‌شناسی می‌پردازد. به منظور اعمال مدلسازی موضوعی روی این نشریه، ابتدا نیاز است که تعداد موضوعات (T) تخمین زده شود. معیار انسجام را به ازای تعداد موضوعات مختلف روی مقالات مختلف این نشریه اعمال کردیم که نتیجه حاصل در شکل ۸ مشخص شده است.



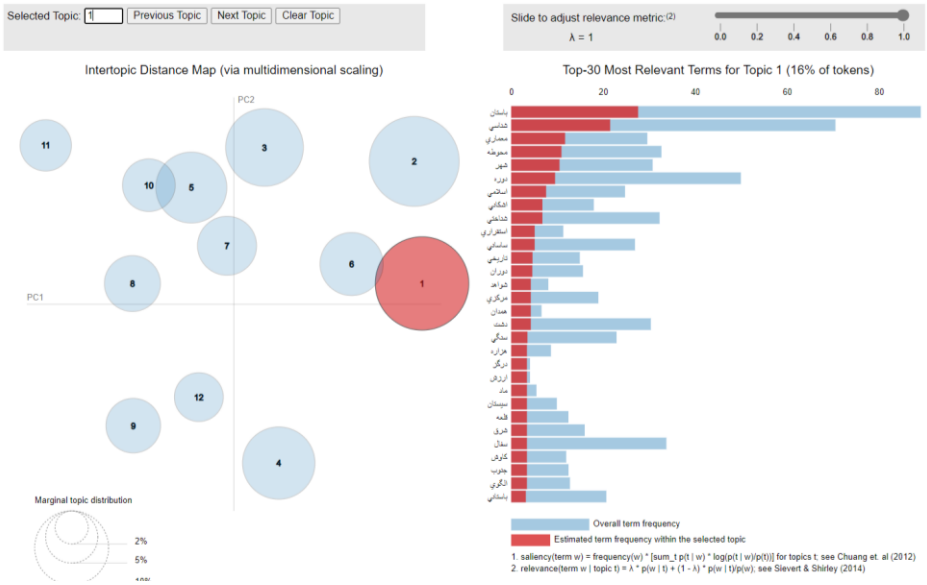
شکل ۸: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات باستان‌شناسی

اعمال روش بازبهنجاری روی داده‌های این نشریه در شکل ۹ نشان داده شده است. محور افقی تعداد موضوعات و محور عمودی مقادیر آنتروپی رونو را به ازای تعداد موضوعات مختلف نشان می‌دهد. در این شکل، آنتروپی رونو برای هر موضوع به صورت جداگانه محاسبه شده است. سپس جفت موضوعات با کوچکترین مقدار آنتروپی رونو انتخاب شده و ترکیب می‌شوند تا در نهایت به دو موضوع برسیم.



شکل ۹: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات باستان‌شناسی

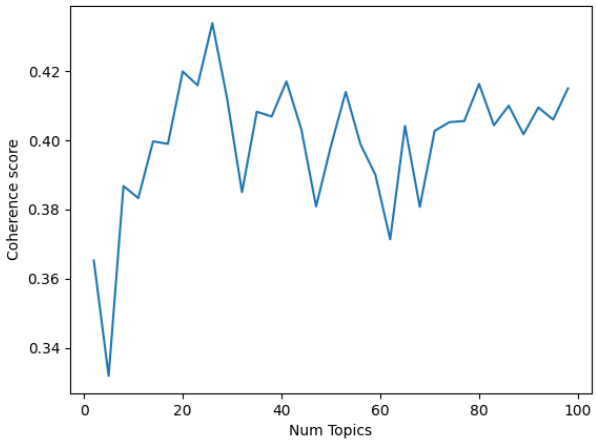
علاوه بر این، نمایش گرافیکی نتیجه مدلسازی روی نشریه مطالعات باستان‌شناسی در شکل ۱۰ نشان داده شده است. دایره‌های موجود در پنل سمت چپ، نشان‌دهنده موضوعات موجود در این نشریه و همچنین میزان اهمیت آنها است. فاصله میان دایره‌ها نیز نشان‌دهنده اختلاف این موضوعات است. البته لازم به ذکر است که این دایره‌ها از فضای چندبعدی به فضای دوبعدی نگاشت شده‌اند و فاصله میان دایره‌ها در واقع فاصله موضوعات در دو بعد نگاشت‌شده می‌باشد.



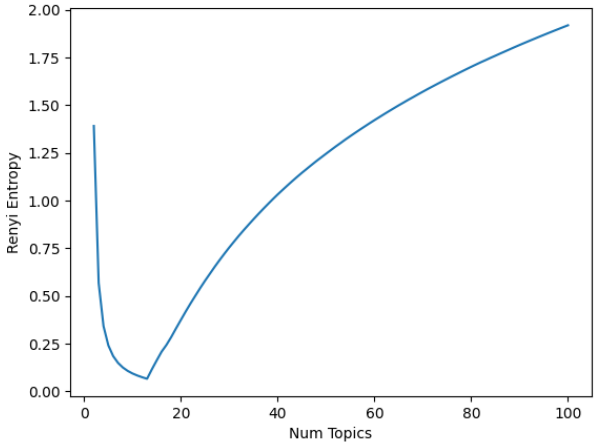
شکل ۱۰: نمای گرافیکی از موضوعات موجود در نشریه مطالعات باستان‌شناسی

نشریه مطالعات مدیریت در حوزه علوم انسانی است و هدف آن انتشار یافته‌های نوین پژوهشی در حیطه علوم ورزشی و کمک به ارتقای سطح دانش علمی و پژوهش در زمینه‌های مختلف علوم ورزشی و نهادینه کردن پژوهش‌های ورزشی در کشور است. مقالات چاپ شده در سال‌های اخیر در این نشریه مورد بررسی قرار گرفت و روی آن، معیار انسجام به ازای تعداد موضوعات مختلف اعمال گردید. نتیجه در شکل ۱۱ مشخص شده است. محور افقی تعداد موضوعات مختلف که از ۲ تا ۱۰۰ تغییر می‌کند و محور عمودی، معیار انسجام را نشان می‌دهد. با افزایش تعداد موضوعات از ۲ تا حدود ۲۲ معیار انسجام افزایش پیدا می‌کند و به ازای مقادیر بزرگتر، شروع به کاهش می‌نماید. البته این معیار، نوساناتی نیز دارد که احتمالا بدلیل قرار گرفتن کلمات نامرتبط به ازای تعداد موضوعات در مدلسازی موضوعی است.

شکل ۱۲ معیار آنتروپی رونو را روی نشریه مطالعات مدیریت نشان می‌دهد. محور افقی تعداد موضوعات و محور عمودی، مقادیر آنتروپی رونو را به ازای تعداد موضوعات مختلف مشخص می‌کند. مقادیر بزرگ آنتروپی رونو نشان‌دهنده ارتباط کم موضوعات موجود در متن است؛ در حالی که مقادیر کمتر آنتروپی رونو، بیشترین ارتباطات اطلاعاتی میان موضوعات را مشخص می‌نماید. با توجه به این نمودار، بهترین حالت که حاوی بیشترین ارتباط اطلاعاتی میان موضوعات است، ۱۱ موضوع می‌باشد.

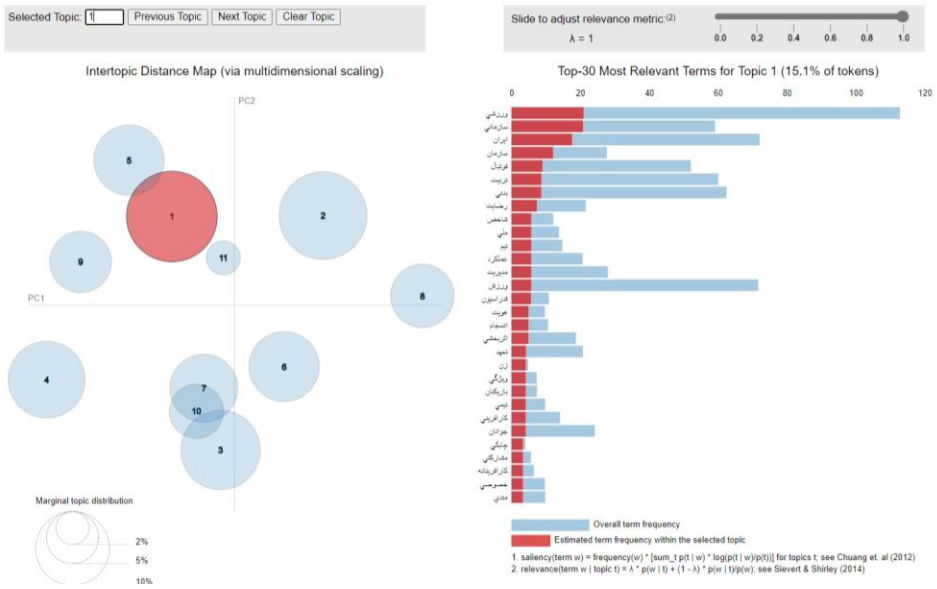


شکل ۱۱: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات مدیریت



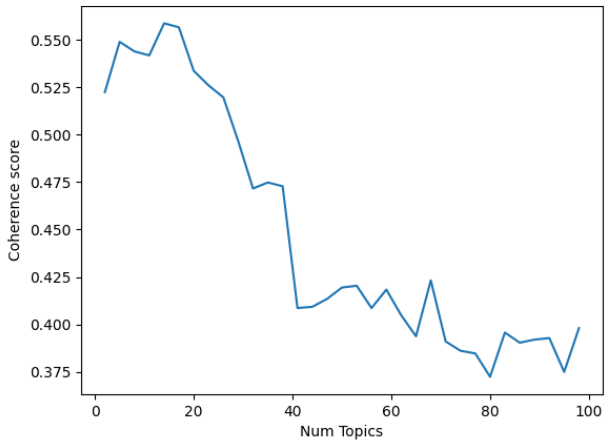
شکل ۱۲: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مطالعات مدیریت

نمای گرافیکی از نمایش نتیجه مدلسازی موضوعی روی نشریه مطالعات مدیریت در شکل ۱۳ مشخص شده است. پنل سمت چپ، موضوعات مختلف در این نشریه را نشان می‌دهد که اندازه هر دایره متناسب با میزان اهمیت هر موضوع است. همانطور که این شکل نشان می‌دهد، موضوع ۱ از اهمیت بالاتری برخوردار است. موضوعات ۳ و ۷ و ۱۰ با یکدیگر همپوشانی دارند. به عبارت دیگر، کلمات مشترکی میان این سه موضوع وجود دارد. موضوع ۱ و ۱۱ به ترتیب، پراهمیت‌ترین و کم‌اهمیت‌ترین موضوعات موجود در این نشریه هستند.

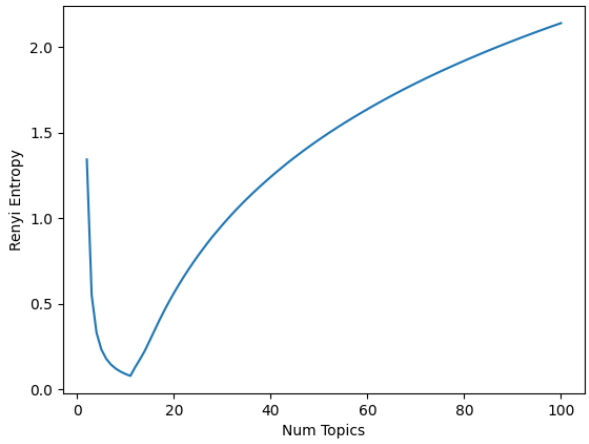


شکل ۱۳: نمای گرافیکی از موضوعات موجود در نشریه مطالعات مدیریت

نشریه فقه و اصول، نشریه دیگری در حوزه علوم انسانی است که به چاپ مقالات در حوزه فقه، اصول فقه، فلسفه فقه و موضوعات میان‌رشته‌ای چون فقه و حقوق می‌پردازد. برای بدست آوردن نتیجه مدلسازی موضوعی روی این نشریه نیز، نیاز به تخمین تعداد موضوعات داریم که برای بدست آوردن این پارامتر از روش گریدی استفاده کردیم که در شکل ۱۴ نشان داده شده است. همانطور که این شکل نشان می‌دهد مقدار بهینه برای این پارامتر حدود ۱۸ می‌باشد. نتایج حاصل از الگوریتم بازبهنجاری در شکل ۱۵ مشخص شده است.

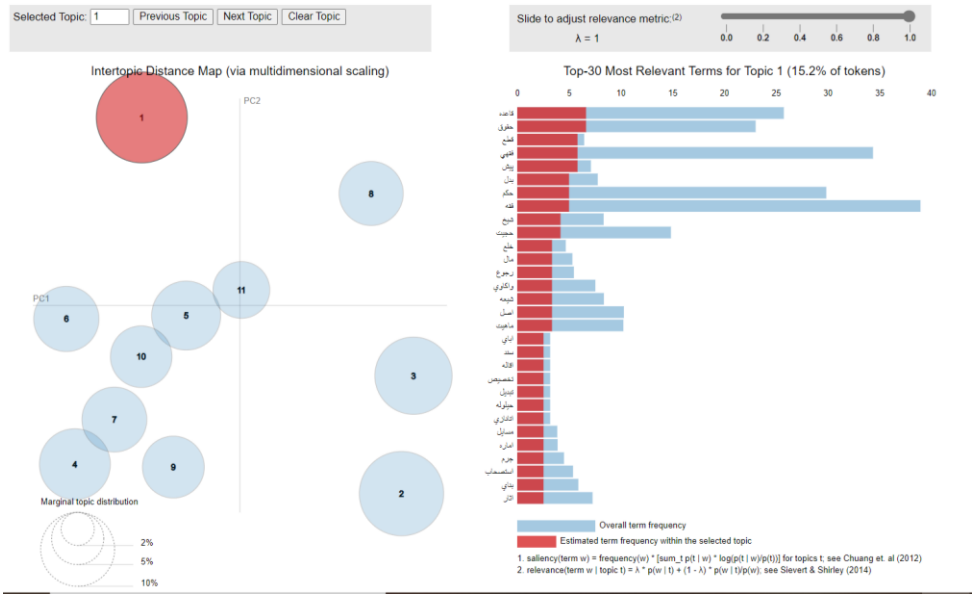


شکل ۱۴: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه فقه و اصول



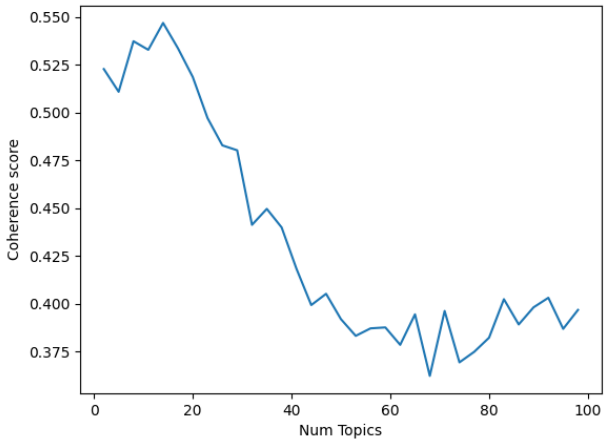
شکل ۱۵: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه فقه و اصول

نمای گرافیکی مدلسازی موضوعی روی نشریه فقه و اصول در شکل ۱۶ نشان داده شده است که می‌توان میزان اهمیت هر موضوع را توسط دایره‌های مختلفی که در پنل سمت چپ است، مشاهده نمود. همچنین تعداد موضوعات را می‌توان به صورت شهودی در این شکل تخمین زد که برابر با ۱۱ می‌باشد.



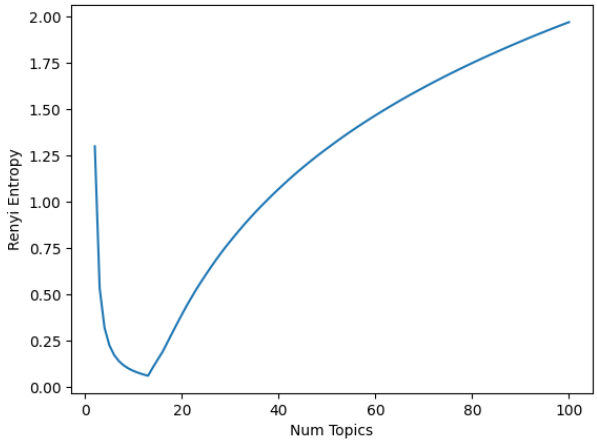
شکل ۱۶: نمای گرافیکی از موضوعات موجود در نشریه فقه و اصول

مهندسی برق و مهندسی کامپیوتر ایران، نشریه دیگری در حوزه فنی و مهندسی است که به معرفی جدیدترین دستاوردهای پژوهشی در زمینه مهندسی برق و کامپیوتر در ایران و جهان و انتشار نتایج پژوهش‌های اصیل متخصصین دانشگاهی و مراکز تحقیقاتی ایران و جهان در زمینه برق و کامپیوتر می‌پردازد. به منظور بدست آوردن تعداد موضوعات در این نشریه، نتایج حاصل از گریدی با توجه به معیار انسجام در شکل ۱۷ مشخص شده است. همانطور که این شکل نشان می‌دهد، از تعداد ۲ تا تقریبا ۱۷ موضوع، این نمودار صعودی است و بعد از آن، نمودار نزولی می‌شود. بالاترین معیار انسجام در این نشریه با تعداد موضوعی تقریبی ۱۷ بدست می‌آید.

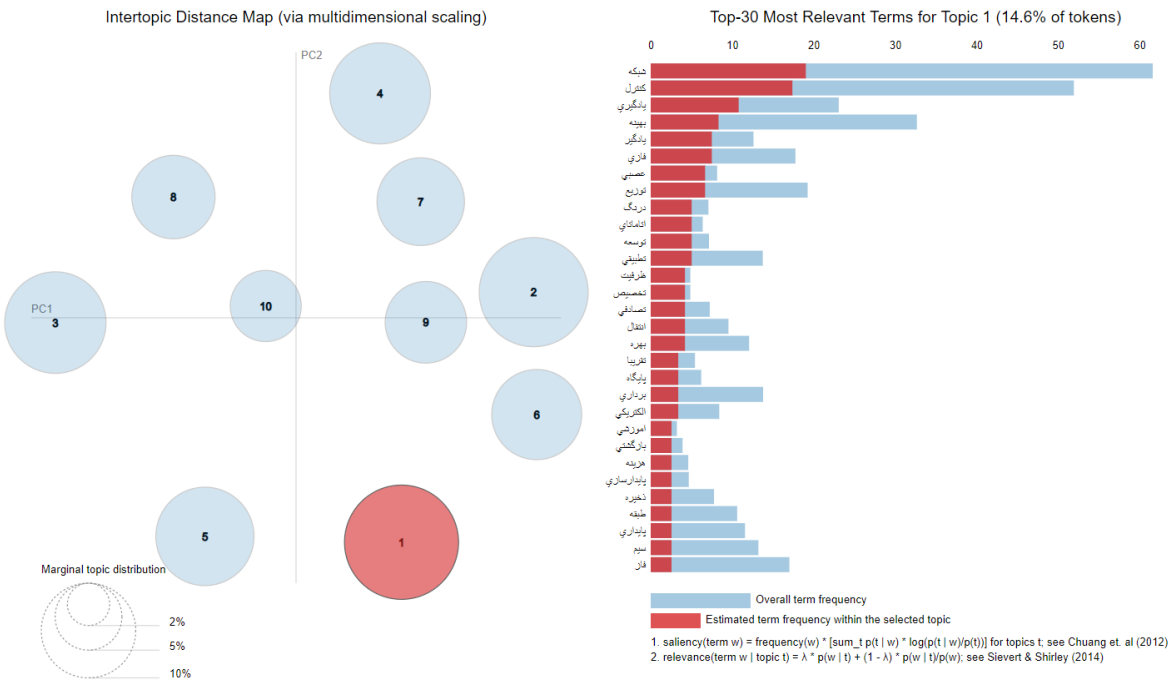


شکل ۱۷: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه مهندسی برق و کامپیوتر ایران

در جهت پاسخ به پرسش اول برای نشریه مهندسی برق و کامپیوتر ایران، علاوه بر روش بالا از الگوریتم بازبهنجاری نیز استفاده شده است که نتایج آن در شکل ۱۸ نشان داده شده است. شکل ۱۹ نمای گرافیکی از مدلسازی روی این نشریه را نشان می‌دهد.

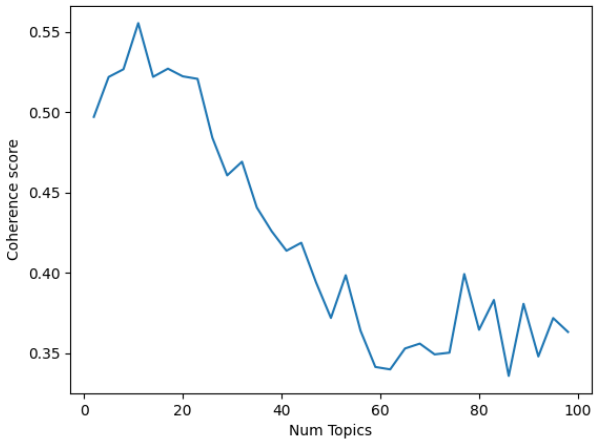


شکل ۱۸: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه مهندسی برق و کامپیوتر ایران

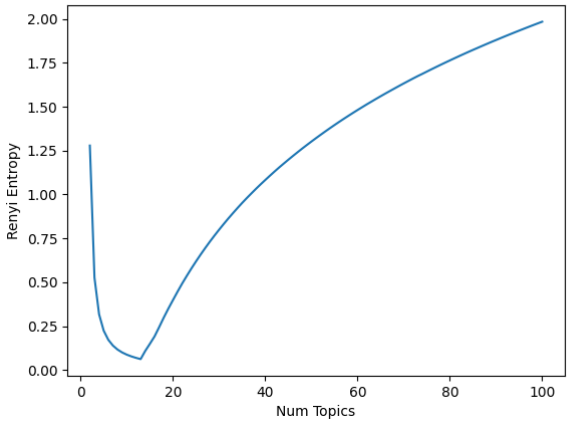


شکل ۱۹: نمای گرافیکی از موضوعات موجود در نشریه مهندسی برق و مهندسی کامپیوتر ایران

نشریه روش‌های عددی در مهندسی در حوزه فنی و مهندسی است که دستاوردهای پژوهشی محققان فارسی زبان را در زمینه‌های مختلف مهندسی که در آن از روش‌های عددی بهره گرفته شده، به چاپ می‌رساند. به منظور پاسخ به پرسش اول پژوهش در جهت بدست آوردن تعداد موضوعات و یا همان ابعاد مساله، از روش گریدی مبتنی بر انسجام (شکل ۲۰) و روش بازبهنجاری (شکل ۲۱) استفاده شده است. در هر دوی این نمودارها، محور افقی تعداد موضوعات می‌باشد که از تعداد موضوع ۲ تا ۱۰۰ تغییر می‌کند. محور عمودی در شکل ۲۰ معیار انسجام و در شکل ۲۱ آنتروپی رونو است. نمودار در شکل ۲۰ تا حدود تعداد ۱۲ موضوع به صورت صعودی و بعد از آن به صورت نزولی است. همین روند در شکل ۲۱ برعکس است؛ یعنی تا حدود تعداد ۱۲ موضوع نمودار نزولی و بعد از آن صعودی است. بالاترین معیار انسجام و کمترین معیار آنتروپی به معنی بهترین تعداد موضوع است.

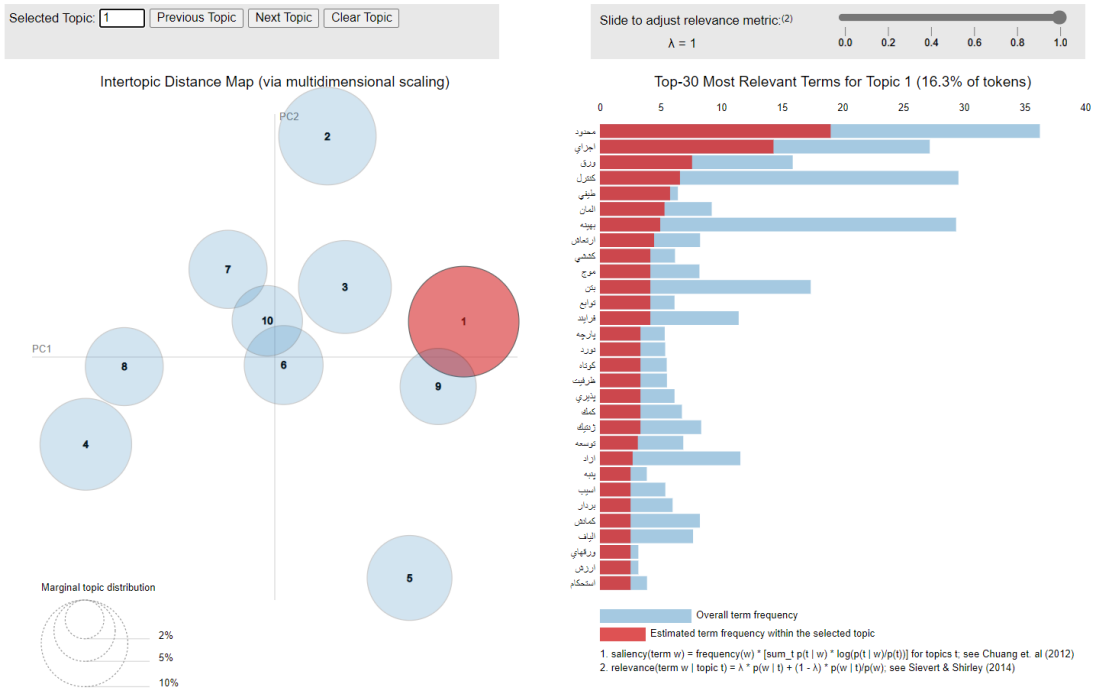


شکل ۲۰: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه روش‌های عددی در مهندسی



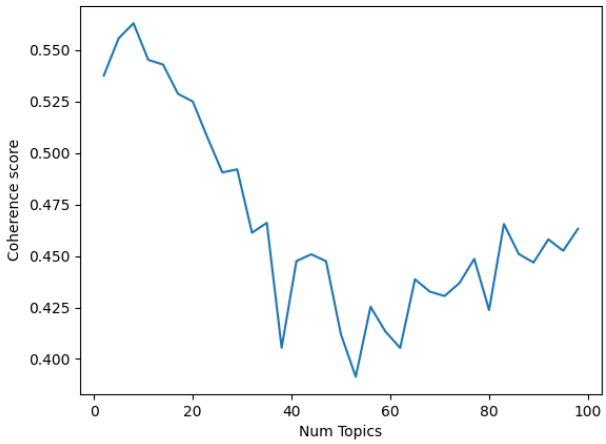
شکل ۲۱: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه روش‌های عددی در مهندسی

نمای گرافیکی از نتیجه مدلسازی موضوعی روی نشریه روش‌های عددی در مهندسی در شکل ۲۲ مشخص شده است. پنل سمت چپ، موضوعات و همچنین میزان اهمیت هر موضوع را نشان می‌دهد. به عنوان مثال موضوع ۱ و ۲ بهترین موضوعات در این نشریه هستند؛ چرا که نسبت به بقیه، ابعاد بزرگتری دارند. با انتخاب هر موضوع در پنل سمت چپ، می‌توان کلمات موجود در آن موضوع را در پنل سمت راست مشاهده نمود. با توجه به این شکل، کلمات "محدود، اجزای، ورق، کنترل، طیفی، المان و بهینه" از مهمترین کلمات در موضوع مشخص شده در پنل سمت چپ است.

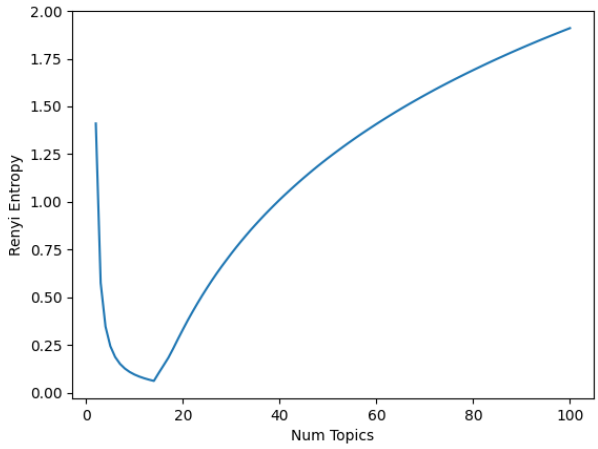


شکل ۲۲: نمای گرافیکی از موضوعات موجود در نشریه روش‌های عددی در مهندسی

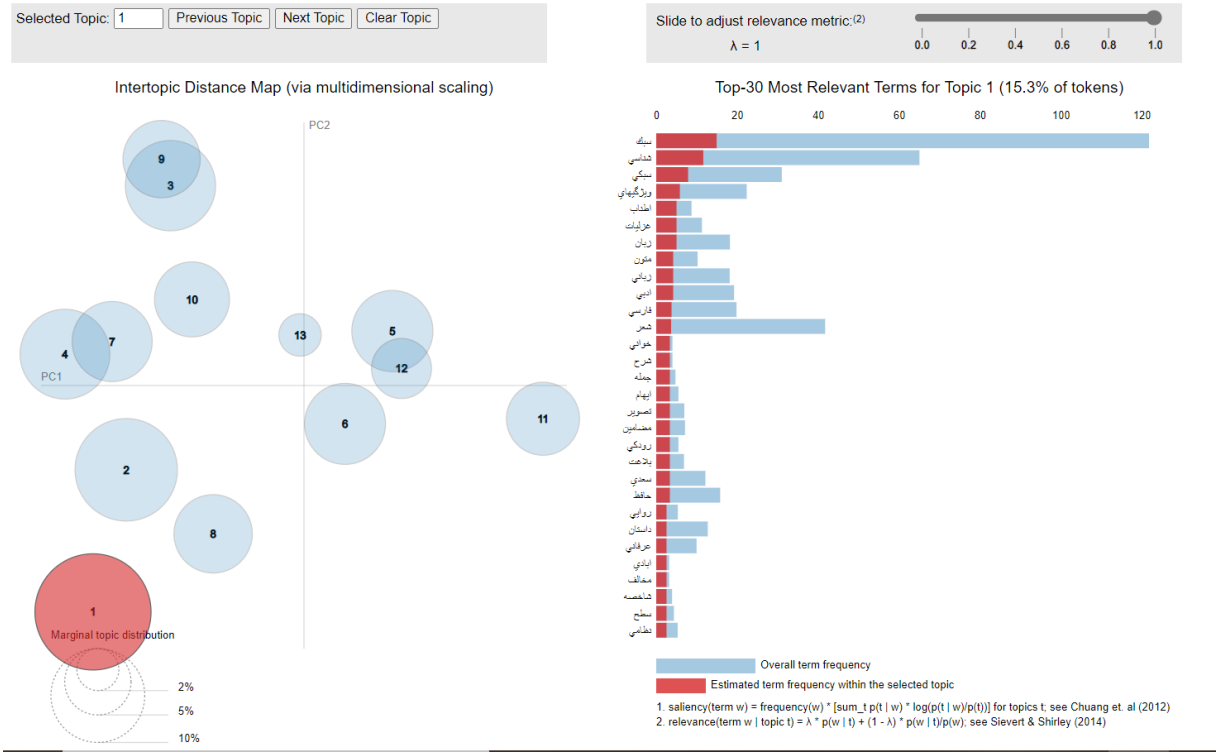
در نشریه سبک‌شناسی نظم فارسی در حوزه علوم انسانی، مقالات در زمینه زبان و ادب فارسی و گسترش این علم در سطح عمومی و دانشگاهی زبان و ادبیات فارسی است. نتیجه حاصل از اعمال روش‌های گریدی و الگوریتم بازبهنجاری به منظور پاسخ به پرسش اول پژوهش در شکل‌های ۲۳ و ۲۴ نشان داده شده است. شکل ۲۵ نمای گرافیکی حاصل از اعمال مدلسازی موضوعی روی این نشریه را نشان می‌دهد.



شکل ۲۳: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه سبک‌شناسی نظم فارسی

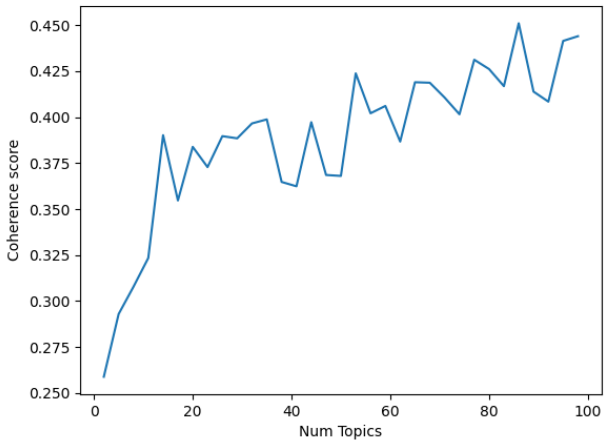


شکل ۲۴: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه سبک‌شناسی نظم فارسی



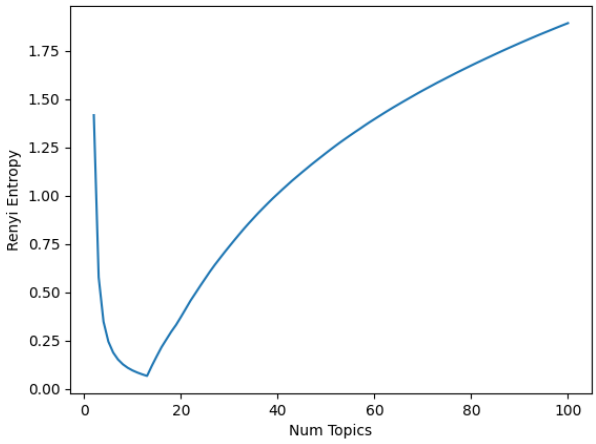
شکل ۲۵: نمای گرافیکی از موضوعات موجود در نشریه سبک‌شناسی نظم فارسی

یکی دیگر از نشریات معتبر در زمینه علوم انسانی، نشریه رهیافتی نو در مدیریت آموزشی است که با هدف ارتقای سطح دانش علمی در حوزه مدیریت آموزشی و ترویج و گسترش مرزهای دانش در حوزه مطالعاتی علوم تربیتی، به چاپ مقالاتی در حیطه مسائل علوم تربیتی و آموزشی می‌پردازد. برای پاسخ به پرسش اول پژوهشی برای این نشریه نیز از روش گریدی استفاده کردیم و نتایج آن در شکل ۲۶ مشخص شده است. بدین صورت که به ازای تعداد موضوعات مختلف که می‌تواند از ۲ تا ۱۰۰ تغییر کند، معیار انسجام را بدست آورده و در نهایت تعداد موضوع با بالاترین معیار انسجام را به عنوان یک تخمین مناسب از تعداد موضوعات و یا همان ابعاد مساله در نظر می‌گیریم. همانطور که در شکل ۲۶ مشخص شده است، در این شکل معیار انسجام به صورت صعودی است و با افزایش تعداد موضوعات، معیار انسجام نیز افزایش پیدا می‌کند. به صورتی که در حدود تعداد ۸۵ موضوع بالاترین معیار انسجام بدست آمده است.



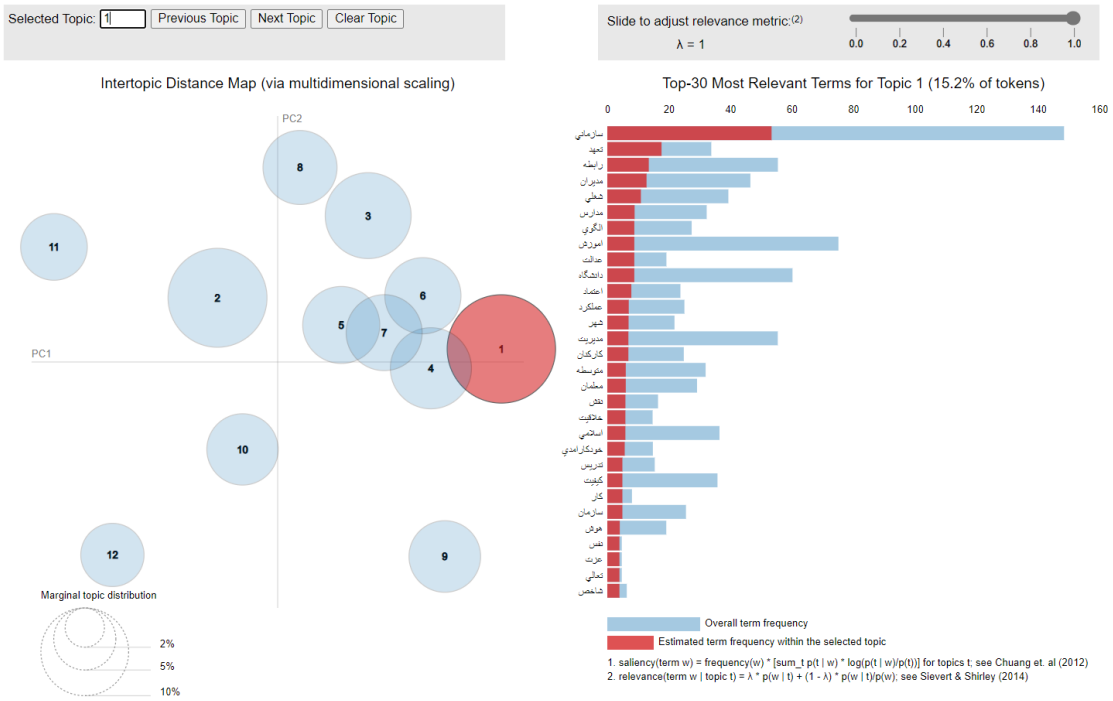
شکل ۲۶: معیار انسجام بر اساس تعداد موضوعات روی داده‌های نشریه رهیافتی نو در مدیریت آموزشی

به منظور بررسی بیشتر و بدست آوردن تعداد موضوعات (ابعاد مساله) الگوریتم بازبهنجاری را نیز روی داده‌های این نشریه اعمال کردیم که نتایج آن در شکل ۲۷ نشان داده شده است. آنتروپی رونو برای هر موضوع به صورت جداگانه محاسبه شده است؛ به صورتی که فقط از احتمالات کلمات در هر موضوع استفاده گردیده است. سپس جفت موضوعات با کوچکترین مقدار آنتروپی رونو انتخاب شده و روند ترکیب انجام می‌گردد.



شکل ۲۷: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه رهیافتی نو در مدیریت آموزشی

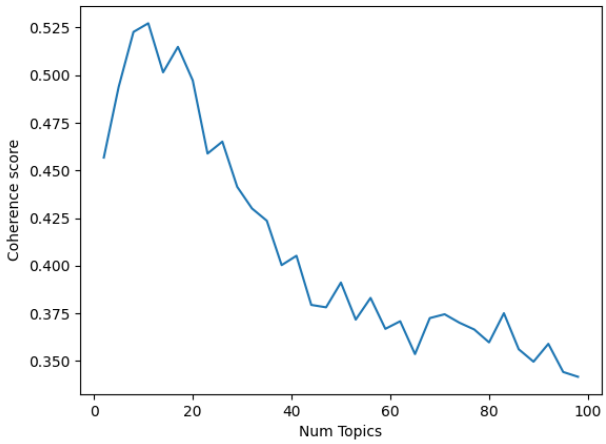
همانطور که این شکل نشان می‌دهد، از تعداد ۲ تا حدود ۱۲ موضوع، نمودار سیر نزولی گرفته و بعد از آن، صعودی است. با توجه به این تعداد، نتایج مدلسازی موضوعی روی نشریه رهیافتی نو در مدیریت آموزشی به صورت زیر بدست آمد و همچنین برای درک شهودی از نتایج مدلسازی موضوعی، شکل ۲۸ نمای گرافیکی را نشان می‌دهد.



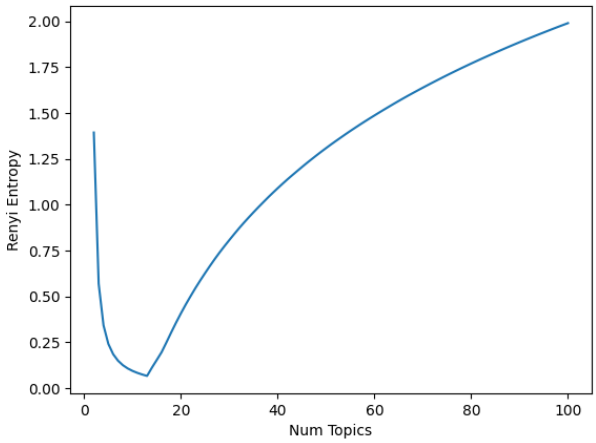
شکل ۲۸: نمای گرافیکی از موضوعات موجود در نشریه رهیافتی نو در مدیریت آموزشی

همانطور که این شکل نشان می‌دهد، تعداد ۱۲ موضوع برای این نشریه مناسب است. در صورتی که تعداد موضوعات را افزایش دهیم، میزان همپوشانی نشریات افزایش پیدا می‌کند. بنابراین در این نشریه، الگوریتم بازبهنجاری توانست به نتایج بهتری نسبت به روش گریدی دست پیدا کند. علاوه بر این، همانطور که جدول ۲ نشان می‌دهد، سرعت اجرای الگوریتم بازبهنجاری نسبت به گریدی به طور قابل توجهی بالاتر است.

آخرین نشریه‌ای که در این پژوهش مورد بررسی قرار می‌گیرد، نشریه صفه در حوزه هنر و معماری است. نشریه صفه به بررسی گذشته و حال معماری و شهرسازی ایران و جهان و مبانی نظری رویدادهای مربوط به حوزه شهرداری پرداخته و به چاپ مقالاتی در این زمینه می‌پردازد. برای این نشریه نیز پرسش‌های پژوهشی مورد بررسی قرار گرفتند. به منظور پاسخ به پرسش اول و یا همان بدست آوردن تعداد موضوعات یا ابعاد مساله، از دو روش، یکی مبتنی بر گریدی و دیگری مبتنی بر الگوریتم بازبهنجاری استفاده شده است. نتایج آن در شکل ۲۹ و ۳۰ نمایش داده شده است. در هر دو شکل، تعداد حدود ۱۱ موضوع بهترین نتیجه را به همراه دارد.



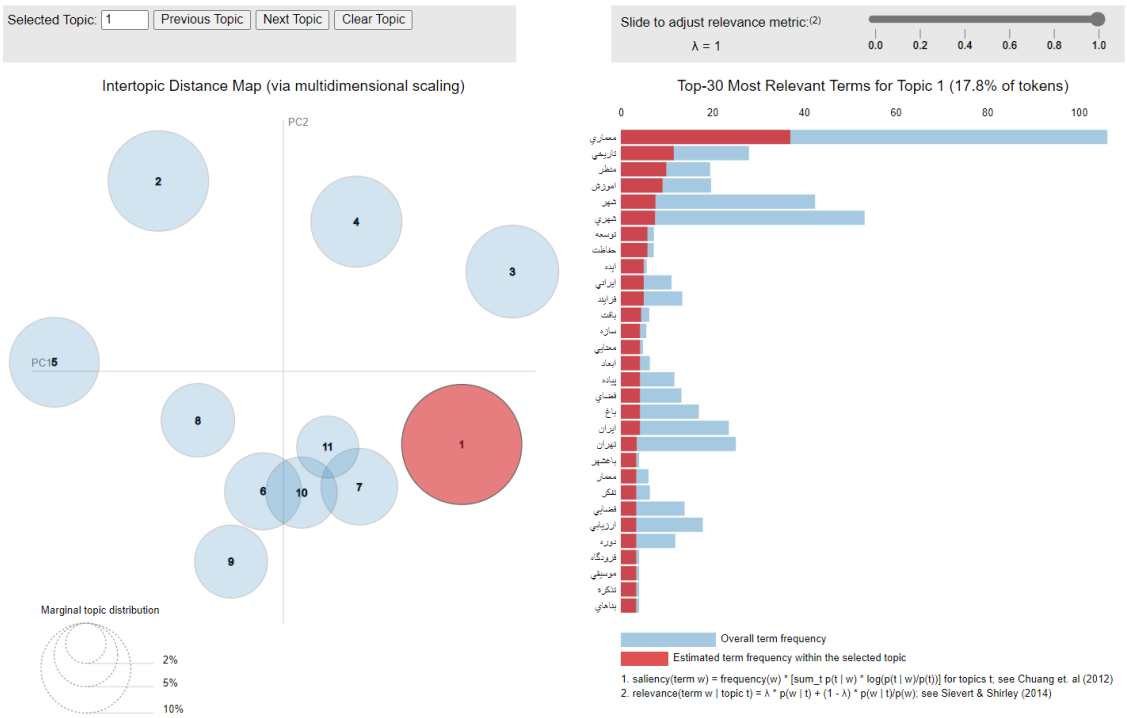
شکل ۲۹: نمای گرافیکی از موضوعات موجود در نشریه صفه



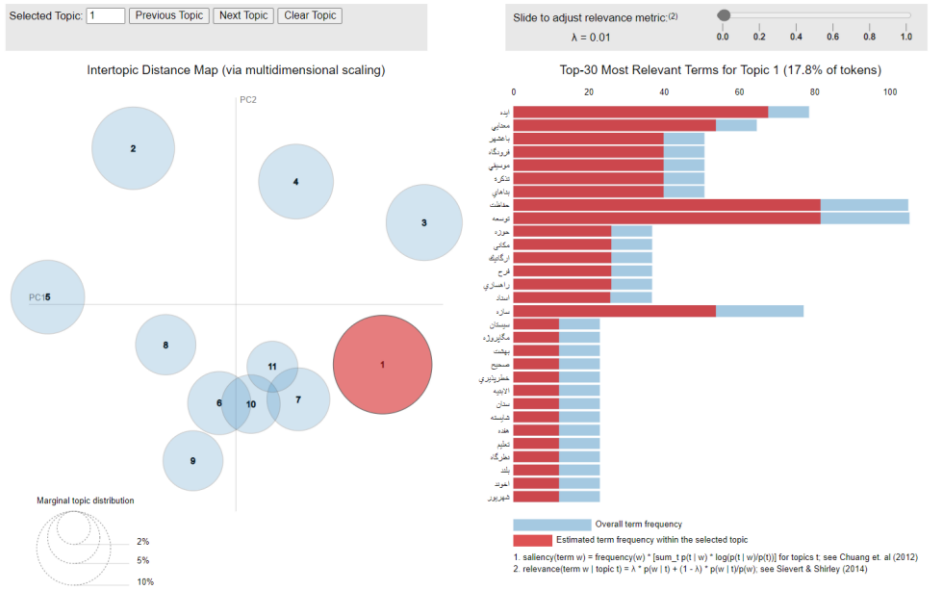
شکل ۳۰: معیار آنتروپی رونو بر اساس تعداد موضوعات روی داده‌های نشریه صفه

شکل ۳۱ نمای گرافیکی از نشریه صفه را نشان می‌دهد. به عنوان نمونه، موضوع ۱ به عنوان مهمترین موضوع در این مدل، انتخاب شده است. در سمت راست، کلمات مرتبط با این موضوع را مشاهده می‌کنید. همانطور که این شکل نشان می‌دهد، مهمترین کلمه در این مجموعه کلمات، "معماری" است که علاوه بر اینکه در مقالات زیادی در این نشریه استفاده شده، در موضوع ۱ نیز تکرار زیادی دارد.

همانطور که قبلا نیز عنوان گردید، یکی از پارامترهایی که در این نمودار مشخص شده است، پارامتر λ است که در این شکل مقدار ۱ گرفته است. این پارامتر میزان اهمیت یک کلمه در موضوع را مشخص می‌کند. هر چقدر احتمال کلمه در یک موضوع بیشتر و آن کلمه در کل مقالات آن نشریه، احتمال کمتری داشته باشد، از درجه اهمیت بالاتری برخوردار است. شکل ۳۲ نمای گرافیکی از مقالات انتخاب شده در نشریه صفه را با توجه به پارامتر دیگری از λ نشان می‌دهد. همانطور که این شکل نشان می‌دهد، با تغییر پارامتر λ، کلمات پراهمیت‌تر در موضوع ۱ نیز تغییر پیدا کرد؛ چرا که با این تغییر پارامتر، بخش دوم فرمول (۴-۳) نقش بیشتری برای تعیین رتبه کلمات گرفته است.



شکل ۳۱: نمای گرافیکی از موضوعات موجود در نشریه صفه



شکل ۳۲: نمای گرافیکی از موضوعات موجود در نشریه صفه با توجه به پارامتر $\lambda=0.01$

به منظور مقایسه زمانی میان دو روش گریدی مبتنی بر معیار انسجام و روش بازبهنجاری، مقایسه‌ای از مدت زمان اجرای دو روش بدست آوردن تعداد موضوعات و یا همان تعداد ابعاد مساله در این پژوهش داشتیم. جدول ۳ مقایسه دو روش گریدی و الگوریتم بازبهنجاری را نشان می‌دهد. همانطور که این جدول نشان می‌دهد، الگوریتم بازبهنجاری به صورت قابل ملاحظه‌ای نسبت به روش گریدی توانسته است تعداد موضوعات را با سرعت بالاتری بیابد.

جدول ۳: مقایسه زمان اجرا (ثانیه) دو روش مختلف برای بدست آوردن تعداد موضوعات

عنوان نشریه	مدت زمان اجرا با استفاده از روش گریدی	مدت زمان اجرا با استفاده از بازبهنجاری و آنتروپی رونو
مکانیک هوافضا	۶۱۹,۰۱۷	۲۱,۲۵
زمین‌شناسی ایران	۱۱۹۲,۴۹	۴۵,۶۵
مطالعات باستان‌شناسی	۹۰۱,۸۵	۴۵,۳
مطالعات مدیریت	۷۹۱,۴۹	۴۲,۸۶
فقه واصول	۱۸۰۸,۲۹	۴۵,۵۱
مهندسی برق و مهندسی کامپیوتر ایران	۲۵۳۴,۹۴	۴۶,۶۴
روش‌های عددی در مهندسی	۸۶۰,۴۴	۴۴,۲۳
سبک‌شناسی نظم فارسی	۶۹۹,۴۶	۴۴,۱۲
رهیافتی نو در مدیریت آموزشی	۱۵۴۶,۶۳	۴۲,۵۵
صفه	۶۹۴,۵۲	۴۶,۵۹

۲_۴_۴ نتایج حاصل از مدلسازی موضوعی روی نشریات فارسی (پاسخ به پرسش دوم)

در این بخش، پرسش دوم مورد بررسی قرار می‌گیرد. هدف این پرسش این است که مشخص کند مدلسازی موضوعی روی نشریات فارسی، چه گروه‌هایی از موضوعات را ایجاد می‌کند. همانطور که قبلا نیز عنوان گردید، قدم اول در مدلسازی موضوعی، بدست آوردن تعداد موضوعات می‌باشد که بخش قبلی به این مهم پرداخت. در این بخش با توجه به نتایج حاصله از بخش قبل، گروه‌های ایجاد شده از موضوعات را روی نشریات منتخب ارائه می‌دهیم.

اولین نشریه، نشریه مکانیک هوافضا می‌باشد که بر اساس نتایج حاصل از بخش قبلی، ۱۱ موضوع از آن استخراج می‌گردد. نتایج مدلسازی موضوعی با توجه به این تعداد موضوعات در زیر نشان داده می‌شود:

```
,۰)]
'۰,۰۱۹*"لوله" + ۰,۰۱۴*"ورق" + ۰,۰۱۳*"استوانه" + ۰,۰۱۳*"کمانش" + ۰,۰۱۱*"عددی" '
'+ ۰,۰۱۱*"پوسته" + ۰,۰۱۰*"محدود" + ۰,۰۱۰*"ارتعاشات" + ۰,۰۱۰*"پیش" + '
'۰,۰۱۰*"فشار"'),
,۱)
'۰,۰۱۷*"تاثیر" + ۰,۰۱۵*"محدود" + ۰,۰۰۸*"تیر" + ۰,۰۰۷*"اسیب" + ۰,۰۰۷*"الیاف" '
'+ ۰,۰۰۷*"مکانیکی" + ۰,۰۰۷*"تقویت" + ۰,۰۰۷*"مرکب" + ۰,۰۰۷*"سوخت" + '
'۰,۰۰۷*"وضعیت"'),
,۲)
'۰,۰۱۹*"حرارتی" + ۰,۰۱۷*"المان" + ۰,۰۱۴*"انتقال_حرارت" + ۰,۰۱۴*"برشی" + '
'۰,۰۱۲*"احتراق" + ۰,۰۱۲*"لغزش" + ۰,۰۱۲*"مواد" + ۰,۰۱۲*"مدرج" + ۰,۰۱۲*"تیوری" '
'+ ۰,۰۱۰*"پراش"'),
,۳)
'۰,۰۱۴*"بهینه" + ۰,۰۱۲*"ابزار" + ۰,۰۱۱*"دستگاه" + ۰,۰۰۹*"فرایند" + '
'۰,۰۰۹*"کمپرسور" + ۰,۰۰۹*"کاری" + ۰,۰۰۹*"ورق" + ۰,۰۰۹*"ارتعاشات" + '
'۰,۰۰۹*"ناوبری" + ۰,۰۰۸*"محدود"'),
,۴)
'۰,۰۱۹*"تیوری" + ۰,۰۱۷*"ورق" + ۰,۰۱۰*"سیگنال" + ۰,۰۱۰*"حرکت" + '
'۰,۰۱۰*"ارتعاشی" + ۰,۰۱۰*"کاری" + ۰,۰۱۰*"هدفمند" + ۰,۰۱۰*"کمانش" + '
'۰,۰۱۰*"بستر" + ۰,۰۰۸*"استاتیکی"'),
,۵)
'۰,۰۱۶*"پیچشی" + ۰,۰۱۶*"محدود" + ۰,۰۱۳*"موتور" + ۰,۰۱۳*"حرارتی" + '
'۰,۰۱۳*"ناپایداری" + ۰,۰۱۱*"عددی" + ۰,۰۱۱*"دینامیکی" + ۰,۰۰۹*"سطح" + '
'۰,۰۰۹*"انتقال_حرارت" + ۰,۰۰۹*"خنک"'),
,۶)
'۰,۰۲۸*"کنترل" + ۰,۰۱۵*"عددی" + ۰,۰۱۵*"ارتعاشات" + ۰,۰۱۴*"ازمایش" + '
'۰,۰۱۴*"متحرک" + ۰,۰۱۲*"غیرخطی" + ۰,۰۱۰*"جایی" + ۰,۰۱۰*"محفظه" + '
'۰,۰۰۹*"کریستال" + ۰,۰۰۹*"پیش"'),
,۷)
'۰,۰۱۸*"بهینه" + ۰,۰۱۶*"کنترل" + ۰,۰۱۱*"جریان" + ۰,۰۱۰*"پایداری" + '
'۰,۰۱۰*"هدایت" + ۰,۰۰۸*"محدود" + ۰,۰۰۸*"فرمان" + ۰,۰۰۸*"لوله" + '
'۰,۰۰۸*"گشتاور" + ۰,۰۰۶*"هواپیمای"'),
,۸)
'۰,۰۲۳*"مواد" + ۰,۰۱۷*"کمانش" + ۰,۰۱۵*"انتقال_حرارت" + ۰,۰۱۵*"استوانه" + '
'۰,۰۱۲*"برشی" + ۰,۰۱۲*"سازه" + ۰,۰۱۲*"ترکیبی" + ۰,۰۱۲*"مرکب" + '
```

```
        '۰,۰۱۲*"غیرخطی" + ۰,۰۰۹*"غلتک"'),
        ,۹)
        '۰,۰۲۰*"معکوس" + ۰,۰۱۶*"کنترل" + ۰,۰۱۳*"خطای" + ۰,۰۱۳*"ربات" + ۰,۰۱۳*"ضریب" + '
        '+ ۰,۰۱۰*"ماشین" + ۰,۰۱۰*"ژنتیک" + ۰,۰۱۰*"حرکت" + ۰,۰۱۰*"سینماتیک" + '
        '۰,۰۱۰*"ساز"'),
        ,۱۰)
        '۰,۰۲۶*"کنترل" + ۰,۰۲۰*"عددی" + ۰,۰۱۶*"فازی" + ۰,۰۱۳*"کاری" + '
        '۰,۰۱۳*"انتقال_حرارت" + ۰,۰۱۳*"موتور" + ۰,۰۱۳*"دنده" + ۰,۰۱۳*"بهینه" + '
        '۰,۰۱۱*"سیال" + ۰,۰۱۰*"جریان"'),
        ,۱۱)
        '۰,۰۱۵*"ارتعاشات" + ۰,۰۱۳*"زاویه" + ۰,۰۱۳*"ورق" + ۰,۰۱۳*"محیط" + '
        '۰,۰۱۳*"ژیروسکوپ" + ۰,۰۱۱*"پانل" + ۰,۰۰۹*"الومینیوم" + ۰,۰۰۹*"مد" + '
    '۰,۰۰۹*"تاثیر" + ۰,۰۰۹*"کسرات"')]
```

گروه‌های موضوعی نشریات از شماره صفر تا ۱۱ مشخص شده‌اند که در هر موضوع، لیستی از کلمات موجود در آن موضوع به همراه یک عدد بین صفر و یک نشان داده شده است. مقدار درج شده در کنار هر کلمه، نشان‌دهنده میزان اهمیت آن کلمه در موضوع است؛ به عنوان مثال، "لوله، ورق، و استوانه" از مهمترین کلمات موجود در موضوع ۱ است. در موضوع ۳، «انتقال حرارت، المان، و حرارتی» از اهمیت بالاتری نسبت به بقیه برخوردار است. تحلیل‌های بدست آمده در بخش قبل مؤید این امر است که تعداد موضوعات موجود در نشریه زمین‌شناسی، ۱۶ موضوع می‌باشد. با توجه به این امر، نتایج مدلسازی موضوعی روی داده‌های نشریه زمین‌شناسی ایران به صورت زیر است:

```
        ,۰)]
        '۰,۰۱۴*"بندی" + ۰,۰۱۴*"زون" + ۰,۰۱۰*"وارون" + ۰,۰۱۰*"منطقه" + '
        '۰,۰۱۰*"ایزوتوپ" + ۰,۰۱۰*"سنگ" + ۰,۰۰۹*"ژیوشیمیایی" + ۰,۰۰۹*"ایران" + '
        '۰,۰۰۷*"تفکیک" + ۰,۰۰۷*"اسماری"'),
        ,۱)
        '۰,۰۱۹*"تنش" + ۰,۰۱۶*"مس" + ۰,۰۱۶*"مخروط" + ۰,۰۱۶*"البرز" + ۰,۰۱۴*"شرق" + '
        '۰,۰۱۱*"جنوب" + ۰,۰۰۹*"غرب" + ۰,۰۰۹*"دیرینه" + ۰,۰۰۹*"مرکزی" + '
        '۰,۰۰۹*"دگرسانی"'),
        ,۲)
        '۰,۰۱۵*"تاقدیس" + ۰,۰۱۵*"کمربند" + ۰,۰۱۵*"ژیوشیمی" + ۰,۰۱۵*"زاگرس" + '
        '۰,۰۱۲*"چین_خوردگی" + ۰,۰۱۲*"استان" + ۰,۰۱۲*"هندسی" + ۰,۰۱۲*"چین" + '
        '۰,۰۱۲*"جنوب" + ۰,۰۱۲*"زمین"'),
        ,۳)
        '۰,۰۲۰*"جنوب" + ۰,۰۱۸*"زمین" + ۰,۰۱۵*"زایی" + ۰,۰۱۳*"مس" + ۰,۰۱۱*"سنگ" + '
        '۰,۰۱۱*"ایران" + ۰,۰۱۱*"کانه" + ۰,۰۱۱*"تنش" + ۰,۰۱۱*"منطقه" + ۰,۰۱۱*"شناسی"'),
```

```
,۴)
'۰,۰۲۵*"کانه" + ۰,۰۲۲*"سیال" + ۰,۰۱۶*"کانی" + ۰,۰۱۵*"زایی" + ۰,۰۱۳*"شناسی" + '
'۰,۰۱۳*"ایران" + ۰,۰۱۳*"سنگ" + ۰,۰۱۲*"کانسار" + ۰,۰۱۲*"غرب" + ۰,۰۱۲*"کوه"'),
,۵)
'۰,۰۱۵*"شمال" + ۰,۰۱۲*"رسوب" + ۰,۰۱۲*"گسل" + ۰,۰۱۰*"غرب" + ۰,۰۱۰*"کانی" + '
'۰,۰۱۰*"عناصر" + ۰,۰۱۰*"استان" + ۰,۰۰۸*"شیمی" + ۰,۰۰۸*"دگرسانی" + '
'۰,۰۰۸*"تالاب"'),
,۶)
'۰,۰۱۳*"اب" + ۰,۰۱۲*"تبدیل" + ۰,۰۱۲*"موجک" + ۰,۰۱۲*"منطقه" + ۰,۰۰۹*"ایلام" + '
'۰,۰۰۹*"چینه" + ۰,۰۰۹*"سطح" + ۰,۰۰۹*"دریا" + ۰,۰۰۹*"تغییرات" + '
'۰,۰۰۸*"پشتیبان"'),
,۷)
'۰,۰۱۶*"زمین" + ۰,۰۱۳*"قم" + ۰,۰۱۳*"شمال" + ۰,۰۱۲*"ایران" + ۰,۰۱۲*"جنوب" + '
'۰,۰۱۱*"برش" + ۰,۰۱۱*"شناسی" + ۰,۰۰۹*"شرق" + ۰,۰۰۹*"لرزه" + '
'۰,۰۰۹*"مغناطیسی"'),
,۸)
'۰,۰۲۴*"زمین" + ۰,۰۲۰*"جنوب" + ۰,۰۲۰*"گسل" + ۰,۰۱۶*"عناصر" + ۰,۰۱۶*"رخساره" '
'+ ۰,۰۱۲*"شرق" + ۰,۰۱۲*"مس" + ۰,۰۱۲*"لرزه" + ۰,۰۱۲*"گروه" + ۰,۰۱۲*"زاگرس"'),
,۹)
'۰,۰۲۸*"زمین" + ۰,۰۱۴*"شمال" + ۰,۰۱۴*"شناسی" + ۰,۰۱۴*"اب" + ۰,۰۱۲*"الودگی" + '
'۰,۰۱۱*"سنجی" + ۰,۰۰۹*"ژیوشیمی" + ۰,۰۰۹*"دگرسانی" + ۰,۰۰۹*"گرانی" + '
'۰,۰۰۹*"جنوب"'),
,۱۰)
'۰,۰۲۴*"سنگ" + ۰,۰۱۹*"ایران" + ۰,۰۱۶*"کانی" + ۰,۰۱۱*"زون" + ۰,۰۱۱*"چینه" + '
'۰,۰۱۱*"مجموعه" + ۰,۰۰۹*"غرب" + ۰,۰۰۸*"طلا" + ۰,۰۰۸*"افیولیتی" + '
'۰,۰۰۸*"توده"'),
,۱۱)
'۰,۰۱۴*"گوگرد" + ۰,۰۱۴*"تجزیه" + ۰,۰۱۴*"اب" + ۰,۰۱۴*"انومالی" + '
"*"پارامترهای" + ۰,۰۰۹*"الودگی" + '0.009 + anomaly" + ۰,۰۱۴*"پتانسیل" + ۰,۰۰۹*"
'۰,۰۰۹*"البرز" + ۰,۰۰۹*"شرایط"'),
,۱۲)
'۰,۰۲۷*"اب" + ۰,۰۲۳*"زیرزمینی" + ۰,۰۱۵*"دگرگونی" + ۰,۰۱۲*"موجک" + '
'۰,۰۱۲*"چاه" + ۰,۰۱۲*"تبریز" + ۰,۰۱۲*"زاگرس" + ۰,۰۰۸*"سطوح" + ۰,۰۰۸*"الوند" '
'+ ۰,۰۰۸*"نقش"'),
,۱۳)
'۰,۰۱۹*"چینه" + ۰,۰۱۶*"امفیبولیت" + ۰,۰۱۴*"غرب" + ۰,۰۱۲*"ایران" + '
'۰,۰۱۱*"برش" + ۰,۰۱۱*"زیرین" + ۰,۰۱۱*"سنگ" + ۰,۰۱۱*"کانی" + ۰,۰۱۱*"کپه_داغ" '
```

```
'+ ۰,۰۱۰*"شرق"'),
,۱۴)
' + '۰,۰۱۸*"تریاس" + ۰,۰۱۵*"راندگی" + ۰,۰۱۵*"جنوبی" + ۰,۰۱۵*"چین"
'۰,۰۱۵*"چین_خوردگی" + ۰,۰۱۵*"گسل" + ۰,۰۱۱*"زمین" + ۰,۰۱۱*"رخساره" + '
'۰,۰۱۱*"تاقدیس" + ۰,۰۱۱*"مغناطیسی"'),
,۱۵)
' + '۰,۰۲۶*"سنگ" + ۰,۰۱۸*"شمال" + ۰,۰۱۷*"منطقه" + ۰,۰۱۵*"شناسی" + ۰,۰۱۲*"شیمی"
'۰,۰۱۰*"پایدار" + ۰,۰۱۰*"مس" + ۰,۰۱۰*"ایزوتوپ" + ۰,۰۱۰*"اهن" + ۰,۰۱۰*"زمین"')]
```

نشریه مطالعات باستان‌شناسی دارای ۱۲ موضوع است که این تعداد با استفاده از تحلیل بازبهنجاری بدست آمده است. نتایج مدلسازی موضوعی روی نشریه مطالعات باستان‌شناسی به صورت زیر است.

```
,۰)]
' + '۰,۰۱۹*"پیش" + ۰,۰۱۸*"تپه" + ۰,۰۱۳*"براق" + ۰,۰۱۳*"نگاره" + ۰,۰۱۳*"مفرغ"
'۰,۰۱۱*"دوره" + ۰,۰۱۰*"اشیا" + ۰,۰۱۰*"تاریخ" + ۰,۰۱۰*"لرستان" + '
'۰,۰۱۰*"تدفین"'),
,۱)
' + '۰,۰۲۴*"ساسانی" + ۰,۰۱۷*"باستان" + ۰,۰۱۶*"معماری" + ۰,۰۱۵*"شهرستان"
'۰,۰۱۵*"نقش" + ۰,۰۱۳*"سنگ" + ۰,۰۱۳*"خلیج_فارس" + ۰,۰۱۲*"دوره" + '
'۰,۰۱۱*"شناختی" + ۰,۰۰۹*"محوطه"'),
,۲)
' + '۰,۰۳۸*"باستان" + ۰,۰۲۸*"شناسی" + ۰,۰۲۰*"مفرغ" + ۰,۰۱۴*"شهر" + ۰,۰۱۱*"هویت"
'+ ۰,۰۱۱*"اسلامی" + ۰,۰۱۱*"سازمان" + ۰,۰۱۱*"نقش" + ۰,۰۱۱*"عصر" + '
'۰,۰۱۱*"شناختی"'),
,۳)
' + '۰,۰۳۰*"سفال" + ۰,۰۲۳*"ایران" + ۰,۰۱۷*"سنگی" + ۰,۰۱۷*"شناسی" + ۰,۰۱۷*"تپه"
'۰,۰۱۴*"دشت" + ۰,۰۰۹*"فرهنگی" + ۰,۰۰۹*"شهرنشینی" + ۰,۰۰۹*"فرایندهای" + '
'۰,۰۰۹*"دوره"'),
,۴)
' + '۰,۰۲۶*"انالیز" + ۰,۰۲۲*"تپه" + ۰,۰۱۵*"پیکسی" + ۰,۰۱۵*"سکه" + ۰,۰۱۵*"عصر"
'۰,۰۱۴*"معماری" + ۰,۰۱۴*"محوطه" + ۰,۰۱۱*"عنصری" + ۰,۰۱۱*"اهن" + '
'۰,۰۱۱*"ایزوتوپ"'),
,۵)
' + '۰,۰۲۰*"هخامنشی" + ۰,۰۲۰*"کاخ" + ۰,۰۱۶*"استقراری" + ۰,۰۱۶*"مس"
'۰,۰۱۶*"شناسی" + ۰,۰۱۴*"الگوی" + ۰,۰۱۲*"عصر" + ۰,۰۱۲*"تخت_جمشید" + '
'۰,۰۱۲*"گل" + ۰,۰۱۲*"ساسانی"'),
,۶)
```

```
      ' + "سفال" + "*۰,۰۱۸" + "شناسی" + "*۰,۰۲۳" + "دوره" + "*۰,۰۲۶" + "باستان" + "*۰,۰۴۱'
      ' + "شهر" + "*۰,۰۱۳" + "ایران" + "*۰,۰۱۴" + "اشکانی" + "*۰,۰۱۵" + "شناختی" + "*۰,۰۱۸'
      '"), "ساسانی" + "*۰,۰۱۲" + "اسلامی" + "*۰,۰۱۳'
      ,۷)
      ' + "دوره" + "*۰,۰۱۶" + "حوزه" + "*۰,۰۱۶" + "باستان" + "*۰,۰۱۹" + "الگوی" + "*۰,۰۱۹'
      ' + "لاران" + "*۰,۰۱۳" + "چهارمحال_بختیاری" + "*۰,۰۱۳" + "رژیم" + "*۰,۰۱۳" + "شناسی" + "*۰,۰۱۶'
      '"), "رود" + "*۰,۰۱۳" + "غذایی" + "*۰,۰۱۳'
      ,۸)
      ' + "شناسی" + "*۰,۰۱۴" + "دوره" + "*۰,۰۱۶" + "اب" + "*۰,۰۱۷" + "باستان" + "*۰,۰۱۸" + "برجسته" + "*۰,۰۲۱'
      ' + "تقارن" + "*۰,۰۱۱" + "رصدخانه" + "*۰,۰۱۱" + "نقوش" + "*۰,۰۱۱" + "فارس" + "*۰,۰۱۱ +'
      '"), "کانال" + "*۰,۰۱۱'
      ,۹)
      ' + "عصر" + "*۰,۰۱۸" + "باستان" + "*۰,۰۱۹" + "ایران" + "*۰,۰۲۴" + "شناسی" + "*۰,۰۲۵" + "تپه" + "*۰,۰۳۰'
      '"), "فرهنگی" + "*۰,۰۱۰" + "شرق" + "*۰,۰۱۰" + "فلات" + "*۰,۰۱۲" + "اهن" + "*۰,۰۱۴" + "اباد" + "*۰,۰۱۸ +'
      ,۱۰)
      ' + "محوطه" + "*۰,۰۱۹" + "معماری" + "*۰,۰۲۰" + "شناسی" + "*۰,۰۳۶" + "باستان" + "*۰,۰۴۷'
      ' + "اشکانی" + "*۰,۰۱۲" + "اسلامی" + "*۰,۰۱۳" + "دوره" + "*۰,۰۱۶" + "شهر" + "*۰,۰۱۸'
      '"), "استقراری" + "*۰,۰۰۹" + "شناختی" + "*۰,۰۱۲'
      ,۱۱)
      ' + "نگاری" + "*۰,۰۱۵" + "دگرگونی" + "*۰,۰۱۵" + "تپه" + "*۰,۰۱۹" + "فرایندهای" + "*۰,۰۲۳'
      ' + "گورستان" + "*۰,۰۱۲" + "دوره" + "*۰,۰۱۲" + "خط" + "*۰,۰۱۲" + "ایزوتوپ" + "*۰,۰۱۲" + "شهر" + "*۰,۰۱۲'
]("هگمتانه" + "*۰,۰۰۸ +'
```

نتیجه مدلسازی موضوعی روی نشریه مطالعات مدیریت، با ۱۱ موضوع به صورت زیر است:

```
      ,۰)]
      ' + "بدنی" + "*۰,۰۲۶" + "ایران" + "*۰,۰۲۶" + "تربیت" + "*۰,۰۲۹" + "ورزشی" + "*۰,۰۴۳'
      ' + "کارافرینی" + "*۰,۰۱۷" + "مدیریت" + "*۰,۰۱۷" + "فوتبال" + "*۰,۰۲۳'
      '"), "عوامل" + "*۰,۰۱۵" + "کیفیت" + "*۰,۰۱۵" + "فناوری_اطلاعات" + "*۰,۰۱۷'
      ,۱)
      ' + "ورزش" + "*۰,۰۲۲" + "تربیت" + "*۰,۰۲۳" + "بدنی" + "*۰,۰۲۳" + "دانش" + "*۰,۰۲۸" + "ورزشی" + "*۰,۰۴۲'
      ' + "عوامل" + "*۰,۰۱۴" + "برنامه" + "*۰,۰۱۴" + "استان" + "*۰,۰۱۶" + "باشگاه" + "*۰,۰۲۰ +'
      '"), "ایران" + "*۰,۰۱۳'
      ,۲)
      ' + "استرس" + "*۰,۰۱۹" + "فوتبال" + "*۰,۰۲۵" + "ایران" + "*۰,۰۳۷" + "ورزش" + "*۰,۰۵۳'
      ' + "کارهای" + "*۰,۰۱۳" + "عوامل" + "*۰,۰۱۶" + "لیگ" + "*۰,۰۱۶" + "انگیزه" + "*۰,۰۱۶'
      '"), "توسعه" + "*۰,۰۱۳" + "اجتماعی" + "*۰,۰۱۳'
      ,۳)
```

```
        '۰,۰۴۹*"سازمانی" + ۰,۰۳۰*"سازمان" + ۰,۰۲۵*"تعهد" + ۰,۰۲۲*"ورزش" + '
        '۰,۰۱۷*"عدالت" + ۰,۰۱۷*"بدنی" + ۰,۰۱۷*"تربیت" + ۰,۰۱۷*"ایران" + '
        '۰,۰۱۵*"ورزشی" + ۰,۰۱۱*"تحول"'),
        ,۴)
        '۰,۰۲۵*"ورزش" + ۰,۰۲۵*"مدیران" + ۰,۰۲۳*"ورزشی" + ۰,۰۱۸*"سازمانی" + '
        '۰,۰۱۷*"ارتباطات" + ۰,۰۱۷*"مهارت" + ۰,۰۱۷*"چالش" + ۰,۰۱۷*"فدراسیون" + '
        '۰,۰۱۷*"ترویج" + ۰,۰۱۷*"امیخته"'),
        ,۵)
        '۰,۰۴۵*"ورزش" + ۰,۰۳۳*"فوتبال" + ۰,۰۳۳*"لیگ" + ۰,۰۳۰*"ایران" + ۰,۰۲۷*"برتر" '
        '+ ۰,۰۲۱*"حرفه" + ۰,۰۲۰*"ورزشی" + ۰,۰۱۸*"جوانان" + ۰,۰۱۶*"سازمانی" + '
        '۰,۰۱۵*"وزارت"'),
        ,۶)
        '۰,۰۳۷*"عملکرد" + ۰,۰۲۹*"ارزیابی" + ۰,۰۲۵*"بدنی" + ۰,۰۲۵*"عدالت" + '
        '۰,۰۲۵*"تربیت" + ۰,۰۲۱*"برنامه" + ۰,۰۱۷*"کیفیت" + ۰,۰۱۷*"ایران" + '
        '۰,۰۱۳*"علوم" + ۰,۰۱۳*"تهران"'),
        ,۷)
        '۰,۰۳۷*"ورزشی" + ۰,۰۲۴*"سازمانی" + ۰,۰۲۰*"یادگیری" + ۰,۰۲۰*"تربیت" + '
        '۰,۰۲۰*"بدنی" + ۰,۰۱۷*"روایی" + ۰,۰۱۴*"اعضای" + ۰,۰۱۴*"مدیران" + '
        '۰,۰۱۴*"علمی" + ۰,۰۱۴*"عوامل"'),
        ,۸)
        '۰,۰۵۸*"ورزشی" + ۰,۰۳۸*"بدنی" + ۰,۰۳۸*"تربیت" + ۰,۰۲۴*"رضایت" + '
        '۰,۰۱۷*"خودکارامدی" + ۰,۰۱۷*"اماکن" + ۰,۰۱۷*"علوم" + ۰,۰۱۷*"رابطه" + '
        '۰,۰۱۷*"تهران" + ۰,۰۱۷*"هییت"'),
        ,۹)
        '۰,۰۴۴*"ورزشی" + ۰,۰۱۹*"فوتبال" + ۰,۰۱۷*"ورزش" + ۰,۰۱۴*"نگرش" + '
        '۰,۰۱۴*"دانشجویان" + ۰,۰۱۴*"سبک" + ۰,۰۱۱*"کننده" + ۰,۰۱۱*"فعالیت" + '
        '۰,۰۱۱*"والیبال" + ۰,۰۱۱*"اعتماد"'),
        ,۱۰)
        '۰,۰۴۱*"ورزشی" + ۰,۰۴۰*"سازمانی" + ۰,۰۳۴*"ایران" + ۰,۰۲۳*"سازمان" + '
        '۰,۰۱۸*"فوتبال" + ۰,۰۱۷*"تربیت" + ۰,۰۱۷*"بدنی" + ۰,۰۱۴*"رضایت" + '
        '۰,۰۱۱*"مدیریت" + ۰,۰۱۱*"تیم"')]
```

نتایج مدلسازی موضوعی روی نشریه فقه و اصول با ۱۱ موضوع به صورت زیر است. در این نشریه، "حجاب، فقه، و عقد" از مهمترین کلمات موجود در موضوع ۱، و "تقلید و ادله" از مهمترین کلمات موجود در موضوع ۲ است.

```
        ,۰)]
        '۰,۰۱۶*"حجاب" + ۰,۰۱۶*"فقهی" + ۰,۰۱۶*"عقد" + ۰,۰۱۶*"قاعده" + ۰,۰۱۰*"مرد" + '
```

```
'۰,۰۱۰*"معصیه" + ۰,۰۱۰*"تعدد" + ۰,۰۱۰*"شهرت" + ۰,۰۱۰*"اجماع" + ۰,۰۱۰*"نکاح"'),
,۱)
'۰,۰۱۸*"تقلید" + ۰,۰۱۶*"ادله" + ۰,۰۱۶*"خیار" + ۰,۰۱۳*"قصاص" + ۰,۰۱۲*"عقد" + '
'۰,۰۱۲*"قاعده" + ۰,۰۱۱*"حکم" + ۰,۰۱۱*"فقهی" + ۰,۰۰۹*"کفار" + ۰,۰۰۹*"فساد"'),
,۲)
'۰,۰۱۶*"اصولیان" + ۰,۰۱۶*"عقلا" + ۰,۰۱۳*"فقه" + ۰,۰۱۳*"اخباریان" + '
'۰,۰۱۳*"احکام" + ۰,۰۱۰*"اخباری" + ۰,۰۱۰*"حجیت" + ۰,۰۱۰*"امامیه" + '
'۰,۰۱۰*"کفر" + ۰,۰۱۰*"ملاک"'),
,۳)
'۰,۰۲۲*"حکم" + ۰,۰۱۹*"سنت" + ۰,۰۱۹*"دلالت" + ۰,۰۱۷*"فقه" + ۰,۰۱۴*"مهر" + '
'۰,۰۱۴*"اهل" + ۰,۰۱۱*"وعده" + ۰,۰۱۱*"نظام" + ۰,۰۱۱*"حرام" + ۰,۰۱۱*"معاملات"'),
,۴)
'۰,۰۲۳*"ضمان" + ۰,۰۱۵*"قاعده" + ۰,۰۱۵*"اعیان" + ۰,۰۱۲*"قتل" + ۰,۰۱۲*"جرایم" '
'+ ۰,۰۱۲*"علیه" + ۰,۰۱۲*"حقوقی" + ۰,۰۱۲*"فقه" + ۰,۰۱۲*"عقد" + ۰,۰۱۲*"عدالت"'),
,۵)
'۰,۰۳۰*"فقه" + ۰,۰۱۷*"اثم" + ۰,۰۱۷*"اعانه" + ۰,۰۱۴*"کلی" + ۰,۰۱۴*"معروف" + '
'۰,۰۱۴*"عقل" + ۰,۰۱۴*"منکر" + ۰,۰۱۴*"مفهوم" + ۰,۰۱۰*"نظریه" + ۰,۰۱۰*"قواعد"'),
,۶)
'۰,۰۲۲*"فقهی" + ۰,۰۲۲*"قصاص" + ۰,۰۱۴*"حکم" + ۰,۰۱۳*"فقه" + ۰,۰۱۳*"قتل" + '
'۰,۰۱۱*"مالکیت" + ۰,۰۱۱*"قوادی" + ۰,۰۱۱*"اعراض" + ۰,۰۱۱*"قانون" + '
'۰,۰۰۹*"خمس"'),
,۷)
'۰,۰۱۹*"حکم" + ۰,۰۱۶*"ضمان" + ۰,۰۱۶*"دیدگاه" + ۰,۰۱۴*"اهل" + ۰,۰۱۳*"محقق" + '
'۰,۰۱۳*"اصل" + ۰,۰۱۰*"پول" + ۰,۰۱۰*"خبر_واحد" + ۰,۰۱۰*"سنت" + ۰,۰۰۸*"فقهی"'),
,۸)
'۰,۰۱۷*"واجب" + ۰,۰۱۷*"شناسی" + ۰,۰۱۵*"فقهی" + ۰,۰۱۱*"علم" + ۰,۰۱۱*"مناط" + '
'۰,۰۱۱*"فقه" + ۰,۰۱۱*"تنقیح" + ۰,۰۱۱*"ارث" + ۰,۰۰۹*"اصل" + ۰,۰۰۹*"نفسی"'),
,۹)
'۰,۰۳۳*"فقه" + ۰,۰۱۸*"اسلامی" + ۰,۰۱۸*"حقوق" + ۰,۰۱۵*"احسان" + ۰,۰۱۲*"فروش" '
'+ ۰,۰۱۲*"مسیولیت" + ۰,۰۱۲*"مدنی" + ۰,۰۰۹*"نقش" + ۰,۰۰۹*"امامیه" + '
'۰,۰۰۸*"حقوقی"'),
,۱۰)
'۰,۰۱۵*"قاعده" + ۰,۰۱۵*"حقوق" + ۰,۰۱۳*"فقهی" + ۰,۰۱۳*"قطع" + ۰,۰۱۳*"پیش" + '
'۰,۰۱۱*"بدل" + ۰,۰۱۱*"حکم" + ۰,۰۱۱*"فقه" + ۰,۰۰۹*"شیخ" + ۰,۰۰۹*"حجیت"')]
```

نتایج مدلسازی موضوعی روی نشریه مهندسی برق و مهندسی کامپیوتر ایران در ذیل آمده است.

```
,۰)]
```

```
' + "*۰,۰۱۹"نیازمندی" + "*۰,۰۱۴"قابلیت_اطمینان" + "*۰,۰۱۲"مبدل" + "*۰,۰۱۲"بهینه" + '
'*۰,۰۰۹"فرکانس" + "*۰,۰۰۹"اطمینان" + "*۰,۰۰۷"مجموعه" + "*۰,۰۰۷"مدیریت" + '
'*۰,۰۰۷"فاز" + "*۰,۰۰۷"یادگیری"'),
,۱)
' + "*۰,۰۲۱"کنترل" + "*۰,۰۱۷"کاهش" + "*۰,۰۱۵"فاز" + "*۰,۰۱۵"اینورتر" + '
'*۰,۰۱۵"جریان" + "*۰,۰۱۳"گشتاور" + "*۰,۰۱۳"ولتاژ" + "*۰,۰۱۱"توزیع" + '
'*۰,۰۱۱"موتور" + "*۰,۰۱۱"خطی"'),
,۲)
' + "*۰,۰۲۲"کنترل" + "*۰,۰۱۷"موتور" + "*۰,۰۱۴"ولتاژ" + "*۰,۰۱۲"تطبیقی" + '
'*۰,۰۱۲"القایی" + "*۰,۰۱۱"گشتاور" + "*۰,۰۰۹"مدولاسیون" + "*۰,۰۰۹"سرعت" + '
'*۰,۰۰۹"بهینه" + "*۰,۰۰۸"شبکه"'),
,۳)
' + "*۰,۰۵۲"شبکه" + "*۰,۰۲۴"حسگر" + "*۰,۰۲۴"سیم" + "*۰,۰۱۸"خوشه" + "*۰,۰۱۸"مسیریابی" '
'+ "*۰,۰۱۴"کوانتومی" + "*۰,۰۱۲"تصویر" + "*۰,۰۱۰"سرویس" + "*۰,۰۱۰"وب" + '
'*۰,۰۱۰"پوشش"'),
,۴)
' + "*۰,۰۱۵"منابع" + "*۰,۰۱۳"تصاویر" + "*۰,۰۱۱"مبدل" + "*۰,۰۱۱"ساز" + "*۰,۰۱۱"انتن" '
'+ "*۰,۰۱۰"ولتاژ" + "*۰,۰۱۰"ذخیره" + "*۰,۰۱۰"شبکه" + "*۰,۰۱۰"تولید_پراکنده" + '
'*۰,۰۱۰"فتوولتاییک"'),
,۵)
' + "*۰,۰۱۳"جریان" + "*۰,۰۱۳"مصرفی" + "*۰,۰۱۳"موازی" + "*۰,۰۱۲"یادگیری" + '
'*۰,۰۰۹"توازن" + "*۰,۰۰۹"جمع" + "*۰,۰۰۹"تقویت" + "*۰,۰۰۹"ولتاژ" + "*۰,۰۰۹"مدار" '
'+ "*۰,۰۰۹"تقویتی"'),
,۶)
' + "*۰,۰۱۶"شبکه" + "*۰,۰۱۲"باند" + "*۰,۰۱۰"پیش" + "*۰,۰۰۹"بهینه" + "*۰,۰۰۸"شارژ" + '
'*۰,۰۰۸"مفاهیم" + "*۰,۰۰۸"ردیابی" + "*۰,۰۰۸"فازی" + "*۰,۰۰۸"ثبت" + "*۰,۰۰۸"وب"'),
,۷)
' + "*۰,۰۱۷"موتور" + "*۰,۰۱۷"بهینه" + "*۰,۰۱۶"کنترل" + "*۰,۰۱۵"پایداری" + '
'*۰,۰۱۵"گشتاور" + "*۰,۰۱۵"برنامه_ریزی" + "*۰,۰۱۱"گیت" + "*۰,۰۰۹"مجدد" + '
'*۰,۰۰۸"ولتاژ" + "*۰,۰۰۸"پاسخ"'),
,۸)
' + "*۰,۰۳۶"شبکه" + "*۰,۰۳۳"کنترل" + "*۰,۰۲۰"یادگیری" + "*۰,۰۱۶"بهینه" + '
'*۰,۰۱۴"فازی" + "*۰,۰۱۴"یادگیر" + "*۰,۰۱۳"عصبی" + "*۰,۰۱۳"توزیع" + "*۰,۰۰۹"درنگ" '
'+ "*۰,۰۰۹"اتاماتای"'),
,۹)
' + "*۰,۰۱۷"کنترل" + "*۰,۰۱۷"ولتاژ" + "*۰,۰۱۴"مبدل" + "*۰,۰۱۴"رسمی" + "*۰,۰۱۴"توصیف" '
"*۰,۰۱۲"قدرت" + "*۰,۰۱۲+ ' dc" + 0.010' + "*۰,۰۱۰"باینری" + "فاز*
```

```
'۰,۰۱۰*"یکپارچه"')]
```

نتایج مدلسازی موضوعی روی نشریه روش‌های عددی در مهندسی به صورت زیر است. همانطور که بخش قبلی نشان داد، تعداد موضوع برای این نشریه، ۱۰ موضوع می‌باشد که با استفاده از نظریه بازبهنجاری بدست آمد.

```
,۰)]
'۰,۰۲۵*"شبکه" + ۰,۰۱۷*"مقاومت" + ۰,۰۱۲*"کنترل" + ۰,۰۱۰*"عصبی" + '
'۰,۰۱۰*"مصنوعی" + ۰,۰۱۰*"بهینه" + ۰,۰۱۰*"توزیع" + ۰,۰۰۹*"وضعیت" + '
'۰,۰۰۸*"قلب" + ۰,۰۰۸*"مغزی"'),
,۱)
'۰,۰۱۹*"جریان" + ۰,۰۱۰*"اجزای" + ۰,۰۱۰*"بتن" + ۰,۰۱۰*"غیرخطی" + '
'۰,۰۰۹*"محوری" + ۰,۰۰۹*"مقاومت" + ۰,۰۰۹*"قطره" + ۰,۰۰۷*"تنش" + '
'۰,۰۰۷*"فرایند" + ۰,۰۰۷*"پایداری"'),
,۲)
'۰,۰۱۵*"کنترل" + ۰,۰۱۵*"عددی" + ۰,۰۱۵*"شبکه" + ۰,۰۱۲*"موج" + ۰,۰۱۲*"فازی" + '
'۰,۰۱۲*"ولتاژ" + ۰,۰۱۲*"کانال" + ۰,۰۰۹*"نانولوله" + ۰,۰۰۹*"محیطی" + '
'۰,۰۰۹*"دسته"'),
,۳)
'۰,۰۱۳*"پرش" + ۰,۰۱۱*"غیرخطی" + ۰,۰۱۱*"محاسباتی" + ۰,۰۱۱*"محدود" + '
'۰,۰۰۹*"پوشش" + ۰,۰۰۹*"جوش" + ۰,۰۰۹*"پردازنده" + ۰,۰۰۹*"ترک" + ۰,۰۰۹*"زبان" '
'+ ۰,۰۰۹*"گرافیکی"'),
,۴)
'۰,۰۱۹*"جریان" + ۰,۰۱۳*"عددی" + ۰,۰۱۳*"بهینه" + ۰,۰۱۱*"کشش" + ۰,۰۱۱*"نخ" + '
'۰,۰۱۱*"استوانه" + ۰,۰۱۱*"ورق" + ۰,۰۱۱*"روزرسانی" + ۰,۰۱۱*"لایه" + '
'۰,۰۱۱*"مودال"'),
,۵)
'۰,۰۱۹*"کنترل" + ۰,۰۱۷*"بهینه" + ۰,۰۱۵*"جریان" + ۰,۰۱۲*"دینامیکی" + '
'۰,۰۱۰*"مونتاژ" + ۰,۰۱۰*"مرحله" + ۰,۰۱۰*"عصبی" + ۰,۰۱۰*"مسایل" + '
'۰,۰۱۰*"ابتکاری" + ۰,۰۱۰*"ماشین"'),
,۶)
'۰,۰۱۶*"محدود" + ۰,۰۱۵*"جریان" + ۰,۰۱۳*"بهینه" + ۰,۰۱۳*"شبکه" + '
'۰,۰۱۳*"انتقال" + ۰,۰۱۱*"پایداری" + ۰,۰۱۱*"جایی" + ۰,۰۱۱*"عددی" + '
'۰,۰۰۹*"محفظه" + ۰,۰۰۹*"اجزای"'),
,۷)
'۰,۰۳۳*"محدود" + ۰,۰۲۵*"اجزای" + ۰,۰۱۳*"ورق" + ۰,۰۱۲*"کنترل" + ۰,۰۱۰*"طیفی" '
'+ ۰,۰۰۹*"المان" + ۰,۰۰۹*"بهینه" + ۰,۰۰۸*"ارتعاش" + ۰,۰۰۷*"کششی" + '
'۰,۰۰۷*"موج"'),
```

```
    ,۸)
    ' + "جریان" + ۰,۰۱۶*"سازه" + ۰,۰۱۵*"اندرکنش" + ۰,۰۱۵*"یابی" + '
    '۰,۰۱۵*"کنترل" + ۰,۰۱۵*"دانایی" + ۰,۰۱۲*"وضعیت" + ۰,۰۱۰*"سطح" + ۰,۰۱۰*"زیست" '
    '+ ۰,۰۱۰*"ازاد"'),
    ,۹)
    '۰,۰۲۵*"صوتی" + ۰,۰۱۳*"بهینه" + ۰,۰۱۲*"انتقال" + ۰,۰۱۲*"شبکه" + '
    '۰,۰۱۲*"انتشار_امواج" + ۰,۰۱۲*"لیتیوم_نیوباته" + ۰,۰۱۲*"بلوره" + '
'۰,۰۱۲*"کوپلینگ_پیزوالکتریکی" + ۰,۰۱۰*"تیتانیم" + ۰,۰۰۹*"بعدی")]
```

نتایج مدلسازی موضوعی روی نشریه سبک‌شناسی نظم فارسی در جهت پاسخ به پرسش دوم را می‌توان در ذیل مشاهده نمود:

```
    ,۰)]
    ' "شعر" + ۰,۰۲۹*"تشبیه" + ۰,۰۱۵*"کودک" + ۰,۰۱۱*"نوجوان" + ۰,۰۱۱*"سبکی" '۰,۰۲۹*
    '+ ۰,۰۱۱*"الطیر" + ۰,۰۱۱*"عناصر" + ۰,۰۱۱*"رساله" + ۰,۰۰۸*"نادره" + '
    '۰,۰۰۸*"وجه"'),
    ,۱)
    '۰,۰۲۹*"تشبیه" + ۰,۰۲۰*"عطار" + ۰,۰۱۵*"زبان" + ۰,۰۱۵*"دیوان" + '
    '۰,۰۱۵*"غزلیات" + ۰,۰۱۵*"سبکی" + ۰,۰۱۵*"شاهنامه" + ۰,۰۱۰*"مرکب" + '
    '۰,۰۱۰*"حافظ" + ۰,۰۱۰*"ابیوردی"'),
    ,۲)
    ' "سبک" + ۰,۰۱۶*"زن" + ۰,۰۱۶*"زبان" + ۰,۰۱۳*"اشعار" + ۰,۰۱۳*"خاقانی" + '۰,۰۱۹*
    '۰,۰۱۳*"روانشناسی" + ۰,۰۱۰*"لیلی" + ۰,۰۱۰*"تداعی" + ۰,۰۱۰*"داستان" + '
    '۰,۰۱۰*"معانی"'),
    ,۳)
    '۰,۰۶۴*"سبک" + ۰,۰۴۵*"شناسی" + ۰,۰۱۵*"فارسی" + ۰,۰۱۱*"زبان" + '
    ' "شاهنامه" + ۰,۰۱۱*"سبکی" + ۰,۰۰۹*"اختر" + ۰,۰۰۹*"نامه" + ۰,۰۰۹*"شعر" '۰,۰۱۱*
    '+ ۰,۰۰۹*"حافظ"'),
    ,۴)
    '۰,۰۸۴*"سبک" + ۰,۰۱۹*"فکری" + ۰,۰۱۵*"شناسی" + ۰,۰۱۱*"ویژگیهای" + '
    '۰,۰۱۱*"خاقانی" + ۰,۰۱۰*"شعر" + ۰,۰۰۸*"محتوایی" + ۰,۰۰۸*"دوگانه" + '
    '۰,۰۰۸*"کنایی" + ۰,۰۰۸*"محتوا"'),
    ,۵)
    '۰,۰۴۱*"سبک" + ۰,۰۱۹*"شعر" + ۰,۰۱۹*"سبکی" + ۰,۰۱۷*"دوره" + ۰,۰۱۴*"امین" + '
    '۰,۰۱۴*"پور" + ۰,۰۱۴*"شناسی" + ۰,۰۱۲*"ویژگی" + ۰,۰۱۰*"غزل" + ۰,۰۱۰*"سعدی"'),
    ,۶)
    '۰,۰۳۱*"سبک" + ۰,۰۲۴*"شناسی" + ۰,۰۱۶*"سبکی" + ۰,۰۱۲*"ویژگیهای" + '
```

```
' + "اطناب" + "*۰,۰۱۰ + "غزلیات" + "*۰,۰۱۰ + "زبان" + "*۰,۰۱۰ + "زبانی" + "*۰,۰۰۹ '
'*۰,۰۰۹ + "ادبی" + "*۰,۰۰۹ + "متون"'),
,۷)
' + "سبک" + "*۰,۰۱۴ + "احمد" + "*۰,۰۱۴ + "سعدی" + "*۰,۰۱۴ + "شناسی" + "*۰,۰۱۴ + "دیوان" + '*۰,۰۳۰
' + "فردوسی" + "*۰,۰۱۲ + "بوستان" + "*۰,۰۱۱ + "زبانی" + "*۰,۰۱۱ + "ملمع" + '*۰,۰۱۱
'*۰,۰۱۱ + "اشعار"'),
,۸)
' + "کاشانی" + "*۰,۰۱۵ + "اثری" + "*۰,۰۱۵ + "فیض" + "*۰,۰۱۵ + "کتاب" + "*۰,۰۱۵ + "گلستان" + '*۰,۰۱۵
'*۰,۰۱۵ + "محتوایی" + "*۰,۰۱۵ + "سیاست" + "*۰,۰۱۵ + "ادبی" + "*۰,۰۱۵ + "عدالت" + '
'*۰,۰۱۵ + "سعدی"'),
,۹)
' + "سبک" + "*۰,۰۲۶ + "داستانی" + "*۰,۰۱۹ + "ادبیات" + "*۰,۰۱۹ + "شعر" + '*۰,۰۱۶
' + "داستان" + "*۰,۰۱۳ + "شگردهای" + "*۰,۰۱۳ + "شناسی" + "*۰,۰۱۳ + "استرابادی" + '*۰,۰۱۰
'*۰,۰۱۰ + "طنز" + "*۰,۰۱۰ + "درونمایه"'),
,۱۰)
' + "سبک" + "*۰,۰۳۱ + "شناسی" + "*۰,۰۲۱ + "خسرو" + "*۰,۰۱۴ + "ساختاری" + "*۰,۰۱۴ + "ادبی" + '*۰,۰۱۰
'*۰,۰۱۰ + "شهراشوب" + "*۰,۰۱۰ + "انواع" + "*۰,۰۱۰ + "ناصر" + "*۰,۰۱۰ + "هندی" + '
'*۰,۰۱۰ + "روایی"'),
,۱۱)
' + "سبک" + "*۰,۰۴۵ + "شناسی" + "*۰,۰۳۲ + "زبانی" + "*۰,۰۱۶ + "نامه" + "*۰,۰۱۶ + "شعر" + '*۰,۰۱۶
' + "فکری" + "*۰,۰۱۳ + "الدین" + "*۰,۰۱۳ + "ویژگیهای" + "*۰,۰۱۳ + "قران" + '*۰,۰۱۰
'*۰,۰۰۷ + "همخوشه"'),
,۱۲)
' + "سبک" + "*۰,۰۴۹ + "شعر" + "*۰,۰۳۵ + "شناسی" + "*۰,۰۳۰ + "قصیده" + "*۰,۰۱۴ + "دیوان" + '*۰,۰۱۱
' + "اشعار" + "*۰,۰۱۱ + "تلمیحات" + "*۰,۰۱۱ + "شخصیت" + "*۰,۰۱۱ + "عشقی" + '*۰,۰۱۱
'*۰,۰۱۱ + "اجتماعی"')]
```

نشریه مدیریت آموزشی، نشریه دیگر در حوزه علوم تربیتی، با هدف ارتقای سطح دانش علمی در حوزه مدیریت آموزشی است. موضوعات استخراج شده از این نشریه که در ذیل آمده است، کاملا متناسب با حوزه نشریه می‌باشد.

```
,۰)]
' + "برنامه" + "*۰,۰۳۰ + "درسی" + "*۰,۰۳۰ + "مدیریت" + "*۰,۰۲۶ + "کنترل" + '*۰,۰۱۹
'*۰,۰۱۹ + "اموزش" + "*۰,۰۱۹ + "تربیت" + "*۰,۰۱۹ + "دانش" + "*۰,۰۱۵ + "سبک" + '
'*۰,۰۱۵ + "کارامدی" + "*۰,۰۱۵ + "معلمان"'),
,۱)
' + "اموزشی" + "*۰,۰۶۹ + "سازمانی" + "*۰,۰۵۰ + "مدیران" + "*۰,۰۳۳ + "اسلامی" + '*۰,۰۲۰
' + "تربیت" + "*۰,۰۱۹ + "رویکرد" + "*۰,۰۱۶ + "ادراک" + "*۰,۰۱۴ + "عدالت" + '*۰,۰۱۳
'*۰,۰۱۹
```

'۰,۰۱۳*"دانشگاه" + ۰,۰۱۱*"رابطه"'),
,۲)
' + '۰,۰۲۹*"متوسطه" + ۰,۰۲۹*"اعتماد" + ۰,۰۲۴*"دانشگاه" + ۰,۰۲۰*"مهارت"
' + '۰,۰۱۸*"اموزش" + ۰,۰۱۶*"فلسفی" + ۰,۰۱۶*"مدیران" + ۰,۰۱۴*"دوره" + ۰,۰۱۳*"هوش"
'+ ۰,۰۱۳*"عاطفی'),
,۳)
' + '۰,۰۲۱*"اموزش" + ۰,۰۲۰*"دانشگاه" + ۰,۰۲۰*"دانش" + ۰,۰۱۸*"کیفیت"
' + '۰,۰۱۸*"علمی" + ۰,۰۱۸*"مدیریت" + ۰,۰۱۶*"مهارت" + ۰,۰۱۶*"شغلی"
'۰,۰۱۴*"ابتدایی" + ۰,۰۱۴*"کشاورزی'),
,۴)
' + '۰,۰۶۲*"سازمانی" + ۰,۰۳۰*"مدیریت" + ۰,۰۲۸*"کیفیت" + ۰,۰۲۳*"فرهنگ"
' + '۰,۰۲۲*"رابطه" + ۰,۰۱۹*"بهره_وری" + ۰,۰۱۹*"اموزش" + ۰,۰۱۹*"دانش"
'۰,۰۱۷*"دانشگاه" + ۰,۰۱۵*"سبک'),
,۵)
' + '۰,۰۵۵*"سازمانی" + ۰,۰۳۰*"اموزش" + ۰,۰۲۵*"هوش" + ۰,۰۲۲*"ابتدایی"
' + '۰,۰۱۹*"مدارس" + ۰,۰۱۹*"ازاد" + ۰,۰۱۹*"دانشگاه" + ۰,۰۱۹*"مدیران"
'۰,۰۱۷*"رابطه" + ۰,۰۱۷*"ارزیابی'),
,۶)
' + '۰,۰۴۰*"دانش" + ۰,۰۲۸*"اثربخشی" + ۰,۰۲۸*"سازمانی" + ۰,۰۲۴*"مدارس"
' + '۰,۰۲۳*"فرهنگ" + ۰,۰۲۳*"مشارکت" + ۰,۰۲۰*"اموزان" + ۰,۰۱۸*"رابطه"
'۰,۰۱۷*"مدیران" + ۰,۰۱۴*"اینترنت'),
,۷)
' + '۰,۰۵۸*"اموزش" + ۰,۰۲۵*"متوسطه" + ۰,۰۲۲*"معلمان" + ۰,۰۲۲*"کیفیت"
' + '۰,۰۲۲*"دانش" + ۰,۰۱۹*"دوره" + ۰,۰۱۹*"ترجمه" + ۰,۰۱۶*"یادگیری" + ۰,۰۱۶*"گرا"
'+ ۰,۰۱۶*"تاثیر'),
,۸)
' + '۰,۰۳۹*"سازمان" + ۰,۰۲۶*"اموزشی" + ۰,۰۲۳*"دبیرستان" + ۰,۰۲۲*"مهارت"
' + '۰,۰۱۸*"روانشناختی" + ۰,۰۱۸*"مدیر" + ۰,۰۱۸*"موفق" + ۰,۰۱۸*"معلمان"
'۰,۰۱۸*"یادگیرنده" + ۰,۰۱۶*"مدیران'),
,۹)
' + '۰,۰۳۵*"سازمانی" + ۰,۰۳۰*"دانشگاه" + ۰,۰۲۲*"اسلامی" + ۰,۰۲۰*"شغلی"
' + '۰,۰۲۰*"ازاد" + ۰,۰۱۹*"مدیریت" + ۰,۰۱۷*"تعهد" + ۰,۰۱۷*"اموزش"
'۰,۰۱۴*"فرهنگ" + ۰,۰۱۴*"دانش'),
,۱۰)
' + '۰,۰۹۲*"سازمانی" + ۰,۰۳۱*"تعهد" + ۰,۰۲۳*"رابطه" + ۰,۰۲۲*"مدیران"
' + '۰,۰۱۹*"شغلی" + ۰,۰۱۵*"مدارس" + ۰,۰۱۵*"الگوی" + ۰,۰۱۵*"اموزش"
'۰,۰۱۵*"عدالت" + ۰,۰۱۵*"دانشگاه'),

```
,۱۱)
'۰,۰۵۶*"سازمانی" + ۰,۰۴۴*"تحصیلی" + ۰,۰۲۶*"پیشرفت" + ۰,۰۲۵*"اجتماعی" + '
'۰,۰۱۸*"سکوت" + ۰,۰۱۸*"دانشگاه" + ۰,۰۱۶*"شغلی" + ۰,۰۱۶*"اموزش" + '
'۰,۰۱۳*"گرایی" + ۰,۰۱۳*"دانشجویان"')]
```

نتایج مدلسازی موضوعی روی نشریه صفه با ۱۱ موضوع به صورت زیر بدست آمد:

```
,۰)]
'۰,۰۲۶*"برنامه_ریزی" + ۰,۰۲۴*"شهری" + ۰,۰۲۳*"شهر" + ۰,۰۲۱*"معماری" + '
'۰,۰۱۶*"های" + ۰,۰۱۴*"پیاده" + ۰,۰۱۳*"تاریخ" + ۰,۰۱۲*"نظریه" + ۰,۰۱۲*"فضایی" '
'+ ۰,۰۱۲*"باغ"'),
,۱)
'۰,۰۲۳*"شهری" + ۰,۰۲۱*"دانش" + ۰,۰۱۹*"های" + ۰,۰۱۹*"شهر" + ۰,۰۱۷*"جهان" + '
'۰,۰۱۵*"مکان" + ۰,۰۱۵*"موزه" + ۰,۰۱۳*"تهران" + ۰,۰۱۱*"شهری" + ۰,۰۱۱*"رشد"'),
,۲)
'۰,۰۲۲*"ساختمان" + ۰,۰۱۹*"طرح" + ۰,۰۱۸*"شهری" + ۰,۰۱۴*"فتوولتاییک" + '
'۰,۰۱۳*"تاریخی" + ۰,۰۱۲*"معماری" + ۰,۰۱۱*"سبز" + ۰,۰۱۱*"یکپارچه" + '
'۰,۰۱۱*"منظر" + ۰,۰۱۱*"فضاهای"'),
,۳)
'۰,۰۵۷*"معماری" + ۰,۰۱۸*"تاریخی" + ۰,۰۱۵*"منظر" + ۰,۰۱۴*"اموزش" + '
'۰,۰۱۲*"شهر" + ۰,۰۱۲*"شهری" + ۰,۰۰۹*"توسعه" + ۰,۰۰۹*"حفاظت" + ۰,۰۰۸*"ایرانی" '
'+ ۰,۰۰۸*"فرایند"'),
,۴)
'۰,۰۲۶*"معماری" + ۰,۰۱۸*"دانشگاه" + ۰,۰۱۶*"منظر" + ۰,۰۱۱*"طراحی" + '
'۰,۰۱۱*"طرح" + ۰,۰۱۱*"ایران" + ۰,۰۱۱*"مسکن" + ۰,۰۰۹*"شهری" + ۰,۰۰۸*"اصفهان" '
'+ ۰,۰۰۸*"کارگاه"'),
,۵)
'۰,۰۴۰*"معماری" + ۰,۰۱۴*"تهران" + ۰,۰۱۲*"ادراک" + ۰,۰۱۰*"اجتماعی" + '
'۰,۰۱۰*"ارزیابی" + ۰,۰۰۹*"نور" + ۰,۰۰۸*"جغرافیایی" + ۰,۰۰۸*"پروژه" + '
'۰,۰۰۸*"مفهوم" + ۰,۰۰۸*"مسجد_جامع"'),
,۶)
'۰,۰۲۹*"معماری" + ۰,۰۲۵*"ایران" + ۰,۰۲۳*"مسکن" + ۰,۰۱۴*"اسلامی" + '
'۰,۰۱۲*"فناوری" + ۰,۰۱۰*"محیط" + ۰,۰۱۰*"کیفیت" + ۰,۰۰۸*"فرایند" + '
'۰,۰۰۸*"مثابه" + ۰,۰۰۸*"مسکونی"'),
,۷)
'۰,۰۲۱*"مکان" + ۰,۰۱۶*"معماری" + ۰,۰۱۴*"شهری" + ۰,۰۱۳*"یادگیری" + '
'۰,۰۱۳*"اموزش" + ۰,۰۱۱*"ارزیابی" + ۰,۰۱۱*"حس" + ۰,۰۱۱*"طراحی" + ۰,۰۱۰*"شهر" '
'+ ۰,۰۰۸*"حرارتی"'),
```

```
        ,۸)
        '۰,۰۲۴*"باغ" + ۰,۰۲۱*"معماری" + ۰,۰۱۸*"شهری" + ۰,۰۱۸*"خانه" + ۰,۰۱۵*"فضا" + '
        '۰,۰۱۲*"دوره" + ۰,۰۱۲*"خیابان" + ۰,۰۱۲*"تاریخی" + ۰,۰۱۲*"جنسیت" + '
        '۰,۰۰۹*"کتیبه"'),
        ,۹)
        '۰,۰۴۲*"شهری" + ۰,۰۲۷*"شهر" + ۰,۰۱۴*"فضاهای" + ۰,۰۱۱*"ساری" + ۰,۰۱۱*"منطقه" '
        '+ ۰,۰۱۱*"خطاهای" + ۰,۰۱۱*"مسکن" + ۰,۰۱۰*"موردی" + ۰,۰۰۹*"زمانی" + '
        '۰,۰۰۹*"تاب"'),
        ,۱۰)
        '۰,۰۷۶*"معماری" + ۰,۰۱۵*"انسان" + ۰,۰۱۱*"بصری" + ۰,۰۱۱*"قرن" + ۰,۰۱۱*"محیط" '
        '+ ۰,۰۱۱*"اموزش" + ۰,۰۰۸*"جمعی" + ۰,۰۰۸*"ازاد" + ۰,۰۰۸*"تهران" + '
'۰,۰۰۸*"کلاسیک"')]
```

فصل پنجم

بحث و نتیجه‌گیری

۵. بحث و نتیجه‌گیری

۵_۱ مقدمه

در بخش نخست این فصل، ابتدا نتیجه‌گیری از یافته‌های فصل قبل ارائه می‌گردد. سپس پیشنهادهای پژوهش در دو بخش پیشنهادهای اجرایی و پیشنهاد برای پژوهش‌های آتی بیان می‌شود.

۵_۲ نتیجه‌گیری

این پژوهش مدلسازی موضوعی را روی مقالات علمی انجام داد که در واقع توزیع احتمالی روی کلمات و یا خوشه‌هایی است که وزن‌ها را برای هر کلمه یا ترم مشخص می‌کنند. به منظور مدلسازی روی متون فارسی که این پژوهش، منحصرا روی مقالات فارسی انجام شده است، چندین مرحله جمع‌آوری داده، پیش‌پردازش، بدست آوردن تعداد موضوعات و مدلسازی موضوعی انجام گرفت. در نهایت نیز به منظور ارزیابی مدل بدست آمده، آن را به صورت کمی و شهودی مورد بررسی قرار دادیم.

تحلیل‌های این پژوهش روی ۲۰۰۰ مقاله از ۱۰ نشریه وزارت علوم انجام شد. در فاز پیش‌پردازش، لیستی از ایست‌واژه‌ها که منحصرا مربوط به مقالات علمی است، به عنوان یک دستاورد از این پژوهش استخراج و ارائه گردید. لازم به ذکر است که این لیست منحصرا مربوط به متون علمی است. به عنوان مثال، کلمه "پژوهش" در

متون عمومی جزو ایست‌واژه‌ها محسوب نمی‌شود؛ ولی در متون علمی، کلمه‌ای است که نقش متمایزکننده‌ای برای متون ندارد؛ چرا که در تمامی مقالات این کلمه وجود دارد.

در این پژوهش از یکی از بهترین الگوریتم‌ها برای مدلسازی موضوعی تحت عنوان LDA استفاده گردید. این روش فرض می‌کند که اسناد، موضوعات متعددی را نمایش می‌دهند. یعنی اسناد از کلماتی تشکیل شده‌اند که هر یک متعلق به یک موضوع است و نسبت موضوعات داخل یک متن با یکدیگر متفاوت می‌باشد. بنابراین با بدست آوردن این نسبت‌ها می‌توان هر سند را در یک موضوع خاص دسته‌بندی نمود. به عبارت دیگر، LDA فرض می‌کند که هر موضوع، توزیعی روی مجموعه‌ای از کلمات است و کلماتی که مربوط به یک موضوع هستند، در آن موضوع دارای احتمال بالایی می‌باشند.

یکی از چالش‌های موجود در مدلسازی توسط LDA، بدست آوردن تعداد موضوعات یا k می‌باشد. روش‌های موجود برای برخورد با این مشکل از روش‌های جستجوی گریدی استفاده می‌کنند. به عنوان مثال، از معیارهای استانداردی مانند سرگشتگی استفاده کرده و مقدار این معیار را برای پارامترهای مختلف مدل بدست می‌آورند. سپس پارامتری که به ازای آن بیشترین مقدار سرگشتگی را نتیجه می‌دهد، در نظر می‌گیرند. یک معیار دیگر که در روش‌های جستجوی گریدی استفاده می‌شود، معیار انسجام معنایی است که نتایجی بهتر از سرگشتگی را به همراه دارد و به قضاوت‌های انسانی شبیه‌تر است. بنابراین در ارزیابی مدل‌های موضوعی هم بهتر است که از معیار انسجام معنایی استفاده شود.

علیرغم اینکه روش‌های جستجوی گریدی در پژوهش‌های مختلفی مورد استفاده قرار گرفته است، ولی یکی از بزرگترین مشکلات آنها، مدت زمان بالای اجرای آن است. به عبارت دیگر، بدست آوردن تعداد موضوعات در هر متن، نیاز به بار محاسباتی بسیار زیادی دارد. بنابراین در این پژوهش، روش مبتنی بر نظریه بازبهنجاری که ساخت یک رویه برای تغییر مقیاس سیستم با حفظ رفتار سیستم است و همچنین آنتروپی رونو به‌کار گرفته شد.

بعضی از نشریات بررسی شده در این پژوهش، عملکرد یکسانی از هر دو روش گریدی و رویکرد مبتنی بر نظریه بازبهنجاری به همراه داشتند که از آن جمله می‌توان به نشریه زمین‌شناسی ایران، مطالعات باستان‌شناسی، مهندسی برق و کامپیوتر ایران، روش‌های عددی در مهندسی، سبک‌شناسی نظم فارسی، و نشریه صفه نام برد. به عبارت دیگر، در این نشریات هر دو روش به‌کار رفته در این پژوهش به منظور تخمین تعداد موضوعات موجود در نشریات، نتیجه یکسانی را به همراه داشت. اگرچه مدت زمان تخمین روش مبتنی بر نظریه بازبهنجاری در این نشریات نسبت به روش گریدی به طرز چشمگیری کمتر است. به عنوان مثال مدت زمان اجرای الگوریتم بازبهنجاری در نشریه زمین‌شناسی ایران ۴۵,۶۵ ثانیه بود؛ درحالیکه اجرای روش گریدی مبتنی بر معیار انسجام روی همین نشریه ۱۱۹۲,۴۹ ثانیه زمان برد.

نتایج حاصل از دو روش تخمین تعداد موضوع در نشریات مکانیک هوافضا، مطالعات مدیریت، و نشریه فقه و اصول اختلاف چندانی نداشت و قابل چشم‌پوشی بود؛ ولی مدت زمان اجرای این دو روش در این نشریات، اختلاف بسیار زیادی داشت. به عنوان مثال در نشریه مکانیک هوافضا، مدت زمان اجرای الگوریتم با روش گریدی، ۶۱۹,۰۱۷ ثانیه و مدت زمان اجرا با استفاده از بازبهنجاری و آنتروپی رونو ۲۱,۲۵ ثانیه بود؛ به عبارت دیگر روش بازبهنجاری حدودا ۲۹ برابر سریعتر از روش مبتنی بر گریدی در این نشریه عمل کرد.

در نشریه رهیافتی نو در مدیریت آموزشی، عملکرد این دو روش در تخمین تعداد موضوعات موجود در نشریه، تفاوت بسیار زیادی داشت؛ به صورتی که روش گریدی در این نشریه ۸۵ موضوع را تشخیص داد و روش مبتنی بر نظریه بازبهنجاری ۱۲ موضوع را در این نشریه تخمین زد. تحلیل نتایج گرافیکی از این نشریه، مؤید درستی روش مبتنی بر نظریه بازبهنجاری بود.

همانطور که در قسمت‌های قبل توضیح داده شد، بدست آوردن تعداد موضوعات موجود در یک متن، یکی از بزرگترین چالش‌های موجود در مدلسازی موضوعی است. این پژوهش نشان داد که روش مبتنی بر نظریه بازبهنجاری می‌تواند با سرعت و دقت بسیار خوبی، تعداد موضوعات موجود در یک متن را تشخیص دهد. سپس با استفاده از الگوریتم‌های تشخیص موضوعات مانند LDA می‌توان موضوعات موجود در نشریات را به صورت توزیعی از کلمات موجود در آن بدست آورد.

در مرحله بعد مدلسازی موضوعی روی تمامی نشریات، انجام گرفته و نتایج آن در این پژوهش ارائه گردید. از این نتایج می‌توان برای تشخیص موضوعات پنهان موجود در نشریات با توجه به مقالاتی که در سال‌های اخیر به چاپ رسانده‌اند، استفاده کرد. این نتایج به صورت گرافیکی هم به نمایش گذاشته شد که می‌تواند به تحلیل‌گر داده در جهت تحلیل متن و شناسایی موضوعات مهم کمک کند.

کلمات مهم موجود در یکی از موضوعات نشریه مکانیک هوافضا با ۱۲ موضوع، شامل "لوله، ورق، استوانه، کمانش، عددی، پوسته، و ارتعاشات" بود. احتمالات هر کلمه در موضوع، اهمیت آن کلمه در آن موضوع را مشخص کرد؛ به عنوان مثال کلمات "لوله و ورق" از مهمترین کلمات موجود در یکی از موضوعات این نشریه بود که کاملا با ماموریت این نشریه همخوانی دارد.

بعضی از کلمات موجود در موضوعات مختلف نشریات ممکن است مشترک باشند که این امر در یافته‌های فصل قبل نیز محقق گردید. به عنوان مثال، در نشریه زمین‌شناسی ایران، کلمه "شمال" در دو موضوع از این نشریه با احتمالات مختلف وجود داشت. در یک موضوع، این کلمه به همراه کلمات "ایران، جنوب، برش، و شرق" دیده شد و در موضوع دیگر کلمه شمال به همراه "آب، آلودگی، ژیوشیمی، و زیرزمینی" وجود داشت. این امر نشان‌دهنده همپوشانی موضوعات موجود در نشریات است.

موضوعات نسبتا مجزایی هم در نشریات وجود داشت که از آن جمله می‌توان به نشریه مطالعات مدیریت اشاره کرد که یکی از موضوعات آن، کلمات "عملکرد، ارزیابی، عدالت، برنامه، کیفیت، و پرسشنامه" را به همراه داشت؛ و دیگری کلمات "ورزش، لیگ، فوتبال، ایران، برتر، و حرفه". البته تعداد این موضوعات به نسبت کم بود؛ چرا که عمدتا کلمات موجود در موضوعات مختلف یک نشریه با یکدیگر همپوشانی دارند. البته هرچقدر تعداد موضوعات موجود و همچنین تعداد کلمات موجود در هر نشریه را کمتر انتخاب کنیم، میزان همپوشانی هم کمتر می‌شود.

۵_۳ پیشنهادهای اجرایی پژوهش

هر نشریه با توجه به اهداف و ماموریتی که عمدتا در قسمت‌های "درباره نشریه" یا "اهداف و چشم‌انداز نشریه" مشخص می‌شود، مقالات مرتبط را پذیرش، بررسی، و در نهایت چاپ می‌کند. با استفاده از یافته‌های این پژوهش، می‌توان به سردبیران نشریات کمک کرد تا میزان تطابق مقالات دریافتی را با اهداف نشریه در زمان اولیه ارسال مقاله توسط نویسندگان، مشخص نمایند.

از دیگر کاربردهای این پژوهش، می‌توان به سنجش میزان تطابق مقالات چاپ شده در نشریات با اهداف آن، در فرآیند ارزیابی نشریات اشاره نمود؛ هر چقدر میزان تطابق مقالات یک نشریه با اهداف و حوزه موضوعی نشریه بیشتر باشد، احتمالا آن نشریه از عملکرد بهتری برخوردار است.

۵_۴ پیشنهاد برای پژوهش‌های آتی

از پژوهش‌های آتی این پژوهش می‌توان به بررسی استفاده از روش‌های برنامه‌نویسی پویا بجای نظریه بازبهنجاری اشاره کرد. همچنین ارائه یک معیار، به منظور بررسی میزان تطابق مقالات چاپ شده در هر نشریه با اهداف آن به صورت خودکار، از دیگر تحقیقاتی است که در ادامه این پژوهش می‌تواند انجام شود.

پیوست‌ها

پیوست ۱

جدول ۱: لیست ایست‌واژه‌های فارسی در مقالات علمی

اولیه	براساس	حل	کاربرد	مقاله
آزاد	بررسی	خواص	کامل	منحنی
آزمایش	بیشتر	داده	کاهش	موجود
ابتدا	پدیده	دقت	کمتر	موثر
اثر	پایین	دلیل	مبتنی	نتیجه
اجرا	پرداخته	راستا	محاسبه	نتایج
اخیر	پژوهش	رفتار	مدل	نرخ
اخیرا	پیشنهاد	روش	مربوط	نرم
ادامه	تابع	روابط	مرتبط	نسبت
ارائه	تجربی	زیاد	مرتبه	یافته
ارتباط	تحت	زمینه	مرزی	برداری
اساس	تحلیل	ساختار	مساله	بلافاصله
استفاده	تخمین	سامانه	متفاوت	حداقل
افزایش	تحقیق	سیستم	مثال	حداکثر
اعمال	تشخیص	شامل	مختلف	علاوه برآن
الگوریتم	تشکیل	شبیه	مقایسه	عنوان
امکان	تطابق	شرط	مقدار	مقدار
انتخاب	تعیین	شکل	مناسب	منطقی
اندازه	تغییر	صورت	منظور	نهایتا
انرژی	توسط	طرف	مورد	همزمان
انجام	توجه	صفحه	میزان	هیچکدام
اهمیت	جدید	طراحی	مشخص	واقعی
ایجاد	جهت	عملکرد	مشخصه	کنونی
باتوجه	حجم	عنوان	مشخصات	یافته

باعث	حد	فرآیند	مطالعه	یکدیگر
بالا	حاصل	فرض	مطلوب	
بسیار	حاضر	قرار	معادله	
بدون	حالت	کار	معرفی	
برابر	حرکت	کارایی	مقادیر	

مراجع

مراجع

آقابخشی، علی‌اکبر (۱۳۸۶). نمایه‌سازی همارا: مفاهیم و روش‌ها. تهران: چاپار: ۱۳۸۶.

دامی، سینا؛ الیکایی، محمدرضا (۱۳۹۶). مدلسازی موضوعی رویدادهای اخبار مبتنی بر یادگیری عمیق افزایشی. چهارمین کنفرانس بین‌المللی مطالعات نوین در علوم کامپیوتر و فناوری اطلاعات.

دامی، سینا؛ طاهرزاده، سیداحمد (۱۳۹۶). شناسایی تهدیدهای امنیتی با استفاده از مدلسازی موضوعی LDA و ماشین بردار پشتیبان. کنفرانس ملی فناوری‌های نوین در مهندسی برق و کامپیوتر.

رحیمی، مرضیه؛ زاهدی، مرتضی؛ مشایخی، هدی (۱۳۹۷). یک مدل موضوعی احتمالاتی مبتنی بر روابط محلی واژگان در پنجره‌های همپوشان. پردازش علائم و داده‌ها. شماره ۴، پیاپی ۳۸، صفحه ۵۷-۷۰.

زمانی، محسن، دیانت، روح الله، صادق‌زاده، مهدی (۱۳۹۳). دسته‌بندی متون فارسی با استفاده از روش آنالیز معنایی پنهان احتمالاتی. اولین همایش ملی کاربرد سیستم‌های هوشمند (محاسبات نرم) در علوم و صنایع.

شکری، سعید؛ معصومی، بهروز (۱۳۹۵). خوشه‌بندی معنایی متن با استفاده از تخصیص پنهان دیریکله و الگوریتم ژنتیک. چهارمین کنفرانس بین‌المللی پژوهش در علوم و تکنولوژی.

عظیمی همت، منیره؛ شمس عزت، فاطمه (۱۳۹۴). مروری بر متن‌کاوی متون فارسی. دومین کنفرانس بین‌المللی و سومین همایش ملی کاربرد فناوری‌های نوین در علوم مهندسی.

فتاحی، رحمت‌الله. (۱۳۸۶) از آرمان ها تا واقعیت: تحلیلی از مهم ترین چالش ها و رویکردهای سازماندهی اطلاعات در عصر حاضر. کتابداری و اطلاع‌رسانی. ۱۰(۴): ۵-۲۶.

گیلوری، عباس (۱۳۷۹). نمایه‌سازی خودکار (گذشته، حال، آینده). تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی (پیام کتابخانه سابق)، زمستان ۱۳۷۹. شماره ۳۹، ص ۱۷-۲۵.

هاشم‌زاده، محمدجواد؛ نخعی، زینب؛ مرادی مقدم، حسین (۱۳۹۲). کاربرد و تعدیل قانون زیف و الگوی بازو در بازشناسی واژه‌های بازدارنده زبان فارسی با استفاده از خوشه زبانی مقالات علمی-پژوهشی رشته کتابداری و اطلاع‌رسانی. پژوهشنامه کتابداری و اطلاع‌رسانی، ۳ (۲)، ۱۹۱-۲۰۸.

یوسفان، ا. (۱۳۸۲). *یک سیستم بازیابی اطلاعات متنی برای زبان فارسی بر پایه نمایه‌گذاری معانی پنهان*. پایان‌نامه کارشناسی ارشد. شیراز: دانشگاه شیراز.

Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text Summarization: A Brief Review. In *Recent Advances in NLP: The Case of Arabic Language* (pp. 1-15). Springer, Cham.

Akturk, E., Bagci, G. B., & Sever, R. (2007). Is Sharma-Mittal entropy really a step beyond Tsallis and Rényi entropies?. *arXiv preprint cond-mat/0703277*.

Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., & Bevacqua, A. (2013). Probabilistic topic models for sequence data. *Machine learning*, *93*(1), 5-29.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, *50*(4), 495-510.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, vector spaces, and information retrieval. *SIAM review*, *41*(2), 335-362.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).

Crain, S. P., Zhou, K., Yang, S. H., & Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In *Mining text data* (pp. 129-161). Springer, Boston, MA.

Davarpanah, M. R., Sanji, M., & Aramideh, M. (2009). Farsi lexical analysis and stop word list. *Library Hi Tech*.

De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (Vol. 6). John Wiley & Sons.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, *97*(457), 77-87.

Feldman, R., & Dagan, I. (1995). KDT-Knowledge Discovery in Texts. Proceddings of the First International Conference on Knowledge Discovery KDD. 112-117.

Ghorab, M. R., Zhou, D., O'connor, A., & Wade, V. (2013). Personalised information retrieval: survey and classification. *User Modeling and User-Adapted Interaction, 23*(4), 381-443.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review, 114*(2), 211.

Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *In Ldv Forum*, Vol. 20, No. 1, pp. 19-62.

Jameel, S., Lam, W., & Bing, L. (2015). Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval Journal, 18*(4), 283-330.

Janani, R., & Vijayarani, S. (2019). Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Systems with Applications, 134*, 192-200.

Kadanoff, L. P. (2000). *Statistical physics: statics, dynamics and renormalization*. World Scientific Publishing Company.

Keyes, A. M. (1995). The Value of the Special Library: Review and Analysis. *Special libraries, 86*(3), 172-87.

Kherwa, P., & Bansal, P. (2017). Latent Semantic Analysis: An Approach to Understand Semantic of Text. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)* (pp. 870-874). IEEE.

Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems, 7*(24).

Koltcov, S. N. (2017). A thermodynamic approach to selecting a number of clusters based on topic modeling. *Technical Physics Letters, 43*(6), 584-586.

Koltcov, S., Ignatenko, V., & Koltsova, O. (2019). Estimating Topic Modeling Performance with Sharma–Mittal Entropy. *Entropy, 21*(7), 660.

Koltcov, S., & Ignatenko, V. (2020, July). Renormalization approach to the task of determining the number of topics in topic modeling. In *Science and Information Conference* (pp. 234-247). Springer, Cham.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556-562).

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272).

Mora, T., & Walczak, A. M. (2016). Renyi entropy, abundance distribution, and the equivalence of ensembles. *Physical Review E*, *93*(5), 052418.

Noji, H., Mochihashi, D., & Miyao, Y. (2013). Improvements to the Bayesian topic n-gram models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1180-1190).

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Sadeghi, M., & Vegas, J. (2014). Automatic identification of light stop words for Persian information retrieval systems. *Journal of information science*, *40*(4), 476-487.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Sato, I., & Nakagawa, H. (2010). Topic models with power-law using Pitman-Yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 673-682).

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry, M. J., & Bialek, W. (2015). Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, *112*(37), 11508-11513.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141-188.

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE international conference on data mining (ICDM 2007)* (pp. 697-702). IEEE.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448-456).

Wilson, K. G., & Kogut, J. (1974). The renormalization group and the ϵ expansion. *Physics reports, 12*(2), 75-199.

Yang, G., Wen, D., Chen, N. S., & Sutinen, E. (2015). A novel contextual topic model for multi-document summarization. *Expert Systems with Applications*, *42*(3), 1340-1352.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015, December). A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics* (Vol. 16, No. 13, pp. 1-10). BioMed Central.

Abstract

With the increasing size of textual data in recent years, it is difficult to get useful information from the huge data. Artificial intelligence by providing different techniques can help to extract valuable information from data. One of the widely used algorithms for analysis of large textual collections is probabilistic topic modeling which discovers the latent themes from a large text corpus. A topic means a set of words that occur together in a collection of documents and suggest a shared theme. In this research, we used the Latent Dirichlet Allocation with Gibbs sampling to extract topic models from the text. This algorithm assumes that any large document collection contains a finite set of topics or latent themes, while each word and each text of such collection belong to each topic with a certain probability.

One of the major challenges of topic modeling in practice is determining the number of topics because this value is not known in advance and the results of topic modeling depend on it. This research employs two algorithms to deal with the problem of determining the number of topics. One algorithm is based on the grid search and the other uses the renormalization theory to estimate the number of topics automatically. The former uses a metric such as perplexity or coherence and calculates the values of these metrics for different values of topic numbers and then chooses the parameter which leads to the best of values. The later uses renormalization theory, which is a mathematical formalism to construct a procedure for changing the scale of the system under which the behavior of the system preserves. Also, the execution time of two algorithms on different Persian journals has been reported. In addition, we propose the results of topic modeling on the academic papers of some Persian journals. As another achievement of this research, a list of Persian stop words for academic articles has been proposed.

Keywords: Topic Modeling; Latent Dirichlet Allocation; Gibs Sampling; Renormalization Theory; Rényi Entropy.

مرکز منطقه ای اطلاع رسانی علوم و فناوری

Regional Information Center for Science and Technology RICeST

Designing and Implementing a Topic Modeling Tool for Persian Text

August 2021