**Human Information Interaction**. Vol. 9, No. 3 (fall2022):1-21.
http://hii.khu.ac.ir
*Original Article*

# Automatic keyword extraction using Latent Dirichlet Allocation topic modeling: Similarity with golden standard and users' evaluation

**Farzaneh Shadanpour,** Ph.D. candidate in Knowledge and Information Science, Kharazmi University, Tehran, Iran.
***Nosrat RiahiNia** (Corresponding author), Professor of Knowledge and Information Science Department, Kharazmi University, Tehran, Iran. riahinia@khu.ac.ir
**Keivan Borna,** Assistant professor of Computer Science Department, Kharazmi University, Tehran, Iran.
**Gholam Ali Montazer,** Professor of Information Technology Engineering, Tarbiat Modares University, Tehran, Iran.
Received: 25/04/2022 Accepted: 19/07/2022

## Abstract
**Purpose:** This study investigates the automatic keyword extraction from the table of contents of Persian e-books in the field of science using LDA topic modeling, evaluating their similarity with the golden standard, and users' viewpoints of the model keywords.
**Methodology:** This is mixed text-mining research in which LDA topic modeling is used to extract keywords from the table of contents of scientific e-books. The evaluation of the used approach has been done by two methods of cosine similarity computing and qualitative evaluation by users.
**Findings:** Table of contents are medium-length texts with a trimmed mean of 260.02 words, about 20% of which are stop-words. The cosine similarity between the golden standard keywords and the output keywords is 0.0932 thus very low. The full agreement of users showed that the extracted keywords with the LDA topic model represent the subject field of the whole corpus, but the golden standard keywords, the keywords extracted using the LDA topic model in sub-domains of the corpus, and the keywords extracted from the whole corpus were respectively successful in subject describing of each document.
**Conclusion:** The keywords extracted using the LDA topic model can be used in unspecified and unknown collections to extract hidden thematic content of the whole collection, but not to accurately relate each topic to each document in large and heterogeneous themes. In collections of texts in one subject field, such as mathematics or physics, etc., with less diversity and more uniformity in terms of the words used in them, more coherent and relevant keywords are obtained, but in these cases, the control of the relevance of keywords to each document is required. In formal subject analysis procedures and processes of individual documents, this approach can be used as a keyword suggestion system for indexing and analytical workforce.
**Keywords**: Keyword extraction, Topic modeling, Latent Dirichlet Allocation (LDA), Similarity evaluation, Users' evaluation.

*Conflicts of Interest: Not reported.*
*Funding: Did not have financial sponsor.*

**How to cite this article**

**APA:** Shadanpour, F., RiahiNia, N., Borna, K., and Montazer, Gh. (2022). Automatic keyword extraction using Latent Dirichlet Allocation topic modeling: Similarity with golden standard and users' evaluation. *Human Information Interaction*, 9 (3);1-21. (Persian)

**Vancouver:** Shadanpour, F., RiahiNia, N., Borna, K., and Montazer, Gh. Automatic keyword extraction using Latent Dirichlet Allocation topic modeling: Similarity with golden standard and users' evaluation. *Human Information Interaction.* 2022; 9 (3);1-21. (Persian)

جلد نهم، شماره سوم (پاییز ۱۴۰۱): ۲۱-۱
http://hii.khu.ac.ir
مقاله اصیل پژوهشی

# استخراج ماشینی کلیدواژه با مدل‌سازی موضوعی ال. دی. اِی.: شباهت‌سنجی با کلیدواژه‌های استاندارد و ارزیابی کاربران

**فرزانه شادان‌پور:** دانشجوی دکتری علم اطلاعات و دانش‌شناسی دانشگاه خوارزمی، تهران، ایران.
***نصرت ریاحی‌نیا** (نویسنده مسئول): استاد گروه علم اطلاعات و دانش‌شناسی، دانشگاه خوارزمی، تهران، ایران. riahinia@khu.ac.ir
**کیوان برنا:** استادیار گروه علوم کامپیوتر، دانشگاه خوارزمی، تهران، ایران.
**غلامعلی منتظر:** استاد گروه مهندسی فناوری اطلاعات، دانشگاه تربیت‌مدرس، تهران، ایران.

## چکیده

نوع مقاله: مقاله پژوهشی
دریافت: ۱۴۰۱/۰۲/۰۵
پذیرش: ۱۴۰۱/۰۴/۲۸

**زمینه و هدف:** هدف این پژوهش، بررسی نتایج استخراج خودکار کلیدواژه از فهرست مندرجات کتاب‌های الکترونیکی فارسی حوزهٔ علوم با استفاده از مدل‌سازی موضوعی ال. دی. اِی.، سنجش شباهت کلیدواژه‌های خروجی با کلیدواژه‌های استاندارد و ارزیابی کاربران از کلیدواژه‌های استخراج‌شده به‌صورت ماشینی است.

**روش پژوهش:** این پژوهش کاربردی، از نوع پژوهش‌های متن‌کاوی و به جنبهٔ روش‌های مورداستفاده در آن پژوهش آمیخته است. از مدل‌سازی موضوعی ال. دی. اِی. برای استخراج کلیدواژه از فهرست‌های مندرجات کتاب‌ها استفاده‌شده و نتایج کاربرد مدل با دو روش سنجش کسینوس شباهت و پژوهش کیفی توسط کاربران مورد ارزیابی قرار گرفته است.

**یافته‌ها:** فهرست‌های مندرجات مورد بررسی با میانگین پیراسته ۲۶۰.۰۲ کلمه از متون با طول متوسط محسوب می‌شوند و حدود ۲۰ درصد از کلمات آن‌ها را ایست‌واژه‌ها تشکیل داده‌اند. میان کلیدواژه‌های استاندارد سرعنوانی و کلیدواژه‌های خروجی مدل ال. دی. اِی. کسینوس شباهت، ۰.۰۹۳۲، بسیار پایین به دست آمد. توافق کامل کاربران نشان داد کلیدواژه‌های خروجی مدل موضوعی ال. دی. اِی. حوزه موضوعی کل پیکره را نشان می‌دهند، اما ازنظر کاربران به ترتیب کلیدواژه‌های سرعنوانی استاندارد، کلیدواژه‌های مستخرج از مدل در زیرحوزه‌های موضوعی و کلیدواژه‌های مستخرج از مدل با کل پیکره در توصیف موضوعات هر تک مدرک موفق‌اند.

**نتیجه‌گیری:** کلیدواژه‌های به‌دست‌آمده از مدل موضوعی ال. دی. اِی. را می‌توان در مجموعه‌های ناشناخته به‌منظور استخراج محتوای موضوعی ناآشکار کل مجموعه به کار برد، اما برای ربط دقیق موضوع به مدرک در پیکره‌های بزرگ با موضوعات ناهمگن و متنوع، نمی‌توان از این روش استفاده کرد. این روش در رویه‌های رسمی توصیف موضوعی تک‌تک مدارک به‌صورت مستقل می‌تواند به‌عنوان یک سیستم پیشنهاددهنده کلیدواژه به نیروی انسانی نمایه‌ساز به کار گرفته شود.

**کلمات کلیدی:** استخراج ماشینی کلیدواژه، مدل‌سازی موضوعی، ال. دی. اِی.، شباهت‌سنجی، ارزیابی کاربر.

تعارض منافع: گزارش نشده است.
منبع حمایت کننده: حامی مالی نداشته است.

**شیوه استناد به این مقاله**

**APA:** Shadanpour, F., RiahiNia, N., Borna, K., and Montazer, Gh. (2022). Automatic keyword extraction using Latent Dirichlet Allocation topic modeling: Similarity with golden standard and users' evaluation. Human Information Interaction, 9 (3);1-21. (Persian)

**Vancouver:** Shadanpour, F., RiahiNia, N., Borna, K., and Montazer, Gh. Automatic keyword extraction using Latent Dirichlet Allocation topic modeling: Similarity with golden standard and users' evaluation. Human Information Interaction. 2022; 9 (3);1-21. (Persian)

## ۱. مقدمه

انسان‌ها در تعامل با محیط و برای رفع نیازهای گوناگون مادی و معنوی خویش به انواع گوناگونی از اطلاعات نیاز دارند که مقدار زیادی از این اطلاعات به‌صورت متن بر انواع متنوع حامل‌های اطلاعاتی و در قالب‌های مختلفی مانند کتاب، نشریه، گزارش علمی و غیره پدید می‌آید. کشف و بازنمایی مباحث و موضوعات موجود در منابع اطلاعاتی، ازجمله متون، مهم‌ترین هدف در نظام‌های توصیف، تحلیل و بازیابی اطلاعات به شمار می‌رود. جستجوهای کاربران در خیل منابع اطلاعاتی – چه متنی باشد یا چندرسانه‌ای – بیشتر جستجوی موضوعی است که از رایج‌ترین و درعین‌حال چالش‌برانگیزترین نوع جستجو توسط کاربران در پایگاه‌های اطلاعاتی و فهرست کتابخانه‌هاست (گلوب[1] و همکاران، ۲۰۱۸) و هنوز هم باوجود پیشرفت‌های چشمگیر فناورانه، اغلب با کلیدواژه صورت می‌گیرد (ریاض[2] ، ۲۰۱۸، ص ۸)؛ بنابراین، مهم است که این کلیدواژه‌ها به‌عنوان نقاط دسترسی موضوعی تعیین و در اختیار کاربر قرار گیرد .

کلیدواژه‌های موضوعی محصول فرایند تحلیل موضوعی و استخراج یا تولید کلیدواژه‌های موضوعی هستند. این فرایند با شناسایی موضوعات و مباحث مطرح در مدرک آغاز می‌شود و سپس واژگانی که فرض می‌شود قادرند به‌درستی موضوع مدرک را بشناسانند، کلیدواژه نامیده می‌شوند و با برداشت مستقیم از متن مدرک یا با انتساب کلماتی از یک بانک واژگان کنترل‌شده استخراج می‌شوند (گو[3] ، ۲۰۱۸) و به‌عنوان واسطه میان کاربر و منبع اطلاعاتی در دسترس قرار می‌گیرند .

در حوزهٔ علم اطلاعات و دانش‌شناسی، بررسی منابع اطلاعاتی از حیث موضوعات مندرج در آن‌ها، کشف موضوعات و تعیین واژگانی که نمایندهٔ آن‌ها باشند، فهرست‌نویسی تحلیلی یا نمایه‌سازی موضوعی (برای مقالات) نامیده می‌شود. اجرای همین فرایند در پردازش زبان طبیعی با استفاده از کامپیوتر و محاسبات در حوزهٔ علوم کامپیوتر و هوش مصنوعی با عبارت «استخراج کلیدواژه» شناخته می‌شود (توشارا[4] و همکاران، ۲۰۱۹). در پردازش زبان طبیعی مدل‌های مختلفی برای استخراج کلیدواژه وجود دارد که دسته‌ای از آن‌ها مدل‌سازی موضوعی[5] نام دارند و روش‌هایی را برای سازماندهی، درک، جستجو و خلاصه‌سازی خودکار منابع الکترونیکی فراهم می‌آورند. این کار با شناسایی موضوعات مدارک و کلمات درون هر موضوع صورت می‌گیرد و ضمن استخراج کلیدواژه‌های موضوعی، می‌توان مدارک مجموعه را به موضوعات مختلف مرتبط کرد (هورتادو، ۲۰۱۶). استفاده از روش‌های گفته‌شده، روند روبه‌رشدی در اقدامات و مطالعات مربوط به تحلیل، پردازش و سازماندهی متون دارد .

کتاب‌های الکترونیکی یکی از انواع اطلاعات متنی هستند که در قالب‌های مختلف و در مکان‌های مختلف، مانند وب‌سایت‌های کتاب‌فروشی‌های اینترنتی، به‌عنوان منابع اطلاعاتی تکمیلی در برخی وب‌سایت‌ها که کارکردهای غیر کتابخانه‌ای دارند و کتابخانه‌های دیجیتالی به‌وفور موجودند. این کتاب‌ها یا از ابتدا در قالب الکترونیکی (دیجیتالی) پدید آمده‌اند، یا با دغدغه حفاظت دیجیتالی و تسهیل و تسریع در دسترسی، بعد از انتشار در قالب چاپی، در قالب الکترونیکی نیز در دسترس قرار گرفته‌اند، اما شکل کاغذی آن‌ها نیز موجود است. برای تحلیل موضوعی این دسته از منابع می‌توان از روش‌های ماشینی تحلیل موضوعی و استخراج کلیدواژه بهره برد. این دسته از منابع، هرچند در قالب دیجیتالی تولید می‌شوند، اما به‌هرحال ساختار کتاب را در خود دارند. ازجمله اجزا آن‌ها فهرست مندرجات است که نمایش ساختار و تقسیمات محتوای کتاب است که توسط پدیدآورنده یا ویراستاران تهیه‌شده است. فهرست مندرجات کتاب‌ها بیان موجز محتوای کتاب‌ها و راهنمایی برای یافتن مطالب در صفحات مربوط است که با حداقل واژگان و اصطلاحات بیشترین بازنمایی محتوای بخش مربوطه را در خود دارد. علاوه بر این توسط پدیدآورنده‌ها تهیه می‌شود که خود از دانش حوزه‌ای خوبی در موضوع کتاب برخوردارند و هم بر واژگان مرتبط و تخصصی آن حوزه تسلط کافی دارند (پوکورنی[6] ، ۲۰۱۸). حتی در پژوهش‌هایی نیز نشان‌داده‌شده است که پرس‌وجوی کاربران در نظام‌های ذخیره و بازیابی اطلاعات بیشتر با فهرست مندرجات کتاب‌ها تطبیق داشته است تا با عبارت‌های نمایه‌ای مانند سرعنوان‌های موضوعی و ضروری است از روش‌های پردازش زبان طبیعی در استخراج کلماتی از فهرست‌های مندرجات که بتوانند بافت معنایی کلمات را نیز نشان دهند، استفاده شود (چوی[7] و همکاران، ۲۰۰۷)؛ بنابراین، باتوجه‌به این‌که نمایه‌سازی تمام متن مستلزم سربار زیاد و مصرف منابع پردازشی در آن بالاست و خواه‌ناخواه بخش مهمی از متن کامل در پیش‌پردازش‌های نمایه تمام متن نیز حذف می‌شود (هایت، ۲۰۲۰، خوشیان و میرزاییان، ۱۳۹۹) و افزایش طول مدارک هم بر تعداد کلیدواژه‌های استخراج‌شده می‌افزاید و هم فضای جستجو

[1] Golube
[2] Riaz
[3] Goh
[4] Tushara
[5] Topic modeling algorithms
[6] Pokorny
[7] Choi

را بزرگ می‌کند (سیدالحسن و نگ[1] ، ۲۰۱۴)، فهرست مندرجات کتاب‌ها می‌تواند به‌عنوان قطعهٔ مهم و برگزیدهٔ محتوای هر کتاب موردتوجه در تحلیل موضوعی قرار گیرد .

علاوه بر اینکه در وضعیت فعلی، تولید محتواهای اطلاعاتی متأثر از تحولات فناورانهٔ دیجیتالی و غیر آن به‌سرعت رو به افزایش است، سازماندهی و توصیف منابع با روش‌های دستی مستلزم صرف وقت و هزینهٔ بالا و عملاً غیرممکن است (حمید، ۲۰۱۶، ص ۱؛ سون[2] و همکاران، ۲۰۲۰)، استانداردهای توصیف کتاب‌شناختی نیز به‌صورت مداوم در شرف پیچیده‌تر شدن و تغییرات هستند که موجب می‌شود به‌روزرسانی و نگهداری ابزارهای جانبی (مانند اصطلاح‌نامه‌ها و سرعنوان‌های موضوعی و سایر بانک‌های مستند) و تطبیق فرایند توصیف منابع با آن‌ها زمان‌بر و پرهزینه شود. در این شرایط اگر انتظار داشته باشیم که نیروی انسانی محدود، منابع بی‌شماری را پردازش کنند، عملاً همواره منابع اطلاعاتی فراوانی را شاهد خواهیم بود که در صف توصیف مانده‌اند، و این مانعی در دسترسی به‌موقع کاربران به منابع موردنیاز آن‌هاست. راه چاره همگامی با تحولات فناورانه و استفاده از امکانات بالقوهٔ همین تحولات در یافتن راه‌حل‌هایی است که بتوانند فرایندها را ماشینی کرده، نیروی انسانی را در مسیر کنترل کیفیت، تکامل، بهینه‌کردن ماشین‌ها به کار بگیرند (شورت[3] (۲۰۱۹)، یونگر[4] (۲۰۱۸) و تچوا[5] (۲۰۱۹).

ماشینی کردن قسمت‌هایی از تحلیل موضوعی کتاب‌های الکترونیکی با روش‌های ماشینی که منجر به تولید کلیدواژه‌های باکیفیتی شود که بازنمون صحیحی از موضوعات کتاب‌ها باشند، می‌تواند نقش مهمی در کاهش هزینه‌های توصیف به‌صورت دستی، کوتاه‌کردن زمان سازماندهی و سرعت دسترسی کاربر به منابع داشته باشد. یادآور می‌شود که کتاب‌های الکترونیکی در مجموعه‌های دیجیتالی با استانداردهایی مانند دوبلین کور[6] و متس[7] و غیره توصیف می‌شوند که استفاده از طرح‌های موضوعی مختلف را در ناحیهٔ تحلیل موضوعی بسته به سیاست کتابخانه دیجیتالی مجاز می‌شمارند. کتابخانه‌های بسیاری در جهان سیاست‌گذاری جدیدی در توصیف مجموعه‌هایشان در پیش گرفته‌اند که مبتنی بر ایجاد نقاط دسترسی موضوعی با استفاده از روش‌های ماشینی، به‌عنوان روش اصلی (و نه‌تنها روش) برای انواع منابع، ازجمله کتاب‌های الکترونیکی – و حتی برای منابع چاپی – است (یونگر، ۲۰۱۸). استخراج کلیدواژه با استفاده از روش‌های کارآمد ماشینی متناسب با ویژگی‌های هر دسته از منابع متنی، بخصوص کتاب‌های الکترونیکی که بخش مهمی از کتابخانه‌ها و انباره‌های دیجیتالی را تشکیل می‌دهند، می‌تواند کمکی در کوتاه‌کردن مسیر دسترسی به محتوای آن‌ها باشد. تحلیل موضوعی و استخراج کلیدواژه‌ها با روش‌های خودکار را می‌توان هم به طور مستقل و هم در ترکیب با روش‌های دستی و به‌عنوان کمک به نیروی انسانی به کار برد. این پژوهش در پی پاسخ به این سؤال اصلی است که آیا اگر از روش‌های ماشینی پردازش زبان طبیعی، به طور مشخص مدل موضوعی ال. دی. اِی.[8] ، به‌عنوان تحلیل موضوعی کتاب‌های الکترونیکی بر مبنای فهرست مندرجات آن‌ها استفاده کنیم، کلیدواژه‌هایی به دست خواهیم آورد که ازنظر کاربران به‌خوبی نمایانگر موضوعات متن باشند و این کلیدواژه‌های به‌دست‌آمده با ماشین، چقدر با کلیدواژه‌ها یا عبارت‌های موضوعی انتساب یافته به مدارک توسط انسان قابل‌رقابت‌اند؟

به این سؤال در قالب پاسخ به سؤال‌های جزئی‌تر زیر پاسخ داده می‌شود :

۱. روش‌های مناسب برای پیش‌پردازش فهرست‌های مندرجات کتاب‌های الکترونیکی فارسی حوزهٔ علوم به‌گونه‌ای که موجب بهینه شدن کار مدل موضوعی ال. دی. اِی. شود، چیست؟
۲. تنظیمات مناسب مدل ال. دی. اِی. برای استخراج کلیدواژه‌های موضوعی از فهرست مندرجات کتاب‌های الکترونیکی فارسی حوزهٔ علوم و خروجی‌های آن چگونه است؟
۳. کلیدواژه‌های استخراج‌شده از فهرست‌های مندرجات با کاربرد مدل ال. دی. اِی. و کلیدواژه‌های موضوعی انتساب یافته به‌صورت دستی در فرادادهای استاندارد به چه میزان شباهت دارند؟
۴. مزایا و معایب استفاده از فهرست مندرجات کتاب‌ها در استخراج ماشینی کلیدواژه چیست؟
۵. کاربران موضوعات استخراج‌شده با روش مورداستفاده را در مقایسه با موضوعات انتساب داده شده به‌صورت دستی در فرادادهٔ استاندارد کتاب‌های موردمطالعه چگونه ارزیابی می‌کنند؟

[1] Saidul Hasan & Ng
[2] Sun
[3] Short
[4] Junger
[5] Tchoua
[6] Dublin Core
[7] Metadata Encoding and Transmission Standard (METS)
[8] LDA: Latent Dirichlet Allocation

## ۲. پیشینه پژوهش

استخراج خودکار کلیدواژه‌های موضوعی با روش‌های پردازش زبان طبیعی و یادگیری ماشین در حال حاضر ادبیات پرباری را شامل می‌شود. برخی پژوهش‌های تلفیقی متن‌کاوی، مانند دسته‌بندی، خوشه‌بندی، خلاصه‌سازی متون که در آن‌ها از استخراج کلیدواژه به‌عنوان یکی از مراحل استفاده می‌شود، بر انبوهی این پیشینه‌ها افزوده‌اند. علت کثرت مقالات پژوهشی و پرهیز از ذکر آثاری که دیگر چندان روزآمد تلقی نمی‌شوند، گزیده‌ای از مقالات مرتبط با موضوع استخراج خودکار کلیدواژه با ال. دی. اِی.، کاربرد ال. دی. اِی. در چارچوب پژوهش‌های کلی‌تر، مانند دسته‌بندی و خلاصه‌سازی، نیز در پیشینه‌ها ذکرشده‌اند، اما در پیشینه‌های داخلی شرط فارسی بودن دیتاست در انتخاب مقالات رعایت شد. عسگری و شاپلیه[1] (۲۰۱۳) با استفاده از روش ال. دی. اِی. به تحلیل و استخراج موضوعات و بررسی معناشناختی مجموعه‌ای از ۱۸۰۰۰ شعر فارسی از ۳۰ شاعر متفاوت پرداختند. برای اجرای مدل، واژه‌نامه‌ای که بتواند از فارسی قدیم و جدید پشتیبانی کند تهیه کردند نتایج همبستگی معناداری میان نمره احتمال شرطی شاعران و موضوعات استخراج‌شده و میان موضوعات و طول اشعار را نشان داد. مسعودی و راحتی قوچانی (۱۳۹۴) با استفاده از مدل ال. دی. اِی. و دسته‌بندی بیشینه آنتروپی روشی برای رفع ابهام از واژگان مبهم فارسی پیشنهاد کردند. روش بر پانزده واژه مبهم پرتکرار در زبان فارسی که از پیکره‌ٔ پژوهشکده پردازش هوشمند علائم استخراج شد، اجرا شد. نتایج دقت حدود ۹۷٪ را نشان داده است که بیان‌گر تأثیر این روش در یافتن معنی مناسب واژگان مبهم است. رهگذر[2] (۲۰۲۰) باهدف همراه‌کردن علوم کامپیوتر و ادبیات از طریق طراحی و توسعه ابزارهای ماشینی تحقیق در اشعار فارسی از الگوریتم ال. دی. اِی. به‌عنوان ابزار استخراج ویژگی، یا همان واژه‌های موضوعی، و از اس. وی. ام.[3] برای دسته‌بندی اشعار حافظ استفاده کرده است. هرچند استفاده از مدل‌های موضوعی در ادبیات و شعر هنوز در ابتدای مسیر است، در نتایج این پژوهش گزارش‌شده که روش به‌کاررفته می‌تواند در دسته‌بندی موضوعی اشعار و نشان‌دادن روابط موضوعی میان آن‌ها در هر دسته و درک بهتر موضوعات در اشعار کمک کند. سید و سپرویت[4] (۲۰۱۷) انسجام موضوعی و رتبه‌بندی انسانی موضوعات خروجی مدل ال. دی. اِی. را به‌منظور بررسی تأثیر استفاده از چکیده یا متن کامل متون علمی در موضوعات خروجی مدل ال. دی. اِی. ارزیابی کردند. نتایج پژوهش آن‌ها نشان داد که طول مدرک و اندازه واژگان پیکره بر انسجام موضوعی و رتبه‌بندی انسانی تأثیر دارد و مجموعه‌های بزرگ کمتر از عبارت‌های غلط و اختلال زا تأثیر می‌پذیرند و موضوعات منسجم‌تر و ازنظر انسان بهتر نتیجه می‌دهند. تفاوت عملکرد مدل در چکیده‌ها و متن کامل در مجموعه‌های کوچک نمایان‌تر است؛ به این صورت که موضوعات خروجی مدل با متن کامل تا ۹۰ درصد و با چکیده تا ۵۰ درصد باکیفیت ارزیابی‌شده‌اند. میر[5] و همکاران (۲۰۱۸) باهدف آشناکردن پژوهشگران حوزه ارتباطات با روش ال. دی. اِی. این روش را در دیتاستی ساخته‌شده از بیش از ۳۰۰۰۰۰ صفحه وب در موضوع امنیت غذایی برای استخراج موضوعات به کار بردند. در نتایج آن‌ها اگرچه مدل‌سازی موضوعی ال. دی. اِی. از حیث خروجی همراه با عدم قطعیت همراه دانسته شده، از حیث بازنمایی موضوعات موجود در پیکره‌های بزرگ روشی مناسب و کاربردی معرفی‌شده است. وانگ[6] و همکاران (۲۰۱۸) باتوجه‌به نقش دیدگاه‌های مشتریان در رقابت محصولات، به‌جای روش‌های سنتی مبتنی بر پیمایش پرسش‌نامه‌ای از مدل موضوعی ال. دی. اِی. برای تحلیل دیدگاه‌های مشتریان و استخراج موضوعات مهم در آن‌ها درباره دو محصول رقابتی موشواره بی‌سیم استفاده و برتری‌ها و ضعف‌های رقابتی هر دو محصول را شناسایی کردند. تفاوت‌های موجود میان دودسته موضوعات، خروجی بیانگر مزایا و معایب دو گروه محصول بوده و اجرای این روش پیامدهای مدیریتی ارزشمندی را برای طراحان محصول و شرکت‌های تجارت الکترونیکی فراهم کرده است. یائو[7] و همکاران (۲۰۱۸) یک روش ال. دی. اِی. مشترک برای مدل‌سازی برچسب‌های شبکه‌های اجتماعی پیشنهاد کردند که دو عامل علاقه کاربر و عامل موضوع پنهان شیء را در رویه تولید برچسب‌ها به طور مشترک وارد می‌کند که هم نظرات پنهان کاربران (کاربر – موضوع) و هم موضوعات موجود در موجودیت وبی را که با تگ‌ها توصیف می‌شود (موجودیت وبی – موضوع) در تولید تگ لحاظ می‌کند. نتایج کاربرد مدل در چهار دیتاست از تگ‌های شبکه‌های اجتماعی در مقایسه با عملکرد پنج مدل‌سازی موضوعی دیگر (pLSA، LSA، QUAR-RTM، CONV-RTM، Gaussianال. دی. اِی. (.نشان داد که مدل مذکور منجر به عملکرد بهتر در مقایسه با این ۵ مدل و استخراج موضوعات

[1] Asgari & Chapelier
[2] Rahgozar
[3] SVM: Support Vector Machine
[4] Syed & Spruit
[5] Maier
[6] Wang
[7] Yao

بهتری در مقایسه با مدل ال. دی. اِی. رایج می‌شود. اسموسن و مولر[1] (۲۰۱۹) چارچوب و مجموعه کدهایی را برای کاربرد مدل‌سازی موضوعی در مجموعه‌های بزرگ مقالات به‌منظور مرور اکتشافی پیشینه‌ها ارائه کردند که شامل مراحل پیش‌پردازش، مدل‌سازی موضوعی با ال. دی. اِی. و پس‌پردازش است. ایم[2] و همکاران (۲۰۱۹) کوشیدند تا کاربرد هر دو الگوریتم ال. دی. اِی. و تجزیه‌وتحلیل احساسات[3] را بر اساس سطح عاطفی متن ارزیابی کنند. یافته نشان داد که خروجی ال. دی. اِی. از حیث کلمات حاوی بار عاطفی پایین نمره اعتماد درک شده بالاتری از داوران گرفته و نمره اعتماد درک شده دو الگوریتم تحلیل احساسات و ال. دی. اِی. درباره سطوح عاطفی متون مورد بررسی تفاوت معناداری باهم نداشتند. همچنین ال. دی. اِی. برای یافتن مباحث در متونی که عمدتاً حاوی اطلاعات عینی است، مؤثرتر است. پیچ و لسمن (۲۰۱۹) باهدف تحلیل موضوعات مطرح‌شده در پاسخ‌های مشتریان به سؤالات باز که از آن‌ها درباره محصولات و خدمات پرسیده می‌شود روشی بی نظارت با استفاده از چهار الگوریتم مدل‌سازی موضوعی بی. تی. ام.[4] و دبلیو. ان. تی. ام.[5] ، ال. دی. اِی. و ال. اف.[6] را به کاربردند که غیر از ال. دی. اِی. سه مورد دیگر برای تحلیل متون کوتاه مناسب‌اند. نتایج نشان دادند که مدل‌های موضوعی برای کاوش داده‌ها و همچنین رتبه‌بندی موضوعات بسیار مفید هستند. به‌خصوص مدل‌های متن کوتاه اختصاصی بی. تی. ام. و دبلیو. ان. تی. ام. نتایج بهتری نسبت به ال. دی. اِی. و ال. اف. به دست می‌آورند، اما درمجموع مدل‌سازی موضوعی را نمی‌توان به‌صورت ربط مستقیم هر موضوع به یک مدرک مورداستفاده قرارداد. سفاکاکیس[7] و همکاران (۲۰۱۹) روش‌شناسی ماشینی برای تحلیل موضوعی ارائه کرده‌اند که هم شامل تعریف موضوع و دربارگی مدرک است و هم ترجمهٔ مفاهیم مرتبط به اصطلاحات موجود در اصطلاح‌نامه یوروووک[8] ، آن‌ها روش جاسازی کلمات[9] که معنای کلمه را در موقعیت‌های مختلف متن در نظر می‌گیرد، و الگوریتم ال. دی. اِی. را برای مدل‌سازی موضوعی پیکره‌ای حاوی مقالات مربوط به حوزهٔ ارزیابی کتابخانهٔ دیجیتالی به کار گرفتند. برای ارزیابی اینکه موضوعات خروجی مدل در اصطلاح‌نامه موجودند از روش شباهت کسینوسی استفاده کردند و مفاهیمی را که بیشترین شباهت را با کلمات یک موضوع داشتند (حد آستانه ۸۴ درصد) باز نمایش مناسبی برای آن موضوعات محسوب کردند. یافته‌های آن‌ها نشان داد در صورت تعریف گردش کاری مناسب مدل آن‌ها می‌تواند به‌عنوان ابزار کمکی برای تحلیل موضوعی منابع به کار گرفته شود. شورت (۲۰۱۹) به بیان تجربه‌های متن¬کاوی در دانشگاه ایلینوی شمالی باهدف بهبود دقت و صحت فهرست‌نویسی با تمرکز بر تحلیل موضوعی ۵۵۰۰۰ کتاب داستان زرد که بین سال‌های ۱۸۶۰ تا ۱۹۱۵ در آمریکا رایج بودند و در این دانشگاه آن‌ها را دیجیتالی می‌کنند، پرداخته است. از دسته‌بندی، استخراج کلیدواژه، شناسایی موجودیت‌های نام، خوشه‌بندی و مدل‌سازی موضوعی با ال. دی. اِی. در کنار سرعنوان‌های موضوعی برای کمک به بازیابی بهتر توسط کاربران استفاده شد. نتایج گزارش نشان داد با متن¬کاوی می‌توان ابعاد متون را کاهش داد تا فهرست‌نویسان وقت کمتری برای مطالعه کل کتاب و انتساب موضوعات صرف کنند. با این روش‌ها می‌توان مشکل عدم همگونی تحلیل موضوعی میان فهرست‌نویسان را کنترل کرد. همچنین استفاده از روش‌های ماشینی تحلیل موضوعی می‌تواند به‌عنوان عملکرد تکمیلی در فهرست‌نویسی کتاب‌ها بسیار مفید باشد، ولی نمی‌تواند جایگزین آن شود. وانگ و تیلور[10] (۲۰۱۹) به‌منظور تشخیص موارد اضطراری شهری ازجمله بلایای طبیعی و فجایع انسانی و سایر موارد اضطراری روش فضایی و مبتنی بر داده‌های موجود در پلتفرم توییتر با استفاده از مدل ال. دی. اِی. پیشنهاد کردند که با استفاده از یک ماژول تشخیص موضوع – جغرافیا به ابعاد جغرافیایی و معنایی رویدادها پرداخته و سطح بحران آن‌ها را بر اساس شدت احساسات منفی از طریق یک ماژول رتبه‌بندی ارزیابی می‌کند. استخراج موضوعات با ال. دی. اِی. صورت‌گرفته و روش ارائه‌شده با موفقیت موارد اضطراری از انواع مختلف را در بین تمام موضوعات کاندیدا شناسایی کرده است. دینگ[11] و همکاران (۲۰۲۰) مدل تکامل‌یافته‌ای از ال. دی. اِی. به نام ای. تی. ام.[12] را توسعه دادند که بتواند در پیکره‌های بزرگ موضوعات تفسیرپذیرتی را در مقایسه ال. دی. اِی. رایج استخراج کند. در این روش تکامل‌یافته، مدل رایج ال. دی. اِی. با روش جاسازی کلمهٔ ترکیب و الگوریتم استنتاج نیز اصلاح‌شده است. مدل تکامل‌یافته در پیکره‌های مختلف با مقادیر بزرگ واژگان عملکرد بهتری از حیث کیفیت موضوعات و عملکرد پیش‌بینی داشته است. سون[13] و همکاران (۲۰۲۰) در پژوهش خود از مدل

[1] Asmussen, & Møller
[2] Im
[3] Sentiment Analysis (SA)
[4] BTM: Biterm Topic Model
[5] WNTM: Word Network Topic Model
[6] LF: Latent Featur
[7] Sfakakis
[8] EuroVoc
[9] Word Embedding
[10] Wang & Taylor
[11] Dieng
[12] ETM: Embedded Topic Model
[13] Sun

موضوعی ال. دی. اِی. برای استخراج موضوعات و نظرکاوی در بحث‌های آنلاین در داده‌های واقعی گروه‌های بحث استفاده کردند. آن‌ها از ساختار درختی رشته بحث‌ها[1] در مدل استفاده کردند. مدل سنجه «محبوبیت»[2] را برای تعیین تعداد پاسخ‌هایی که یک دیدگاه دریافت می‌کند، ابداع کردند که این عامل به عامل بسامد وقوع کلمه افزوده می‌شود. علاوه بر این، مدل ال. دی. اِی. به‌دست‌آمده که مدل آگاه از ساختار مکالمه[3] نام دارد، با ویژگی دیگری به نام «قابلیت انتقال»[4] می‌تواند موضوعات استخراج‌شده را به دیدگاه مربوط آن متصل کند. از روش با نظارت برای ارزیابی استفاده کردند و سه کدگذار داده‌ها را برچسب‌گذاری کردند. برچسب‌هایی مورداستفاده قرار گرفت که حداقل دو نفر از کدگذاران بر آن‌ها توافق داشتند. برای نشان‌دادن عملکرد بهبودیافته مدل در استخراج موضوع از مقیاس‌های ارزیابی انسجام و دقت در انتساب موضوعات استفاده شد. سبالچیرو و ادر[5] (۲۰۲۰) باهدف بررسی مشکلات روش‌شناختی در کاربرد مدل‌های موضوعی، نتایج چند تجربه را حیطه تنظیم پارامتر تعداد موضوع در مدل ال. دی. اِی. هنگام تحلیل متون طولانی رمان‌های انگلیسی به اجرا درآوردند. هدف یافتن رابطه بین طول متن و تعداد مطلوب موضوعات در تنظیم پارامترها و ساخت مدل بود. نتایج تجربه آن‌ها رابطه عکس میان طول متن و تعداد مناسب موضوعات را نشان داد، با این توضیح که قطعات متن کوتاه‌تر با تعداد موضوع بالا، منتهی موضوعات با کلمات تکراری و جزئی می‌شود و تعداد کمتر موضوع در قطعات طولانی‌تر متون منتهی به موضوعات کلی‌تر می‌شود و دراین‌خصوص تفاوتی میان متون علمی و ادبی وجود ندارد. آن‌ها این رابطه را به‌صورت معادله‌ای ریاضی نشان دادند و اشاره کردند که کاربرد مدل می‌تواند برای هر دو نوع متن کوتاه و طولانی به‌شرط تنظیم صحیح تعداد موضوعات مفید باشد .

از مجموع پیشینه‌ها چنین برمی‌آید که مدل ال. دی. اِی. در پژوهش‌های متنوعی برای استخراج موضوعات و کلیدواژه‌های موضوعی به‌کاررفته و عموماً عملکردی قابل‌قبول ولی قابل ارتقا داشته است. تفاوت این پژوهش با پیشینه‌ها کاربرد مدل ال. دی. اِی. در جامعه متفاوت (فهرست‌های مندرجات)، در حوزه موضوعی متفاوت (علوم) و در زبان فارسی است که از حیث ارزیابی نیز تلفیقی از روش‌های مورداستفاده در پیشینه‌ها را به کار برده است.

## ۳. روش‌شناسی پژوهش

این پژوهش کاربردی، از نوع پژوهش‌های متن¬کاوی است که برای تحلیل متن و استخراج کلیدواژه به کار می‌رود و از جنبهٔ روش‌های مورداستفاده در آن پژوهش آمیخته است که در ادامه توضیح داده می‌شوند.

### ۳-۱. داده‌ها و آماده‌سازی آن‌ها

پیکره متنی مورداستفاده متشکل است از فهرست مندرجات ۲۰۰۰ کتاب در حوزه موضوعی علوم پایه (شماره‌های ۵۰۰–۵۹۹ در رده‌بندی دیویی) که از وب‌سایت‌های تأمین و مدرک و فروش به دست آمد. علوم پایه به لحاظ عینیت موضوعات، ساده و همه‌فهم بودن زبان مورداستفاده و عاری بودن از ایهام، تمثیل، کنایه و اشاره برای پردازش‌های ماشینی با فناوری‌های ماشینی فعلی، به‌ویژه در زبان فارسی، مناسب‌تر تشخیص داده شدند. به علت اینکه یکی از پرسش‌های پژوهش میزان شباهت و همخوانی کلیدواژه‌های موضوعی استخراج‌شده در این پژوهش با سرعنوان‌های موضوعی است که در کتابشناسی ملی ایران برای کتاب‌ها ذکرشده، این موضوعات نیز برای هر کتاب استخراج و در فایل اکسل دیتاست درج شد. این موضوعات در حقیقت نماینده تحلیل موضوعی انسانی هستند .

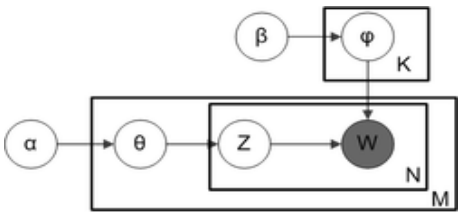
برای توصیف پیکره از آمار توصیفی و استنباطی و ضریب همبستگی اسپیرمن برای برآورد ضریب همبستگی میزان کلمات موجود در هر یک از دو عنصر اصلی پیکره استفاده شد.

به‌منظور اجرای مدل پس از گردآوری و تشکیل پیکره، بر اساس روش‌های پردازش زبان طبیعی، مراحلی برای آماده‌سازی داده‌ها و اجرای مدل صورت می‌گیرد که به مجموعه آن‌ها پیش‌پردازش گفته می‌شود. پیش‌پردازش‌ها به‌شدت به محتوای پیکره و ویژگی‌های آن وابسته‌اند. ازآنجاکه مدل‌های موضوعی به حجم پیکره حساس‌اند، توصیه‌شده که بهتر است پیش‌پردازش‌ها به‌صورت ملایم‌تر انجام گیرند تا از کیفیت موضوعات کاسته نشود (سکوفیلد، و همکاران، ۲۰۱۷). در این پژوهش از جعبه‌ابزار پارسیور[6] برای بخشی از پیش‌پردازش‌های پیکره استفاده‌شده است. برای تعیین ایست‌واژه‌ها از روش بسامد کلمه (زیف) در مجموعه مدارک و تلفیق نتایج با فهرست ایست‌واژه‌های فارسی عمومی موجود در ابزار ان. ال. تی. کا.[7] استفاده شد.

[1] Discussion thread tree structure
[2] Popularity
[3] CSATM: Conversational Structure Aware Topic Model
[4] Transitivity
[5] Sbalchiero, & Eder
[6] Parsivar
[7] Nltk: Natural Language Toolkit

### ۳-۲. مدل موضوعی ال. دی. اِی.

استخراج کلیدواژه با استفاده از الگوریتم مدل‌سازی ال. دی. اِی. صورت گرفت. این مدل بیشترین کاربرد را در مدل‌سازی موضوعی دارد (میر و همکاران، ۲۰۱۸). از خانوادهٔ روش‌های بیزی ناپارامتری[1] و تکامل‌یافتهٔ روش تحلیل معنای نهفته[2] است که با ورود مدل احتمالاتی در آن تبدیل به تحلیل معنای نهفته احتمالاتی[3] و بعد با درنظرگرفتن توزیع احتمال پیشین دیریکله به ال. دی. اِی. تکامل یافت و البته تنوعی از آن در ترکیب با الگوریتم‌های دیگر بسته به موضوع و اهداف تحلیل ارائه‌شده است (هورتادو، ۲۰۱۶). ایدهٔ اصلی در ال. دی. اِی. این است که مدارک ترکیب‌های تصادفی از موضوعات نهفته هستند و هر کلمه احتمال دارد به یک یا چند موضوع تعلق داشته باشد. ال. دی. اِی. موضوعاتی را که یک مدرک ذیل آن‌ها قرار می‌گیرد بر اساس کلمات خود مدرک می‌یابد. هر مدرک یک کیسه از کلمات[4] است و ویژگی‌های دستوری، مانند اسم و فعل بودن یا ترتیب کلمات (دستور زبان) مهم نیست. سه پارامتر اصلی این مدل عبارت‌اند از: ۱. تعداد موضوعات؛ ۲. تعداد کلمات در هر موضوع؛ ۳. تعداد موضوع در هر سند (رورت[5]، ۲۰۱۹). هدف این است که بر اساس کلمات موجود در هر مدرک تعیین شود که آن مدرک ذیل چه موضوعاتی قرار می‌گیرد و میزان ربط موضوعات به مدارک چه اندازه است. در اینجا ربط به معنی توزیع‌های احتمال هر موضوع نسبت به هر مدرک و هر کلمه نسبت به موضوعی است که درون آن قرار دارد (سبالچیرو، و ادر، ۲۰۲۰). درنتیجه، احتمال توزیع موضوعات در مدارک با مجموعه‌ای از واژه‌های محتمل مشخص می‌شود (بلی و همکاران، ۲۰۰۳). نمایش گرافیکی این مدل در شکل ۱ نشان‌داده‌شده است. در مدل ال. دی. اِی. و در این شکل:




شکل ۱. نمایش گرافیکی ال. دی. اِی.صفحه‌ها نشان‌دهنده تکرارند (برگرفته از بلی و همکاران ۲۰۰۳)

- α پارامتر پیشینی دیریکله[6] و پارامتری تراکمی است. α توزیع موضوعات را در هر مدرک – موضوع[7] یا چگالی مدرک موضوع را نشان می‌دهد. مقدار بالای α منتهی به تعداد موضوع بیشتری در مدارک می‌شود و مقدار پایین، برعکس، تعداد کمتری موضوع به هر مدرک منتسب می‌کند.
- β پارامتر تراکمی پیشینی دیریکله و نشان‌دهنده توزیع کلمات در هر موضوع یا چگالی موضوع – کلمه[8] است. مقدار بالای β نشان‌دهنده این است که تعداد کلمه بیشتری برای مدل‌کردن موضوع مورداستفاده قرار می‌گیرد، پس موضوعات کلمات بیشتری را در خود دارند و منتهی به توزیع بیشتر کلمه در موضوع می‌شود. مقدار پایین آن منجر به این می‌شود که با تعداد کمتری از کلمات موضوعات ساخته شوند و موضوعات حاوی تعداد کمتری از کلمات خواهند بود
- M تعداد مدارک پیکره، N تعداد کلمات در یک مدرک، توزیع کلمات در یک موضوع، θ نماد توزیع موضوع در مدرک،Z نشان‌دهنده موضوع منتسب شده به یک کلمه، W کلمه در مدرک و k تعداد موضوعات است که از قبل برای الگوریتم تعیین می‌شود که چه تعداد موضوع بیابد.

ذکر این نکته لازم است که در این روش همه مدارک مجموعه یکسانی از موضوعات را در اختیار دارند؛ ولی هر مدرک این موضوعات را با نسبت‌های متفاوتی در خوددارند. در فرایند مولد بالا فقط کلمات موجود در مدارک متغیر مشاهده‌شده هستند، حال‌آنکه سایر متغیرهای پنهان پارامترهای ($\varphi$ و $\theta$) و ابرپارامترهای ($\alpha$ و $\beta$) هستند. احتمال کلی مدل برای یک مجموعه‌داده (همان پیکره)، با داشتن احتمال تقریبی هر سند، به‌صورت زیر محاسبه می‌شود:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{\mathcal{D}} \int p(\theta_d \mid \alpha) \left[\prod_{n=1}^{N_i} \sum p(z_{d_i} \mid \theta_d) p(w_{d_i} \mid z_{d_i}, \beta)\right] d\theta_d \quad (۱)$$

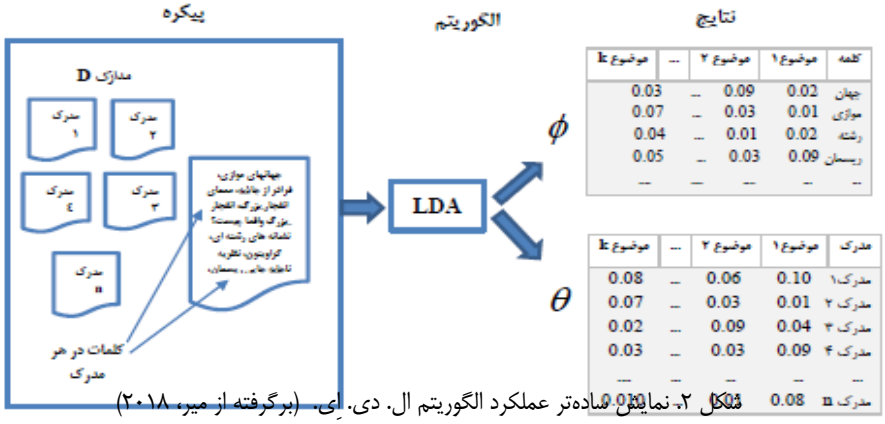
فرمول بالا احتمال مدرک را نشان می‌دهد. بعد از علامت مساوی چهار عبارت وجود دارد. اولی و سومی برای یافتن موضوعات و دومین و چهارمین برای یافتن کلمات در مدارک هستند. دو جمله

[1] Bayesian non- parametric algorithms
[2] LSA: Latent Semantic Analysis
[3] pLSA: Probabilistic Latent Semantic Analysis
[4] Bag of Word
[5] Revert

[6] توزیع دیریکله در نظریه احتمال و آمار یک توزیع پیوسته است. این توزیع به‌طورکلی حالت گسترش‌یافته توزیع بتا برای توابع چندمتغیره است. معمولاً از توزیع دیریکله به‌عنوان توزیع پیشین در مدل‌سازی بیزی استفاده می‌شود.
[7] Document-topic density
[8] Topic- word

آخر توزیع دیریکله و بقیه جملات بعد از علامت مساوی توزیع چندجمله‌ای هستند. خروجی کار الگوریتم ترکیبی از موضوعاتی است که هر مدرک در پیکره دارد. ما تنها کلمات را درون مدارک می‌بینیم و نیاز به استنباط ساختار پنهان داریم. این ساختار پنهان همان سهم موضوعات در هر مدرک است. به‌این‌ترتیب توزیع پسینی

حوزه موضوعی خاصی هستند، رخ می‌دهد (بلیگا[5] و همکاران، ۲۰۱۵). همچنین معیارهایی مانند جامعیت، مانعیت، دقت و سنجه اف.[6] به علت نیاز به مجموعه بزرگی از داده‌های مناسب که بتوان بازیابی را در آن اندازه‌گیری کرد و لزوم گروه‌بندی صحیح مدارک که به معنی مقایسه نتایج با کدگذاری دستی است در ارزیابی مدل‌های موضوعی زیر سؤال است (اسموسن و مولر، ۲۰۱۹).




شکل ۲. نمایش ساده‌تر عملکرد الگوریتم ال. دی. اِی. (برگرفته از میر، ۲۰۱۸)

باهمان شرطی که ساختار پنهان را در مدرک موردنظر به دست می‌آورد، با معادله زیر محاسبه می‌شود:

$$p(\theta, z \mid w, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (۲)$$

این توزیع پسین دیریکله برای استنتاج دقیق غیرقابل‌حل است و چندین الگوریتم استنتاج تقریبی مانند تقریب لاپلاس[1]، زنجیره مارکوف مونت‌کارلو[2] (مانند نمونه‌گیری گیبس[3]) برای به‌دست‌آوردن آن وجود دارد (بلی و همکاران، ۲۰۰۳). شکل ۲ به‌صورت ساده‌تری عملکرد الگوریتم را نشان می‌دهد.

اگرچه ال. دی. اِی. در اساس روشی بی‌نظارت است، به‌صورت با نظارت نیز برای داده‌های برچسب خورده مورداستفاده قرار گرفته است که همین نیاز به داده‌های برچسب خورده مهم‌ترین نقطه‌ضعف این توسعه‌های ال. دی. اِی. به شمار می‌رود (ممتازی[4]، ۲۰۱۸). دو مشکل عمده در روش‌های با نظارت وجود دارد که یکی لزوم تهیه داده یادگیری با کلیدواژه‌های استخراج‌شده به‌صورت دستی و توسط نیروی انسانی است که به‌ویژه در داده‌های با اندازه بزرگ عملاً غیرممکن است؛ و دیگر مسئله تورش است که در فرایند یادگیری داده‌هایی که مرتبط با

نکته دیگر که لازم است درباره مدل‌سازی موضوعی و یا اصولاً روش‌های یادگیری با ماشین ذکر شود، روایی و پایایی آن‌هاست. پایایی معمولاً با شباهت خروجی‌ها در تنظیمات مشابه مدل حاصل می‌شود (لوی و فرانکلین[7]، ۲۰۱۴) که در این پژوهش نیز از همین روش استفاده و خروجی مدل با تنظیمات مشابه بر دیتاست تکرار و کسینوس شباهت به دست آمد. همچنین ارائه‌شدن مدل توسط توسعه‌دهندگان معتبر، پذیرفته‌شدن مدل در جوامع علمی، کاربرد متعدد آن در پژوهش‌های گوناگون تأییدی بر روایی آن به شمار می‌رود.

## ۳-۳. ارزیابی‌های پیش‌بینی‌شده و ویژگی‌های روش‌شناختی آن‌ها

برای تعیین روش‌های ارزیابی و طراحی آن‌ها، ترسیم خطوط اصلی جهت‌گیری پژوهش ضروری است. اولاً، نگاه کلی پژوهش خودکارسازی استخراج کلیدواژه‌های موضوعی به‌عنوان بخشی از فرایند توصیف کتاب‌ها بر ای جایگزینی نیروی انسانی یا کمک به تسریع فرایندهای دستی است. دیگر اینکه، در پارادایم‌های توصیف منابع، سهولت دسترسی کاربر نهایی به محتوای منابع راهنمای عمل در همه فرایندهاست؛ بنابراین ارزیابی کاربران از

[1] Laplace approximation
[2] Markov chain Monte Carlo
[3] Gibbs sampling
[4] Momtazi
[5] Beliga
[6] F-measure
[7] Levy, & Franklin

خروجی الگوریتم در مقایسه با خروجی نمایه‌سازی انسانی عامل مهم در مطلوبیت کلیدواژه‌های موضوعی است. ارزیابی خروجی مدل مورداستفاده در سه دسته صورت گرفت :

- ارزیابی خروجی و عملکرد مدل با مقیاس انسجام موضوعی: انسجام موضوعی مقیاسی موجود در پیاده‌سازی‌های مدل است و تلفیقی از روش‌های مختلف ارزیابی در یک چارچوب ارزیابی انسجام میان موضوعات استنباط و استخراج‌شده توسط مدل بوده و با نمرهٔ انسجام موضوعی[1] با داور انسانی همبستگی دارد (یان[2] و همکاران، ۲۰۱۳). انسجام موضوعی چندین مدل دارد که در این پژوهش از مدل C_v(رودر و همکاران، ۲۰۱۵) به علت نزدیک‌تر بودن با داوری انسانی، برای سنجش کیفیت موضوعات استفاده می‌شده است .
- ارزیابی شباهت کلیدواژه‌های خروجی مدل با استاندارد طلایی: در این مرحله، مانند اونال سوزک[3] (۲۰۱۷) و وانگ و تیلور (۲۰۱۹) شباهت کلیدواژه‌های به‌دست‌آمده توسط مدل و کلیدواژه‌های داده شده به مدارک پیکره موجود در استاندارد طلایی (توسط متخصص موضوعی) با استفاده از کسینوس شباهت سنجیده می‌شود که بر اساس پژوهش‌ها نتایج این روش شباهت بیشتری با درک انسانی از شباهت مدارک دارد (تاونی[4] و همکاران، ۲۰۱۶). در اندازه‌گیری کسینوس شباهت تعداد هر داده (کلمه) نقشی در تعیین میزان شباهت ندارد، بلکه نوع داده‌ها (کلمات) باهم مقایسه می‌شوند .
- ارزیابی انسانی توسط داوران درباره کلیدواژه‌های خروجی مدل در مقایسه با استاندارد طلایی: سنجش کیفیت خروجی مدل‌های موضوعی و داوری انسانی درباره آن‌ها از موضوعات مهم همچنان موضوع بررسی در مدل‌سازی‌های موضوعی و روش‌های متن‌کاوی ماشینی است. هنگامی‌که مدل‌های موضوعی به‌عنوان مرحله‌ای در پژوهش‌های دیگری مانند دسته‌بندی و خلاصه‌سازی به کار می‌روند، ارزیابی کارایی آن‌ها به‌صورت بیرونی و با روش‌های نظارتی میسر است. ولی خروجی این مدل‌ها، به‌عنوان کلیدواژه برای استفاده انسانی در نظام‌های بازیابی اطلاعات و به‌صورت بی¬نظارت مسئله چالش‌برانگیز تفسیرپذیری برای انسان را پیش می‌آورد و سنجش کیفیت و تفسیرپذیری این خروجی‌ها توسط داور انسانی از روش‌های مهم ارزیابی آن‌ها در پیکره‌های گوناگون است (زینگ[5] و همکاران، ۲۰۱۹) و متخصص موضوعی برای تفسیر خروجی‌ها لازم است (دی ماجیو[6] و همکاران، ۲۰۱۳). این نوع داوری‌ها معمولاً در چارچوب ارزیابی‌های کیفی قرار می‌گیرند و روش‌های اجرا و آماری نزدیک به هم دارند. در این پژوهش از روش‌های ذکرشده در بالاتر و نیز از روش پیچ و لسمن (۲۰۱۸) با تغییراتی متناسب ولی منطبق و سازگار با اهداف و پرسش‌های این پژوهش و پیکره مورداستفاده در آن، استفاده‌شده است. با این پیش‌فرض که کلیدواژه‌های موضوعی محصول نمایه‌سازی انسانی از درجه خوبی از کارآمدی برخوردارند، ولی قرار است در رقابت با کلیدواژه‌های خروجی مدل قرار گیرند و هدف یافتن گروه کلیدواژه موضوعی است که ازنظر کاربران «بهتر» نمایانگر محتوای مدارک هستند و با محتوای مدارک، آن‌چنان‌که در فهرست‌های مندرجات آن‌ها نشان‌داده‌شده است، سازگاری و تناسب بیشتری دارند، ارزیابی خروجی مدل در مقایسه با کلیدواژه‌های استاندارد در قالب فرم‌های ارزیابی طراحی شد و به اجرا درآمد. هرچند ارزیابی از نوع کیفی است، معالوصف به‌منظور کیفیت بهتر و رعایت معیار اعتبارپذیری در یک پژوهش کیفی، فرم‌ها توسط پنج‌تن از اعضای هیئت‌علمی در رشته‌های علوم پایه، کتابداری و علوم کامپیوتر که تحصیلاتشان مرتبط با حیطه‌های این پژوهش بود، بررسی و مورد تأیید قرار گرفت. در فرم‌های ارزیابی دو گروه کلی سؤال در نظر گرفته شد:

- در سؤال اول همه کلیدواژه‌های خروجی مدل در معرض دید کاربران قرار گرفت و از آن‌ها خواسته شد که با مرور کلیدواژه‌ها حدس بزنند متون پیکره در کدام‌یک از رشته‌های دانش (فقط یک‌رشته که وجه غالب در

[1] Topic coherence
[2] Yan
[3] Onal Suzek
[4] Towne
[5] Xing
[6] Di Maggio

کلیدواژه‌ها دارد) قرار می‌گیرند. ده حوزه کلی دانش برگرفته از تقسیمات نظام دسته‌بندی دیویی در ادامهٔ سؤال ذکر شد که کاربر می‌بایست یکی از آن‌ها را که از دید وی با کلیدواژه‌ها مرتبط‌تر بود برگزیند .

- در بخش دوم، ۱۰۰ فهرست مندرجات (معادل ۵درصد از کل پیکره) و سه گروه کلیدواژه مربوط به آن‌ها شامل کلیدواژه¬های استاندارد (گروه الف)، کلیدواژه‌های خروجی مدل با کل پیکره (گروه ب)، و کلیدواژه¬های خروجی مدل با اجرا بر مدارک هر زیررده (گروه ج) توسط دو گروه ده‌نفری از کاربران متخصص در حوزه موضوعی (مجموعاً بیست کاربر) که با فراخوان دعوت به همکاری در گروه‌های دانشجویان تحصیلات تکمیلی در رشته‌های علوم پایه دعوت و انتخاب شدند، ارزیابی‌شده است. تقسیم کاربران به دو گروه به‌منظور رعایت معیار قابلیت اطمینان و تأییدپذیری در ارزیابی و امکان مقایسه بوده است که در پژوهش کیفی، هم‌ارز پایایی محسوب می‌شود. سقف ۱۰۰ نمره برای هر گروه کلیدواژه در نظر گرفته شد. نمره‌بندی تا ۱۰۰ برای هر دسته کلیدواژه باوجود کیفی بودن پژوهش، به‌منظور عینی‌تر کردن ارزیابی‌ها و قابلیت مقایسه بوده است تا بتوان میزان توافق دو گروه داور را بادقت بیشتری بررسی و تفسیر کرد. توافق میان دو گروه داوران درباره هر یک از سه گروه کلیدواژه برای هر فهرست مندرجات با ضریب همبستگی اسپیرمن برآورد شد. این ضریب مقدار و معنی‌داری بین نمره را در دو متغیر (دو کاربر) اندازه می‌گیرد در حالتی که مفروضه‌های آزمون‌های پارامتری در آن‌ها رعایت نشده باشد (گرین و دی الیویرا، ۱۳۹۴، ص ۹۸ و ۱۶۴). استفاده از ضریب همبستگی اسپیرمن به این دلیل بود که کاربران کمتر از ۳۰ مورد و نمره مای به‌دست‌آمده دارای چولگی و کشیدگی بوده‌اند و این ضریب به داده‌های پرت حساس نیست. همچنین مقیاس داده‌ها (۱۰۰) نیز به عدد و کمّی به نظر می‌رسد، ولی ذاتاً عددی نبوده و در حقیقت یک معیار برای رتبه‌بندی است، بنابراین می‌توان آن را از نوع ترتیبی محسوب کرد.

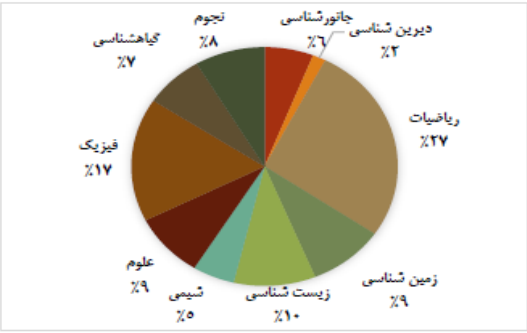
### ۳-٤. ابزارهای تحلیل داده‌ها

ابزارهای پیش‌بینی‌شده برای گردآوری و اجرای پژوهش عبارت‌اند از :

- نرم‌افزار اکسل نسخه (۲۰۱۳)؛
- زبان برنامه‌نویسی و کتابخانه‌های متنوع پایتون، در محیط اجرای ژوپیتر نوت‌بوک[1] و کوندا[2] و کتابخانه‌های مختلف در پایتون، مانند سایکیت- لرن[3] ، پانداس[4] و نامپای[5] ، جعبه‌ابزار پارسیور و کتابخانه جنسیم[6] و زیرمجموعه‌های آن؛
- نرم‌افزار اس. پی. اس. اس.[7] نسخه ۲۵ برای محاسبات آمار توصیفی و سنجش توافق میان کاربران.

## ٤. بحث و نتیجه‌گیری

داده‌های اصلی این پژوهش فهرست‌های مندرجات کتاب‌ها و سرعنوان‌های موضوعی مربوط به هر عنوان کتاب برگرفته از کتابشناسی ملی ایران به‌عنوان کلیدواژه‌های موضوعی استاندارد، بودند که زیرحوزه موضوعی ریاضیات با۴۵۰ عنوان کتاب (۲۷درصد از کل) بیشترین تعداد و دیرین‌شناسی با ۳۲ مورد (۱.۶درصد) کمترین تعداد کتاب را در کل پیکره دارا بودند (نمودار ۱).




نمودار ۱. درصد فراوانی مدارک پیکره در زیررده دیویی

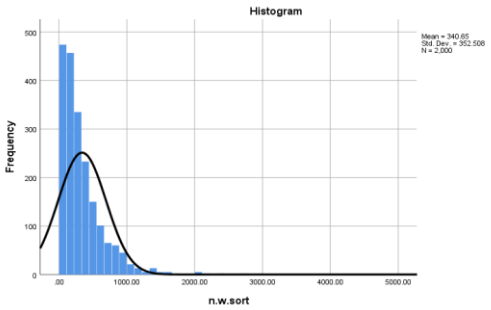
ویژگی مهم دیگر یک پیکره در روش‌های ماشینی پردازش زبان طبیعی طول مدارک (به معنی تعداد کلمات تشکیل‌دهنده مدرک)

[1] Jupyter notbook
[2] Conda
[3] Scikit-learn
[4] Pandas
[5] Numpy
[6] Gensim
[7] spss

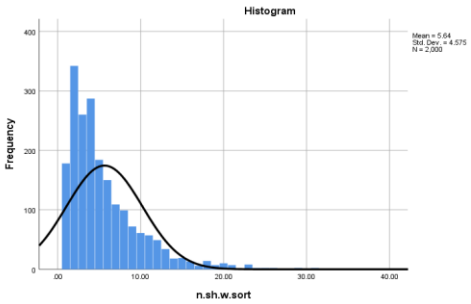
هر پیکره است. تعداد کل کلمات فهرست‌های مندرجات پیش از اعمال هرگونه پیش‌پردازش ۷۳۱۹۱۰ کلمه (شامل کلمات، اعداد و علائم سجاوندی) بود که پس از اعمال بخشی از پیش‌پردازش‌ها شامل حذف علائم سجاوندی، اعداد و فضاهای خالی تعداد ۶۸۱۳۰۶ کلمه باقی ماند که شامل کلمات انگلیسی نیز هستند. تعداد کلمات فهرست‌های مندرجات از ۶ کلمه (کمینه) تا ۴۲۳۷ کلمه (بیشینه) در تغییر است. میانگین طول فهرست‌های مندرجات برابر با ۳۴۰.۶۵۳ کلمه و انحراف معیار ۳۵۲۵۰۷۶۸۶ است. آمار توصیفی کلمات در فهرست‌های مندرجات و سرعنوان‌های موضوعی در جدول ۱ نشان‌داده‌شده است.

| آمار توصیفی کلمات در فهرست‌های مندرجات | | آمار توصیفی کلمات در سرعنوان‌های موضوعی | |
|---|---|---|---|
| تعداد اولیه (کلمات، علائم، اعداد) | ۷۳۱۹۱۰ | تعداد اولیه (کلمات، علائم، اعداد) | ۱۳۴۵۶ |
| تعداد کلمات (بعد از حذف علائم و اعداد) | ۶۸۱۳۰۶ | تعداد کلمات (بعد از حذف علائم و اعداد) | ۱۱۲۸۷ |
| تعداد سطرها | ۲۰۰۰ | تعداد سطرها | ۲۰۰۰ |
| کمینه (کمترین تعداد) | ۶ | کمینه (کمترین تعداد) | ۱ |
| بیشینه (بیشترین تعداد) | ۴۲۳۷ | بیشینه (بیشترین تعداد) | ۳۵ |
| میانگین | ۳۴۰.۶۵۳ | میانگین | ۵.۶۴ |
| انحراف معیار | ۳۵۲.۵ | انحراف معیار | ۴.۵۷ |
| چارک اول (۲۵٪) | ۱۱۸ | چارک اول (۲۵٪) | ۲ |
| چارک دوم (۵۰٪) | ۲۴۲ | چارک دوم (۵۰٪) | ۴ |
| چارک سوم (۷۵٪) | ۴۴۶ | چارک سوم (۷۵٪) | ۷ |
| میانه | ۲۴۲ | میانه | ۴ |

تحلیل داده‌ها در نرم‌افزار اس. پی. اس. اس. نیز چولگی به میزان ۳.۳۷۶ واحد و کشیدگی به میزان ۲۱.۰۰۵واحد را نشان داد که حاکی از آن است که توزیع کلمات در فهرست‌های مندرجات توزیعی نرمال نیست. این یافته با این اصل آماری پذیرفته‌شده در متن‌کاوی که «در متون معمولاً توزیع بسامد کلمات دارای چولگی است» (صادقی و وگاس[1] ، ۲۰۱۴) همخوانی دارد؛ بنابراین، میانگین پیراسته تعداد کلمات فهرست‌های مندرجات با ۲۰درصد حذف داده‌های پرت، پایین‌ترین حد و بالاترین حد برابر با ۲۶۰.۰۲محاسبه شد. همچنین ۲۵ درصد فهرست‌های مندرجات کمتر از ۱۱۸ کلمه دارند و بقیه طولانی‌ترند. معمولاً اگر متن حاوی کمتر از ۳۰ کلمه باشد آن را متن کوتاه قلمداد می‌کنند (زون و دیگران، ۲۰۱۶). در پیکره ۶۲ فهرست مندرجات (۳.۱ درصد کل مدارک) کمتر از ۳۰ کلمه داشته‌اند و جزء متون کوتاه، مانند توییت‌ها و دیدگاه‌ها در شبکه‌های اجتماعی قرار می‌گیرند. در کلمات سرعنوان‌های موضوعی دامنه تغییرات از ۱(کمینه) تا ۳۵ (بیشینه) کلمه است. ، ۲۵ درصد از سرعنوان‌های موضوعی کمتر از ۲ کلمه، ۵۰ درصد حاوی کمتر از ۴ کلمه و ۷۵٪ از فهرست‌های مندرجات کمتر از ۷ کلمه دارند. در سرعنوان‌های موضوعی نمره میانگین برابر با ۵.۶۴ بوده، انحراف معیار (۴.۵۷) و تفاوت آن با میانگین کمتر از ۲ واحد است. پس می‌توان نتیجه گرفت که توزیع کلمه در سرعنوان‌های موضوعی مربوط به مدارک، داده پرت ندارد و با نمره چولگی (۱.۹۷۷) تقریباً نرمال ولی دارای کشیدگی به میزان بیش از ۳ واحد است. که بازهم در زمره توزیع‌های نرمال قرار نمی‌گیرد؛ ولی در مقایسه با توزیع کلمات در فهرست‌های مندرجات از وضعیت بهتری برخوردار است (نمودار ۲).




[1] Sadeghi & Vegas




نمودار ۲. توزیع کلمات در فهرست‌های مندرجات (بالا) و سرعنوان‌های موضوعی (پایین)

تعداد کلمات موجود در فهرست‌های مندرجات هر کتاب و کلمات موجود در سرعنوان‌های موضوعی مربوط به هر کتاب را می‌توان متغیرهایی از آن‌ها در نظر گرفت که بین این دو متغیر نمره ۰.۹۹۴و در سطح معنی‌داری ۰.۰۵ برابر با ۰۱/ ۰۰-p) (value=0.01<0.05 رابطه همبستگی مثبت به‌دست‌آمده است؛ بدین معنی که با افزایش تعداد کلمات در هریک در دیگری نیز افزایش تعداد ملاحظه می‌شود.

در پاسخ به پرسش اول، باتوجه‌به محتوای فهرست‌های مندرجات مراحل پیش‌پردازش زیر اجرا شد:

- نرمال‌سازی به‌منظور تبدیل کاراکترهای متن به یونی‌کد (utf=8) حذف اعراب، تشدید و نویسه «–»، تبدیل همزه به شکل اصلی «ی»، «و»، و «ا»، تبدیل فاصله کامل در بعضی کلمات مرکب به نیم‌فاصله انجام شد. پسوندهای «تر» و «ترین» به کلمات پیش از خود متصل و «ب» چسبیده به کلمات (مانند «بموقع») به «به» تغییریافته است. معالوصف، ازآنجاکه هنوز ابزارهای فارسی برای پردازش متن در مسیر تکامل قرار دارند، پس از انجام نرمال‌سازی، مواردی از علائم جمع، مانند «مای»، «ها»، «تر»، «ترین» در متن پردازش‌شده ملاحظه شدند که به‌منظور حذف در فهرست ایست واژگان قرار گرفتند. همچنان که در فهرست ایست‌واژه‌های فارسی ان. ال. تی. کا. نیز این علائم گنجانده شده است. همچنین ملاحظه شد که در مواردی، به علت مسائل حروف‌چینی اگر فاصله دو جزء یک کلمه بیش‌ازحد معمول باشد، یا از قلم‌های خاصی که در یونی‌کد موجود نیستند استفاده‌شده باشد، در پیش‌پردازش اصلاحات موردنظر صورت نمی‌گیرد. این مرحله با ابزار پارسیور اجرا شد.
- طی فرایند پاک‌سازی حروف انگلیسی، اعم از بزرگ یا کوچک؛ حروف یونانی مانندγ ،Ω، ω، ϑو غیره؛ علائم سجاوندی، مانند نقطه، دو نقطه، خط تیره، اسلش، پرانتز و غیره؛ اعداد؛ فاصله‌ها و ایست‌واژه‌ها از متن فهرست‌های مندرجات حذف شدند. هرچند به علت رعایت الزامات رسم‌الخط استاندارد فارسی در سرعنوان‌های موضوعی فارسی که توسط کتابخانه ملی صورت می‌گیرد، میزان این اشکالات بسیار کمتر از متن فهرست‌های مندرجات است. در پیش‌پردازش سرعنوان‌های موضوعی، به‌عنوان استاندارد طلایی، پاک‌سازی ایست‌واژه‌ها صورت نگرفت و فقط سایر موارد بالا در مورد آن‌ها اعمال گردید. پاک‌سازی‌ها با ابزار ان. ال. تی. کا. و برخی کدهای خاص در پایتون اجرا شد .
- باتوجه‌به اینکه در این پژوهش فقط کلمات فارسی موردنظر بودند، به عملکرد تبدیل حروف بزرگ انگلیسی به حروف کوچک نیازی نبوده و این حروف در مرحله پاک‌سازی حذف شدند. این پیش‌پردازش بر سرعنوان‌های موضوعی نیز صورت گرفت تا بتوان در بخش‌های مقایسه آن‌ها را نیز در الگوریتم وارد کرد. هرچند به علت رعایت الزامات رسم‌الخط استاندارد فارسی در سرعنوان‌های موضوعی فارسی که توسط کتابخانه ملی صورت می‌گیرد، میزان این اشکالات بسیار کمتر از متن فهرست‌های مندرجات است.
- ایست‌واژه‌ها با رویکردی تلفیقی تعیین شدند که مبتنی بر محاسبه بسامد واژگان در مجموعه کل مدارک (برگرفته از قانون زیف) و مقایسه فهرست به‌دست‌آمده با فهرست‌های عمومی موجود در زبان فارسی برای تعیین ایست‌واژه‌ها است.
- توکن‌بندی یا تقطیع متن به کلمات تشکیل‌دهنده آنکه هرچند در ابزار پارسیور نیز به‌خوبی صورت می‌گیرد با کتابخانه جنسیم به‌عنوان یکی از فرایندهای مدل‌سازی موضوعی با ال. دی. اِی. انجام شد.

- استخراج ان‌تایی‌ها (در این پژوهش، دوتایی‌ها و سه‌تایی‌ها) با کتابخانه جنسیم استخراج‌شده‌اند. تعداد کلمات در ترکیب تا سه کلمه در نظر گرفته شد، چون در زبان فارسی تعداد ترکیب‌هایی که بیش از سه کلمه در خود داشته باشند و یک معنا را افاده کنند اندک بوده و اجرای چندتایی‌ها تا ترکیب ۵ کلمه نیز خروجی خاصی نداشت. در مدل ال. دی. اِی. هم¬وقوعی کلمات از مبانی محاسبات برای تعیین کلیدواژه‌ها به شمار می‌رود. به همین علت در استخراج چندتایی‌ها به‌خوبی عمل می‌کند که در این پژوهش که با پیکره فارسی‌زبان صورت گرفت نیز مشهود بود.
- در این پژوهش دو مرحله معمول لم‌گیری و ریشه‌گیری انجام نشد، اولاً به دلیل ضعف ابزارهای در دسترس فارسی است و دیگر اینکه برگرداندن کلمات به ریشه آن‌ها لزوماً آن‌ها را پرمعناتر و قابل‌درک‌تر نمی‌کند. به‌عنوان‌مثال، تبدیل «رفته است» به «رو» در مجموعه کلیدواژه‌ها و خارج از بافت جمله اصلی ابهام‌برانگیز است. علاوه بر این شکل‌های مختلف یک کلمه نهایتاً در یک موضوع جای می‌گیرند. در پژوهش سید و سیپرویت (۲۰۱۸) و وانگ و دیگران (۲۰۱۸) نیز به این چند علت در عدم لم‌گیری و ریشه‌گیری اشاره‌شده است .
- مجموع کل توکن‌های متن پس از پاک‌سازی اولیه (شامل حذف علائم سجاوندی، فضاهای خالی و اعداد) به‌دست‌آمده تا پیش از حذف ایست واژگان ۴۸۰۰۴ بوده است. پس از حذف ایست‌واژه‌ها تعداد کلمات باقی‌مانده ۳۸۴۹۵ و تعداد کلمات یکتای باقی‌مانده در پیکره ۷۵۲۸ عدد است (نمودار ۳). این بدان معنی است که تعداد ۶۴۲۸۱۱ کلمه و علامت ناخواسته در فرایند پالایش – ازجمله با حذف ایست‌واژه‌ها– از فهرست‌های مندرجات موجود در پیکره حذف‌شده‌اند. اما تعداد ایست‌واژه‌های حذف‌شده ۹۵۰۹ کلمه بوده که حدود ۲۰ درصد است که با مقداری که در گزارش‌ها ذکرشده (شوبل[1] ، ۱۹۹۷) مبنی بر این بین ۳۰ تا ۵۰ درصد توکن‌ها ایست‌واژه هستند همخوانی ندارد و این می‌تواند به علت ویژگی فهرست‌های مندرجات باشد که در آن‌ها جملات کوتاه‌تر و معمولاً فاقد افعال است.

در پاسخ به پرسش دوم باید اشاره کرد که ازآنجا هر چه تعداد موضوعات بیشتر باشد، مقدار بیشتری از کلمات موجود در متون شانس کلیدواژه قرارگرفتن و (به‌ویژه در حجم‌های بالای داده) دیده‌شدن خواهند داشت. با تعیین تعداد ۹۸، ۸۲، ۷۴، ۵۰، ۳۲ و ۲۰ موضوع از مدل خروجی گرفته شد و عدد ۶۸ با نمره انسجام ۰/۴۲۹۸ در معیار *c-v* به دست آمد. تنظیمات مدل در این پژوهش بر اساس توصیه‌های کتابخانه جنسیم عبارت‌اند از :

*(num_topics=68, random_state=1, update_every=1, chunksize=2000, passes=20, alpha='auto', eta='auto', per_word_topics=True, iterations=100)*

ازآنجاکه مدل ال. دی. اِی. مدلی احتمالاتی است، وجود فرایندهای تصادفی در الگوریتم باعث جلب‌توجه به مسئله پایایی می‌شود؛ بنابراین، باوجود علم به‌احتمال وجود تفاوت‌های اندک در خروجی‌های مدل با پارامترهای مشابه، در این پژوهش برای سنجش پایایی، مدل با پارامترهای مشابه تکرار و نتایج آن، یعنی موضوعات و کلیدواژه‌های آن‌ها با روش کسینوس شباهت ارزیابی شد و عدد ۰/۹۲۹۸ از یک به دست آمد که نشان از پایایی قابل‌قبول برای مدل با پارامترهای ذکرشده در بالاتر است و این نتایج میر و دیگران (۲۰۱۸) مبنی بر اینکه تکرار مدل با پارامترهای یکسان موجب اخذ نتایج مشابه می‌شود را تأیید می‌کند. همچنین اجرای مدل با تعداد مختلف موضوع تا سقف ۱۰۰ موضوع نشان‌دادن که با افزایش تعداد موضوعات کلمات مکرر آن‌ها بیشتر و کلیدواژه‌های بیشتری در جهت جزئیات بیشتر استخراج می‌شوند و با کاهش تعداد موضوعات، کلیدواژه‌ها تعداد کمتری خواهند بود؛ ولی لزوماً به معنای استخراج کلمات حاوی


نمودار ۳. تغییرات تعداد کلمات و واحدهای زبانی در هر مرحله از پیش‌پردازش پیکره

[1] Schauble

معانی کلی‌تر نیست. قسمت اول یافته با بخشی از یافته‌های پژوهش مسعودی و راحتی قوچانی (۱۳۹۷) مبنی بر اینکه زیادشدن تعداد موضوع به موضوعات با جزئیات بیشتر منتهی می‌شود، هماهنگ است.

باتوجه‌به دستور خروجی تنها ۱۰ کلمه با بالاترین نمره در هر موضوع، ۶۸۰ کلیدواژه از عملکرد الگوریتم بر کل پیکره با تنظیمات گفته‌شده به دست آمد که در شکل ۳ تعداد ۵ موضوع اول برای رعایت اختصار نشان‌داده‌شده است.

```
[(0,
  '0.021*"گروه"*۰,۰۱۷ + "آزمون"*۰,۰۱۸ + "تعریف"*۰,۰۲۱ + "رده" + '
  '0.010*"انقراض"*۰,۰۰۸ + "ساختار"*۰,۰۰۹ + "لایه"*۰,۰۰۹ + "معادلات"
+ '
  '0.008*"جبر"*۰,۰۰۸ + "احیای"'),
 (1,
  'گروه" + "جذب"*۰,۰۱۵ + "ضرب"*۰,۰۱۵ + "افزایش"*۰,۰۱۲ + "تست" + '
  '0.010*"رصد"*۰,۰۰۹ + "یادداشت"*۰,۰۱۰ + "قدرت"*۰,۰۱۰ + "میدان"*۱۲" + '
  '0.008*"انتخاب"'),
 (2,
  '0.030*"آزمون"*۰,۰۱۲ + "شناسایی"*۰,۰۱۵ + "تصادفی"*۰,۰۲۷ + "دماوند"
. + "تحلیل"*۰,۰۱۱ + "پیکرشناسی_دره"*۰,۰۱۱ + "ماکزیمال"*۰,۰۱۲ + "ن"
ورزش"*۰,۰۰۹ + "فرض"*۰,۰۰۹ + "ماتریس"*۰,۰۱۰"),
 (3,
  '0.023*"موجو"*۰,۰۱۳ + "انرژی"*۰,۰۱۴ + "ماتریسی"*۰,۰۱۴ + "شناسایی"
. + "جذب"*۰,۰۱۱ + "معادلات"*۰,۰۱۰ + "فضا"*۰,۰۱۱ + "د"
آزمون"*۰,۰۰۹ + "کمپلکس"*۰,۰۹"),
 (4,
  '0.028*"تحلیل"*۰,۰۱۲ + "انرژی"*۰,۰۱۴ + "نرمال"*۰,۰۲۷ + "جهش" + '
  '0.012*"سنج_سید"*۰,۰۱۰ + "انتخاب"*۰,۰۱۱ + "روابط"*۰,۰۱۲ + "ساختار"
م" + '
  '0.010*"ناحیه"*۰,۰۰۹ + "ستاره"'),
 (5,
  '0.028*"توزیع"*۰,۰۱۱ + "انتخاب"*۰,۰۱۲ + "مغناطیسی"*۰,۰۱۴ + "سیستم"
. + "جذب"*۰,۰۰۹ + "رده"*۰,۰۰۹ + "خمشی"*۰,۰۱۰ + "معادلات"*۰,۰۱۱ + "ع"
مبدل"*۰,۰۰۷ + "ساختار"*۰,۰۰۷"),
```

شکل ۳. پنج موضوع اول و کلیدواژه‌های تشکیل‌دهنده آن‌ها

با اجرای مدل بر مدارک متعلق به ۱۰ زیرحوزه موضوعی پیکره، کلیدواژه‌هایی به دست آمد که به حوزه موضوعی اختصاص کامل دارند؛ بنابراین اجرای مدل در سطح پیکره با مجموعه‌ای متنوع از موضوعات، کلیدواژه‌هایی به دست می‌آید که از کل پیکره گرفته‌شده‌اند و برخی از آن‌ها ممکن است در متن فهرست مندرجات مربوط به آن وجود نداشته باشند. به تعبیر دیگر، کلیدواژه‌های استخراج‌شده از کل پیکره ممکن است به‌خوبی بازنمای محتوای تک مدرکی که به آن مربوط است نباشند؛ هرچند این مسئله در عدد مربوط به‌احتمال هر کلیدواژه نمایان است. بنابراین در صورت وجود متون با موضوعات گوناگون و تنوع واژگانی بالا در پیکره مدل ال. دی. اِی. موضوعاتی نتیجه می‌دهد که بازگوکننده معانی کل پیکره هستند، معانی‌ای که با عنوان یا واژگان صرفاً پربسامد نمی‌توان آن‌ها را دریافت. این یافته هماهنگ با یافته‌های پژوهش میر و همکاران (۲۰۱۸) و اسموسن و مولر (۲۰۱۹) و پیچ و لسمن (۲۰۱۸) است که ال. دی. اِی. رویکردی قدرتمند در شناسایی خوشه‌های موضوعی اصلی متن است که می‌تواند ارتباط پنهان معنایی کلمات را آشکار کند، حتی اگر آن‌ها هرگز در یک مدرک باهم دیده نشوند. اما برای ربط موضوع به مدرک در پیکره‌های بزرگ با موضوعات و ناهمگن متنوع، به‌گونه‌ای که موضوع کاملاً منطبق با محتوای همان مدرک باشد مناسب نیست و چالش‌برانگیز است. در مجموعه متون یک حوزه موضوعی، مانند ریاضیات یا فیزیک و غیره، با تنوع کمتر و یکدست‌تر از حیث واژگان به‌کاررفته در آن‌ها، کلیدواژه‌های منسجم‌تر و مرتبط‌تری به دست می‌آید، ولی در این موارد نیز کنترل ربط با تک مدرک ضروری است.

در پاسخ به پرسش سوم، کلیدواژه‌های سرعنوان‌های موضوعی و خروجی مدل، شباهت کسینوسی در سطوح دیگر پیکره با استفاده از ابزار کاونت وکتورایزر[1] و کتابخانه‌های سایکیت– لرن و پانداس صورت گرفت. شباهت دو دسته کلیدواژه‌های خروجی مدل (با ۶۸ موضوع) با کلیدواژگان سرعنوانی – به‌عنوان استاندارد طلایی پذیرفته‌شده و مورداستفاده – با روش کسینوس شباهت و ۰/۰۹۳۲ معادل ۹/۳ درصد به دست آمد که میزان پایینی از شباهت را نشان می‌دهد (شباهت کامل = ۱، عدم شباهت =۰). ولی با نتایج پژوهش خطیر و گنجه فر (۱۳۹۷) که همپوشانی کلمات کلیدی و توصیفگرها (توسط نمایه‌ساز) را در مقالات مورد بررسی ۸درصد ذکر کرده‌اند بسیار نزدیک است .

هرچند از نمره شباهت نمی‌توان در مورد کیفیت کلیدواژه‌های خروجی مدل نتیجه‌گیری کرد، اما باتوجه‌به اینکه کلیدواژه‌های مدل برگرفته از متون پیکره هستند، کلیدواژه‌هایی که مدل از متن می‌گیرد لزوماً در قسمت‌هایی از متن رخ نمی‌دهند که به‌آسانی در معرض دید قرار می‌گیرند، مانند عنوان، صفحه حقوق و غیره. بلکه مفاهیم و معانی پنهان و کمتر رؤیت‌پذیر نیز با این مدل نشان داده می‌شوند. هدف توسعه مدل ال. دی. اِی. نیز همین بوده است .

[1] Count Vectorizer

پرسش چهارم را چنین می‌توان پاسخ داد که اصلی‌ترین مزیت فهرست‌های مندرجات در مقایسه با متن کامل کتاب‌ها برای استخراج کلیدواژه و نمایه‌سازی، خلاصه بودن در مقایسه با متن اصلی است که پردازش‌های کامپیوتری را به علت کمی حجم در مقایسه با متن کامل کتاب‌ها آسان‌تر می‌کند. کلمات به‌کاررفته در فهرست‌های مندرجات کتاب‌ها دقیقاً برخاسته از بدنه آن‌ها و طرح‌واره‌ای از محتوای کتاب‌ها هستند. حتی چکیده‌های مقالات پژوهشی نیز از این مزیت به طور کامل برخوردار نیستند، چراکه نویسنده یا چکیده‌نویس چکیده مقالات را بر مبنای متن مقاله می‌نویسند و آن‌ها کاملاً سیاهه برداری از متن مقالات نبوده، حاوی کلمات ربط، قید، افعال و ایست‌واژه‌ها هستند؛ نکته‌ای که همراهی چکیده‌ها را با کلیدواژه‌های برای مقالات و در ساختار آن‌ها باهدف تلخیص بیشتر محتوا، ضروری می‌کند. از معایب فهرست‌های مندرجات در استخراج ماشینی کلیدواژه این است دسترسی به نسخه‌های قابل کپی‌کردن در فرمت‌های مناسب برای پردازش ماشینی برای عموم بسیار اندک است هرچند کتابخانه‌ها و مراکز بزرگ و نیز ناشران کتاب‌های الکترونیکی از این امر تقریباً مستثنا هستند. در فهرست‌های مندرجات، به‌ویژه در حوزه علوم و فنون، عدم یکدستی در کاربرد زبان مشاهده می‌شود. کاربرد کلمات لاتین و علائم یونانی و لاتینی، به‌عنوان‌مثال در رده‌بندی‌های گیاهی و جانوری، فرمول‌های شیمیایی و ریاضیات، از مسائلی است که عملکرد و نتایج را تحت‌تأثیر قرار می‌دهد. در کتاب‌های علمی، بخش زیادی از فهرست‌های مندرجات ازاین‌دست کلمات هستند که اگر محقق تصمیم بگیرد آن‌ها را نگه دارد، عدم یکدستی در خروجی پیش خواهد آمد، و اگر حذف کند، بخش مهمی از معنای مدارک از بین می‌رود؛ چنان‌که در این پژوهش نیز حذف کلمات انگلیسی و لاتینی حذف شد. ناکارآمدی «او سی آر» مای فارسی از دیگر نقاط ضعفی است که شاید مستقیماً و تنها مربوط به فهرست‌های مندرجات کتاب‌ها نیست؛ ولی همانند سایر بخش‌های کتاب‌ها یا سایر منابع متنی که در قالب‌های غیر از تکست و ورد تهیه می‌شوند، مشکل تبدیل فرمت را در زبان فارسی پیش می‌آورد. مسئله دیگر در استفاده از فهرست‌های مندرجات به‌عنوان پایه‌ای برای پردازش متن و استخراج کلیدواژه موضوعی از متن، عدم قطعیت در میزان تفصیل و توضیح محتواست. فهرست‌های مندرجات قطعاً – و بخصوص در کتاب‌های علمی– گرفته‌شده از متن و نمایانگر سرفصل‌ها و تیترهای قطعات محتوایی کتاب‌ها هستند، اما نویسنده و ناشر الزامی برای رعایت جزئیات و میزان تفصیل دراین‌خصوص ندارند و تا حدودی سلیقه‌ای است. این مسئله هرچند در پردازش ماشینی کل پیکره مسئله مهمی نیست، اما در تحلیل موضوعی هر عنوان کتاب و ایجاد ارتباط آن با نمایه به‌منظور امکان بازیابی مشکل‌ساز است، چراکه مندرجات بسیار کوتاه و موجز اطلاعات بسیاری را برای پردازش فاقدند و از داده‌های تُنُک محسوب می‌شوند. برای پردازش ماشینی تلفیق فهرست مندرجات، عنوان و صفحات مقدمه، پیشگفتار و نتیجه‌گیری توصیه می‌شود .

برای پاسخ به پرسش پنجم پژوهش، نتیجه ارزیابی نشان داد که همه کاربران بر اساس کلیدواژه‌های خروجی مدل به‌درستی حوزه کلی مدارک (علوم پایه) را برگزیدند؛ بنابراین، می‌توان نتیجه گرفت که خروجی مدل از حیث معرفی و بازنمایی محتوای کل پیکره، کارآمد بوده است. در بخش دیگر ارزیابی، کلیدواژه‌های سرعنوانی مربوط به هر فهرست مندرجات از کاربران گروه اول به طور میانگین ۸۷.۲۲ نمره از ۱۰۰ و از کاربران گروه دوم به طور میانگین ۸۳.۶۱ نمره از ۱۰۰ را کسب کردند و بهتر از دو گروه دیگر ارزیابی شدند. نمره ضریب همبستگی اسپیرمن میان نمرات ارزیابی دو گروه کاربران که میزان توافق آن‌ها محسوب می‌شود، ۰/۴۸۲ و همبستگی در سطح ۰/۰۱ > ۰/۰۰۰ =p معنادار است، بدین معنی که میان کاربران در حد متوسط توافق وجود دارد .

کلیدواژه‌های مستخرج از اجرای مدل در فهرست‌های مندرجات زیرحوزه‌های موضوعی از کاربران گروه اول به طور میانگین ۶۴.۳۹ نمره از ۱۰۰ و از کاربران گروه دوم به طور میانگین ۶۴.۹۵نمره از ۱۰۰ را کسب کردند. نمره ضریب همبستگی اسپیرمن میان نمرات ارزیابی دو گروه کاربران برای این گروه از کلیدواژه‌ها ۰/۷۳۸ و همبستگی در سطح ۰/۰۱ > ۰/۰۰۰ =p معنادار است، بدین معنی که میان کاربران در حد زیاد توافق وجود دارد.کلیدواژه‌های مستخرج از اجرای مدل با کل پیکره کاربران گروه اول به طور میانگین ۵۸/۹۳ نمره از ۱۰۰ و از کاربران گروه دوم به طور میانگین ۶۰/۳۴ نمره از ۱۰۰ را کسب کردند که پایین‌تر از دو گروه دیگر و با فاصله کم از گروه کلیدواژه‌های مستخرج از زیررده‌ها قرار گرفتند. نمره ضریب همبستگی اسپیرمن میان نمرات ارزیابی دو گروه کاربران ۰/۷۹۷ و همبستگی در سطح ۰/۰۱ > ۰/۰۰۰ =p معنادار است، بدین معنی که میان کاربران در حد زیاد توافق وجود دارد. این پژوهش نیز همانند لوی و فرانکلین (۲۰۱۴) و دی ماجیو و دیگران (۲۰۱۳) و شورت (۲۰۱۹) نیز اشاره کرده‌اند هنوز نمی‌توان باتکیه‌بر روش‌های موجود تحلیل موضوعی را به‌صورت تماماً خودکار انجام داد و تنها به‌صورت کمک به کار نمایه‌ساز و فهرست‌نویس برای صرفه‌جویی در وقت و منابع می‌تواند مورداستفاده قرار گیرد.

پرسش اصلی پژوهش را بدین‌گونه می‌توان پاسخ داد که فهرست‌های مندرجات منبع خوبی برای استخراج کلیدواژه محسوب می‌شوند که ترکیب آن‌ها نه با متن کامل کتاب، بلکه با

اجزاء خلاصه ولی گویای دیگر کتاب‌ها مانند عنوان، مقدمه و نتیجه‌گیری می‌تواند بر کارآمدی تحلیل خودکار بیفزاید و ال. دی. اِی. به‌عنوان یک روش پردازش زبان طبیعی می‌تواند در مواجهه با مقادیر زیاد اطلاعات متنی نقش یاریگری کارآمد را برعهده بگیرد، ولی تا خودکار شدن فرایند تحلیل موضوعی و استخراج کلیدواژه به‌گونه‌ای که رضایتمندی کاربر انسانی را برآورده سازد، تلاش‌های بسیاری لازم است. مدل‌سازی‌های موضوعی و استخراج کلیدواژه با روش‌های ماشینی پردازش زبان را می‌توان در مجموعه‌های توصیف نشده و ناشناخته به‌منظور استخراج محتوای موضوعی ناآشکار کل مجموعه به کار برد. اما در روندها و رویه‌های رسمی توصیف موضوعی تک‌تک مدارک می‌توانند به‌عنوان یک سیستم پیشنهاددهنده کلیدواژه به نیروی انسانی نمایه‌ساز و تحلیلگر موضوعی به کار برده شوند.

## ۵. محدودیت‌های پژوهش و پیشنهاد برای پژوهش‌های آتی

- از محدودیت‌های گردآوری داده‌ها در این پژوهش عدم دسترسی به فایل قابل کپی‌برداری کتاب‌ها و فهرست‌های مندرجات آن‌ها بود که گردآوری خودکار را برای پژوهشگر میسر نکرد.
- در دوره گردآوری داده‌ها صفحات موردنیاز در وب‌سایت «خانه کتاب» از دسترس خارج شد که روند پیش‌بینی‌شده برای گردآوری را مختل کرد؛ چراکه پیش‌بینی‌شده بود فهرست‌های مندرجات و کلیدواژه‌های استاندارد طلایی هر دو از یک منبع، وب‌سایت «خانه کتاب»، برداشت شوند.
- داده‌های مربوط به بعضی از کتاب‌ها، ازجمله آی. اس. بی. ان.[1]، رده‌بندی کتاب، و موضوعات انتساب داده شده صحیح نبودند که البته در روند کلی کار اختلال زیادی ایجاد نکرد، اما کنترل با چند بانک اطلاعاتی را ضروری نمود که وقت‌گیر بوده است.
- موضوعات سرعنوانی انتساب داده شده به تعداد کمی از کتاب‌ها در کتابشناسی ملی با داده‌های خانه کتاب متفاوت بود که در این موارد محقق بنا بر تشخیص خود موضوعات یکی از دو بانک اطلاعاتی را که صحیح‌تر به نظر می‌آمد برگزیده است .

ابزارهای فارسی برای پیش‌پردازش زبان فارسی انگشت‌شمارند و اغلب همه زبان‌های برنامه‌نویسی را پوشش نمی‌دهند. این مسئله موجب کاهش تنوع عملکردها و پردازش‌ها، بخصوص برای بهره‌برداران غیرمتخصص، می‌شود و درنهایت بر کیفیت تحلیل‌ها نیز اثر می‌گذارد. این پژوهش نیز با این محدودیت مواجه بود و برای اجرای پیش‌پردازش‌ها محقق مجبور به استفاده از کدهای کتابخانه‌های پراکنده شد.

برای پژوهش‌های آتی موارد زیر پیشنهاد می‌شوند :

- پژوهش با تلفیق عناوین، فهرست‌های مندرجات، صفحات پیشگفتار، مقدمه و نتیجه‌گیری یا هر بخشی که مؤخره کتاب‌ها محسوب می‌شود، اجرا شود.
- به علت عدم دسترسی به فایل تمام متن کتاب‌ها امکان بررسی نمایه‌های آخر کتاب و مقایسه آن‌ها با کلیدواژه‌های خروجی مدل میسر نشد که می‌تواند دست‌مایه پژوهش‌های آتی قرار گیرد.
- بازنمایی اهمیت کلمات و عبارت‌ها، عموم و خصوص و جزئیت و کلیت در فهرست‌های مندرجات که معمولاً به شکل تورفتگی و کاربرد قلم درشت‌تر یا ایتالیک رعایت و نمایانده می‌شود، در پردازش‌ها و پیش‌پردازش‌های متن‌کاوی به‌صورت تأثیر در وزن‌دهی کلمات و اصطلاحات در نظر گرفته‌شده و تأثیر آن در خروجی بررسی شود.
- پژوهش با متون حوزه‌های دیگر علم و انواع ادبی دیگر اجرا شود.
- پرس‌وجوی کاربر در سیستم بازیابی اطلاعات که از نیاز اطلاعاتی و فضای واژگان و دانش او نشئت می‌گیرد، مهم‌ترین مسئله در سیستم‌های تحلیل، نمایه‌سازی و بازیابی اطلاعاتی است که تحقیقات سامان‌یافته‌تری را می‌طلبد. مطالعات مشابهی که ازنظر بازیابی اطلاعات، تطبیق کلیدواژه‌های پرس‌وجوی کاربران با کلیدواژه‌های به‌دست‌آمده از مدل‌های پردازش زبان طبیعی مطلوب خواهد بود .
- فقدان یک پایگاه‌داده تمام متن از کتاب‌های فارسی که دست‌کم حاوی فهرست‌های مندرجات و صفحات اول کتاب‌ها باشد، و با کمک آن بتوان به ارزیابی توان بازیابی مدارک توسط کلیدواژه‌های خروجی مدل ال. دی. اِی.

[1] ISBN

پرداخت، منجر به طراحی این پژوهش به‌صورت بی‌نظارتی شد. در صورت دسترسی به چنین پایگاه داده‌ای می‌توان از روش‌های جامعیت، مانعیت، دقت و سنجه اف. در ارزیابی خروجی استفاده کرد .

- استخراج کلیدواژه با چندین مدل در پیکره واحد صورت گیرد و نتایج مقایسه و ارزیابی شود؛
- استخراج کلیدواژه با کاربرد هستی‌شناسی، جاسازی کلمات صورت گیرد و شباهت‌سنجی با فناوری سافت کوساین[1] و فناوری‌های مرتبط اجرا شود.

## یادداشت

این مقاله مستخرج از رساله دکتری است.

## تقدیر و تشکر

بدین‌وسیله از کارکنان محترم خانه کتاب ایران، به‌ویژه سرکار خانم سمانه نادری که فهرست کتاب‌های حوزه علوم منتشرشده در کشور را در اختیار این‌جانب قراردادند، سپاسگزاری می‌کنیم. همچنین قدردان اعضای محترم هیئت‌علمی و داورانی محترمی هستیم که فرم‌های ارزیابی این پژوهش را در رشته‌های مرتبط داوری کردند.

## References

Asgari, E., Chappelier, J.-C. (2013). Linguistic resources & topic models for the analysis of Persian poems. In Proceedings of the Second Workshop on Computational Linguistics for Literature ( pp. 23–31), Atlanta, Georgia, June 14, 2013. Association for Computational Linguistics .

Asmussen, C. B., & Mller, Ch. (2019). Smart literature review: A practical topic modeling approach to exploratory literature review. Journal of Big Data, 6(93). DOI: 10.1186/s40537-019-0255-7

Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2015). An overview of graph-based keyword extraction methods and approaches. Journal of Information and Organization Sciences, 39(1), 1-20. Retrieved from https://jios.foi.hr/index.php/jios/article/view/938

Blei, Ng, and Jordan. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022. DOI: 10.5555/944919.944937

Choi, Y., Hsieh-Yee, I., & Kules, B. (2007). Retrieval effectiveness of table of contents and subject headings. JCDL '07 June 18–23, 2007, Vancouver, British Columbia, Canada (pp.103-104). DOI:10.1145/1255175.1255195

Dieng, A. B., Ruiz, F. J. R., Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439–453. DOI: 10.1162/tacl a 00325

Di Maggio, P., Nag, M., Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding, Poetics, 41(6), 570-606. DOI: 10.1016/j.poetic.2013.08.004.

Goh, R. (2018). Using Named Entity Recognition for Automatic Indexing. Paper presented at the IFLA WLIC, 2018, Kuala Lumpur, Malaysia

Golube, K, Hagelbach, J., & Ardo, A. (2018). Automatic classification using DDC on the Swedish Union Catalogue. CEUR-WS.org/vol-2200/paper1.pdf

Hamid, F. (2016). Evaluation techniques and graph-based algorithm for automatic summarization and keyphrase extraction. (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No. 10307512)

Hoyt, B. (2020). Best practices for content manager ondemand full-text search. Retrieved from https://www.ibm.com/support/pages/sites/default/files/inline-files/Best%20practices%20for%20Using%20Full%20Text%20Searching%20with%20Content%20Manager%20OnDemand-4-22-2020.pdf

Hurtado, J. L. (2016). Text mining and topic modeling for social and medical decision support. (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No. 10583055)

Im, Y., Park, J., Kim, M., & Park, K. (2019). Comparative study on perceived trust of topic modeling based on affective level of educational text. Appl. Sci, 9(21), 4565. DOI: 10.3390/app9214565

[1] Soft Cosine

Junger, U. (2018). Automated first- The subject cataloguing policy of the Deutsche Nationalbibliothek. Paper presented at IFLA WLIC 2018- Kuala Lumpur, Malaysia- Transform Libraries, Transform Societies in Session 115- Subject Analysis and Access. Retrieved from http://library.ifla.org/2213/1/115-junger-en.pdf

Khoshian, Nahid, and Mirzaeian, Vahidreza (2020). The Most Widely Used Functions of Natural Language Processing in the Field of Library Science and Information Science. Knowledge Retrieval and Semantic Systems, 6(23), 117-151. DOI: 10.22054/jks.2020.44502.1238. (Persian)

Levy, K. E. C., & Franklin, M. (2014). Driving regulation: Using topic models to examine political contention in the U.S. trucking industry. Social Science Computer Review, 32(2), 182–194. DOI: 10.1177/0894439313506847

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam,A. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures, DOI: 10.1080/19312458.2018.1430754

Mas'oudi, B., & Rahati Ghochani S. (2016). Farsi word sense disambiguation with LDA Topic model . JSDP, 12 (4), 117-125. Retrieved from http://jsdp.rcisp.ac.ir/article-1-58-fa.html. (Persian)

Momtazi, S. (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. Information Processing and Management, 54,380–393. DOI: 10.1016/j.ipm.2018.01.001

Onal Suzek, T. (2017). Using latent semantic analysis for automated keyword extraction from large document corpora. Turkish Journal of Electrical Engineering & Computer Sciences, 25, 1784-1794. DOI: 10.3906/elk-1511-203

Pietsch, A.-S., & Lessmann, S. (2018) Topic modeling for analyzing open-ended survey responses. Journal of Business Analytics, 1(2), 93-116. DOI: 10.1080/2573234X.2019.1590131

Pokorny, J. (2018). Automatic subject indexing and classification using text recognition and computer based analysis of the table of contents. In Chau, L.; & Mounier, P. ELPUB 2018. June 2018, Toronto, Canada. DOI: 10.4000/proceedings.elpub.2018.19.

Rahgozar, A. (2020). Automatic poetry classification and chronological semantic analysis. (Doctoral dissertation). The University of Ottawa. Canada. Retrieved from https://ruor.uottawa.ca/bitstream/10393/40516/3/Rahgozar_Arya_2020_thesis.pdf

Revert, F. (2019). An Overview of Topics Extraction in Python with Latent Dirichlet Allocation. Retrieved from https://www.kdnuggets.com/2019/09/overview-topics-extraction-python-latent-dirichlet-allocation.html

Riaz, K. H. (2018). Improving search via named entity recognition in morphologically rich languages – A case study in Urdu (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No. 10747478)

Risch, J. (2016). Detecting Twitter topics using Latent Dirichlet Allocation. (Master's Thesis). Retrieved from http://uu.diva-portal.org/smash/get/diva2:904196/FULLTEXT01.pdf

Roder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In The Eighth ACM International Conference on Web Search and Data Mining WSDM'15, February 2–6, Shanghai, China (pp. 39– 408). ACM. DOI: 10.1145/2684822.2685324

Sadeghi, M., & Vegas, J. (2014). Automatic identification of light stop words for Persian information retrieval systems. Journal of Information Science, 40, 476 - 487. DOI: 10.1177/0165551514530655

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts, and the best number of topics: Some Problems and solutions. Quality & Quantity, 54, pp. 1095–1108. DOI: 10.1007/s11135-020-00976-w

Saidul Hasan, K., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 23-25, 2014. Pp. 1262-1273. DOI: 10.3115/v1/p14-1119

Schauble, P. (1997). Multimedia information retrieval: Content-based information retrieval from large text and audio databases. New York: Springer Science+Business Media .

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In Proceedings

of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, April 2017 (pp. 432–436). Association for Computational Linguistics.https://www.aclweb.org/anthology/E17-2069.pdf

Sfakakis, M., Zoutsou, K., Papachristopoulos, L., Tsakonas, G., & Papatheorodu, Ch. (2019, August). Between two worlds: harmonizing automated and manual term labeling. Paper presented at IFLA WLIC 2019 - Athens, Greece - Libraries: dialogue for change in Session S02 - Knowledge Management with Digital Humanities/Digital Scholarship. In: Artificial Intelligence (AI) and its impact on libraries and librarianship, 22 August 2019, Corfu, Greece. Retrieved from http://library.ifla.org/2759/1/s02-2019-sfakakis-en.pdf

Short, M. (2019). Text mining and subject analysis for fiction; or, using machine learning and information extraction to assign subject headings to dime novels. Cataloging and Classification Quarterly, 57(5), 315-336. DOI: 10.1080/01639374.2019.1653413

Sun, Y., Loparo, K., & Kolacinski, R. (2020). Conversational Structure Aware and Context Sensitive Topic Model for Online Discussions. 2020 IEEE 14th International Conference on Semantic Compting(ICSC),(pp.8592).DOI:10.1109/ICSC.2020.00019

Tchoua, R. B. (2019). Hybrid human-machine scientific information extraction. (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No. 13904924)

Sun, Ch., Hu, L., Li, Sh., Li,T., Li, H., & Chi, L. (2020). A Review of Unsupervised Keyphrase Extraction

Methods Using Within-Collection Resources. Symmetry, 12(1864). DOI:10.3390/sym12111864

Syed, Sh., and Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 2017 IEEE International Conference on Data Science and Advanced

Analytics (DSAA), Tokyo, Japan, 2017, pp. 165-174. doi: 10.1109/DSAA.2017.61

Tushara, M. G., Mownika, T., & Mangamuru, R. (2019). A comparative study on different keyword extraction algorithms. In Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), Erode. India, 2019. Pp 969-973; DOI: 10.1109/ICCMC.2019.8819630

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using Latent Dirichlet Allocation. Electronic Commerce Research and Application, 29, 142-156. DOI:10.1016/j.elerap.2018.04.003

Wang, Y., & Taylor, J. E. (2019). DUET: data-driven approach based on Latent Dirichlet Allocation topic modeling. Journal of Computing in Civil Engineering, 33(3), 04019023.

Xing, L., Paulz, M. J., & Carenini, G. (2019). Evaluating Topic Quality with Posterior Variability. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International

Joint Conference on Natural Language Processing, Hong Kong, China, November 3–7, 2019 (pp. 3471–3477). Association for Computational Linguistics. DOI: 10.18653/v1/D19-1349

Yan, Y., Guo, J., Lan, Y., & Cheng, X. (2013). A Biterm topic model for short texts. WWW2013, May, 13-17,2013, Rio de Janeiro, Brazil. DOI: 10.1145/2488388.2488514

Yao, J., Wang, Y., Zhang, Y., Sun, J., & Zhou, J. (2018). Joint Latent Dirichlet Allocation for social tags. IEEE Transactions on Multimedia, 20(1). DOI: 0.1109/TMM.2017.2716829