Comparison of the Performance of Approaches in Discovering **and Extracting E-book Topics**

Fatemeh Zarmehr
PhD Candidaie in Knowledge and Information Science; Department of Knowledge and Information Science; University of Isfahan; Isfahan, Iran Email: fatemezarmehr99@gmail.com

Ali Mansouri*
PhD in Knowledge and Information Science; Associate Professor; Department of Knowledge and Information Science; University of Isfahan; Isfahan, Iran Email: a.mansouri@edu.ui.ac.ir

Hossein Karshenas Najafabadi
PhD in Artificial Intelligence; Assistant Professor; Faculty of Computer Engineering; University of Isfahan; Isfahan, Iran; Email: h.karshenas@eng.ui.ac.ir

Received: 25, May 2022 **Accepted: 04, Sep. 2022**

Abstract: Keyword extraction is one of the most important issues in text processing and analysis and provides a high-level and accurate summary of the text. Therefore, choosing the right method to extract keywords from the text is important. The aim of the present study was to compare the performance of three approaches in discovering and extracting the subject keywords of e-books using text mining and machine learning techniques. In this regard, three experimental approaches have been introduced and compared including the successive implementation of the clustering process, improving the quality of clusters in terms of semantics and enriching the stop words of a specific field, use of specialized keyword template, finally, the use of important parts of the text in discovering and extracting key words and important topics of the text. The statistical population includes 1000 e-book titles from the subject fields of library and information science based on the congress classification system. Bibliographic information of e-books was obtained from the Congress Library database, then the original text was prepared. The extraction of topic keywords and clustering of training data was performed using the non-negative matrix factorization algorithm with three experimental approaches. The quality and performance of the subject clusters resulting from the implementation of three approaches in the automatic classification of experimental data were compared using a support vector machine. The findings showed that the Hamming loss (0.020) and in other words the error rate in the correct classification of experimental texts in the third approach is far less than the other

* Corresponding Author

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 4 | pp. ??-??
Summer 2023
https://doi.org/jipm.38.4

two approaches. Also, the F1 score (0.82), which is the average of the two criteria of precision (0.87) and recall (0.78) and is a reflection of the correct performance of the classification process in topic labeling of texts, is better in the third approach than the other two approaches. The results showed that the quality and semantic coherence of the subject clusters obtained from the third approach, i.e. the use of important parts of the text in discovering and extracting the subject, was better compared to other two approaches. In this approach, by focusing on the main parts of the data, which represent the main content and theme of the text, more meaningful topic clusters were obtained. In addition, the keywords obtained from the topic cluster of the third approach can be used in unspecified and unknown collections in order to extract the unknown thematic content of the whole collection. The results of third approach also was better in terms of accuracy and readability (0.79) and the rate of classification error (0.020) of texts, in comparison of other two approaches.

Keywords: E-book, Extraction, Subject Keywords, Text Mining, Subject Modeling

مقایسه عملکرد رویکردهای کشف و استخراج موضوعات کتاب‌های الکترونیک

فاطمه زرمهر
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛ دانشگاه اصفهان؛ اصفهان، ایران؛
Fatemezarmehr99@gmail.com

علی منصوری
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ گروه علم اطلاعات و دانش‌شناس؛ دانشکده علوم تربیتی و روان‌شناسی؛ دانشگاه اصفهان؛ اصفهان، ایران؛
پدیدآور رابط a.mansouri@edu.ui.ac.ir

حسین کارشناس نجف‌آبادی
دکتری هوش مصنوعی؛ استادیار؛ گروه هوش مصنوعی؛ دانشکده مهندسی کامپیوتر؛ دانشگاه اصفهان؛ اصفهان، ایران h.karshenas@eng.ui.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۱/۰۳/۰۴ **پذیرش:** ۱۴۰۱/۰۶/۱۳ مقاله برای اصلاح به مدت ۱۵ روز نزد پدیدآوران بوده است.

چکیده: استخراج کلمات کلیدی از مسائل مهم در زمینه پردازش و تحلیل متن بوده و خلاصه‌ای سطح بالا و دقیق از متن ارائه می‌دهد. بنابراین، انتخاب روش مناسب برای استخراج کلمات کلیدی متن حائز اهمیت است. هدف پژوهش حاضر، مقایسه عملکرد سه رویکرد درکشف و استخراج کلیدواژه‌های موضوعی کتاب‌های الکترونیک با استفاده از تکنیک‌های متن‌کاوی و یادگیری ماشین است. در این راستا سه رویکرد آزمایشی شامل، ۱) اجرای متوالی فرایند خوشه‌بندی، ارتقای کیفیت خوشه‌ها از نظر معنایی و غنی‌سازی کلمات توقف حوزه خاص، ۲) استفاده از الگوی کلیدواژه‌های تخصصی، ۳) استفاده از بخش‌های مهم متن در کشف و استخراج واژگان کلیدی و موضوعات مهم متن معرفی و مقایسه شده است. جامعه آماری شامل ۱۰۰۰ عنوان کتاب الکترونیک از زیرشاخه‌های موضوعی حوزه علم اطلاعات و دانش‌شناسی بر اساس نظام رده‌بندی کنگره است که بعد از کسب اطلاعات کتابشناختی آن از پایگاه کتابخانه کنگره، اقدام به تهیه متن اصلی گردید. استخراج کلیدواژه‌های موضوعی و خوشه‌بندی داده‌های آموزش به کمک الگوریتم تجزیه نامنفی ماتریس و با سه رویکرد آزمایشی انجام شد و کیفیت و عملکرد خوشه‌های موضوعی حاصل از اجرای سه رویکرد در بخش دسته‌بندی خودکار داده‌های آزمایشی به کمک ماشین بردار پشتیبان مقایسه شد.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ٤ | ص‌ص ۱۳۶۹-۱۳۹٤
تابستان ۱۴۰۲
https://doi.org/jipm.38.4

یافته‌ها نشان داد که افت همینگ (۰/۰۲۰) یا میزان خطا در دسته‌بندی صحیح متون آزمایشی در رویکرد سوم یعنی بهره‌گیری از بخش‌های مهم متن در استخراج کلیدواژه‌های موضوعی، از دو رویکرد دیگر کمتر است. همچنین امتیاز F1 (۰/۸۲) که میانگین دو معیار دقت (۰/۸۷) و بازخوانی (۰/۷۸) و بازتابی از عملکرد درست فرایند دسته‌بندی در برچسب‌گذاری موضوعی متون است، در رویکرد سوم بهتر از نتایج دو رویکرد دیگر است. نتایج تحلیل‌ها نشان داد که کیفیت و انسجام معنایی خوشه‌های موضوعی حاصل از رویکرد سوم، یعنی استفاده از بخش‌های مهم متن در کشف و استخراج موضوع، در مقایسه با دو رویکرد دیگر بهتر بود. افزون بر این، کلیدواژه‌های به‌دست‌آمده از خوشه‌های موضوعی رویکرد سوم را می‌توان در مجموعه‌های توصیف‌نشده و ناشناخته به‌منظور استخراج محتوای موضوعی ناآشکار کل مجموعه به‌کار برد.

کلیدواژه‌ها: کتاب الکترونیک، استخراج، کلیدواژه‌های موضوعی، متن‌کاوی، مدل‌سازی موضوعی

۱. مقدمه

در توصیف منابع به‌منظور سازماندهی و ذخیره‌سازی، تحلیل موضوعات مندرج در آن‌ها و استخراج و بازنمایی این موضوعات در قالب عباراتی که آن‌ها را «کلیدواژه یا توصیفگر» می‌نامند، اهمیت بسیاری دارد. این عبارات یا به‌طور مستقیم برگرفته از محتوای منبع هستند، و یا از مجموعه واژگانی کنترل‌شده انتخاب می‌شوند. این کلیدواژه‌ها در بانک اطلاعاتی با استفاده از روش‌های فنی به مدارکی که حاوی همان موضوعات باشند، تخصیص داده می‌شوند. به این کلیدواژه‌ها نقطه دسترسی یا شناسه نیز می‌گویند (Wang, Zhang & Klabjan 2020). به‌منظور تعیین محتوای منابع و بازیابی بهینه منابع اهمیت دارد که این کلیدواژه‌ها به‌عنوان نقاط دسترسی موضوعی تعیین و در اختیار کاربر قرار گیرند. کلیدواژه‌های موضوعی محصول فرایند تحلیل موضوعی منابع و استخراج یا تولید کلیدواژه‌های موضوعی است. کلیدواژه‌های موضوعی ابزار مهمی برای ثبت مهم‌ترین اطلاعات در یک سند متنی هستند. این واژگان در سازماندهی (Onan 2018)، بازیابی اطلاعات (Berger & Lafferty 2017)، خلاصه‌سازی متن، و شناسایی موضوعات پرداخته‌شده در رسانه‌های اجتماعی (Onan 2017) استفاده می‌شوند. در حوزه علم اطلاعات و دانش‌شناسی، بررسی منبع از حیث موضوعات پرداخته‌شده در آن، کشف و استخراج موضوعات و تعیین واژگانی که بازنمون واضحی از این موضوعات برای کتاب‌ها هستند، فهرست‌نویسی تحلیلی، و برای مقالات و سایر انواع منابع اطلاعاتی نمایه‌سازی موضوعی نامیده می‌شود. فرایند استخراج عبارات کلیدی که به‌عنوان «استخراج خودکار کلیدواژه‌های موضوعی» نامیده می‌شود، در دو دهه گذشته توجه بسیاری را به خود جلب کرده است (Basaldella et al. 2018) و در بسیاری از عمکردها و خروجی پردازش زبان طبیعی مانند خلاصه‌سازی متن (Zhang 2004)، خوشه‌بندی اسناد (Gers 2002) یا وظایفی غیر از پردازش زبان

طبیعی مانند تحلیل شبکه اجتماعی (Graves & Schmidhuber 2005) یا مدل‌سازی کاربر (De Nart 2015) و بازیابی اطلاعات در بایگانی اسناد دیجیتال (Kaur & Chopra 2016) به‌کار گرفته شده‌اند. رویکردهای استخراج کلیدواژه‌های موضوعی را می‌توان به رویکردهای آماری ساده، رویکردهای زبانی، رویکردهای یادگیری ماشین و سایر رویکردها تقسیم نمود (Zhang 2008). رویکردهای آماری ساده نیازی به آموزش ندارند و کلیدواژه‌های موضوعی را بر اساس تجزیه‌وتحلیل آماری و محاسبه فراوانی کلمات، احتمالات و سایر ویژگی‌ها استخراج می‌کنند؛ مانند روش بسامد واژه–بسامد معکوس مدرک[1] (Salton & Buckley 1988)، بسامد کلمه (Wang et al. 2017)، تکنیک N-gram (Cohen 1995)، وقوع کلمه (Matsuo & Ishizuka 2004). غالب رویکردهای آماری از یادگیری بدون نظارت بهره می‌گیرند. مکانیسم این روش‌ها اختصاص وزن به هر کلمه است. در رویکردهای زبانی، ویژگی‌های زبانی کلمات، جملات و کل مدرک، مانند روش بهره‌گیری از کلمه (Ercan & Cicekli 2007) بهره‌گیری از نحو (Hulth 2003) محتوا برای تجزیه‌وتحلیل (Dennis 1967; Salton & Buckley1991) و یافتن کلمات کلیدی موضوع در اسناد (Zhang 2008) لحاظ می‌شود. لازم به ذکر است که استفاده از رویکردهای زبانی مستلزم تجزیه‌وتحلیل بسیار پیچیده بوده و به مهارت‌های زبان‌شناسی نیاز دارد. کشف و استخراج کلیدواژه‌های موضوعی متن بر اساس روش‌های متن‌کاوی و الگوریتم‌های مختص این روش، رویکردی جدید است که در پژوهش‌های مختلف سعی شده است جنبه‌هایی از نیاز کتابداران و متخصصان موضوعی را به‌منظور شناسایی موضوعات و تفسیر منابع و کارکرد آن در بازیابی بهینه منابع مورد بررسی قرار دهند (Onan 2020). از آنجا که بیشتر اطلاعات (بیش از ۸۰درصد) به‌صورت متن ذخیره شده‌اند، و حاوی اطلاعات ارزشمند و نهفته‌ای هستند، اعتقاد بر این است که متن‌کاوی ارزش بالقوه بالایی دارد (Beliga, Meštrović & Martinčić-Ipšić 2015). از جمله کاربردهای متن‌کاوی می‌توان به دسته‌بندی، خوشه‌بندی، خلاصه‌سازی و یافتن روابط میان مفاهیم در متون و موضوعات اصلی متون اشاره کرد. روش‌های متن‌کاوی و الگوریتم‌های مرتبط با کشف و استخراج کلیدواژه‌های موضوعی این قابلیت را دارد که افزون بر کشف موضوعات پنهان و ارتباط آن با موضوع‌های آشکار، موضوع‌های کمتر شناخته‌شده یا کمتر پرداخته‌شده را شناسایی کند (باغ محمد، منصوری و چشمه‌سهرابی ۱۳۹۹، ۳۰۰). بر اساس همین ویژگی و کارکرد اصلی روش متن‌کاوی، در سال‌های اخیر متون و منابع حجیم از قبیل کتاب و متن مقالات با استفاده از فنون متن‌کاوی مورد تحلیل قرار می‌گیرد. از سوی دیگر، در عصر فناوری‌های اطلاعاتی و با رشد سریع داده‌های متنی حجیم، بازنگری در نحوه پردازش، کشف موضوعات، بازیابی و استفادهٔ بهینه از این متون عظیم را اجتناب‌ناپذیر کرده

1. term frequency- inverse document frequency (TF-IDF)

است (Onan ۲۰۱۸). کتاب‌های الکترونیک از جمله منابع اطلاعاتی هستند که هر روز بر مجموعه‌های کتابخانه‌ها و انباره‌های دیجیتال افزوده می‌شوند. از ویژگی‌های اصلی این منابع اطلاعاتی بدون ساختار بودن، طولانی بودن متن و قالب اطلاعاتی ارائه‌شده به‌صورت فایل‌های پی‌دی‌اف یا فایل‌های کتاب الکترونیک و چندرسانه‌ای است. روش‌های متعددی برای استخراج کلیدواژه‌های موضوعی این دسته از منابع وجود دارد. در وضعیت فعلی که تولید محتواهای اطلاعاتی متأثر از تحولات فناورانه دیجیتال و غیر آن به‌سرعت روبه‌افزایش است، سازماندهی و توصیف منابع با روش‌های دستی مستلزم صرف وقت و هزینهٔ بالا بوده و هم مانعی است در دسترسی به‌موقع کاربران به منابع مورد نیاز آن‌ها که همین امر ماشینی کردن همه یا بخشی از روندهای شناسایی محتوا و تخصیص کلیدواژه به منابع الکترونیک را به ذهن متبادر می‌سازد. افزون بر این، مشکلاتی در حوزه سازماندهی و توصیف منابع، به‌ویژه در کتابخانه‌ها و انباره‌های دیجیتال وجود دارد که نمی‌توان آن‌ها را نادیده انگاشت. به‌عنوان مثال، هر چند کتاب الکترونیک پرکاربردترین منبع اطلاعاتی در کتابخانه‌های دیجیتال است، اما به‌علل گوناگون از جمله تعدد ابزارهای تحلیل موضوعی و انتساب کلیدواژه (سرعنوان‌های موضوعی، اصطلاحنامه‌ها و نمایه‌سازی تمام‌متن)، جست‌وجو و بازیابی در این کتابخانه‌ها با مسائل عدیده‌ای چون عدم موفقیت در بازیابی و در نتیجه، عدم رضایت کاربران روبه‌روست. به همین خاطر، استفاده از ابزارها و روش‌هایی که در سریع‌ترین زمان ممکن و با مناسب‌ترین نتیجه محتوا را شناسایی و کلیدواژه‌های مناسب را به منابع الکترونیک تخصیص دهد، ضروری است و از جمله این روش‌ها متن‌کاوی و یادگیری ماشینی است.

به‌منظور تعیین و شناسایی محتوای موضوعی منابع الکترونیک، بر اساس تکنیک متن‌کاوی و یادگیری ماشینی پژوهش‌های مختلفی انجام شده است. البته، بیشتر پژوهش‌های انجام‌شده در این زمینه در رابطه با متن‌های کم‌حجم همچون مقالات، توئیت‌ها، گزارش ثبت اختراع و نظرات کاربران بوده است، اما رویکرد مؤثری در استخراج کلیدواژه‌های موضوعی کتاب‌های الکترونیک به‌عنوان یکی از انواع متون حجیم معرفی و بررسی نشده است. چالش اصلی در رابطه با استخراج کلیدواژه‌های موضوعی متون حجیم همچون کتاب‌ها مواجه شدن با ابعاد گسترده ویژگی (کلیدواژه‌ها) و به‌دنبال آن، تعدد موضوعات پرداخته‌شده در متن است که همین امر تحلیل و دسته‌بندی موضوعی آن‌ها را در مقایسه با سایر داده‌های متنی پیچیده‌تر می‌کند. بنابراین، بهره‌گیری از رویکردی مؤثر در زمینه کاهش ابعاد ویژگی‌های کتاب‌ها و استخراج کلیدواژه‌های موضوعی نقش به‌سزایی در بازنمایی محتوا و درون‌مایه متن خواهد داشت. با توجه به مشکلات مذکور، پژوهش حاضر در پی پاسخ به این سؤال است که کدام‌یک از سه رویکرد مورد آزمایش در این پژوهش، یعنی اجرای متوالی فرایند خوشه‌بندی، ارتقای کیفیت خوشه‌ها از نظر معنایی و غنی‌سازی کلمات توقف حوزه خاص، استفاده از الگوی کلیدواژه‌های تخصصی، و استفاده

از بخش‌های مهم متن در کشف و استخراج واژگان کلیدی و موضوعات مهم متن منجر به استخراج کلیدواژه‌های موضوعی معنادارتر و مرتبط‌تر از کتاب‌های الکترونیک به‌عنوان یکی از انواع متن‌های بدون ساختار و حجیم خواهد شد و خوشه‌های موضوعی به‌دست‌آمده از این رویکردها، تا چه اندازه در دسته‌بندی موضوعی خودکار داده‌ها موفق خواهد بود؟ انتظار می‌رود پاسخ به این پرسش‌ها این امکان را برای پژوهشگران ایجاد کند که ضمن آشنایی با سه رویکرد مورد بحث در این پژوهش، با عملکرد هر کدام در شناسایی محتوا و تعیین کلیدواژهٔ مناسب برای توصیف محتوای کتاب‌های الکترونیک آشنا شده و در استفاده از رویکرد مناسب برای توصیف محتوای کتاب‌های الکترونیک آگاهانه اقدام نمایند.

۲. تکنیک‌های متن‌کاوی و یادگیری ماشین و استخراج کلیدواژه‌های متن

یادگیری ماشین کاربرد الگوریتم‌هایی است که حجم عظیمی از داده را به دانش تبدیل می‌کنند. یادگیری ماشین تلاش می‌کند که برنامه‌هایی را طراحی کند که از داده‌ها آموزش ببینند و در تصمیم‌گیری بعدی اقدام مناسب داشته باشند. به‌عبارت دیگر، این الگوریتم‌ها این قابلیت را به ماشین می‌دهند که رفتارش را متناسب با داده یا به‌عبارتی الگوها تغییر دهد. الگوریتم‌های یادگیری ماشین یک مدل ریاضی را بر اساس داده‌های نمونه، معروف به «داده‌های آموزش» برای پیش‌بینی یا تصمیم‌گیری بدون دستورالعمل‌های صریح برای انجام کار تولید می‌کنند. یادگیری ماشین دارای سه روش یادگیری «بدون نظارت»، «با نظارت» و «تقویتی» است (Allahyari et al. 2017) که با توجه به روش مورد استفاده در این پژوهش به توصیف دو روش اول اکتفا می‌شود.

یادگیری «بدون نظارت» روش‌هایی برای پیدا کردن ساختار پنهان از داده‌های بدون برچسب است. خوشه‌بندی و مدل موضوعی از جمله الگوریتم‌های مورد استفاده در یادگیری بدون نظارت است. در خوشه‌بندی متون، هدف افزایش هرچه بیشتر شباهت اسناد داخل یک خوشه با هم و کاستن از شباهت اسناد یک خوشه با خوشه‌های دیگر است. البته، در اینجا تأکید بر روی شباهت مفهومی و معنایی است (Han & Kamber 2006). در بیشتر پژوهش‌ها بر پایهٔ خوشه‌بندی، از الگوریتم‌های متمایزکننده مثل تجزیه‌وتحلیل معنایی پنهان[1] تا مدل‌های مولد، مثل تجزیه‌وتحلیل معنایی پنهان احتمالاتی[2]، تخصیص دیریکله پنهان[3]، مدل‌سازی موضوعی همبسته[4] و تجزیه ماتریس نامنفی[5] استفاده شده است (Mouhoub & Helal 2018).

1. latent semantic analysis (LSA)
2. probability latent semantic analysis (PLSA)
3. latent dirichlet allocation (LDA)
4. correlated topic modeling (CTM)
5. nonnegative matrix factoriztion (NMF)

از آنجا که حجم داده‌ها (اسناد یا متون) و نیز ابعاد (کلمات) بسیار زیاد است، از تکنیک‌های کاهش بُعد یا تخمین داده‌ای نظیر تجزیه ماتریس نامنفی در خوشه‌بندی متون الکترونیک استفاده می‌شود. تجزیه ماتریس نامنفی، برای اولین بار به‌وسیله Paatero and Tapper (1994) معرفی شد، اما به‌وسیله پژوهش‌های Lee & Seung (1999) معروفیت پیدا کرد. تجزیه ماتریس نامنفی به‌عنوان یک شیوه «نمایش بر مبنای اجزا»[1] مطرح شده است. از این نمایش داده‌ای می‌توان به‌طور مستقیم خوشه‌های متنی را به‌دست آورد و نیازی به سایر روش‌های متعارف خوشه‌بندی نیست. به‌عبارت دیگر، این روش برای نشان دادن ویژگی‌های معنایی در فعالیت‌هایی همچون خوشه‌بندی متون، برای کاهش بعد، تحلیل داده‌ها، تشخیص الگو، کشف اطلاعات معنادار در متون و خوشه‌بندی و دسته‌بندی آن‌ها به‌کار می‌رود (Langville & Albright 2014).

یادگیری «با نظارت» در متن‌کاوی نیز نتایج خوبی در دسته‌بندی موضوعی به همراه دارد (Gupta, Kumar & Pant 2018)، ولی نیازمند مجموعه آموزش‌های برچسب‌خورده[2] وسیعی هستند که فراهم‌کردن این حجم از داده برای حوزه‌های خاص موضوعی بسیار دشوار است. الگوریتم ماشین بردار پشتیبان[3] یکی از محبوب‌ترین الگوریتم‌های دسته‌بندی متون است. مبنای کار این الگوریتم دسته‌بندی خطی داده‌هاست (Li, Shang & Yan 2016). ماشین بردار پشتیبان بر پایه نظریه یادگیری محاسباتی توسعه یافته و بر اصل «حداقل‌سازی خطای ساختاری» تکیه دارد.

بر مبنای بررسی متون در رابطه با کشف و استخراج کلیدواژه‌های موضوعی متون با استفاده از تکنیک‌های یادگیری ماشین و متن‌کاوی، روش تجمیعی از واژگان معنایی اولین‌بار توسط Blei & Jordan (2003) ارائه شد. این روش یک شیوه بدون نظارت جهت استخراج موضوعات موجود در متن است و با هدف یافتن مفاهیم معنایی پنهان در متن به‌کار می‌رود (Chien, Lee & Tan 2018). در این روش با استفاده از روش تخصیص دیریکله پنهان[4] و توزیع چندجمله‌ای، به تعیین میزان تأثیر هر واژه از پیکره متنی در هر موضوع پنهان و همچنین توزیع احتمال رخداد موضوعات پنهان در متون پرداخته می‌شود. در صورت کاربرد این روش در تحلیل متن فرض می‌شود که هر متن دارای ترکیبی از چند موضوع پنهان است و هر موضوع پنهان نیز دارای توزیع احتمالی روی کلمات آن پیکره متنی است (Wilson & Chew 2010).

---

1. parts based representation
2. tagged
3. support vector machine (SVM)
4. latent drichelet allocation

تخصیصی دیریکله پنهان روشی جدید برای خوشه‌بندی معنایی است. این روش فرض می‌کند که اسناد موضوعات متعددی را نمایش می‌دهند. یعنی از کلماتی تشکیل شده است که هر یک متعلق به یک موضوع است و نسبت موضوعات داخل یک متن با هم متفاوت است.

«چن و لی» در پژوهش خود با توجه به کم‌تکرار بودن کلمات کلیدی در متن‌های خبری و وجود معانی مختلف از کلمات و همچنین مبهم بودن کاربرد برخی از کلمات که دارای بیش از یک معنا هستند، از الگوریتم «الدی‌ای»[1] برای شناسایی مفاهیم موجود در متن جهت سازماندهی متون استفاده کردند (Chen & Li 2016). ایده اصلی آن‌ها ایجاد ارتباط بین موضوع و درون‌مایه متن و فراوانی تکرار واژگان موجود در آن بود. در پژوهش Casalino et al. (2018) به‌منظور ارائه چارچوبی برای تجزیه هوشمند داده‌های توییتر از تجزیه ماتریس نامنفی استفاده کردند. این چارچوب به کاربران اجازه می‌داد با استخراج کلیدواژه‌های موضوعی به جست‌وجو و کشف توییت‌های مرتبط به‌صورت خودکار بپردازند.

در بیشتر پژوهش‌های مرتبط با استخراج کلیدواژه‌های موضوعی، پس از فاز پیش‌پردازش، حد بالا و پایینی از فراوانی تکرار واژه برای انتخاب ویژگی‌های اولیه (کلیدواژه‌های موضوعی) را در نظر می‌گیرند و واژگانی را که دارای فراوانی تکرار کمتر از حد آستانه پایین و بیشتر از حد آستانه بالا باشند، از مجموعه لغات استخراج‌شده از متن حذف می‌کنند (انبایی فریمانی، طباطبائی و کفاشان کاخکی ۱۳۹۸، ۱۸۸۸). همچنین، از فراوانی تکرار واژه‌ها به‌عنوان وزن مؤثر آن‌ها استفاده شده و دلیل حذف واژه‌ها به‌علت لحاظ نشدن این مسئله در الگوریتم‌های اجراست. در برخی از پژوهش‌های پیشین (مانند Wang, Zhang & klabjan 2020) پژوهشگران نشان دادند که انتخاب و حذف کلمات توقف عمومی و حوزه خاص از موارد تأثیرگذار در استخراج ویژگی است و می‌تواند به کاهش ابعاد ویژگی بدون ایجاد خدشه در بازنمایی محتوا و درون‌مایه متن منجر شود. همچنین، برخی از پژوهش‌های مرتبط دیگر به بهره‌گیری از الگوسازی موضوعات با استفاده از هستی‌شناسی‌ها تأکید دارد. در روش پیشنهادی Zhao & Mao (2018) برای ارائه یک متن، پس از استخراج واژه‌های موجود در متن، گروهی از آن‌ها بر اساس فراوانی تکرار به‌عنوان واژه‌های پایه انتخاب شده و برای هر واژهٔ پایه، خوشه‌ای از واژه‌هایی که ارتباط معنایی بین آن‌ها با استفاده از ویکی‌ها تعیین گردیده، استخراج شد. در روش پیشنهادی Elhadad, Badran & Salama (2017) با بهره‌گیری از ساختار سلسله‌مراتبی هستی‌شناسی واژه‌هایی که با سایر دسته‌های نحوی ارتباطی ندارند، از مجموعه ویژگی‌های استخراج‌شده حذف می‌گردد.

در ادامه، با استفاده از معیارهای شباهت Meng, Huang & Gu (2013) به بررسی شباهت معنایی و نحوی واژه‌های استخراج‌شده پرداخته شده و تنها واژه‌هایی با میزان شباهت بیشتر انتخاب می‌گردد. همچنین، در برخی پژوهش‌ها همچنون پژوهش Choi, Hsieh-Yee & Kules (2007) با بهره‌گیری از بخش‌های مهم متن همچون فهرست مندرجات، عنوان و مقدمه ضمن کاهش ابعاد ویژگی‌ها به استخراج

1. latent Dirichlet allocation (LDA)

کلماتی که نماینده درون‌مایه و محتوای متن است، اشاره گردیده است. نتایج بررسی‌ها نشان می‌دهد که صحت دسته‌بندی یا سازماندهی به نحوه استخراج مفاهیم مرتبط با واژه‌ها وابسته است و یکی از چالش‌های اصلی در این زمینه مواجه شدن با ابعاد زیاد ویژگی‌ها (کلمات متن) است. ابعاد بزرگ فضای کلیدواژه‌ها در دسته‌بندی متون به‌طور معمول، دردسرساز است. در حقیقت با بزرگ شدن فضای کلیدواژه‌ها، تعداد ویژگی‌ها نیز افزایش می‌یابد که از طرفی باعث پیچیدگی بیشتر، صرف هزینه زمانی و فضای حافظهٔ بیشتر گردیده و پردازش و تحلیل‌های بعدی را سخت‌تر می‌کند (Hoyt 2020). بنابراین، دقت و عملکرد انتخاب کلمات کلیدی تأثیر به‌سزایی در کاهش ابعاد (ویژگی‌ها) دارد. در پژوهش حاضر تلاش بر این است که با تعدیل چالش مذکور یعنی وسعت کلیدواژه‌ها در متن، در راستای کاهش ابعاد و استخراج کلیدواژه‌های موضوعی موجود در کتب الکترونیک، سه رویکرد (شامل غنی‌سازی ایست‌واژه‌های دامنه خاص، بهره‌گیری از الگوی کلیدواژه‌های تخصصی، و استفاده از بخش‌های مهم متن در استخراج موضوع و مفاهیم) معرفی و مقایسه شود. هدف از این بررسی بهره‌گیری از رویکردی مؤثر جهت استخراج کلیدواژه‌های موضوعی معنادارتر و مرتبط‌تر از کتاب‌های الکترونیک به‌عنوان یکی از انواع متن‌های بدون ساختار و حجیم است که به دسته‌بندی موضوعی بهتر آن‌ها به‌صورت خودکار منجر خواهد شد.

آنچه که موجب تمایز این پژوهش از پژوهش‌های دیگر است، این است که تلاش‌هایی که برای ماشینی کردن استخراج کلیدواژهای موضوعی شده، بیشتر در حوزه منابع غیرکتابی مانند مقالات، منابع وبی، پایان‌نامه‌ها و از این قبیل بوده است. در تحقیقات و پیشینه‌های موجود، استخراج کلیدواژه‌های موضوعی منطبق با رویکردهای پیشنهادی در این پژوهش و به‌خصوص بر روی کتاب‌های الکترونیک در یک حوزه موضوعی خاص انجام نشده است، که این پژوهش در صدد انجام آن است. کاربرد مدل موضوعی تجزیه نامنفی ماتریس برای تحلیل موضوعات در سه رویکرد مذکور نیز وجه تمایز دیگر این پژوهش با تحقیقات موجود است. کاربرد مدل‌ها در انواع منابع و در زبان‌های مختلف و حتی حوزه‌های موضوعی متفاوت، نتایج و چالش‌های خاص خود را داراست که تجربه و آزمون در این خصوص می‌تواند به رشد دانش و اصلاح رویکردهای موجود کمک کند.

۳. روش پژوهش

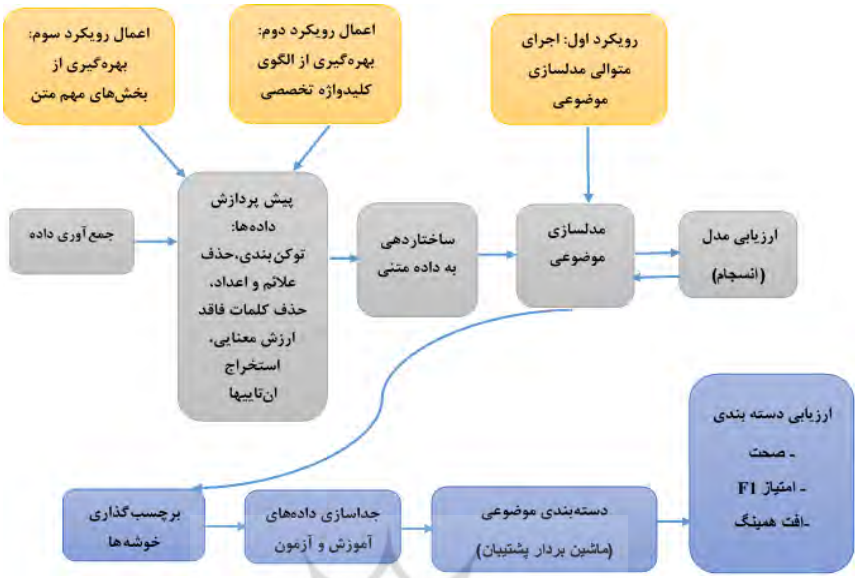
پژوهش حاضر از لحاظ هدف، کاربردی، و از نوع پژوهش‌های آمیخته متن‌کاوی است که در آن از فنون پردازش زبان طبیعی، مدل‌سازی موضوعی و یادگیری ماشین استفاده شده است. این پژوهش شامل سه گام اصلی است که عبارت‌اند از: پیش‌پردازش داده‌ها، خوشه‌بندی به کمک الگوریتم تجزیه نامنفی ماتریس و دسته‌بندی به کمک ماشین بردار پشتیبان. لازم به ذکر است که سه رویکرد آزمایشی اجرای

متوالی فرایند خوشه‌بندی و غنی‌سازی کلمات توقف دامنه خاص، بهره‌گیری از الگوی کلیدواژه‌های تخصصی، و استفاده از بخش‌های مهم متن جهت استخراج کلیدواژه‌های موضوعی در مرحله پیش‌پردازش و خوشه‌بندی اعمال می‌شود و به‌منظور بررسی اثربخشی رویکردهای پیشنهادی، در بخش دسته‌بندی عملکرد این سه رویکرد به کمک معیارهای ارزیابی صحت، دقت، بازخوانی، امتیاز F و افت همینگ[1] تحلیل شد. کلیه مراحل تحقیق در شکل ۱، قابل مشاهده است.

جامعه آماری پژوهش حاضر شامل متن کامل ۱۰۰۰ عنوان کتاب الکترونیک در حوزه علم اطلاعات و دانش‌شناسی است. به‌منظور جمع‌آوری داده‌های پژوهش، ابتدا با بهره‌گیری از اوپک کتابخانه کنگره، اطلاعات کتابشناختی ۱۰۰۰ عنوان از کتاب‌های حوزه علم اطلاعات و دانش‌شناسی مشخص شد و سپس، متن کامل آن‌ها به زبان انگلیسی و در قالب الکترونیک تهیه گردید. بعد از گردآوری داده‌ها، فرایند ورود و خوانش آن‌ها و تبدیل متن به رشته‌ای از کلمات و تشکیل پیکره متنی برای تحلیل‌های بعدی انجام شد.

کلیه مراحل کار در این پژوهش با اجرای کتابخانه‌های زبان برنامه‌نویسی «پایتون» مثل کتابخانه‌های «اسکیت لرن»[2]، «جنسیم»[3] و «ان‌ال‌تی کِی»[4] در محیط «گوگل کلب»[5] (آزمایشگاه گوگل) اجرا شده است.

1. Hamming loss
2. Scikit learn
3. Gensim
4. NLTK
5. Google colaboratory



شکل ۱. چارچوب کلی روش تحقیق

۳-۱. مراحل اجرای روش پژوهش

گام اول: پیش‌پردازش

پس از جمع آوری داده‌های مورد نیاز، لازم است عملیات پیش‌پردازش به‌منظور آماده‌سازی داده‌های خام جهت ورود به مرحله متن کاوی و مدل‌سازی موضوعی با الگوریتم تجزیه نامنفی ماتریس انجام گیرد. اقداماتی که در این مرحله انجام می‌شود، شامل تکه‌تکه کردن متن به اجزای آن (توکن‌سازی)، ریشه‌یابی لغات و حذف کلمات توقف برای محاسبه و تعیین وزن کلمات در یک متن است. همچنین، بعضی از کلمات ممکن است در غالب متن‌ها و همچنین جامعه آماری پژوهش حاضر به فراوانی یافت شود که به کلمات مشترک یا رایج مشهور است و برای شناسایی محتوای متن مفید نیستند و بنابراین باید مانند کلمات توقف حذف شوند. مقدار TF-IDF (بسامد معکوس) معیار تصمیم‌گیری در این رابطه است. در نهایت، کلمات اصلی یا کلیدی باقی می‌مانند که برای ورودی تجزیه‌وتحلیل مورد استفاده قرار می‌گیرند (Wang, Zhang & Klabjan 2020). سپس، ماتریس کلمه–متن تولید می‌شود که هر ردیف در آن به یک متن مربوط است و هر ستون به یک کلمه اختصاص دارد و سلول‌های ماتریس نیز میزان اهمیت کلمات در جملات را نمایش می‌دهد که در این پژوهش بر اساس روش TF-IDF به‌دست آمده است.

گام دوم: خوشه‌بندی

در این مرحله الگوریتم مدل‌سازی موضوعی مورد نظر (تجزیه نامنفی ماتریس) بر روی ماتریس کلمه-متن اعمال می‌شود تا موضوعات و در نتیجه، خوشه‌ها مشخص گردند. یافتن تعداد مناسب خوشه‌ها یکی از چالش‌های بسیار مهم در این زمینه است. بدین منظور از میان سه معیار رایج ارزیابی کیفیت تعداد خوشه‌های موضوعی، یعنی «انسجام»[1]، «احتمال»[2] و «عدم یکدستی»[3]، در این پژوهش از معیار انسجام استفاده شده است؛ زیرا معیار انسجام موضوعات، ابزار مناسبی است که بر اساس میزان تفسیرپذیری موضوعات از دیدگاه کاربر بنا شده است (Newman & et al. 2010). روش‌های ارزیابی انسجام موضوعات کمک می‌کنند که موضوعات خوب از بد تمیز داده شوند و نشان می‌دهد که موضوعات به‌دست‌آمده تا چه حدی منسجم، قابل فهم و معنادار هستند و کلماتی که در یک خوشه موضوعی قرار می‌گیرند تا چه میزان به هم مرتبط هستند. هرچه موضوع منسجم‌تر باشد، برای کاربر قابل فهم‌تر و مفیدتر است. روش C_V یکی از بهترین روش‌های ارزیابی خودکار انسجام معرفی شده که ارتباط قوی با ارزیابی کاربران دارد و در این پژوهش استفاده شده است. محاسبه C_V طی چهار مرحله صورت می‌گیرد: ۱) قطعه‌بندی داده‌ها به جفت کلمات، ۲) محاسبه احتمالات کلمه یا جفت کلمات، ۳) محاسبه کمی شدت تأیید یک مجموعه از کلمات توسط یک مجموعه دیگر از کلمات، و ۴) جمع نمره‌های تأیید در یک نمره انسجام کلی (Syed & Spruit 2017).

سه رویکرد مورد تحلیل در پژوهش حاضر (اجرای متوالی فرایند خوشه‌بندی و غنی‌سازی کلمات توقف دامنه خاص، بهره‌گیری از الگوی کلیدواژه‌های تخصصی و استفاده از بخش‌های مهم متن) در این مرحله بررسی می‌شود. هدف از اجرای این سه رویکرد مختلف، تلاش برای به‌دست آوردن کلیدواژه‌های موضوعی معنادارتر و مرتبط‌تر و قضاوت درباره مرتبط بودن کلیدواژه‌ها و خوشه‌های موضوعی هر کدام. این سه رویکرد به‌شرح زیر است:

رویکرد اول: همان‌طور که در مقدمه اشاره شد، برای بهبود کیفیت موضوعات، محققان به‌طور معمول، به پردازش‌های پیچیده قبل و بعد از عملیات مدل‌سازی، مثل ایجاد لیست کلمات توقف و آموزش مجدد مدل‌های تجزیه نامنفی ماتریس بدون این کلمات می‌پردازند. به‌طور کلی، کلمات توقف به دو دسته تقسیم می‌شوند. کلمات توقف متعارف و معمول مثل (\the, \and) و کلمات توقف حوزه خاص مثل

1. coherence
2. likelihood
3. perplexity

(\child, \son) در یک پیکره متنی[1] مربوط به بچه‌ها). کلمات توقف متعارف و معمول اغلب با ارجاع به لیست‌های استاندارد از پیش‌آماده حذف می‌شود. با وجود این، کلمات توقف حوزه خاص امری جزئی و بدیهی نیست و محقق با اجرای مکرر خوشه‌بندی و بررسی تک‌تک خوشه‌ها، این لیست را تهیه می‌کند. بنابراین، در این پژوهش نیز به‌منظور تهیه کلمات توقف حوزه خاص در این حوزه موضوعی (اطلاعات و دانش‌شناسی)، بعد از اولین اجرای الگوریتم خوشه‌بندی، به بررسی تک‌تک خوشه‌ها پرداخته شد و کلمات فاقد بار معنایی از آن‌ها استخراج و به لیست کلمات توقف اضافه و مجدداً اجرای مراحل مدل‌سازی انجام گردید. هدف از این کار، افزون بر پاک‌سازی خوشه‌های موضوعی از کلمات نامرتبط و بدون معنا در هر اجرا و دستیابی به خوشه‌های موضوعی مرتبط‌تر و معنادارتر، غنی‌سازی کلمات توقف حوزه خاص (در این پژوهش علم اطلاعات و دانش‌شناسی) است.

رویکرد دوم: از آنجا که کتاب‌ها از نظر طول متن بسیار بزرگ‌تر از سایر قالب‌های اطلاعاتی مکتوب هستند، بنابراین یافتن کلمات کلیدی که نمایانگر دقیق محتوا و درون‌مایه متن باشد، پیچیده‌تر خواهد بود. با به‌کارگیری تکنیک پاک‌سازی کلمات به کمک وزن‌دهی و حذف کلمات توقف، ممکن است هنوز کلیدواژه‌هایی در متن باشند که به‌رغم معنادار بودن، نقشی در بازنمایی محتوا و درون‌مایه و تحلیل موضوع اصلی کتاب نداشته باشند. بنابراین، در رویکرد دوم سعی شد از یک الگوی کلیدواژه‌های تخصصی حوزه علم اطلاعات و دانش‌شناسی به‌منظور استخراج کلمات کلیدی استفاده شود. به همین منظور، فهرست جامعی از کلیدواژه‌های موضوعی حوزه علم اطلاعات و دانش‌شناسی از پایگاه چکیده‌های علوم و فناوری اطلاعات[2] و «لیزا»[3] استخراج گردید و از آن به‌عنوان الگوی استخراج کلمات کلیدی داده‌های مورد آزمایش استفاده شد. در توضیح چگونگی اجرا در این مسیر می‌توان گفت که درست عکس عملکرد کلمات توقف، از این لیستِ کلیدواژه‌های تخصصی به‌منظور بازیابی کلمات کاملاً مرتبط و وزن‌دار بهره گرفته شد.

رویکرد سوم: در این روش به نشانه‌گذاری بخش‌های مهم تک‌تک مدارک پرداخته شد. منظور از بخش‌های مهم، عنوان، فهرست مندرجات، مقدمه، پیشگفتار و پارگراف ابتدایی هر فصل و سپس نمایه انتهای کتاب بود که از نظر پژوهشگران و متخصصان علم اطلاعات جزء مهم‌ترین بخش‌های کتاب جهت تصمیم‌گیری راجع به محتوا و درون‌مایه کتاب محسوب می‌شود (Miner et al. 2012, 959-966). این

1. corpus
2. Information Science & Technology Abstract (ISTA)
3. Library and Information Science Abstract (LISA)

بخش‌ها بیان موجز محتوای کتاب‌ها و راهنمایی برای درک مطالب مندرج در متن است که با حداقل واژگان و اصطلاحات بیشترین بازنمایی محتوای متن را در خود دارد (Pokorny 2018). سپس، همین بخش‌ها به مدل داده شد و استخراج کلمات و خوشه‌بندی موضوعی تنها از همین بخش‌ها انجام گردید.

در ادامهٔ گام دوم بعد از خوشه‌بندی، فرایند برچسب‌گذاری خوشه‌های موضوعی انجام گرفت. موضوعاتی که از مدل موضوعی تجزیه نامنفی ماتریس حاصل می‌شود، به‌صورت مجموعه‌ای از کلمات هستند و برچسب موضوع برای آن‌ها مشخص نیست. برای مثال، در خوشه‌های موضوعی جدول ۱، برچسبی برای خوشه‌ها تعیین نشده و در واقع، الگوریتم تجزیه نامنفی ماتریس توانایی تعیین برچسب برای دسته‌بند را ندارد؛ اما این مسئله برای اجرای گام سوم تحقیق، یعنی دسته‌بندی یا طبقه‌بندی مشکل‌زا بود؛ چون هدف نهایی در کشف و استخراج موضوع کمک به دسته‌بندی اسناد است و در این عمل بهتر است یک برچسب به‌عنوان خروجی برای هر خوشه تعیین شود. به این منظور، در این مرحله برای هر یک از خوشه‌های موضوعی استخراجی یک برچسب به‌عنوان معرف موضوع توسط متخصصان موضوعی انتخاب گردید.

گام سوم: دسته‌بندی

هر سند یا مدرک صرفاً یک موضوع ندارد، بلکه ترکیبی از چند موضوع است. برای تعیین موضوعات یک سند وقتی اطلاعاتی از آن در دست نیست، باید بر اساس تکنیک یادگیری ماشین، سیستم به‌طور خودکار، کار دسته‌بندی کلمات و تعیین موضوع را انجام دهد. در مجموعه‌های بزرگ اطلاعات که برچسب اولیه وجود ندارد، می‌توان با استفاده از روش خوشه‌بندی، این برچسب اولیه را فراهم و در اختیار دسته‌بند قرار داد. در پژوهش حاضر، ماشین بردار پشتیبان به‌عنوان الگوریتم دسته‌بندی در مدل پیشنهادی استفاده شد. از موضوعات استخراج‌شده توسط مدل موضوعی «تجزیه نامنفی ماتریس» به‌عنوان مجموعه آموزش اولیه استفاده گردید. به این ترتیب که هر مجموعه کلمه‌های استخراجی از تجزیه نامنفی ماتریس که معرف یک موضوع است، به همراه وزن‌های اختصاصی به هر کلمه به‌عنوان یک آموزش اولیه به ماشین بردار پشتیبانی داده شد و سپس، عمل آموزش و یادگیری با نظارت به‌وسیله این مجموعه کلمه‌ها شکل گرفت. در ادامه، هر متن بر مبنای شباهت به مجموعه کلمه‌های موضوعی داخل خوشه‌ها به یک موضوع نسبت داده شد و از برچسب اختصاص‌یافته برای آن خوشه موضوعی به‌عنوان برچسب موضوعی متن استفاده گردید. لازم به ذکر است که مدل موضوعی هر سه رویکرد در این مرحله به‌صورت جداگانه مورد استفاده واقع شد.

در پایان، به‌منظور بررسی عملکرد دسته‌بندی با توجه به سه رویکرد آزمایشی استفاده‌شده در این تحقیق از معیارهای ارزیابی زیر استفاده شد:

صحت[1]: معیار صحت برابر است با تعداد مواردی که درست پیش‌بینی شده «که آن را مثبت صحیح[2] می‌نامند)، تقسیم بر تعداد کل پیش‌بینی‌هایی که انجام شده است.

اُفت همینگ: اُفت همینگ درصد برچسب‌هایی را که به‌درستی دسته‌بندی نشده‌اند، محاسبه می‌کند؛ یعنی نمونه‌ای که به برچسب نادرست نسبت داده شده یا برچسب صحیحی که به یک نمونه متعلق بوده، اما پیش‌بینی نشده باشد.

دقت[3]: درصد متونی را که دسته‌بندی کننده به‌درستی برچسب‌گذاری نموده از تعداد کل متن‌هایی که برای هر موضوع پیش‌بینی کرده است، محاسبه می‌کند.

بازخوانی[4]: این معیار درصد متون پیش‌بینی شده مدل برای هر موضوع از کل تعداد متن‌هایی را که باید برای آن موضوع پیش‌بینی کرد، مشخص می‌کند و در واقع، تعداد پیش‌بینی‌های موارد مثبت را که از تمام مثال‌های مثبت در مجموعه داده انجام شده است، تعیین می‌کند.

بعد از جمع‌آوری داده‌ها، با استفاده از کتابخانه «پانداس»[5] و «نامپی»[6] و با استفاده از زبان برنامه‌نویسی «پایتون» در محیط «گوگل کلب»، فراخوانش داده‌ها انجام شد. برای تکمیل مرحله پیش‌پردازش، افزون بر حذف لغات توقف مثل حروف ربط و اعداد و علائم، کلمات و اصطلاحات و لغات دیگری که با توجه به حوزه موضوعی داده‌های مورد تحقیق اهمیتی نداشته و فاقد بار معنایی بوده‌اند نیز حذف شد.

از آنجا که جامعه پژوهش، یعنی کتاب‌های الکترونیک، دارای حجم قابل توجهی از کلمات و عبارات بود و با یک مرتبه پالایش کلمات از لغات و اصطلاحات فاقد بار معنایی به خوشه‌های موضوعی حاوی کلمات کاملاً مرتبط و معنادار منجر نمی‌شد، این بود که با اِعمال رویکرد اول یعنی اجرای متوالی (سه مرتبه) در بخش خوشه‌بندی موضوعی، هر مرتبه خوشه‌های موضوعی حاصل مورد مشاهده و بررسی قرار گرفته و کلمات بی‌معنا و نامرتبط از داخل خوشه‌ها حذف و به لیست کلمات ایستا جهت حذف در اجرای دوم و سوم الگوریتم خوشه‌بندی موضوعی اضافه شد. لازم به ذکر است که این عمل صرفاً در رویکرد اول که تمام ویژگی‌های استخراج‌شده از متن –به‌جز کلمات توقف عمومی حذف‌شده در مرحله پیش‌پردازش– در فرایند خوشه‌بندی لحاظ گردیده و به‌عبارتی با ابعاد بزرگ‌تری از ویژگی در مقایسه با

1. accuracy
2. true positive
3. precision
4. recall
5. Pandas
6. Numpy

دو رویکرد بعدی مواجه هستیم، استفاده شده است و در رویکرد دوم و سوم به لحاظ استراتژی به‌کاررفته در کاهش ابعاد، لازم به اعمال این روش نبود. با این توضیح که در رویکرد دوم به‌دلیل استفاده از الگوی کلیدواژه‌های تخصصی و در رویکرد سوم به‌دلیل تمرکز صرف بر بخش‌های مهم متن، کاهش ابعاد و ویژگی‌ها به‌صورت مؤثری انجام گرفت و نیاز به اجرای متوالی خوشه‌بندی و پاک‌سازی خوشه‌ها از کلمات و مفاهیم نامرتبط نبود و فقط کلمات توقف حذف شدند. همچنین با در نظر گرفتن معیار انسجام در هر اجرا تعداد خوشه‌ها بر اساس این معیار و با توجه به تعداد کلیدواژه‌های موضوعی متفاوت بود (جدول ۱). نمونه‌ای از نتایج خوشه‌بندی متوالی و حذف کلمات فاقد بار معنایی برای سه خوشه موضوعی در همین جدول آورده شده است.

جدول ۱. نمونه‌ای از خوشه‌های موضوعی حاصل از سه اجرای متوالی خوشه‌بندی در رویکرد اول

نمونه سه خوشه موضوعی حاصل از اجرای متوالی الگوریتم خوشه‌بندی موضوعی	تعداد خوشه	اجراها
topic 1: Networking- sit, trailer, glog, quikqr, profi, poem-refl-ect, jaycut, animoto topic 2: child, booktalk, eect, prot, adult, book talk, overdu, issn, patron, privacy topic 3: signic, eect, prot, inuenti, transnat, tempor, pronounc, dirent, Interpol, briey	۲۴۰	اجرای اول
Topic 1: smsa, netbook, openoffic, hardback, district, couthership, multiset Topic 2: speech, death, adventur, comic, disput, refus,prose, leafpag, watercolor, eiusdem Topic 3: Zealand, placeworkspac, schirmach, subdivis, floor spac, derek, datanet	۱۸۰	اجرای دوم
Topic 1: ontology, weed, librarianship, staff, patron, developn, undergradu, organis, digit, librari Topic 2: cart, printer, booktalk, disposit, watermark, grammar, martial, commonplac, heckler, magazin Topic 3: classif, cascadelink, trfn, ware, subclass, parchment, district, program, acquisit, center	۱۲۰	اجرای سوم

در رویکرد دوم به‌منظور ایجاد خوشه‌های موضوعی حاوی کلمات کاملاً معنادار و مرتبط با حوزه علم اطلاعات و دانش‌شناسی، از کلمات تخصصی پایگاه اطلاعاتی علوم و فناوری اطلاعات (لیستا) و (لیزا) به‌عنوان الگوی استخراج کلیدواژه‌های موضوعی داده‌ها استفاده گردید. به همین منظور لیست جامعی از کلیدواژه‌ها و عبارت‌های تخصصی این پایگاه‌ها جمع‌آوری شد که تعداد آن‌ها ۳۷۱۱ کلمه و عبارت واحد بود. همین کلیدواژه مبنای استخراج کلمات موجود در داده‌های مورد تحقیق قرار گرفت و سپس الگوریتم خوشه‌بندی بر روی همین کلیدواژه‌ها صورت گرفت. تعداد خوشه‌های موضوعی حاصل از این رویکرد با در نظر گرفتن معیار انسجام (شکل ۲) تعداد ۸۰ خوشه موضوعی بود.

در رویکرد سوم تنها با تکیه بر بخش‌های مهم متن که با نشانه‌گذاری از سایر بخش‌های متن تفکیک شده بود، استخراج کلیدواژه‌های موضوعی تنها از همین بخش‌ها صورت گرفت. تعداد خوشه‌های موضوعی حاصل از این رویکرد با در نظر گرفتن انسجام (شکل ۲) تعداد ۹۰ خوشه موضوعی بود. لازم به ذکر است که برای تعیین تعداد خوشه‌های موضوعی در هر رویکرد، ابتدا با نظر متخصصان موضوعی و بر اساس تعداد رده‌های موضوعی در رده‌بندی کنگره، تعداد ۱۲۰ موضوع به‌عنوان موضوع‌های اصلی حوزه علم اطلاعات و دانش‌شناسی تعیین شد. در مرحله بعد به‌منظور تأیید تعداد حوزه موضوعی مذکور بر اساس داده‌های گردآوری‌شده، معیار انسجام برای تعیین بهترین تعداد موضوع‌ها مورد استفاده قرار گرفت. به همین منظور، تعداد دسته‌های ۵۰ تا ۱۴۰تایی در الگوریتم مذکور برای هر رویکرد استفاده شد. همان‌طور که در شکل ۲، مشخص است، مقدار انسجام خوشه‌های موضوعی در رویکرد اول برای ۱۲۰ خوشه، در رویکرد دوم برای ۸۰ خوشه و در رویکرد سوم برای ۹۰ خوشه با بالاترین مقدار برآورد شد.

شکل ۲. نمودارهای مربوط به محاسبه معیار انسجام رویکردها

در جدول ۲، نمونه‌ای از ۳ خوشه موضوعی نزدیک به هم و ۱۰ کلمه اول موجود در هر خوشه را که حاصل اجرای مدل با سه رویکرد آزمایشی مد نظر در این پژوهش بوده، نمایش داده شد.

جدول ۲. سه نمونه از خوشه‌های موضوعی مشابه و کلمات موجود در آن بر اساس سه رویکرد آزمایشی

برچسب موضوعی مرتبط با کلمات درون خوشه	رویکردها	نمونه‌ای از خوشه‌های موضوعی نزدیک به هم در سه رویکرد
Bookbinding technique	۱	Books, yarn, glow, cotton, wrappers, oriental binding, designed book, sewing support, modern machine, packed, flexible, raised cord
	۲	book covers, book size, Bookbinding machinery, bookbinding repairing, bookbinding, books, books binders, book spine, publishers binding paperback publishing, book jackets
	۳	Book, Sewing, hardcover, section binding, Coptic stitch, spiral bind, case bound, soft cover, glow, inner pages
Printing history	۱	Technology spread, woodblock, books, colonial publisher, Europe machine, text printing, publishing date, china originated, civilization, certify
	۲	Preprint, paper making, paperbacks, book size, paperback publishing, publishing history, publishers binding
	۳	History, print, book publish, press, global spread, screen printing, mass producing, movable type, Gutenberg, lithograph
Book selling	۱	Wholesale, bulk,barcode, new books, collectible, popular book, cash, jacket, standard book, online shop, ISBN
	۲	book advertising, book size, book sellers, book collecting, collector book, book donation, book jacket, book list, book industry
	۳	Book selling, Amazon, common books, book store, list, section, collections, children books, ebooks, demanded books, rare list, book title

همان‌طور که پیش‌تر ذکر شد، هدف از اجرای متوالی خوشه‌بندی و غنی‌سازی کلمات توقف حوزه خاص در رویکرد اول کاهش ابعاد ویژگی و رسیدن به خوشه‌های موضوعی معنادار بود. اگرچه هدف مذکور تا حدودی حاصل شد، اما با وجود اصلاحات پی‌درپی در خوشه‌های موضوعی به‌منظور حذف کلمات بی‌ربط، به‌دلیل حجم زیاد کلمات، بعد از اجرای متوالی، کلمات بی‌ربط یا فاقد معنا هنوز در خوشه‌ها مشاهده می‌شود؛ برای مثال، کلمه "flexible" در خوشه موضوعی اول یا "packed" در خوشه موضوعی ۳ که در جدول ۲، قابل مشاهده است.

در رویکرد دوم، به‌دلیل محدود کردن کلمات به کلیدواژه‌های تخصصی از پیش انتخاب‌شده، کلمات کاملاً معنادار و مفهومی در خوشه‌ها مشاهده می‌شود، اما به‌دلیل این محدودیت اعمال‌شده در زمینه انتخاب کلمات، نه‌تنها تعداد خوشه‌های موضوعی- در قیاس با رویکرد اول -کاهش یافته، بلکه

به نظر می‌رسد کلمات به‌رغم معنادار بودن، در خوشه‌های مختلف یکسان است و تنوع کلمات به‌طور قابل توجهی کاهش یافته است و گمان می‌رود که بعضی از کلمات تخصصی و معنادار در متن به‌دلیل محدودیت ایجادشده حذف گردیده است. به‌عنوان مثال، در هر سه نمونه خوشه موضوعی حاصل از رویکرد دوم کلیدواژه «booksize»یا «book list» قابل مشاهده است. در حالی که خوشه‌های موضوعی رویکرد سوم از انسجام معنایی خوبی برخوردار است و کلماتی که بی‌معنا یا فاقد بار اطلاعاتی هستند، مشاهده‌نمی‌شود.

در این رویکرد، برای استخراج کلیدواژه‌های موضوعی تنها بخش‌های مهم متن همچون عنوان، مقدمه، فهرست مندرجات، بخش‌های ابتدایی هر فصل و نمایه استفاده گردید و کلمات به‌کاررفته در این بخش‌ها عمدتاً برخاسته از بدنه اصلی متن و محتوای کتاب است و می‌توان گفت که در مقایسه با متن کامل کتاب میزان کلمات ربط، قید، افعال و کلمات توقف و بدون معنا و نامرتبط با موضوع اصلی در این بخش‌ها کمتر دیده می‌شود که با اعمال پیش‌پردازش، پاک‌سازی به‌طور مؤثرتری صورت گرفته و در نتیجه، خوشه‌های موضوعی حاصل از این رویکرد انسجام معنایی بهتری در بر داشت؛ به‌طوری که با مشاهده هر خوشه موضوعی، معنای واضحی به ذهن متبادر شده که همین امر برچسب‌گذاری موضوعی هر خوشه را نیز راحت‌تر می‌کند.

به‌منظور ارزیابی کیفیت خوشه‌های موضوعی به‌دست‌آمده از مرحله مدل‌سازی موضوعی، در بخش دوم آزمایش و در قالب دسته‌بندی بر اساس مدل یادگیری، وضعیت خوشه‌های موضوعی نمونه‌های جدید داده‌ها بر اساس برچسب موضوعی نمونه‌های قبلی پیش‌بینی و تعیین شد. در این مرحله به‌منظور تعیین داده‌های آموزش و آزمون از روش K-fold استفاده شد. در این روش مجموعه داده آزمون به‌صورت تصادفی به K گروه افراز می‌شود. سپس، K-1 گروه از داده‌ها برای آموزش و یک گروه برای آزمون الگوریتم به‌کار می‌روند. تعداد دفعات اجرای این فرایند K است و K مدل با تخمین‌های مختلف به‌دست می‌آید و در نهایت، از کارایی مدل‌ها میانگین‌گیری می‌شود. در این پژوهش نیز داده‌ها به پنج قسمت تقسیم گردید و در هر مرتبه اجرا (پنج مرتبه) یک پنجم از آن‌ها (۲۰ درصد) مورد آزمون قرار گرفت و بقیه داده‌ها به‌عنوان داده‌های آموزش انتخاب شد و سپس، از نتایج، میانگین گرفته شد. کیفیت طبقه‌بندیِ صورت‌گرفته بر اساس سه معیار دقت، بازخوانی، امتیاز F 1، و اُفت همینگ به‌شرح جدول ۳، بود.

جدول ۳. نتیجه ارزیابی کیفیت دسته‌بندی بر اساس معیارهای دقت و بازخوانی، امتیاز F، صحت و اُفت همینگ در سه رویکرد آزمایشی

رویکردها	اُفت همینگ	صحت	امتیاز F1	بازخوانی	دقت
۱	۰/۰۲۲۹	۰/۲۱۹	۰/۷۸	۰/۷۷	۰/۸۰
۲	۰/۰۲۴۶	۰/۲۲۶	۰/۷۴	۰/۷۴	۰/۷۵
۳	۰/۰۲۰	۰/۲۵	۰/۸۲	۰/۷۸	۰/۸۷

بر اساس نتایج ارزیابی در جدول ۳، مشاهده می‌شود که درصد خطا یا اُفت همینگ (۰/۰۲۰) و به‌عبارتی، میزان خطا در دسته‌بندی صحیح متون آزمایشی در رویکرد سوم به مراتب از دو رویکرد دیگر کمتر است. همچنین، امتیاز F1 (۰/۸۲) که میانگین دو معیار دقت (۰/۸۷) و بازخوانی (۰/۷۸) و بازتابی از عملکرد درست فرایند دسته‌بندی در برچسب‌گذاری موضوعی متون است، در رویکرد سوم بهتر از نتایج دو رویکرد دیگر است.

٤. بحث و نتیجه‌گیری

با توجه به خوشه‌های موضوعی حاصل از اعمال رویکردها و معیارهای ارزیابی در فرایند دسته‌بندی موضوعی متون آزمایشی، این نتیجه حاصل شد که رویکرد سوم یعنی بهره‌گیری از بخش‌های مهم متن در استخراج کلیدواژه‌های موضوعی، در کشف و استخراج موضوع نتیجه بهتری دربرداشته و دسته‌بندی بهتری از کتاب‌های الکترونیک این حوزه ارائه داد. در توجیه این نتایج می‌توان گفت که اگرچه در رویکرد اول سعی شده بود کلمات غیرمعنایی استخراج و به لیست کلمات توقف اضافه شود تا خوشه‌های موضوعی معنادارتری به‌دست آید، اما همچنان در بعضی از خوشه‌های موضوعی حاصل از این رویکرد کلمات غیرتخصصی و یا فاقد بارمعنایی مشاهده شد. افزون بر این، همسو با نتایج پژوهش Wang, Zhang & klabjan (2020) به نظر می‌رسد که تلاش برای غنی‌سازی لیست کلمات توقف دامنه خاص با اجرای مکرر الگوریتم خوشه‌بندی احتمال، سوگیری و یا تعصب اجراکننده را به همراه خواهد داشت.

در رویکرد دوم، برای رفع نواقص نتایج رویکرد اول، فهرستی جامع از کلیدواژه‌های موضوعی حوزه علم اطلاعات و دانش‌شناسی ایجاد و سپس، به‌عنوان الگوی کشف و استخراج موضوع در داده‌های تحقیق مورد استفاده قرار گرفت. نتایج حاصل از بررسی خوشه‌های موضوعی تشکیل‌شده در این رویکرد نشان داد که اگرچه با این روش، خوشه‌های موضوعی حاوی کلمات دارای معنای مرتبط و تخصصی هستند، اما به‌دلیل ایجاد محدودیت در بازیابی کلیدواژه‌های از پیش تعریف‌شده احتمال می‌رود، بازنمایی مضمون و محتوای داده‌ها به‌درستی انجام نشده باشد و پاره‌ای از کلمات و موضوعات کلیدی متن که

منطق با الگوی کلیدواژه‌ای نبوده، حذف شده باشد. نتایج پژوهش Elhadad, Badran & Salama (2017) نیز نشان داد که روش‌های مبتنی بر الگوی کلیدواژه‌های تخصصی یا هستی‌شناسی‌ها در مجموعه‌ای از متون گسترده و بزرگ ناکارآمد هستند، زیرا بسیاری از کلیدواژه‌های موجود در متن به لحاظ بهره‌گیری از الگوی مشخص حذف و نادیده گرفته می‌شود. بر این اساس، کاربرد الگوی کلیدواژه‌های تخصصی و یا هستی‌شناسی‌ها در فرایند کاهش بعد و استخراج مفاهیم مستتر در متون وبی گوناگون و متنوع با حجم بزرگ را باید با احتیاط بیشتری مورد توجه قرار داد.

در حالی که در رویکرد سوم با تمرکز بر بخش‌های اصلی متن که نماینده محتوای اصلی و درون‌مایه متن است، خوشه‌های موضوعی معنادارتری به‌دست آمد. در این رویکرد برای کاهش ابعاد ویژگی و غلبه بر حجم گسترده متن، تنها بخش‌های مهم کتاب‌های الکترونیک مانند عنوان، فهرست مندرجات، پیشگفتار، مقدمه، پارگراف اول هر بخش و نمایه مورد استفاده قرار گرفت. این بخش‌ها خلاصه‌ای از کل مطالب مهم و اصلی در متن بوده و بنابراین، می‌تواند شامل تمام موضوعات اصلی و پرداخته‌شده در متن باشد. گفته می‌شود که این بخش‌ها بیان موجز محتوای کتاب‌ها و راهنمایی برای یافتن مطالب در صفحات مربوط است که با حداقل واژگان و اصطلاحات بیشترین بازنمایی محتوا را در خود دارد. افزون بر این، این بخش‌ها 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. به‌علاوه، دسته‌بندی انجام‌گرفته به کمک خوشه‌های موضوعی حاصل از این رویکرد نیز نتایج بهتری دربرداشت که گواه آن امتیاز کسب‌شده حاصل از محاسبه معیارهای صحت، دقت، بازیابی، امتیاز F و اُفت همینگ بود.

رویکرد پیشنهادی نه‌تنها در کشف و استخراج کلیدواژه‌های موضوعی متون حجیم مؤثر خواهد بود، بلکه به دسته‌بندی کتاب‌های الکترونیک حوزه علم اطلاعات و دانش‌شناسی نیز با گستردگی زیادی کمک خواهد کرد؛ زیرا در به‌کارگیری روش‌های دسته‌بندی خودکار، در آغاز مجموعه داده برچسب خورده اولیه برای آموزش دسته‌بندها وجود ندارد و همین مسئله سبب دشواری و کاهش دقت دسته‌بندی در منابع الکترونیک می‌شود. رویکردهای پیشنهادی و مدل حاصل از آن در این پژوهش، با بهره‌گیری از دانش موجود در موضوعات استخراجی از تجزیه نامنفی ماتریس، این مشکل را تا حد زیادی برطرف می‌سازد. همچنین، در زمان و هزینه تجزیه‌وتحلیل داده‌های متنی صرفه‌جویی شده و اطلاعات دسترس‌پذیر می‌شود. توسعه پژوهش‌هایی از این دست می‌تواند تسهیل‌کننده رفع چالش‌هایی مانند تعدد ویژگی‌های متنی یا خلاصه‌سازی داده‌های حجیم در راستای کشف و استخراج موضوعات آشکار و پنهان در متون عظیم وب‌پایه باشد.

این پژوهش به‌عنوان نمونه‌ای کوچک به‌منظور به تصویر کشیدن کارکردهای متن‌کاوی در کشف

و استخراج خودکار موضوعات اصلی متون منتشرشده در محیط وب طراحی گردید. بر این پایه، با توجه به قابلیت‌های انجام و تکرار پژوهش در محیط‌های مشابه پیشنهاد می‌گردد که از روش‌های به‌کاررفته در این پژوهش در کشف و استخراج مفاهیم و موضوعات اصلی پرداخته‌شده در منابع الکترونیک سایر حوزه‌های موضوعی استفاده شود و تأثیر آن در سازماندهی و بازیابی اطلاعات به‌طور گسترده‌تری مورد بررسی قرار گیرد. این امر نه‌تنها به مقایسه و بازآزمون رویکردهای پیشنهادی پژوهش حاضر منجر می‌شود، بلکه زمینه غنای ادبیات نظری و پژوهش‌های عملیاتی بیشتر در این حوزه را با رویکردی بین رشته‌ای در پی خواهد داشت.

فهرست منابع

انایی فریمانی، سعیده، حمید طباطبائی، و مجتبی کفاشان کاخکی. ۱۳۹۸. جستاری بر فرایند سازماندهی و بازیابی متون وبی مبتنی بر تجمیع مفاهیم معنایی. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۴ (۴): ۱۸۷۹–۱۹۰۴. https://jipm.irandoc.ac.ir/article-1-4151-fa.html (دسترسی در ۱۴۰۱/۲/۱۳)

باغ محمد، مریم، علی منصوری، و مهرداد چشمه‌سهرابی. ۱۳۹۹. بررسی توسعه و روند موضوعی حوزه علم اطلاعات و دانش‌شناسی بر اساس مدل موضوعی LDA. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۶ (۲): ۲۹۷–۳۲۸. https://jipm.irandoc.ac.ir/article-1-4480-fa.html (دسترسی در ۱۴۰۱/۳/۵)

References

Allahyari, M., P. Pouriyeh, M. Assefi, A. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. e-prints. arXiv: 1707.02919.

Ardimento, P., M. Bilancia, S. Monopoli. 2016. Predicting bug-fix time: Using standard versus topic-based text categorization techniques. In International Conference on Discovery Science. Cham, Germany: Springer. 167–182. doi:10.1007/978-3-319-46307-0_11

Basaldella, M., E. Antolli, G. Serra, & C. Tasso. 2018. Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction. Springer International Publishing. doi:10.1007/978-3-319-73165-0_18

Beliga S., A. Meštrović, & S. Martinčić-Ipšić. 2015. An overview of graph-based keyword extraction methods and approaches. *J Inform Organ Sci.* 39 (1):1–20. https://www.researchgate.net/publication/280092953_An_Overview_of_Graph- Based_Keyword_Extraction_Methods_and_Approaches (Accessed Jul. 20, 2021)

Berger A, & J. Lafferty. 2017. Information retrieval as statistical translation. ACM SIGIR Forum. 51 (2): 219–26. https://doi.org/10.1145/3130348.3130371.

Blei, D. M., A. Y. Ng, & M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993-1022.

Casalino, G., C. Castiello, N. Del Buono, & C. Mencar. 2018. A framework for intelligent Twitter data analysis with non-negative matrix factorization. *Int. J. Web Inf. Syst*. 14: 334–356. https://doi.org/10.1108/IJWIS-11-2017-0081

Chen, Y., & S. Li. 2016. Using latent Dirichlet allocation to improve text classification performance of support vector machine. Paper presented at the 2016 IEEE Congress on Evolutionary computation (CEC). Vancouver, BC, Canada. doi:10.1109/CEC.2016.7743935.

(Accessed June 2, 2022)

Chien, J.-T., C.-H. Lee, & Z.-H. Tan. 2018. Latent Dirichlet mixture model. eurocomputing12-22 :278 .. doi:https://doi.org/10.1016/j.neucom.2017.08.029

Cohen, J. D. 1995. Highlights: Language-and domain-independent automatic indexing terms for abstracting. *J. Am. Soc. Inf. Sci.* 46: 162–174.

Choi, Y., I. Hsieh-Yee, & B. Kules. 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

De Nart, D., D. Degl'Innocenti, A. Pavan, M. Basaldella, & C. Tasso. 2015. Modelling the User Modelling Community (and Other Communities as Well). In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds) User Modeling, Adaptation and Personalization. UMAP 2015. Lecture Notes in Computer Science (), vol 9146. Springer, Cham. https://doi.org/10.1007/978-3-319-20267-9_31

Dennis, S. F. 1967. The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text. In Information Retrieval: A Critical Review. Washington, DC, USA: Thompson Book Company. 67–94.

Elhadad, M. K., K. Badran, & G. I. Salama. 2017. A novel approach for ontology-based dimensionality reduction for web text document classification. Paper presented at the 2017 IEEE/ ACIS 16th International Conference on Computer and Information Science (ICIS). Wuhan, China.

Ercan, G., & I. Cicekli. 2007. Using lexical chains for keyword extraction. Inf. Process. Manag. 43: 1705–1714. https://doi.org/10.1016/j.ipm.2007.01.015.

George, S., & V. Srividhya. 2020. Comparison of LDA and NMF Topic Modeling Techniques for restaurant reviews. *Indian Journal of Natural Sciences* 10 (62): 28210–28216. https://www.researchgate.net/publication/350236296_Comparison_of_LDA_and_NMF_Topic_Modeling_Techniques_for_Restaurant_Reviews (Accessed Oct. 15, 2021)

Gers, F. A. 2002. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research* 3: 115-143. https://doi.org/10.1162/153244303768966139

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

Gupta, S., V. Kumar, & B. Pant. 2018. Classification of Textual Data in Distributed Environment. 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), Allahabad, India. 120-124.

Graves, A and J. Schmidhuber.2005. "Framewise phoneme classification with bidirectional LSTM networks," Proceedings. IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 2005, 2047-2052, 4, doi:10.1109/IJCNN.2005.1556215.

Han. J., and M. Kamber. 2006. Data mining: concepts and techniques, 2nd ed. San Francisco: Morgan Kaufmann, 227-228, 615-631.

Hjørland, B. 2019. Indexing: Concepts and theory. In ISKO Encyclopedia of Knowledge Organization, eds. Birger Hjørland, coed. Claudio Gnoli. http://www.isko.org/cyclo/indexing (Accessed March,18 2021)

Holley R. M., & D. N. Joudrey. 2020. Aboutness and conceptual analysis: A review. *Cataloging & Classification Quarterly* doi:10.1080/01639374.2020.1856992

Hoyt, B. 2020. Best practices forlti, l kdsj. Gxhtjk il fhan. gxht content manager on demand full text search. https://www.ibm.com/support/pages/sites/default/files/inlinefiles/Best%20practices%20for%20Using%20Full%20Text%20Searching%20with%20Content%20Manager%20OnDemand-4-22-2020.pdf (Accessed July 9, 2021)

Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Strauzburg, PA, USA, 216–223.

Kaur, A, D. and Chopra. 2016. "Comparison of text mining tools,". 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2016, 186-192, doi:10.1109/ICRITO.2016.7784950.

Langville, A. N., C. D. Meyer, R. Albright, I. Cox, and D. Duling. 2014. Algorithms Initialization ad Convergenece for non-negative matrix factorization. In proceedings of 12th ACM SIGKDD international conference of knowledge and description & data mining. SAS Technical Report. 54-91. doi:10.48550/arXiv.1407.7299.

Lee, D., & H. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791.https://doi.org/10.1038/44565.

Li. Z, W. Shang, & M. Yan. 2016. News text classification model based on topic model, 2016 IEEE/ ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE, 1–5. Okayama, Japan, 2016, 1-5, doi:10.1109/ICIS.2016.7550929.

Maguduru, N. 2003. Text Mining With Support Vector Machins and Non negative matrix factorization. *Linear Multilinear Algebra* 51 (1): 83-95.

Matsuo, Y., & M. Ishizuka. 2004. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on artificial intelligence tool*. 13 (1): 157–169. https://doi.org/10.1142/S0218213004001466.

Meng, L., R. Huang, & J. Gu. 2013. A review of semantic similarity measures in WordNet. *Journal of Intelligent & Fuzzy Systems* 36 (4): 3045-3059

Miner, G., IV. Elder J., F. Fast,T. Hill, R. Nisbet, & Delen. 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. USA:Academic Press.

Mouhoub. M., and M. Al Helal. 2018. Topic Modelling in Bangla Language: An LDA Approach to Optimize Topics and News Classification. *Compute. Inf. Sci*. 11 (4): 77–83.

Newman, D., J. H. Lau, K. Grieser, & T. Baldwin. 2010. Automatic Evaluation of Topic Coherence. Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Los Angeles, California, USA.

Onan, A. 2017. Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes* 46 (2): 330–48.

_____. 2018a. An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science. 44 (1): 28–47. https://doi.org/10.1177/01655515166779*

_____. 2018b. Sentiment analysis on twitter based on ensemble of psychological and linguistic feature sets. *Balkan J Electr Comput Eng*. 6: 1–9.

_____, & S. Korukoglu. 2017. A feature selection model based on genetic rank aggregation for text sentiment classification. *J Journal of Information Science* 43 (1): 25–38.

_____, & M. A. Toçoğlu. 2020. Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts. *Computer Applications in Engineering Education* 29: 675 - 689.

Paatero. P, & U. Tapper. 1994. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics* 1: 111-126.

Pokorny, J. 2018. Automatic subject indexing and classification using text recognition and computer based analysis of the table of content. In Chau, L. & Mounier, P. ELPUB 2018. June 2018, Tornto, Canada. doi:10.4000/proceedings.elpub.2018.19

Salton, G., & C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag*. 24: 513–523.

_____. 1991. *Automatic Text Structuring and Retrieval-Experiments in Automatic Encyclopaedia Searching*. Ithaca, NY, USA: Cornell University.

Srivastava, S. Singh, & J. S. Suri. 2019. Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm. *Computer methods and Programs in Biomedicine* 172: 35–51. https://doi.org/10.1016/j.cmpb.2019.01.011.

Syed, Sh., and M. Spruit. 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

Thiyagarajan, D., & N. Shanthi. 2017. A modified multi objective heuristic for effective feature selection in text classification. *Cluster Computing journal* 11: 1-11, doi/10.1007: s10586-017-1150-7.

Wang, L &, S. Li. 2017. Keyphrase Extraction with Model Ensemble and External Knowledge. Proceedings of the 15th International Workshop on Semantic Evaluations. Association for Computational Linguistics. 934–937. doi:10.18653/v1/S17-2161.

_____, X., X. Tang, W. Qu, & M. Gu. 2017. Word sense disambiguation by semantic inference. In Proceedings of the 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC), Krakow, Poland, 1–6.

_____, X., L. Zhang, D. Klabjan2020 .. Keyword-based Topic Modeling and Keyword Selection. Computer Science, Mathematics.2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA. 1148-1154, doi:10.1109/BigData52589.2021.9671416.

Wilson, A. T., & P. A. Chew. 2010. Term weighting schemes for Latent Dirichlet Allocation. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California.

Zhang, Y. A.-H. 2004. World wide web site summarization. *Web Intelligence and Agent Systems* 2 (1): 39- 53

_____, C. 2008. Automatic keyword extraction from documents using conditional random fields. *J. Comput. Inf. Syst*. 4: 1169–1180.

_____, Q., Y. Wang, Y. Gong, & X. Huang. 2018. Key phrase Extraction Using Deep Recurrent Neural Networks on Twitter. Proceedings of the 8102Conference on Empirical Methods in Natural Language Processing. macao, china. 836-845. doi:10.18653/v1/D16-1080

Zhao, R., & K. Mao. 2018. Fuzzy Bag-of-Words Model for Document Representation. IEEE Transactions on Fuzzy Systems 26 (2): 794-804. doi:10.1109/TFUZZ.2017.2690222.

فاطمه زرمهر

دانشجوی دکتری علم اطلاعات و دانش‌شناسی، گرایش ذخیره و بازیابی اطلاعات از دانشگاه اصفهان است.

مدیریت اطلاعات و متن کاوی از جمله علایق پژوهشی وی است.

علی منصوری

دارای مدرک دکتری در رشتهٔ علوم کتابداری و اطلاع‌رسانی از دانشگاه شهید چمران اهواز است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه اصفهان است.

علم‌سنجی، فناوری‌سنجی، مدیریت اطلاعات و داده‌کاوی از جمله علایق پژوهشی وی است.

حسین کارشناس نجف‌آبادی

دارای مدرک تحصیلی دکتری در رشته هوش مصنوعی از دانشگاه پلی‌تکنیک مادرید - اسپانیا است. ایشان هم‌اکنون استادیار گروه هوش مصنوعی دانشگاه اصفهان است.

هوش محاسباتی، تحلیل و مدل‌سازی داده‌ها، یادگیری ماشین و سامانه‌های هوشمند سلامت از جمله علایق پژوهشی وی است.

پژوهش‌نامه
پردازش و مدیریت اطلاعات