مجله دستاوردهای نوین در مطالعات علوم انسانی
سال ششم، شماره ۶۳، مرداد ۱۴۰۲
ISSN: ۲۵۸۸-۶۹۶۷
www. jonahs.ir

متن‌کاوی: مفاهیم، الگوریتم‌ها و ابزارها
با نگاهی به کاربرد آن در علم اطلاعات و دانش شناسی

مهدی کریمی[۱] ، فائزه دلقندی[۲]
[۱] دانشجوی دکتری تخصصی بازیابی اطلاعات و دانش، دانشگاه پیام‌نور، ایران
رئیس بخش سازماندهی و توسعه منابع دیجیتال سازمان کتابخانه ها، موزه ها و مرکز اسناد آستان قدس رضوی (نویسنده مسئول)
[۲] استادیار گروه علم اطلاعات و دانش‌شناسی، دانشگاه پیام‌نور، ایران

چکیده

متن‌کاوی فرآیند و تکنیکی برای جستجو، بازیابی و استخراج اطلاعات مفید و هدفمند از اقیانوس داده‌ها و اطلاعات طبقه‌بندی نشده است که در قالب متون نوشته‌شده به زبان طبیعی استفاده می‌شود؛ این حوزه یک زمینه تحقیقاتی نوظهور است که به آن مهندسی متن، داده‌کاوی متن یا تجزیه‌وتحلیل متن نیز گفته می‌شود. متن‌کاوی از تکنیک‌های هوش مصنوعی بهره می‌گیرد ازاین‌رو، این سیستم در مقایسه با توانایی‌های انسان ازنظر محدودیت زمانی و شمارش کلمات که مستلزم دقت است، بسیار کارآمد است، متن‌کاوی درزمینه علم اطلاعات و دانش شناسی چشم‌اندازهای زیادی برای ارائه دارد؛ این تکنیک می‌تواند در مدیریت اطلاعات در حال رشد در هر زمینه دانش بسیار مفید باشد؛ همچنین می‌تواند برای تحقیقات این حوزه نیز مورداستفاده قرار گیرد؛ پژوهشگران و کتابداران ممکن است تلاش در این زمینه را برای سازگاری بهتر با آینده مفید بدانند، این مقاله که باهدف بیان مفاهیم کلی متن‌کاوی و کاربردهای آن در حوزه علم اطلاعات و دانش شناسی تدوین‌شده است، جهت ایجاد انگیزه محققان این رشته در مشارکت در پروژه‌های متن‌کاوی باهدف کمک به دانش جدید در بهبود حرفه و خدمات کتابخانه‌ای مفید است، ازآنجایی‌که متن‌کاوی یک زمینه نوظهور در تحقیقات است، ادبیات کمتری به‌ویژه مرتبط با علم اطلاعات و دانش شناسی در دسترس است.

واژه‌های کلیدی: تحلیل داده، متن‌کاوی، استخراج داده، داده‌کاوی، مدیریت دانش، دانش شناسی، علم اطلاعات

مقدمه

امروزه یکی از بزرگ‌ترین چالش‌ها برای سازمان‌ها استخراج دانش نهفته از منابع منتشرشده برای کاربردهای تجاری و تحقیقاتی است؛ کاربرد متن‌کاوی در بین محققین محبوبیت پیداکرده است و کاربردهای آن در حوزه‌های مختلف تحقیقاتی به‌طور تصاعدی در حال رشد است [۱].

اکثر داده‌هایی که روزانه با آن‌ها مواجه می‌شویم، به‌صورت متن بدون ساختار یعنی به‌صورت کتاب، ایمیل، روزنامه و صفحات وب است، متن‌کاوی منابع ساختار نیافته را به یک فرم قابل‌خواندن توسط ماشین یا به‌عبارت‌دیگر، یک قالب ساختاریافته تبدیل می‌کند و سپس دانش را از آن استخراج می‌کند، این تکنیک از فناوری‌های هوش مصنوعی مانند یادگیری ماشینی و پردازش زبان طبیعی استفاده می‌کند؛ متن‌کاوی، استفاده از روش‌های خودکار برای بهره‌برداری از حجم عظیم دانش متنی موجود در اسناد متنی است. متن‌کاوی نشان‌دهنده یک گام به جلو از بازیابی متن است. متن‌کاوی که گاهی به‌عنوان داده‌کاوی متن نامیده می‌شود، به‌طورکلی به فرآیند استخراج اطلاعات باکیفیت بالا از متن اشاره دارد. محققانی از قبیل فلدمن و داگان[۱] (۱۹۹۵) و هرثت[۲](۱۹۹۹) اشاره کردند که متن‌کاوی به‌عنوان "داده‌کاوی متنی " و "کشف دانش در پایگاه‌های داده متنی " نیز شناخته می‌شود. تفاوت بین داده‌کاوی معمولی و متن‌کاوی در این است که در متن‌کاوی، الگوها از متون زبان طبیعی استخراج می‌شوند ولی در داده‌کاوی الگوها از پایگاه داده‌های ساخت‌یافته استخراج می‌شود.

متن‌کاوی یک حوزه بین‌رشته‌ای مرتبط با علوم اطلاعات و دانش شناسی، علوم کامپیوتر، ریاضیات و زبان‌شناسی محاسباتی است [۴] و با بازیابی اطلاعات، داده‌کاوی، یادگیری ماشین و آمار نیز مرتبط است [۵]. کتابداران در این سناریو نقش سه‌گانه‌ای دارند، یکی استفاده از این تکنیک برای بهبود خدمات کتابخانه‌ای به‌عنوان یک سیستم بازیابی اطلاعات کارآمد و طبقه‌بندی و غیره؛ دوم اینکه کتابداران می‌توانند نقش خود را با کمک به پژوهشگرانی که مایل به تلاش در کار متن‌کاوی هستند، تقویت کنند؛ سوم اینکه متن‌کاوی یک زمینه تحقیقاتی درحال‌توسعه در حوزه علوم اطلاعات و دانش شناسی در سراسر جهان است.

۲. متن‌کاوی چیست؟

تاکنون تعارف مختلفی از متن‌کاوی توسط صاحب‌نظران و محققان این حوزه ارائه‌شده است، ماینر، دلن، الدر، فست، هیل و نیسبت[۳](۲۰۱۲) توضیح می‌دهند که «متن‌کاوی و تجزیه‌وتحلیل متن اصطلاحات گسترده‌ای هستند که طیفی از فناوری‌ها را برای تجزیه‌وتحلیل و پردازش داده‌های متنی نیمه ساختاریافته و بدون ساختار توصیف می‌کنند، موضوع وحدت‌بخش پشت هر یک از این فناوری‌ها، نیاز به «تبدیل متن به اعداد» است تا بتوان الگوریتم‌های قدرتمندی را در پایگاه‌های داده اسناد بزرگ اعمال کرد. طبق نظر فلدمن و سانگر[۴](۲۰۰۷) "متن‌کاوی را می‌توان به‌طورکلی به‌عنوان یک فرآیند دانش فشرده که در آن کاربر با یک مجموعه اسناد در طول زمان با استفاده از مجموعه‌ای از ابزارهای تجزیه‌وتحلیل تعامل می‌کند" تعریف کرد، جو[۵] (۲۰۱۹) متن‌کاوی را به‌عنوان "فرایند استخراج دانش ضمنی از داده‌های متنی" تعریف می‌کند، به‌طور مشابه، کوارتلر[۶] (۲۰۱۷) اشاره می‌کند که «متن‌کاوی فرآیند استخراج بینش‌های عملی از متن است»، لیدی[۷](۲۰۰۰) می‌گوید «متن‌کاوی فرآیند تجزیه‌وتحلیل متن طبیعی برای کشف و ضبط اطلاعات معنایی به‌منظور درج و ذخیره‌سازی در ساختار سازمان دانش

---

[۱] Feldman & Dagan
[۲] Hearst
[۳] Miner, Delen, Elder, Fast, Hill, & Nisbet,
[۴] Feldman & Sanger
[۵] Jo
[۶] Kwartler
[۷] Liddy

(KOS) باهدف نهایی کشف دانش از طریق هر متن یا دسترسی بصری برای استفاده در طیف وسیعی از برنامه‌های کاربردی مهم است».

۳. تاریخچه متن‌کاوی

به اعتقاد ماینر و همکاران (۲۰۱۲) حداقل سه دلیل وجود دارد که باید تاریخچه متن‌کاوی را دانست، یکی دیدن مسیرهای توسعه تکنیک‌های متن‌کاوی، دوم اینکه ببینیم چگونه می‌توان تکنیک‌های متن‌کاوی را در آینده گسترش داد و بهبود بخشید، سوم اینکه از گذشته درس بگیریم و از تکرار اشتباهات خودداری کنیم. متن‌کاوی ریشه در سه فرآیند بازیابی اطلاعات، استخراج و خلاصه‌سازی دارد، ماینر و همکاران استدلال می‌کنند که همان‌طور که نمی‌توانیم تاریخ کامپیوترها را بدون درک کامل کار چارلز بابیج در مورد ماشین محاسبه‌گر درک کنیم، به همین ترتیب نمی‌توانیم فرآیند متن‌کاوی را بدون درک ریشه‌های آن درک کنیم.

تحقیقات در زمینه متن‌کاوی از اواسط ۱۹۸۰ میلادی توسط پروفسور دان سوانسون[۸]استاد دانشگاه ایالات‌متحده انجام‌شده است؛ او در تحقیقات خود متوجه شد که با ترکیب برش اطلاعاتی از مقالات پزشکی به‌ظاهر نامرتبط، می‌توان فرضیه‌های جدیدی را استنباط کرد (جو سو، الفوره؛[۹] ۲۰۱۲) در سال‌های اولیه تحقیق متن‌کاوی، سیستم‌های متن‌کاوی متخصصان اطلاعات را هدف قرار می‌دادند و آن‌ها معمولاً به ترکیبی از تخصص متن‌کاوی و رایانه نیاز دارند. امروزه کار بر روی متن‌کاوی توسط محققان مختلف انجام می‌شود [۱۰]. متن‌کاوی با فناوری‌ها و برنامه‌های کاربردی دیگر و افزایش اطلاعات متنی در جهان توسعه پیداکرده است.

به گفته ماینر و همکاران (۲۰۱۲) سه زمینه‌ای که در آن روش‌های دسترسی به اطلاعات متنی توسعه‌یافته است، علم کتابداری، علم اطلاعات و پردازش زبان طبیعی است.

علم کتابداری: فهرست کتابخانه اولین نمونه از خلاصه‌سازی متن است، اولین فهرست توسط توماس هاید[۱۰]در سال ۱۸۶۴ برای کتابخانه بودلیان[۱۱]در دانشگاه آکسفورد ایجاد شد.

علم اطلاعات: قبل از ظهور رایانه، اطلاعات باید فهرست‌نویسی و به‌صورت فهرست برگه ارائه می‌شدند و کاربران کتابخانه باید کار پرزحمت یافتن اطلاعات موردنیاز را انجام می‌دادند.

پردازش زبان طبیعی: این اصطلاح اغلب به‌عنوان مترادف با زبان‌شناسی محاسباتی استفاده می‌شود، پیشرفت اصلی در پردازش زبان طبیعی،[۱۲]فناوری یادگیری ماشینی برای ایجاد الگوریتم‌های تجزیه است که کلمات را به نشانه‌ها تقسیم می‌کند و الگوریتم‌های بنیادی که کلمات را به ریشه تبدیل می‌کنند، خوشه‌بندی یک فرآیند خودکار است که اسناد را بر اساس شباهت‌ها خوشه‌بندی می‌کند.

۴. اهداف متن‌کاوی

برخی کارشناسان (ماینر و همکاران، ۲۰۱۲؛ ییننگ،[۱۳]۲۰۱۲) متن‌کاوی را به هفت حوزه عملی زیر تقسیم می‌کنند:

ذخیره‌سازی و بازیابی اطلاعات: این شامل ذخیره‌سازی و بازیابی اطلاعات از اسناد، مانند موتورهای جستجو و جستجوی کلمات کلیدی است؛

خوشه بندی اسناد: تکنیک خوشه‌بندی اسناد برای گروه‌بندی و دسته‌بندی اصطلاحات، پاراگراف‌ها و اسناد استفاده می‌شود؛

---

[۸] Prof Don Swanson
[۹] Jusoh & Alfawareh
[۱۰] Thomas Hyde
[۱۱] Bodleian Library
[۱۲] Natural Language Processing
[۱۳] Ying

طبقه‌بندی اسناد: روش‌های طبقه‌بندی مبتنی بر مدل‌های نمونه‌های برچسب‌گذاری شده برای گروه‌بندی و دسته‌بندی اصطلاحات، پاراگراف‌ها و اسناد استفاده می‌شود؛

وب کاوی: با استفاده از اینترنت با تمرکز ویژه‌بر مقیاس و به هم‌پیوستگی وب انجام می‌شود؛

استخراج اطلاعات[۱۴]: برای تبدیل متن بدون ساختار و نیمه ساختاریافته به متن ساختاریافته برای شناسایی و استخراج حقایق و روابط مرتبط استفاده می‌شود؛

پردازش زبان طبیعی: پردازش کامپیوتری سطح پایین برای تعامل با زبان انسان است؛

استخراج مفهوم: فرآیند گروه‌بندی کلمات و عبارات به گروه‌های مرتبط معنایی است.

۵. روش‌شناسی متن‌کاوی

ماینر و همکاران (۲۰۱۲) روش‌شناسی را به‌عنوان "فرایند مستند و تا حدودی استانداردشده برای اجرا و مدیریت پروژه‌های پیچیده که شامل بسیاری از وظایف مرتبط با یکدیگر با استفاده از انواع روش‌ها، ابزارها و تکنیک‌ها است" تعریف می‌کنند، آن‌ها همچنین توضیح می‌دهند که متن‌کاوی یک تکنیک نسبتاً جدید و غیراستاندارد است و هیچ روش پذیرفته‌شده‌ای برای متن‌کاوی وجود ندارد؛ آن‌ها مراحل متن‌کاوی را به شکل زیر معرفی کردند:

۵-۱ **تعیین هدف مطالعه**: متن‌کاوی مانند هر پروژه دیگری با تعیین هدف مطالعه شروع می‌شود، درک موضوع و تعیین اهداف ضروری است، این هدف می‌تواند از طریق بیان مسئله و همچنین از طریق مذاکره با کارشناسان موضوعی حوزه به دست آید.

۵-۲ **بررسی دسترس پذیر بودن و ماهیت داده‌ها**: پس از تدوین اهداف، مرحله بعدی، بررسی در دسترس بودن داده‌ها برای فرآیند است، برخی از وظایف درگیر می‌تواند شامل شناسایی منابع برای به دست آوردن داده‌های متنی، ارزیابی دسترسی به داده‌ها و قابلیت استفاده از آن، جمع‌آوری و سپس بررسی غنای آن، ارزیابی نهایی در مورد کیفیت و کمیت داده‌ها باشد، پس از رسیدن به نتایج مثبت، فرآیند بعدی جمع‌آوری داده‌ها و ادغام آن‌ها برای استفاده بیشتر است.

۵-۳ **تهیه داده‌ها و توسعه و ارزیابی مدل‌ها**: این مرحله شامل سه فرآیند اساسی است، ایجاد مجموعه، پیش‌پردازش داده‌ها و استخراج دانش.

۵-۴ **ارزیابی نتایج**: پس از توسعه مدل‌ها و ارزیابی دقت و کیفیت، لازم است تمامی فعالیت‌های مربوطه تأیید و اعتبارسنجی شوند، این فرآیند تکرار کل فرآیند و بررسی اعتبار آن است، تنها در این صورت می‌توان به سمت استقرار نتایج حرکت کنیم، این مرحله به حذف احتمال خطا که می‌تواند منجر به‌اشتباه در فرآیند تصمیم‌گیری شود، کمک می‌کند.

۵-۵- **استقرار نتایج**: هنگامی‌که فرآیند، مرحله ارزیابی را پشت سر گذاشت، آماده پیاده‌سازی است؛ این مرحله از فرآیند می‌تواند ارائه گزارش یافته‌هایی باشد که منجر به تصمیم‌گیری بهتر و یا توسعه و تولید یک سیستم اطلاعات جدید شود.

۶. تکنیک‌های مصورسازی داده‌ها

کوارتلر (۲۰۱۷) تکنیک مصورسازی ابرهای کلمه را برای ارائه داده پیشنهاد می‌دهد، ابر کلمه یکی دیگر از روش‌های مصورسازی داده‌ها است، در این روش فرکانس با اندازه کلمه نشان داده می‌شود، پرتکرارترین کلمات در اندازه‌های بزرگ ظاهر می‌شوند؛ ضمن اینکه روش دیگر استفاده از رنگ‌ها یا گروه‌بندی آن‌ها است.

۷. کاربردهای متن‌کاوی

پنج نوع اساسی از کاربردهای متن‌کاوی وجود دارد که برای مسائل تحلیل متن به کار می‌روند: [۱۲].

[۱۴] information extraction

استخراج معنا: شامل استخراج معنا از داده‌های بدون ساختار و درک مضامین اصلی و پیام‌های مرتبط بدون خواندن متن است؛

دسته‌بندی خودکار متن: این‌یک روش عالی برای پردازش پایین‌دستی با طبقه‌بندی خودکار متن است؛

بهبود دقت پیش‌بینی: در مدل‌سازی پیش‌بینی کننده و مدل‌سازی بدون نظارت، روشی کارآمد برای دستیابی به‌دقت با ترکیب داده‌های بدون ساختار با اطلاعات عددی ساخت‌یافته است؛

شناسایی سند خاص: این‌یک وظیفه مهم در بازیابی اطلاعات برای استخراج کارآمد اسناد مشابه یا مرتبط در یک موضوع خاص است؛

استخراج اطلاعات خاص: استخراج اطلاعات خاص مانند نام‌ها، مناطق جغرافیایی و غیره روشی کارآمد برای ارائه داده به تصمیم‌گیرندگان است.

۸. تکنیک‌های متن‌کاوی

انواع مختلفی از تکنیک‌ها وجود دارد که با استفاده از آن تجزیه‌وتحلیل متن انجام می‌شود که پنج تکنیک اساسی آن به شرح زیر استفاده می‌شود[۱۳].

استخراج اطلاعات: این تکنیک از متن بدون ساختار برای شناسایی عبارات کلیدی و روابط درون‌متن استفاده می‌کند که شامل «توکن‌سازی، شناسایی موجودیت‌های نامدار، تقسیم‌بندی جملات و تخصیص بخشی از گفتار» است. هدف روش‌های استخراج اطلاعات (IE) استخراج اطلاعات مفید از متن از قبیل استخراج موجودیت‌ها، رویدادها و روابط است که از متن نیمه ساختاریافته یا بدون ساختار به دست می‌آید.

دسته‌بندی[۱۵]دسته‌بندی یک روش یادگیری تحت نظارت می‌باشد زیرا بر اساس نمونه‌های ورودی-خروجی در طبقه‌بندی اسناد جدید است، فرآیند دسته‌بندی متن معمولی شامل پیش‌پردازش، نمایه‌سازی، کاهش ابعادی و رده‌بندی است. دسته‌بندی شامل شناسایی موضوعات اصلی یک سند با قرار دادن سند در مجموعه‌ای از موضوعات از پیش تعریف‌شده است[۱۴].

خوشه‌بندی[۱۶]خوشه‌بندی فرآیندی بدون نظارت است که از طریق آن اشیا به گروه‌هایی به نام خوشه طبقه‌بندی می‌شوند. خوشه‌بندی در طیف گسترده‌ای از زمینه‌های تجزیه‌وتحلیل داده‌ها ازجمله داده‌کاوی، بازیابی اسناد، تقسیم‌بندی تصویر و طبقه‌بندی الگوها مفید است. در بسیاری از این مشکلات، اطلاعات قبلی کمی در مورد داده‌ها در دسترس است و تصمیم‌گیرنده باید تا حد امکان فرضیات کمتری در مورد داده‌ها داشته باشد در چنین مواردی روش خوشه‌بندی مناسب‌تر است [۲]. یکی دیگر از مزایای خوشه‌بندی این است که اسناد می‌توانند در چندین موضوع فرعی ظاهر شوند، بنابراین اطمینان حاصل می‌شود که یک سند مفید از نتایج جستجو حذف نمی‌شود [۱۴]. خوشه بندی یکی از موضوعات جالب و مهم در متن‌کاوی است که هدف آن یافتن ساختارهای ذاتی در اطلاعات و مرتب کردن آن‌ها در زیرگروه‌های مهم برای مطالعه و تحلیل بیشتر است[۱۵].

مصورسازی[۱۷]این تکنیک از پرچم‌های متنی (رنگ‌های تراکم) برای کشف دسته‌بندی سند یا اطلاعات مرتبط استفاده می‌کند و می‌تواند برای اسناد فردی یا گروهی اعمال شود، مصورسازی اطلاعات در سه مرحله انجام می‌شود: (۱) آماده‌سازی داده: یعنی تعیین و به دست آوردن داده‌های اصلی. (۲) تجزیه‌وتحلیل و استخراج داده‌ها: یعنی تجزیه‌وتحلیل و استخراج داده‌های بصری موردنیاز از داده‌های اصلی و (۳) مصورسازی [۱].

خلاصه‌سازی[۱۸]این تکنیک برای کاهش اسناد طولانی به شکل خلاصه‌شده با حفظ نکات کلیدی و مفهوم مشترک استفاده می‌شود. عملیات پیش‌پردازش و پردازش بر روی متن خام برای خلاصه‌سازی انجام می‌شود. یک فرآیند خلاصه‌سازی خودکار را

[۱۴] Classification
[۱۵] Clustering
[۱۶] visualization
[۱۷] summarization

می‌توان به سه مرحله تقسیم کرد: (۱) در مرحله پیش‌پردازش، یک نمایش ساختاریافته از متن اصلی به دست می‌آید (۲) در مرحله پردازش، یک الگوریتم باید ساختار متن را به یک ساختار خلاصه تبدیل کند و (۳) در مرحله تولید خلاصه نهایی از ساختار خلاصه به دست می‌آید [۱۴].

۹. ابزارهای متن‌کاوی

برای اجرای متن‌کاوی نیاز به ابزارهایی است که در ادامه به بیان مختصر برخی از مورداستفاده محققان پرداخته می‌شود:

۹-۱ بسته نرم‌افزاری R

بسته نرم‌افزاری R یکی از محبوب‌ترین نرم‌افزارهای متن‌باز برای تحلیل متن است. این بسته شامل مجموعه‌ای از بسته‌های کتابخانه‌ای برای انجام عملیات پردازش زبان طبیعی[۱۹]تجزیه‌وتحلیل احساسات، مدل‌سازی موضوع، تعیین فرکانس کلمات و تشکیل ابرهای کلمه است. این بسته‌هایی مانند NLTK را با ویژگی‌هایی برای توکن‌سازی، واژه‌سازی، ریشه‌یابی، پیش‌پردازش متن و پیاده‌سازی برای استقرار، LDA، قوانین ارتباط، الگوریتم‌های طبقه‌بندی و سایر الگوریتم‌ها ارائه می‌کند. این نرم‌افزار نتایج کمی را نشان می‌دهد که در نمودارهای آماری داده‌های متنی ارائه‌شده‌اند [۱].

۹-۲ سیستم تجزیه‌وتحلیل آماری[۲۰](SAS)

یک نرم‌افزار اختصاصی برای پردازش داده‌ها و تجزیه‌وتحلیل آماری است. در واقع یک نرم‌افزار برنامه‌نویسی است که می‌تواند حجم بالایی از داده‌ها را برای انجام تحلیل متن پردازش کند. ساس در درجه اول برای بازیابی داده‌ها، سازماندهی داده‌ها و انجام تجزیه‌وتحلیل متن توسعه داده شد. این ابزار پیشرو در صنعت برای تجزیه‌وتحلیل تجاری است که به تجزیه‌وتحلیل آماری کمک می‌کند [۱۷].

۹-۳ پایتون

پایتون یکی از سریع‌ترین زبان‌های برنامه‌نویسی است که دارای کتابخانه‌ها و بسته‌های داخلی متعدد برای انجام استخراج متن است. پایتون یک بسته NLTK شامل روش‌های NLP اعدادی مانند نشانه‌گذاری جملات و کلمات، بخشی از برچسب‌گذاری گفتار، تقسیم‌بندی و طبقه‌بندی دارد .Natural Language Toolkit مجموعه‌ای از کتابخانه‌ها و برنامه‌ها برای پردازش زبان طبیعی آماری برای زبان‌های برنامه‌نویسی پایتون است. بسیاری از سازمان‌ها و شرکت‌ها از این ابزار برای مطالعات متعدد مانند مدل‌سازی موضوع، پردازش زبان طبیعی، توسعه وب، ساخت نرم‌افزار، هوش مصنوعی، تجزیه‌وتحلیل تجاری، تحلیل بازار و محاسبات علمی استفاده می‌کنند. بسته‌هایی مانند جنسیم[۲۱]و اسپیسی[۲۲]ازجمله بسته‌های موجود در پایتون هستند که برای برنامه‌های پیشرفته متن‌کاوی پیاده‌سازی شده‌اند.

۹-۴ رپید ماینر[۲۳]

رپید ماینر یک ابزار علم داده است که در درجه اول توسط شرکت‌ها برای یادگیری ماشین و وظایف داده‌کاوی استفاده‌شده است. رپید ماینر یک رابط کاربری "کشیدن و رها کردن" برای طراحی گردش کار فرآیند تجزیه‌وتحلیل ارائه می‌دهد. این نرم‌افزار می‌تواند به مجموعه داده‌های ذخیره‌شده در پایگاه داده‌های رابطه‌ای یا صفحات گسترده یا فرمت‌های فایل خاص بسته آماری متصل شود. رپید ماینر از XML برای توصیف فرآیند کشف دانش با بسیاری از الگوریتم‌های یادگیری از WEKA استفاده می‌کند. این نرم‌افزار به وظایف داده‌کاوی برای حل مشکلات تجاری و پیش‌بینی روندهای داغ مانند تجزیه‌وتحلیل احساسات، تشخیص تقلب و ... کمک می‌کند [۱].

---

[۱۹] NLP
[۲۰] Statistical Analysis System(SAS)
[۲۱] Gensim
[۲۲] Spacy
[۲۳] RapidMiner

۹-۵ مالت[۲۴]

مالت یک نرم‌افزار متن‌باز است که به زبان جاوا نوشته‌شده است و دارای چندین نرم‌افزار است که برنامه‌های کاربردی یادگیری ماشین برای متن‌کاوی مانند استخراج متن، طبقه‌بندی، خوشه‌بندی، مدل‌سازی موضوع، پردازش زبان طبیعی و کارهای دیگر را شامل می‌شود. مالت الگوریتم‌های مختلفی مانند بیز ساده[۲۵]حداکثر آنتروپی و درخت‌های تصمیم را پیاده‌سازی می‌کند [۱۷].

۹-۶ رمزگذار KH

KH Coder نرم‌افزار رایگانی است که برای تجزیه‌وتحلیل متن و سایر عملیات زبان‌شناسی محاسباتی استفاده می‌شود. این نرم‌افزار انواع مختلف جستجو و تجزیه‌وتحلیل آماری ابزارهای پشتیبان مانند گلوله برفی را باقابلیت اتصال به MYSQL و R ارائه می‌دهد. همچنین دارای قابلیت تجزیه‌وتحلیل اسناد از زبان‌های مختلف مانند ژاپنی، انگلیسی، فرانسوی، آلمانی، ایتالیایی، پرتغالی و اسپانیایی است [۱۹].

۱۰. الگوریتم‌های متن‌کاوی

الگوریتم‌های متن‌کاوی با بهره‌گیری از یادگیری ماشینی برای حجم‌هایی از داده‌های متنی چندضلعی که از زبان طبیعی به شکل ساختاری و عددی تبدیل‌شده‌اند اعمال می‌شود [۵]. همه این تکنیک‌ها اهداف مشترکی برای کشف اطلاعات پنهان، روندها، الگوها برای کمک به تصمیم‌گیری دارند. در ادامه به معرفی مختصر برخی از این الگوریتم‌ها پرداخته می‌شود:

۱۰-۱ تخصیص دیریکله نهفته[۲۶](LDA)

LDAیک تکنیک یادگیری بدون نظارت است که برای مدل‌سازی موضوع استفاده می‌شود. مدل‌سازی موضوع یک فرآیند آماری برای شناسایی موضوعات از مجموعه‌های متنی است. همبستگی بین موضوعات و کلمات موجود در مجموعه اسناد را ترکیب می‌کند.

تخصیص دیریکله پنهان (LDA) یک مدل احتمالی برای شناسایی موضوعات معنایی نهفته از مجموعه‌های متنی گسترده است. احتمال هر سند را برای مجموعه‌ای از موضوعات بر اساس مدل بیزی سلسله مراتبی سه سطحی فراهم می‌کند. با استفاده ازLDA ، موضوعات اساسی یک سند به‌طور خودکار بر اساس محتوای اسناد قابل‌شناسایی هستند [۲۰].

این الگوریتم در زمینه‌های مختلف از قبیل پردازش زبان طبیعی، تشخیص گفتار، فیلتر هرزنامه، وب کاوی و تجزیه‌وتحلیل ویدئو کاربردهای گسترده‌ای دارد [۷].

تمایز LDA از مدل خوشه‌بندی چندجمله‌ای دیریکله بسیار مهم است. یک مدل خوشه‌بندی کلاسیک شامل یک مدل دوسطحی است که در آن‌یک دیریکله یک‌بار برای یک پیکره نمونه‌برداری می‌شود، یک متغیر خوشه‌بندی چندجمله‌ای یک‌بار برای هر سند در پیکره انتخاب می‌شود و مجموعه‌ای از کلمات برای سند مشروط به خوشه انتخاب می‌شود. از سوی دیگر، LDAشامل سه سطح است و به‌ویژه، گره موضوعی به‌طور مکرر در سند نمونه‌برداری می‌شود. تحت این مدل، اسناد می‌توانند با موضوعات متعدد مرتبط شوند [۲۰].

LDAدر زمینه‌های دیگر مانندNLP ، تشخیص گفتار، فیلتر هرزنامه، وب کاوی و تجزیه‌وتحلیل ویدئو کاربردهای گسترده‌ای دارد [۷].

۱۰- ۲ دسته‌بندی بیز ساده(NBC)

---

[۲۴] Mallet

[۲۵] Naïve Bayes

[۲۶] Latent Dirichlet Allocation(LDA)

الگوریتم بیز روشی برای دسته‌بندی پدیده‌ها، بر اساس احتمال وقوع یا عدم وقوع یک پدیده است. این روش یکی از ساده‌ترین الگوریتم‌های پیش‌بینی در جهان به شمار می‌رود و نکته مهم در مورد این الگوریتم این است که در عین سادگی، دقت قابل قبولی هم دارد که هر دو از مزیت‌های آن به شمار می‌روند. دقت این الگوریتم را می‌توان با استفاده از برآورد چگالی کرنل به‌صورت قابل‌توجهی بالا برد. شیوه یادگیری در روش بیز ساده از نوع یادگیری با نظارت است. این روش در دهه ۱۹۶۰ در میان دانشمندان بازیابی اطلاعات توسعه یافت و هنوز هم از روش‌های محبوب در دسته‌بندی اسناد به شمار می‌آید. این الگوریتم کاربردهای مختلفی مانند یافتن اسناد مرتبط، طبقه‌بندی، دسته‌بندی، شناسایی سن/جنسیت، تشخیص زبان و تحلیل روند دارد [۴۰]. بااین‌حال، در درجه اول برای طبقه‌بندی و دسته‌بندی متن استفاده می‌شود [۱۶].

۱۰-۳ ماشین بردار پشتیبان(SVM)

ماشین بردار پشتیبان یک "الگوریتم یادگیری ماشینی نظارت‌شده" است که برای طبقه‌بندی و تجزیه‌وتحلیل رگرسیون مانند تشخیص هرزنامه، تجزیه‌وتحلیل احساسات و طبقه‌بندی اسناد به دسته‌هایی مانند اخبار، ایمیل‌ها، مقالات و صفحات وب استفاده می‌شود [۲۲].

این الگوریتم، خطی را ترسیم می‌کند که به‌عنوان "هیپرپلان" شناخته می‌شود. هدف اصلی این الگوریتم تقسیم دقیق مجموعه داده‌ها و یافتن "حداکثر حاشیه ابر صفحه " است [۲۲].

مدل‌های SVM همچنین برای طبقه‌بندی مجموعه داده‌های بدون برچسب مانند طبقه‌بندی متن و تصویر، تشخیص دست‌نویس، تجزیه‌وتحلیل تشخیص احراز هویت بیومتریک استفاده می‌شوند.

۱۰-۴ قوانین انجمنی[۲۷]

قوانین انجمن مبتنی بر روش‌های یادگیری ماشینی است و برای یافتن الگوی جالب و همبستگی بین تعداد زیادی متغیر در یک مجموعه داده اعمال می‌شود. کاربردهای گسترده قوانین انجمنی عبارت‌اند از تجزیه‌وتحلیل سبد، خوشه‌بندی، بازاریابی متقابل، طبقه‌بندی، فهرست و طراحی [۳].

با استفاده از این الگوریتم متن‌کاوی، می‌توان الگوهای جالبی را کشف کرد و دانش عمیقی در مورد داده‌های انتخاب‌شده در مجموعه گسترده‌ای از مجموعه داده‌ها به دست آورد.

۱۰-۵ کا نزدیک ترین همسایه[۲۸]

کا نزدیک‌ترین همسایه یک "الگوریتم طبقه‌بندی نظارت‌شده" است که برای دسته‌بندی اسناد جدید در میان مجموعه داده‌ها استفاده می‌شود، کا نزدیک‌ترین همسایه، یک الگوریتم ساده و دارای فرآیند ناپارامتریک است که به‌خودی‌خود از تجربیات گذشته می‌آموزد. علاوه بر این، این الگوریتم نقاط جدید را بر اساس مفاهیم شباهت و ارزش اقلام عملیات کای داده‌ها طبقه‌بندی می‌کند. هدف اصلی این الگوریتم، جستجوی اسناد مشابه در میان مجموعه داده‌های بزرگی از اقلام است [۸].

۱۰-۶ شبکه‌های عصبی[۲۹]

شبکه عصبی یک شبکه عصبی مصنوعی است که از عملکرد مغز بیولوژیکی انسان الهام گرفته‌شده است، در تئوری کلی مانند شبکه‌ای از نورون‌ها عمل می‌کند. یک شبکه عصبی دارای گره‌های مختلفی است که به یکدیگر متصل هستند و دارای سه لایه است: "لایه ورودی"، "لایه میانی" و "لایه بیرونی"؛ لایه ورودی سیگنال‌ها را از لایه بیرونی و همچنین یک‌لایه میانی شامل نورون‌ها دریافت می‌کند و خروجی می‌دهد. تکنیک شبکه عصبی متن را به‌صورت خودکار طبقه‌بندی می‌کند. این شبکه‌های مصنوعی به‌ویژه برای مدل‌سازی پیش‌بینی و تحلیل آماری مورداستفاده قرار می‌گیرند که در آن از طریق مجموعه داده‌ها

[۲۷] Association Rules
[۲۸] K-Nearest Neighbors
[۲۹] Neural Networks(NN)
[۳۰] Artificial Neural Networks - ANN

آموزش داده می‌شوند. این مدل شبکه عصبی چندین مشکل را در زمینه‌هایی مانند فناوری، آمار و اقتصاد حل کرده است که داده‌ها را در یک‌زمان ثبت می‌کند و با مقایسه طبقه‌بندی آن‌ها با منابع معتبر یاد می‌گیرد [۲۳].

۱۰- ۷ درخت تصمیم[۳۱]

درخت تصمیم نقشه‌ای از نتایج احتمالی یکسری از انتخاب‌ها یا گزینه‌های مرتبط به هم است به‌طوری‌که به یک فرد یا سازمان اجازه می‌دهد تا اقدامات محتمل را ازلحاظ هزینه‌ها، احتمالات و مزایا بسنجد. از درخت تصمیم می‌توان برای پیشبرد اهداف و برنامه‌های شخصی و غیررسمی و همچنین ترسیم الگوریتمی که بر اساس ریاضیات بهترین گزینه را پیش‌بینی می‌کند، استفاده کرد. یک درخت تصمیم‌گیری به‌طورمعمول با یک نُود اولیه شروع می‌شود که پس‌ازآن پیامدهای احتمالی به‌صورت شاخه‌هایی از آن منشعب شده و هرکدام از آن پیامدها به نُودهای دیگری منجر شده که آن‌ها هم به‌نوبه خود شاخه‌هایی از احتمالات دیگر را ایجاد می‌کنند که این ساختار شاخه‌شاخه سرانجام به نموداری شبیه به یک درخت مبدل می‌شود. این مدل‌سازی پیش‌بینی در آمار، متن‌کاوی و همچنین در یادگیری ماشینی کاربرد دارد. درختان تصمیم به‌عنوان یک روش طبقه‌بندی نیز شناخته می‌شوند [۲۴].

۱۱. محدودیت‌ها و توانمند سازها در متن‌کاوی

ماینر و همکاران (۲۰۱۲) محدودیت‌های مختلفی را برای فرآیند متن‌کاوی از قبیل مسائل مربوط به حریم خصوصی/دسترسی، محدودیت‌های نرم‌افزاری، محدودیت‌های سخت‌افزاری، چالش‌های زبانی، ارائه کرده‌اند؛ درحالی‌که متن‌کاوی دارای چندین عامل و توانمندی نیز است که عبارت‌اند از بهره‌گیری از روش‌های پردازش زبان طبیعی، ابزارهای نرم‌افزاری، تخصص دامنه و رایانه‌های سریع.

۱۲. متن‌کاوی در علم اطلاعات و دانش شناسی

کاربردهای عمده متن‌کاوی در خدمات کتابخانه‌ای در طبقه‌بندی، استخراج کلمات کلیدی، شناسایی موجودیت‌های نامدار، مدل‌سازی موضوع و خوشه‌بندی است که می‌تواند کتابخانه‌ها را در مدیریت اطلاعات، بازیابی اطلاعات و تصمیم‌گیری کمک نماید.

ژانگ و گو[۳۲] (۲۰۱۱) توضیح می‌دهند که ازآنجایی‌که درصد بالایی از دانش جهان به شکل بدون ساختار است، بنابراین احتمالاً از فناوری به‌عنوان متن‌کاوی برای استخراج اطلاعات منابع به‌طور کارآمد استفاده می‌شود؛ کار آن‌ها این است که برای تعیین نام‌های مناسب، تغییرات و دسته‌بندی‌های آن‌ها، از شناسایی موجودیت‌های نامدار بهره می‌گیرند.

کونگ[۳۳] (۲۰۱۷) پیشنهاد می‌کند که با توجه به رشد مستمر قیمت کتاب و کمبود بودجه برای خرید کتاب، لازم است بودجه برای خرید کتاب‌های موردنیاز صرف شود؛ برای این منظور، اولویت‌های مطالعه خوانندگان موردبررسی قرار گرفت و مشخص شد که خوانندگان کتاب‌های خاصی را برای مطالعه ترجیح می‌دهند.

الدیهانی و آبراهامز[۳۴] (۲۰۱۶) مجموعه داده‌ها را از حساب‌های توییتر ده کتابخانه دانشگاهی جمع‌آوری کردند، تک‌کلمه‌ای که بیشترین فراوانی را دارد "باز" است، کلمه دومی که بیشترین فراوانی را نشان می‌دهد "مجموعه‌های ویژه" و کلمه سوم "ذخیره داده‌ها" بود؛ رایج‌ترین دسته «منابع» بود، یافته‌های این مطالعه اهمیت تجزیه‌وتحلیل کل داده‌های کتابخانه‌های دانشگاهی در حساب‌های اجتماعی را برای کمک به تصمیم‌گیری و بهبود برنامه‌ریزی برای خدمات به مراجعان برجسته می‌کند.

---

[۳۱] Decision Tree
[۳۲] Zhang & Gu
[۳۳] Cong
[۳۴] Al-Daihani & Abrahams

اوکرسون[۳۵] (۲۰۱۳) دو حوزه اصلی فعالیت را برای کتابداران در برآوردن نیازهای متن‌کاوی در مؤسسات تحقیقاتی پیشنهاد می‌کند، یکی توسعه یک‌زبان دارای مجوز است، بسیاری از کتابداران در توسعه زبان مجوز و مجوز برای متن‌کاوی فعال هستند که برای این فرآیند ضروری است، فرصت دوم حمایت از پژوهشگران است، کتابداران می‌توانند به محققان برای انجام پروژه‌های متن‌کاوی کمک کنند. او همچنین توضیح می‌دهد که لازم است با معرفی بهتر خدمات کتابخانه به جامعه بفهمانیم که ارائه ابزار برای متن‌کاوی کار یک کتابدار است.

به‌طور مشابه، هیگینز[۳۶] (۲۰۱۴) به دلایلی برای مشارکت کتابداران در تکنیک‌های متن‌کاوی اشاره می‌کند، به‌عنوان‌مثال، برای حمایت از آخرین تحقیقاتی که در علوم انسانی به کار گرفته‌شده‌اند، متن‌کاوی این پتانسیل را دارد که کتابداران و محققان را به علایق متقابل خود نزدیک کند. کتابداران و دانشمندان علوم انسانی در توجه به ارزش و مفاهیم داده‌های متنی یکسان هستند، ابتکارات متن‌کاوی می‌تواند در غنی‌سازی مجموعه‌های دیجیتالی کتابخانه کمک کند، متن‌کاوی می‌تواند به تولید فراداده بسیار توصیفی که زمینه اصلی نگرانی کتابداران است کمک کند، روش‌های متن‌کاوی به جمع‌آوری موجودیت‌های نامدار کمک می‌کند، کتابداران این فرصت رادارند که توضیحاتی را برای داده‌ها با محتوای موضوعی گسترده‌تر تولید کنند، مدل‌سازی موضوعی را می‌توان هم برای تجمیع و هم برای متمایز کردن منابع استفاده کرد.

اندرسون و کریگلو[۳۷] (۲۰۱۷) مراحل متن‌کاوی را در کتابخانه‌های دانشگاهی توضیح می‌دهند، این‌یک دیدگاه جزئی است از اینکه چگونه کتابداران می‌توانند با کمک به اعضای هیئت‌علمی به تسهیل تحقیق متن‌کاوی کمک کنند، کتابداران می‌توانند در شناسایی منابع، صدور مجوز داده‌ها، استخراج داده‌ها، جمع‌آوری داده‌ها، ابداع مدل‌ها، نگهداری و حفظ و انتشار آن‌ها کمک کنند.

علم اطلاعات و دانش شناسی از فناوری متن‌کاوی برای تقویت تحقیقات در حوزه خود و ایجاد بینش متفکرانه استفاده می‌کند، همان‌طور که بسیاری از رشته‌های دیگر و حوزه نوظهور علوم انسانی دیجیتال نیز این کار را انجام می‌دهند، ناگارکار و کومبهار[۳۸] (۲۰۱۵) یک مطالعه تحلیلی از تحقیقات منتشرشده در علم اطلاعات و کتابداری از سال ۱۹۹۹ تا ۲۰۱۳ انجام دادند، آن‌ها رشد زمانی ادبیات پژوهشی را در مورد متن‌کاوی، کشورهای پیشرفته، مؤسسات، بخش‌ها و افرادی که به‌طور فعال در متن نقش دارند، تجزیه‌وتحلیل کردند، برای این منظور از Pajek و VoSviewer استفاده کردند

تیماکوم، کیم و سونگ[۳۹] (۲۰۲۰) ساختار دانش علم اطلاعات و دانش شناسی را با استفاده از مقالات مجلات متن کامل تحلیل کردند، آن‌ها از تکنیک‌های متن‌کاوی، تحلیل هم کلمه، خلاصه‌سازی متن و مدل‌سازی موضوع استفاده کردند، آن‌ها از تکنیک‌های بصری متن‌کاوی نیز برای ترسیم یافته‌های خود استفاده کردند و متوجه شدند که مدیریت اطلاعات دیجیتال اصلی‌ترین حوزه توسعه‌یافته تحقیق است

در کره، لی، مون و کیم[۴۰] (۲۰۰۷) از متن‌کاوی برای بررسی ساختار فکری مدیریت سوابق و علم آرشیو استفاده کردند. در مطالعه دیگری، لی، اچ کیم و پی جی کیم[۴۱] (۲۰۱۰) از تحلیل دامنه با متن‌کاوی برای مطالعه روندهای پژوهشی در تحقیقات کتابخانه دیجیتال استفاده کردند، در چین، ینگ (۲۰۱۲) با استفاده از نرم‌افزار متن‌کاوی ساتی[۴۲] مطالعه‌ای در زمینه استخراج رکوردهای کتابشناختی انجام داد، او برای ترسیم نقشه دانش و نمودار استراتژیک نتایج، تجزیه‌وتحلیل خوشه‌ای و تحلیل مقیاس چندبعدی را ارائه نمود.

---

[۳۵] Okerson
[۳۶] Higgins
[۳۷] Anderson & Craiglow
[۳۸] Nagarkar & Kumbhar
[۳۹] Timakum & Kim & Song
[۴۰] Lee & Moon & Kim
[۴۱] Lee, J, & Kim, H., & Kim, P.G
[۴۲] SATI

۱۳. نتیجه‌گیری

متن‌کاوی به‌عنوان داده‌کاوی متن یا کشف دانش در متن (KDT) نیز شناخته می‌شود، به‌طورکلی به فرآیند استخراج اطلاعات و دانش از متن بدون ساختار اشاره دارد [۳۱]. متن‌کاوی یک حوزه بین‌رشته‌ای مرتبط با علوم اطلاعات و دانش شناسی، علوم کامپیوتر، ریاضیات و زبان‌شناسی محاسباتی است [۴] و با بازیابی اطلاعات، داده‌کاوی، یادگیری ماشین و آمار نیز مرتبط است. ازآنجایی‌که بیشتر اطلاعات (بیش از ۸۰٪) به‌صورت متن ذخیره می‌شود، اعتقاد بر این است که متن‌کاوی ارزش تجاری بالایی دارد. دانش ممکن است از بسیاری از منابع اطلاعاتی کشف شود، بااین‌حال، متون بدون ساختار بزرگ‌ترین منبع دانش هستند. متن‌کاوی در حوزه علم اطلاعات و دانش شناسی کاربردهای فراوان دارد. بهره‌گیری از این تکنیک برای چکیده‌نویسی منابع اطلاعاتی، بازیابی اطلاعات، تولید فراداده، معرفی بهتر منابع اطلاعاتی به پژوهشگران برخی از کاربردهای آن در این علم است. متن‌کاوی کمک می‌کند که کتابخانه‌ها خدمات خود را افزایش دهند و کتابداران یا متخصصان علم اطلاع‌رسانی و دانش شناسی می‌توانند این تکنیک را در حوزه تحقیقات کتابخانه‌ای و اطلاعاتی به‌کارگیرند. ضمن اینکه به پژوهشگران حوزه متن‌کاوی کمک نمایند و مشارکت فعالانه در فرایند متن‌کاوی آن‌ها داشته باشند.

منابع و مراجع

۱. Kwartler, T.(2017). Text mining in practice with R, New Jersey: John Wiley & Sons.

۲. Feldman, R., & Dagan,I.(1995). “Knowledge discovery in textual databases (kdt),” in Proceedings of the Conference on Knowledge Discovery and Data Mining, 112– 117.

۳. Hearst, M. A.(1999). “Untangling text data mining,” in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 3–10.

٤. Jo, T.(2019). Text mining: Concepts, implementation and big data challenge, Seoul: *Springer* International Publishing.

٥. Nagarkar, S., & Kumbhar, R.(2015). Text mining: An analysis of research published under the subject category ‘information science and library science’ in web of science database during 1999-2013, Library Review, 64(3). 248-262.

٦. Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. A.(2012). Practical text mining and statistical analysis for non-structured text data applications, Amsterdam: Academic Press.

۷. Feldman, R ،& Sanger, J(2007). The text mining hand book: Advanced approaches in analyzing unstructured data, New York: Cambridge University Pres.

۸. Liddy, E. D.(2000).Text mining, Bulletin of the American Society for Information Science,13-14. Retrieved from: https://asistdl, onlinelibrary, wiley, com/doi/full/10, 1002/bult, 184

۹. Jusoh, S., & Alfawareh, H.M.(2012). Techniques , Applications and Challenging Issue in Text Mining.

۱۰. Kalra, V., & Aggarwal, R.(2017). Importance of text data preprocessing and implementation in RapidMiner. In Proceedings of the First International Conference on Information, ICITKM, New Delhi, 71–75.

۱۱. Ying, L, Q, Y.(2012). A study on mining bibliographic records by designed software-SATI: Case study on library and information science. *Journal of Information ResourceManagement*, 1, 35-67.

۱۲. Timakum, T., Kim, G., & Song, M.(2020). A data driven analysis of the knowledge structure of library science with full text journal articlesm. *Journal of Librarianship and Information Science*. 52(2), 345-365.

۱۳. Gupta, V., & Lehal, G. S.(2009). A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence.1(1),60-76.

۱٤. Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. Journal of Business Economics, 87(3), 317–335. doi:10.1007/s11573-016-0822-8

۱٥. Diane, H., & Lawrence, S.(2009). Latent Dirichlet allocation for text, images, and music (p. 1). San Diego: Department of Computer Science University of California.

۱٦. Behera, S., & Kumar, N. V.(2015). Filtering of unstructured text. *International Journal of Engineering Research and Development*.12(11), e2278-067X.

۱۷. Lee, J., Moon, J., & Kim, H.(2007). Examining the intellectual structure of records management and archival science in Korea with text mining. *Journal of the Korean Society for Library and Information Science*, 41(1), 345-372.

۱۸. Karanikas, H., & Theodoulidis, B.(2001). "Knowledge Discovery in Text and Text Mining Software". Centre for Research in Information Management, UK.

۱۹. Zhang, H.(2005) . Exploring conditions for the optimality of naïve bayes, International Journal of Pattern Recognition and Artificial Intelligence,19(2), 183–198. doi:10.1142/ S0218001405003983

۲۰. Blei, D. M., Ng, A. Y., & Jordan, M. I..(2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022

۲۱. Brereton, R. G., & Lloyd, G. R.(2010).Support vector machines for classification and regression. Analyst, 135(2), 230-267.

۲۲. Brown, M., & Lewis, H.(1999).Support vector machines and linear spectral unmixing for remote sensing. In S. Singh (Ed.). International Conference on Advances in Pattern Recognition ,395–404. London: Springe

۲۳. Talib, R. , Hanif, M. K., Ayesha, S., & Fatima, F.(2016). Text mining: Techniques, applications and issues, International, Journal of Advanced Computer Science and Applications, 11(7), 414-418.

۲٤. Zhang, Y., & Gu, H.(2011). Text mining with application to academic libraries. Retrieved from

https://link,springer,com/chapter/10,1007/978-3-642-22694-6_28

۲٥. Cong, D.(2017).Application of text mining in library book procurement. Retrieved from https://www‹researchga,net/publication/314783043_Application_of_text_mining_,in_library_book_procurement

۲۶. Al-Daihani, S, H., & Abrahams, A.(2016). text mining analysis of academic libraries' tweets, The *Journal of Academic Librarianship*,42(2) 135-143.

۲۷. Okerson, A.(2013). Text and data mining: A librarian overview. Retrieved from

https://www,fosteropenscience,eu/sites/default/files/original/3628,pdf

۲۸. Higgins, D.(2014). Reading and non-reading: Text mining in critical practice, In K, J, Varnum (Ed ،), The top technologies every librarian needs to know: A LITA guide (pp), Chicago, IL: ALA Techsource, 85-99.

۲۹. Anderson, C, B., & Craiglow, H, A.(201۷). Text mining in business libraries ،*Journal of Business & Finance Librarianship*, 22(2), 149-165.

۳۰. Lee, J, Y., Kim, H., & Kim, P.J.(2010). Domain analysis with text mining: Analysis of digital library research trends using profiling methods, *Journal of Information Science*, 36, 144-161.

۳۱. Khusbu, T, & Vinit, K.(2022). Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools, New Review of Academic Librarianship, 28(3).279-302.DOI: 10, 1080/13614533, 2021, 1918190