وزارت بهداشت، درمان و آموزش پزشکی
دانشگاه علوم پزشکی و خدمات بهداشتی و درمانی گناباد
معاونت تحقیقات و فناوری

متن کاوی انتشارات جهانی کروناویروس

کد اخلاق طرح:

IR.GMU.REC.1398.189
۱۳۹۸/۱۲/۲۷

پژوهشگران

میثم داستانی، فرشید دانش، محمد قربانی

بهار ۱۳۹۹

فهرست مندرجات

چکیده

مقدمه: پاندمی چالش‌برانگیز و جهانی کروناویروس، موجب شده پژوهشگران مطالعات گسترده‌ای را در خصوص کشف راه‌های پیشگیری و درمان انجام داده و نتایج کسب‌شده را در قالب مقالات علمی در مجلات منتشر نمایند. بنابراین تحلیل روندها و موضوعات انتشارات علمی کروناویروس دارای اهمیت راهبردی است. این پژوهش مدل‌سازی موضوعی انتشارات جهانی کروناویروس در پنجاه سال گذشته را مطالعه کرده است.

روش بررسی: این پژوهش از نوع کاربردی است که با استفاده از فنون متن‌کاوی و با رویکرد تحلیلی انجام گردیده است. جامعه آماری این پژوهش تمامی انتشارات جهانی کروناویروس است. به‌منظور استخراج داده‌ها از پایگاهWeb of Science Core Collection در بازه زمانی ۱۹۷۰ -۲۰۲۰ استفاده‌شده است. به‌منظور تعیین کلیدواژه‌های اصلی جهت طراحی راهبرد جستجو از MESH و به‌منظور تجزیه تحلیل داده‌ها و اجرای الگوریتم‌های متن‌کاوی از قبیل مدل‌سازی موضوعی Latent Dirichlet Analysis (LDA) و زبان برنامه‌نویسی Python به‌کاررفته است.

یافته‌ها: یافته‌های نشان داد که واژگان sars، science، protein، mers، veterinary، cell، human، rna، medicine، virology مهم‌ترین واژگان در انتشارات جهانی کروناویروس بوده است. همچنین با اجرای الگوریتم مدل‌سازی موضوعی، هشت موضوع مهم در انتشارات جهانی کروناویروس شناسایی شد. این موضوعات به ترتیب بیشترین انتشار عبارت‌اند از:

“structure and proteomics”, “Cell signaling and immune response”, “clinical presentation and detection”, “Gene sequence and genomics”, “Diagnosis tests”, “vaccine and immune response and outbreak”, “Epidemiology and Transmission” and “gastrointestinal tissue” .

نتیجه‌گیری: این پژوهش به‌طور شفاف قلمروهای موضوعی انتشارات جهانی کروناویروس را نشان داد. نتایج حاکی از آن است که با توجه به پاندمی شدن کروناویروس و نگرانی بین‌المللی و آسیب‌ها و زیان‌های شدید در حوزه‌های بهداشت و درمان، اقتصاد، اجتماع و سیاست؛ انتشارات جهانی کروناویروس عمدتاً بر روی محورهای پیشگیری، تشخیص و درمان به‌موقع متمرکز است.

واژگان کلیدی: کروناویروس؛ کووید-۱۹؛ متن‌کاوی؛ مدل‌سازی موضوعی؛ علم‌سنجی؛ وب‌گاه علم

فصل اول: کلیات پژوهش

مقدمه

طبق اعلام سازمان جهانی بهداشت(WHO) ، بیماری‌های ویروسی همچنان پدیدار می‌شوند و مسئله‌ای جدی برای بهداشت عمومی هستند. در بیست سال گذشته ، چندین بیماری همه‌گیر ویروسی مانند سندرم حاد تنفسی شدید کروناویروس (SARS-CoV) در سال‌های ۲۰۰۲ تا ۲۰۰۳ و آنفولانزای H1N1 در سال ۲۰۰۹ ثبت‌شده است. اخیراً، سندرم تنفسی خاورمیانه MERS-CoV اولین بار در سال ۲۰۱۲ در عربستان سعودی شناسایی شد(۱).

در یک بازه زمانی که به امروز می‌رسد، اپیدمی مواردی از عفونت‌هایی با درگیری تنفسی کم (خفیف) نامشخص در وهان، بزرگ‌ترین منطقه شهری استان هوبئی چین، برای نخستین بار در تاریخ ۳۱ دسامبر ۲۰۱۹ به دفتر WHO در کشور چین گزارش شد. ازآنجاکه قادر به شناسایی عامل ایجادکننده نبودند ، این افراد به‌عنوان "ذات‌الریه با علت ناشناخته" طبقه‌بندی شدند. مرکز کنترل و پیشگیری از بیماری چین (CDC) و مرکز کنترل و پیشگیری‌های محلی یک برنامه پژوهش گسترده شیوع بیماری سازمان‌دهی کردند. اتیولوژی این بیماری اکنون به ویروس جدیدی مربوط به خانواده کروناویروس COVID-19 نسبت داده‌شده است. در ۱۱ فوریه سال ۲۰۲۰ ، مدیرکل WHO ، دکتر Tedros Adhanom Gbrebreyesus، اعلام کرد که بیماری ناشی از این کروناویروس جدید یک "COVID-19" است ، که مخفف "بیماری کروناویروس ۲۰۱۹" است.

با توجه به مطالعات و شواهد موجود ویروس COVID-19 بسیار مسری بوده و سرعت انتشار بالایی در سطح جهانی دارد. در جلسه قوانین سلامت بین‌الملل (IHR.2005) در تاریخ ۳۰ ژانویه ۲۰۲۰ شیوع بیماری از طریق WHO به‌عنوان یک اورژانس نگران‌کننده سلامت عمومی اعلام گردید چراکه آن به ۱۸ کشور از طریق ۴ کشور از طریق تماس انسان به انسان منتشرشده بود (۱).

سالانه در سرتاسر کشورهای جهان انتشارات علمی فراوانی توسط استادان، پژوهشگران و دانشجویان تحصیلات تکمیلی دانشگاه‌ها و سازمان‌های پژوهشی مختلف منتشر می‌شود که انتشارات علمی مذکور دربرگیرنده یافته‌های مهم و کاربردی بوده و می‌توانند در پیشرفت و توسعه ملی و بین‌المللی نقش بسزایی ایفاء کنند. اثربخشی قابل‌توجهی که انتشارات علمی بر پیشرفت و توسعه کشورها دارند، ضرورت تحلیل و طبقه‌بندی انتشارات علمی را بیش‌ازپیش آشکار می‌سازد؛ به‌گونه‌ای که بتوان بین انواع مدارک نمایه شده در پایگاه‌های استنادی بین‌المللی (ازلحاظ کیفیت محتوایی) به بررسی و متن‌کاوی پرداخت. تولیدات علمی و پروانه ثبت اختراعات، دانش پیشرفته اکتشافات علمی و همچنین توسعه‌های فنّاورانه را ثبت می‌کنند، بنابراین، تحلیل و تطبیق وضعیت آن‌ها موجب شناسایی شکاف‌های علمی شده و فرصت‌های فنّاورانه بالقوه‌ای

را بررسی می‌نماید (۲). با توجه به اینکه انواع انتشارات علمی معتبر نمایه شده در پایگاه‌های استنادی شاخه‌ای از پیشینه پژوهش بوده و نتایج علمی قلمروهای گوناگون موضوعی را شرح داده و ابزاری به‌منظور بررسی مسائل پژوهشی و تبادلات علمی را ارائه می‌نماید. ازآنجاکه آخرین اطلاعات فناوری‌ها اغلب در مقالات علمی موردبحث و تبادل‌نظر قرار می‌گیرد، انتشارات علمی یک حامل مهم اطلاعات در مورد فناوری و یک منبع مهم داده برای مطالعه توسعه و تغییر فناوری است (۳). با افزایش روزافزون کمیت تولیدات علمی، شناسایی موضوعات و پژوهش‌های مطرح و مهم علمی یک حوزه خاص از ضرورت‌های علمی است، همچنین با این حجم عظیم از مقالات منتشرشده در سراسر جهان، ارزیابی و بررسی تک‌تک مقالات تقریباً کاری غیرممکن یا بسیار سخت و زمان‌بر است، لذا در اینجا استخراج خودکار دانش از این حجم عظیم مهم و ضروری می‌نماید. بنابراین تجزیه‌وتحلیل این منابع می‌تواند از اهمیت بسرایی برخوردار باشد. ارائه ابزارهایی که با بررسی متون بتواند تحلیلی روی این‌ها انجام دهند منجر به شکل‌گیری حوزه‌ی متن‌کاوی شده است. این حوزه تمامی فعالیت‌هایی که به‌نوعی به دنبال استخراج دانش از متن هستند را شامل می‌گردد (۴). همچنین تشخیص روندهای در حال ظهور[1] یک مسئله و مبحث مهم تجزیه‌وتحلیل متن است. شناسایی روندهای نوظهور معمولاً به‌عنوان "زمینه‌های موضوعی‌ای شناخته می‌شود که در طول زمان در حال افزایش علاقه و کاربرد هستند". یکی از وظایف بسیار مهم برای شناسایی روندهای نوظهور ،پیدا کردن روند در حال ظهور پژوهش در مجموعه‌ای از مقالات علمی است (۵). بررسی و شناسایی مباحث و موضوعاتی که اخیراً در یک حوزه علمی موردتوجه قرارگرفته است، برای پژوهشگران از اهمیت ویژه‌ای برخوردار است و سودمندی فراوانی برای آن‌ها دارد. بررسی دستی کلیه مقالات یک حوزه علمی جهت شناسایی موضوعات نوظهور کاری وقت‌گیر و تقریباً غیرممکن است. در این شرایط، تشخیص خودکار روندهای پژوهشی نوظهور می‌تواند پژوهشگران را در درک سریع وقایع و گرایش‌های یک حوزه موضوع یاری رساند، از ین رو استفاده از فنون متن‌کاوی و روش‌های خوشه‌بندی متن می‌تواند در جهت شناسایی روندهای نوظهور انتشارات علمی مورداستفاده قرارگرفته و راهگشا باشد (۶).

بیان مسئله

بررسی و تحلیل متون علمی منتشرشده حوزه‌های موضوعی گوناگون از اهمیت زیادی برای سازمان‌ها، پژوهشگران و سیاست‌گذاران علمی برخوردار است و با توجه به روند رشد سریع تولیدات علمی، استخراج دانش و بررسی دستی چنین

[1] Emerging trend detection

حجم عظیمی از متون علمی امری محال دست‌نیافتنی است. بنابراین راه‌حلی که می‌تواند برای دستیابی به این مشکل مورداستفاده قرار گیرد؛ شناسایی موضوعات (مدل‌سازی موضوعی) و تحلیل کلمات کلیدی انتشارات علمی است. زیرا مدل‌سازی موضوعی و تحلیل کلیدواژه‌های تولیدات علمی موضوع آن‌ها را مشخص می‌کند.

به‌منظور بررسی و درک انتشارات علمی حوزه‌های موضوعی که به‌سرعت در حال رشد هستند؛ استخراج خودکار داده‌های متنی و استفاده از فنون متن‌کاوی ضروری است. متن‌کاوی اکتشاف و استخراج دانش جذاب[۲] و غیر بدیهی[۳] از متن آزاد یا غیر ساختارمند است (۷). متن‌کاوی تمام فعالیت‌هایی که به‌نوعی به دنبال استخراج دانش از متن هستند را شامل می‌گردد(۸). مطالعات انجام‌شده نشان می‌دهد که متن‌کاوی یکی از رشته‌های در حال پیشرفت که در چندین رشته پژوهش‌های زبان‌شناسی محاسباتی[۴]، بازیابی اطلاعات[۵] و داده‌کاوی مورداستفاده قرار می‌گیرد. روش‌های کشف دانش در ابتدا در مورد داده‌های ساختاریافته به کار گرفته شدند و علمی به نام داده‌کاوی[۶] را به وجود آوردند(۹). فنون متن‌کاوی متشکل از زیرمجموعه‌های داده‌کاوی است که هدف آن استخراج دانش و کشف اطلاعات جدید و از پیش ناشناخته، به‌وسیله استخراج خودکار اطلاعات از داده‌های متنی بدون ساختار یا نیمه ساختاری است و کاربردهای گسترده‌ای در تجزیه‌وتحلیل و پردازش مدارک متنی دارد(۱۰،۴) و شناسایی الگوهای و استخراج دانش بالقوه در حجم زیادی داده‌های متنی، یک امر مهم در زمینه‌های علمی مختلف محسوب می‌شود. متن‌کاوی به تحلیل هوشمند متن، داده‌کاوی متنی یا کشف دانش از متن نیز مشهور است. به‌طورکلی به فرایند استخراج دانش و اطلاعات موردعلاقه و مهم از مجموعه متنی غیر ساختاریافته اشاره دارد(۱۱-۱۴). از مهم‌ترین مزایای متن‌کاوی می‌توان به موارد زیر اشاره نمود:

۱- کمک کاملاً مؤثر به استخراج اطلاعات مفید و سودمند از حجم زیادی داده، در مدت‌زمان کوتاه

۲- کمک به پیش‌بینی جنبه‌های آینده بر اساس مهیاکردن آمار و ارقام و مشاهدات

۳- کمک به ساخت و ایجاد الگوهایی از داده‌های در دست، که اطلاعاتی در مورد افزایش و یا کاهش روند (برای مثال در تجارت و اقتصاد) در اختیار قرار می‌دهد

۴- نرم‌افزارهای متن‌کاوی همچنین به سازمان‌های امنیتی، به‌وسیله‌ی مشاهده و آنالیز اطلاعات متنی به‌دست‌آمده از منابع اینترنتی کمک می‌کنند

[2] Interesting
[3] Non-trivial
[4] computational linguistics
[5] Information Retrieval
[6] data mining

دیگر مزیت مربوط به تکنیک‌های متن‌کاوی، به استفاده‌ی آن‌ها در پایگاه داده‌های پزشکی برمی‌گردد، که به پژوهش در ادبیات پزشکی کمک بسیار می‌کند. روش‌های متن‌کاوی، آنالیز، ذخیره‌سازی و در دسترس‌پذیری اطلاعات بر روی وب‌سایت‌های مختلف و موتورهای جست‌وجو را برای پردازش و جست‌وجوی مؤثرتر و با دقت بیشتر، را توسعه می‌دهند. متن‌کاوی همچنین آنالیز لغوی و تشخیص الگو را انجام داده و به مطالعه‌ی توزیع فرکانس کلمات کمک می‌کند (۱۵).

به‌عنوان یک جنبه از روند پژوهش‌های متن‌کاوی، می‌توان به زمینه‌های مختلفی ازجمله اطلاعات مقالات دانشگاهی و اطلاعات مقالات خبری، پژوهش‌هایی با استفاده از فنون استخراج متن که به‌طور فعال انجام می‌شود و استخراج اطلاعات ضمنی از حجم زیادی از داده‌ها اشاره نمود(۱۵, ۱۶). دراین‌باره سالوم [7] و همکاران (۲۰۱۷) نیز تأکید می‌کنند که فنون متن‌کاوی نقش مهمی در تبدیل متن بدون ساختار به دانش اطلاعاتی بازی می‌کنند (۹). همچنین سالوم و همکاران (۲۰۱۸)اهداف پژوهش‌ها متن‌کاوی به شرح زیر بیان می‌کند ۱) استفاده از فنون متن‌کاوی برای شناسایی موضوعات متون علمی و سیر تکاملی این موضوعات ۲) استفاده از ابزارهای بصری سازی[8] برای ارائه هر موضوع و ارتباط میان آن‌ها به‌عنوان روشی مناسب جهت کمک به کاربران برای تعیین موضوعات مربوطه(۱۷).

مدل‌سازی موضوعی نیز یک روش آماری (۱۸) و متن‌کاوی است (۱۹) که به بررسی اسناد برای شناسایی مضامین یا موضوعات آن‌ها می‌پردازد. از نتایج این الگوریتم می‌توان برای تحلیل چگونگی ارتباط مباحث با یکدیگر و چگونگی تکامل آن‌ها باگذشت زمان استفاده کرد (۱۱–۱۳) مدل‌های موضوعی مبتنی بر این ایده هستند که اسناد از مجموعه موضوعات تشکیل می‌شوند، آنجا که یک موضوع به‌عنوان توزیع احتمال بر روی کلمات تعریف می‌شود (۲۰). همچنین تکنیک‌هایی که مباحث و اصطلاحات مرتبط را کشف می‌کند ، مانند مدل‌سازی موضوع ، و ارزیابی‌هایی که اهمیت اصطلاحات موضوعات داده‌شده را باگذشت زمان ردیابی می‌کند، به‌طور بالقوه می‌تواند به تحلیلگران اجازه دهد تا ارتباط و تغییر موضوعات داده‌شده را بهتر درک کنند(۲۱– ۲۳) و کاربرد مدل‌سازی موضوعی پتانسیل بالایی در سیاست‌گذاری‌های پژوهشی و راهبردی دارد (۲۴).

بر همین اساس این پژوهش انتشارات علمی نمایه شده کروناویروس Web of Science Core Collection در پنجاه سال اخیر را بررسی و تحلیل کرده و به شناسایی ساختار موضوعی و محتوایی انتشارات علمی قلمرو موضوعی مذکور پرداخته است، همچنین شناسایی تکامل موضوعی تولیدات علمی باگذشت زمان و شناسایی روندهای نوظهور بر

[7] Salloum
[8] Visualization

اساس محتوای آن‌ها به‌منظور ارائه تصویری روشن از موضوعات پژوهش‌های منتشرشده کروناویروس مسئله پژوهش حاضر بوده و اجرای آن با در قلمرو موضوعی کروناویروس در نیم‌قرن اخیر با روش‌های یادشده بیش‌ازپیش ضروری به نظر می‌رسد.

ضرورت پژوهش

با توجه به شرایط حال حاضر دنیا، در خصوص شیوع بحران همه‌گیری کروناویروس جدید، پژوهشگران حوزه‌های مختلف علمی به دنبال انجام پژوهش‌های در خصوص شیوه‌های پیشگیری، درمان و نیز ساخت واکسن و انواع داروهای مؤثر جهت مقابله و از بین‌بردن این همه‌گیری هستند و نتایج مطالعات آنان در مجلات معتبر منتشر می شود. همچنین با توجه به اینکه اولویت‌های پژوهشی سازمان‌ها و مؤسسات پژوهشی به سمت موضوعات مختلف این بحران هست، روزانه نیز مطالعات و مقالات متعددی در این زمینه منتشر می‌گردد و همواره نیز بر تعداد آن‌ها افزوده خواهد شد. بنابراین تحلیل و ارزیابی مدارک منتشرشده مرتبط با کروناویروس و کووید-۱۹ با استفاده از فنون متن‌کاوی از اهمیت ویژه‌ای برای پژوهشگران، سیاست‌گذاران و برنامه‌ریزان علوم پزشکی در سطح ملی و بین‌المللی برخوردار بوده و ضرورت انجام چنین پژوهشی را بیش‌ازپیش آشکار می‌سازد .

متن‌کاوی یکی از روش‌های ارزشمند استخراج خودکار دانش از میان حجم انبوهی از داده‌های متنی فراهم می‌کند، یکی از الگوریتم‌ها مهم و مفید متن‌کاوی خوشه‌بندی موضوعی یا مدل‌سازی موضوعی است. با انجام عملیات خوشه‌بندی، حیطه گسترده‌ای از داده‌های پراکنده در گروه‌های مدون و سازمان‌یافته قرار می‌گیرند. گروه‌های متعدد ایجادشده با برخورداری از ویژگی‌های مشترک درون هر گروه دارای ارتباط ارگانیک و ساختاری با یکدیگر هستند. همچنین با انجام روش خوشه‌بندی، مقالات درون خوشه‌های واحد قرار می‌گیرند به‌گونه‌ای که مقالات درون هر خوشه دارای حداکثر شباهت با یکدیگر و حداقل شباهت با دیگر خوشه‌ها هستند. بدون شک برای پژوهشگران و مؤسسات پژوهشی این موضوع دارای اهمیت است که بدانند درزمینهٔ علمی کروناویروس چه پژوهش‌هایی و در چه موضوعاتی صورت گرفته است و چه روندی را طی کرده است. نتایج این پژوهش برای پژوهشگران این حوزه در جهت انتخاب موضوعات پژوهش و انجام پژوهش‌های آتی بر اساس اولویت بسیار مفید خواهد بود و می‌تواند از انتخاب پژوهش‌های موازی و هدر رفت زمان و بودجه خودداری نماید. همچنین برای سیاست‌گذاران این حوزه در جهت تعیین اولویت‌های پژوهشی مفید و کارآمد واقع شود.

اهداف و پرسش‌های پژوهش

الف) هدف اصلی پژوهش

هدف اصلی پژوهش شناسایی مدل های موضوعی انتشارات جهانی کروناویروس با روش متن کاوی در ۵۰ سال اخیر است.

ب) اهداف اختصاصی پژوهش

۱- شناسایی مهمترین واژگان[9] بکار گرفته شده در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم؛

۲- شناسایی روند تغییرات واژگان بکار گرفته شده در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم؛

۳- مدل‌سازی موضوعی در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم؛

۴- شناسایی روند تغییرات موضوعی در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم در طول زمان،

۵- شناسایی موضوعات نوظهور در در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم؛

ج) هدف کاربردی

دانشمندان علوم پزشکی متخصص در زمینه ویروس شناسی یا بیماری های واگیردار با استفاده از نتایج از روند تغییرات موضوعی کرونا ویروس و مهمترین انتشارات هر موضوع در ۵۰ سال اخیر آگاه شده‌اند.

د) محل کاربست نتایج طرح:

تمامی متخصصان و پژوهشگران علوم پزشکی می توانند از نتایج این طرح بهره مند شده‌اند.

ه) پرسش‌های پژوهش

- مهمترین واژگان[10] بکار گرفته شده در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم کدامند؟

[9] keywords
[10] keywords

۲- روند تغییرات واژگان بکار گرفته شده در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم به چه صورت هست؟

۳- موضوعات به کار گرفته شده (مدلسازی موضوعی) در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم چگونه است؟

۴- تغییرات موضوعی در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم در طول زمان به چه صورت است؟

۵- موضوعات نوظهور در در انتشارات علمی جهانی کرونا ویروس در ۵۰ ساله اخیر در وب گاه علم چیستند؟

تعریف واژگان (تعریف نظری و عملیاتی)

متن‌کاوی

تعریف مفهومی: متن‌کاوی فنی میان‌رشته‌ای است که این امکان را فراهم می‌کند تا بتوان از طریق شناسایی و اکتشاف الگوهایی در داده‌های متنی، زمینه‌ی استفاده‌ی مفید از این داده را فراهم نمود(۴۰). همچنین متن‌کاوی می‌تواند در جهت درک بهتر اطلاعات موجود در اسناد به کار گرفته شود. درواقع متن‌کاوی امکان بازیابی و درک اطلاعات پنهان در متن‌ها را فراهم می‌کند و برای کشف ساختار، الگوها و دانش در مجموعه‌های متنی بزرگ بکار می‌رود (۴۱).

تعریف عملیاتی: در این پژوهش از فنون متن کاوی جهت دست یابی به اهداف پژوهش استفاده می شود.

خوشه‌بندی

تعریف مفهومی: به فرایندی اطلاق می‌شود که در آن به سازمان‌دهی مفاهیم در قالب گروه‌های مختلف پرداخته می‌شود، یا به‌عبارت‌دیگر گروه‌بندی داده‌ها یا متن‌ها در قالب زیرگروه‌های به نام خوشه می‌باشد. خوشه‌بندی ابزار مهمی در متن‌کاوی به‌منظور تشخیص توزیع داده‌ها و الگوهای موجود درون آن‌هاست. هدف از خوشه‌بندی، که به‌عنوان تجزیه و تحلیل خوشه نیز شناخته می‌شود کشف گروه‌بندی طبیعی (خوشه‌بندی) مجموعه‌ای از نقاط، الگوها و اشیاء است(۴۲).

تعریف عملیاتی: در این پژوهش به منظور دسته بندی موضوعی مقالات منتشر شده حوزه ویروس کرونا از الگوریتم های خوشه بندی استفاده می شود.

مدل سازی موضوعی:

تعریف مفهومی: مدل سازی موضوعی به عنوان یک ابزار متن کاوی برای پردازش ، سازماندهی ، مدیریت و استخراج دانش از مقدار زیادی از داده های متنی موجود در پایگاه های داده مختلف عمل می کند(۴۳). الگوریتم های مدل سازی موضوع از تکنیک های آماری تشکیل شده است که هدف آنها توصیف موضوعات مورد بحث در اسناد مربوط به مجموعه مستندهای خاص است. این الگوریتم ها مبتنی بر تجزیه و تحلیل آماری کلمات موجود در اسناد ، شناسایی خوشه های کلمات ، روابط بین آن مباحث و تکامل آنها در طول زمان هستند(۴۴). همچنین الگوریتم تخصیص پنهان دیریکله ، یک تکنیک مدل سازی موضوعی است که به طور گسترده مورد استفاده قرار می گیرد(۴۳). براساس این ایده، که یک سند می تواند ترکیبی از تعداد محدودی از موضوعات در نظر گرفته شود ، و هر کلمه در آن سند می تواند با یک موضوع خاص در ارتباط باشد. بنابراین ، LDA می تواند برای شناسایی مجموعه ای از مباحث ، مرتبط کردن مجموعه ای از کلمات با یک موضوع و تعیین ترکیب موضوعات در هر سند مورد استفاده قرار گیرد(۴۵).

تعریف عملیاتی: در این پژوهش از الگوریتم مدل سازی موضوعی تخصیص پنهان دیریکله جهت بدست آوردن خوشه های موضوعی استفاده شد.

روند های نوظهور

تعریف مفهومی: زمینه های موضوعی ای که در طول زمان در حال افزایش علاقه و کاربرد هستند(۴۶).

تعریف عملیاتی: شناسایی موضوعاتی که در سال های اخیر، بیشتر مورد علاقه پژوهشگران و نویسندگان مقالات بوده است.

فصل دوم

مبانی نظری و مرورپیشینه پژوهش

مقدمه

در این فصل به بیان مبانی نظری و پیشینه پژوهش پرداخته می‌شود. به‌عبارت بهتر، ابتدا مبانی نظری پژوهش در مباحث داده‌کاوی و متن‌کاوی پرداخته شده است. سپس، پژوهش‌های مرتبط با موضوع پژوهش حاضر مرور شده است.

مبانی نظری

داده‌کاوی

پیشرفت‌های به‌وجود آمده در جمع‌آوری داده‌ها و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از اطلاعات روبرو شویم. داده‌کاوی کوششی برای به‌دست‌آوردن اطلاعات مفید از میان این داده‌هاست و رشد بی‌رویهٔ داده‌ها در سطح جهان اهمیت داده‌کاوی را دوچندان کرده است. تکنولوژی مدیریت پایگاه داده‌های پیشرفته انواع مختلفی از داده‌ها را می‌تواند در خود جای دهد، درنتیجه تکنیک‌های آماری و ابزار مدیریت سنتی برای آنالیز این داده‌ها کافی نیست و استخراج دانش از این مقدار حجیم یک چالش بزرگ تلقی می‌شود.

برای داده‌کاوی تعاریف متعددی وجود دارد. برخی از این تعاریف عبارت‌اند از: فرایند به‌خدمت گرفتن یک روش‌شناسی رایانه‌ای که با استفاده از فنون مختلف دانش را مستقیم از داده‌ها استخراج می‌کند (۴۷). داده‌کاوی جستجویی است برای اطلاعات جدید و نوین از میان مقادیر بزرگ داده‌ها و فرآیندی مشارکتی میان انسان و کامپیوتر (۱۸). داده‌کاوی فرایند اکتشاف و تحلیل به‌وسیله ابزار خودکار و نیمه‌خودکار مقادیر زیاد داده‌ها به‌منظور اکتشاف الگوهای معنی‌دار و قواعد است (۲۳).

اصولاً فناوری داده‌کاوی، پایگاه‌های داده‌ی بزرگ را به‌عنوان منبع بالقوه‌ای از دانش ارزشمند برای تصمیم‌گیری در نظر می‌گیرد. در فرایند داده‌کاوی بهترین نتیجه زمانی حاصل می‌شود که دانش یک فرد خبره درخصوص یک مسئله با توانایی‌های کامپیوتر ترکیب شود (۴). فقط در این صورت است که می‌توان بدون نفی توانایی‌های فرد خبره، سیستم‌های کامپیوتری را در خدمت نیرومندتر ساختن او قرارداد.

اهداف داده‌کاوی

عملاً دو هدف اساسی فناوری داده‌کاوی، پیش‌بینی[11] و توصیف[12] است. در عملیات پیش‌بینی بعضی از متغیرها یا حوزه‌هایی از مجموعه‌های داده به‌منظور پیش‌بینی ارزش ناشناخته یا ارزش آینده داده‌های دیگر مورد استفاده قرار می‌گیرد. از سوی دیگر تشریح، بر یافتن الگوهای تشریحی داده‌ها که می‌توانند به‌وسیله انسان تعبیر شوند، تمرکز می‌کند. در نتیجه داده‌کاوی را می‌توان در یکی از دو گروه زیر جای داد:

- **داده‌کاوی پیش‌بینی‌کننده**: این روش با استفاده از مجموعه داده‌ها، مدل‌هایی را برای توضیح سیستم تولید می‌کند که با استفاده از آن‌ها می‌توان عملکرد متغیرهای مختلف را پیش‌بینی کرد.
- **داده‌کاوی توصیفی**: اطلاعات جدید را براساس مجموعه‌های داده در دسترس تولید می‌کند که الگوهای رفتاری متغیرها را تشریح می‌کند.

هدف از داده‌کاوی پیش‌بینی‌کننده تولید مدلی است که با استفاده از یک کد اجرایی وظایفی چون پیش‌بینی، دسته‌بندی، تخمین مقدار، تخمین عملکرد و غیره را انجام دهد. هدف از داده‌کاوی تشریحی دستیابی به درکی کامل از سیستم تحت بررسی با استفاده از الگوهای پنهان در آن و روابط درون مجموعه‌های داده (۱۳).

ریشه‌های داده‌کاوی

پایه‌های اصلی داده‌کاوی بر دو اصل آمار و یادگیری ماشین[13] استوار است. آمار نیز ریشه در ریاضیات و منطق داشته و بنابراین داده‌کاوی نیز علاوه بر آمار ریشه در این دو علم دارد. در مقابل یادگیری ماشینی نیز علمی کامپیوتری است که اصول آن‌را در هوش مصنوعی می‌توان یافت. تضادی که در اینجا رخ می‌نماید این است که آمار به دلیل طبیعت ریاضی خود متمایل به فرموله کردن مسائل و مدل‌سازی است، درحالی‌که یادگیری ماشینی مسائل را با استفاده از الگوریتم‌ها حل می‌کند. اینجاست که باید نسبت به ترکیب این دو علم برای استفاده آن‌ها در داده‌کاوی اقدام کرد. داده‌کاوی رویه‌های تحلیلی را در زمینه‌های آمار، ریاضیات، تجارت و نظریه اقتصاد پیوند می‌زند. داده‌کاوی علاوه بر علوم فوق به‌خاطر استفاده از اصول اساسی مدل‌سازی از نظریه

[11] Prediction
[12] Description
[13] Machine Learning

کنترل نیز استفاده می‌کند. این نظریه عموماً در سیستم‌های مهندسی و فرآیندهای صنعتی مورداستفاده قرار می‌گیرد.

بنابراین داده‌کاوی یک فناوری میان‌رشته‌ای است و برای استفاده مؤثر از آن باید از علوم تشکیل‌دهنده آن شناخت کافی داشت. البته زمانی که بخواهیم از داده‌کاوی برای مقاصد نوآورانه و خلاقانه‌تر استفاده کنیم، نیاز به این شناخت به‌مراتب عمیق‌تر می‌شود.

با وجود ارتباط میان داده‌کاوی و آمار، تفاوت‌های اساسی میان این دو علم وجود دارد. آمار، علمی تأییدی[14] است؛ یعنی کوشش دارد مفروضاتی را با استفاده از فنون مختلف تصدیق یا رد کند، درحالی‌که داده‌کاوی یک علم اکتشافی[15] است، بدین معنی که سعی به کشف الگوهای دانشی از داده‌های موجود دارد. از سوی دیگر آمار استنتاجی از نمونه‌های کوچک و بسط آن‌ها به جامعه استفاده می‌کند و ماهیتاً توان پردازش نمونه‌های بزرگ را ندارد در حالی‌که در داده‌کاوی از نمونه‌های بسیار بزرگ و حتی خود جامعه استفاده می‌شود زیرا این فناوری از روش‌های پیشرفته کامپیوتری استفاده می‌کند که توان پردازش بالایی را در اختیار آن قرار می‌دهد و نهایتاً آمار فقط می‌تواند نمونه را به جامعه‌ای که از آن انتخاب شده بسط دهد، درحالی‌که در داده‌کاوی نمونه‌ها به دسته‌ای از جوامع بسط داده می‌شود (۱۵).

کاربردهای داده‌کاوی

تاریخچه کشف دانش در پایگاه‌های اطلاعاتی که امروزه به داده‌کاوی مشهور است، قدمت چندانی ندارد. در اوایل دهه ۱۹۹۱، هنگامی‌که اصطلاح کشف دانش در پایگاه‌های اطلاعاتی برای نخستین‌بار مطرح شد، هجمه‌ای همگانی به‌سمت طراحی الگوریتم‌های داده‌کاوی صورت پذیرفت (۲۲) و این زمانی بود که شرکت‌ها اقدام به ذخیره‌سازی مقادیر عظیم داده‌ها کرده و به دنبال روش‌هایی برای بهره‌وری از این انبار داده‌ها بودند (۹). باتوجه‌به توان تحلیل بالای فناوری داده‌کاوی و با وجود قدرت پردازش بی‌نظیر آن، از این فناوری می‌توان برای حل مسائل بی‌شماری در دنیای واقعی استفاده کرد.

برخی از کاربردهای داده‌کاوی عبارت‌اند از:

[14] Confirmatory
[15] Exploratory

- تحلیل رفتار افراد و گروه‌ها (۲۲)؛

- پردازش اطلاعات پزشکی (۳۰)؛

- تشخیص الگوهای رفتاری مصرف‌کنندگان (۱۷)؛

- یافتن پروتئین‌های مختلف از نقشه ژنی[16] موجودات زنده (۱۲)؛

- هوشمندی کسب‌وکار و کاهش ابهامات ناشی از محیط (۱۴)؛

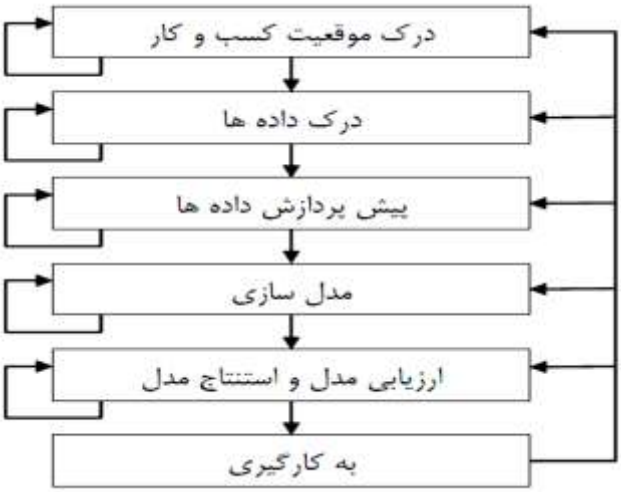
- مبارزه با جرم و جنایت و تشخیص الگوهای رفتاری گروه‌های تروریستی (۳۳) و

- بهینه‌سازی تصمیمات و تخمین‌ها در بازارهای مالی (۳۲).

فرایند داده‌کاوی

باتوجه‌به این امر که داده‌کاوی فرایند اکتشاف مدل‌های گوناگون، خلاصه‌ها و ارزش‌های نشئت‌گرفته از مجموعه خاصی از داده‌ها است، برای پیاده‌سازی چنین فرآیندی باید از روش‌شناسی خاصی استفاده کرد. در این راستا روش‌شناسی فرایند استاندارد میان صنعتی داده‌کاوی[17] (CRISP) به‌وسیله تحلیل نمایندگی‌های دایملر کرایسلر ایجاد شد (۴۰) این متدولوژی توانمند و منعطف جهت ارتقای شایستگی داده‌کاوی در حل مسائل سازمانی است. براساس این روش، یک پروژه داده‌کاوی مبتنی بر چرخه عمر و متشکل از شش گام است و این گام‌ها به‌صورت مستمر و تکراری در تمام فرایند داده‌کاوی به کار گرفته می‌شود. گام‌های روش‌شناسی داده‌کاوی CRISP به شرح زیر است (۴۰؛ ۱۸):

[16] Genomic Map
[17] Cross Industry Standard Process



شکل ۱. مراحل داده‌کاوی CRISP

متن‌کاوی

پیشرفت پیوسته در فناوری باعث افزایش شدیدی در حجم اطلاعات، خصوصاً در ادبیات علمی و تکنیکی شده است. در نتیجه، پیگیری اطلاعات یک وظیفهٔ چالش‌برانگیز برای دانشمندان شده است؛ بنابراین، راهبردهای متن‌کاوی برای کمک‌کردن به دانشمندان در زمینهٔ کسب اطلاعات سودمند از حجم بسیار زیادی از اطلاعات، موردنیاز است. متن‌کاوی و یا کشف دانش از متن اشاره به فرآیندی می‌کند که موجب به‌دست‌آوردن الگوهای غیربدیهی، جالب و با کیفیت بالا و همچنین اطلاعات و دانش از اسناد متنی ساختارنیافته می‌شود. متن‌کاوی (که به‌عنوان داده‌کاوی متن[18] نیز شناخته می‌شود) به جست‌وجو در میان داده‌های متنی برای استخراج اطلاعات مفید می‌پردازد که معمولاً طبیعتی ساختارنیافته دارند (۳۱). هدف اولیه متن‌کاوی، بازیابی اطلاعات از متون ساخت‌نیافته و همچنین ارائه دانش به‌صورت خالص برای کاربران در یک شکل چکیده است (۲۹). متن‌کاوی هدفش قادر ساختن، استفاده‌کنندگان برای جمع‌آوری، ذخیره، تفسیر و کشف دانش مورد نیاز برای تحقیق و آموزش مؤثر و نظام‌مند است. متن‌کاوی شامل سه فعالیت بزرگ است: بازیابی اطلاعات که بازیابی متون مربوط به سؤال استفاده‌کنندگان است؛ خلاصه اطلاعات که شناختن و استخراج نکات ریز متون که مربوط به سؤال هستند؛ داده‌کاوی که رابطه مستقیم یا غیرمستقیم بین قسمت‌های اطلاعات استخراجی از متون را پیدا می‌کند. متن‌کاوی شاخه‌ای از داده‌کاوی یا همان کشف دانش است (۳۳).

[18] Text Data Mining

هدف متن‌کاوی، کشف اطلاعات از قبل ناشناخته است که هنوز کسی نمی‌داند و بنابراین مستند نشده است (۱۹). ازآنجاکه بسیاری از اطلاعات به شکل متن ذخیره شده‌اند، متن‌کاوی، ارزش اقتصادی بسیار بالایی در پی خواهد داشت. دانش ممکن است از منابع گوناگون اطلاعاتی به‌دست‌آمده باشد، اما متون ساختارنیافته، بیشترین منابع دانش در دسترس را تشکیل می‌دهند. مسئله کشف دانش از متون، استخراج مفاهیم صریح و نیز غیرصریح و روابط معنایی میان مفاهیم با استفاده از فنون پردازش زبان طبیعی است. هدف استخراج دانش، به‌دست‌آوردن بصیرت‌هایی درباره داده‌های متنی عظیم است. کشف دانش از متن ریشه در پردازش زبان طبیعی دارد؛ اما روش‌هایی از آمار، یادگیری ماشینی، استدلال استخراج اطلاعات، مدیریت دانش و دیگر رشته‌های مرتبط برای فرایند کشف خود، وام‌گرفته است. کشف دانش از متن، نقش فزاینده و مهمی در ظهور برنامه‌هایی مانند فهم متن ایفا می‌کند. متن‌کاوی با استفاده از پردازش پیچیده زبان طبیعی، کاربردپذیری کشف دانش از داده‌ها را به‌طور خیره‌کننده‌ای افزایش داده است. این بدان معناست که نیازی نیست فرایند کشف دانش از داده‌ها را تنها به آن دسته از اطلاعات موجود در پایگاه‌های ساختاریافته محدود کنیم. باتوجه‌به اینکه بیشتر اطلاعات ارزشمند برای استخراج هم‌اکنون در متون زبان طبیعی وجود دارد، پردازش زبان طبیعی می‌تواند فنون موردنیاز برای متن‌کاوی را فراهم کرده و دانش را به‌طور خودکار از این متون استخراج کند. متن‌کاوی درباره جستجوی انگاره‌ها در متن زبان طبیعی است. آن عبارت است از فرایند تحلیل متن به‌منظور استخراج اطلاعات از حجم عظیمی از متون غیر ساختاریافته. متن‌کاوی یک نوع فناوری است که امکان کشف انگاره‌ها و گرایش‌ها را به‌طور نسبتاً خودکار از درون متن غیر ساختاریافته آزاد فراهم می‌کند. همان‌طور که گفته شد، متن‌کاوی به‌عنوان امتداد طبیعی داده‌کاوی به‌شمار می‌رود. آن به منزله استفاده از همان فناوری‌های داده‌کاوی در قلمرو اطلاعات متنی به کار می‌رود. متن‌کاوی در مقایسه با داده‌کاوی فرایند پیچیده‌تری است، زیرا با آن دسته از داده‌های متنی سروکار دارد که غیر ساختاریافته هستند. متن‌کاوی یک قلمرو چندرشته‌ای است و شامل بازیابی اطلاعات، تحلیل متن، دسته‌بندی اطلاعات، طبقه‌بندی اطلاعات، دیداری‌سازی اطلاعات، فناوری پایگاه‌های اطلاعاتی، یادگیری ماشینی و داده‌کاوی است. علاوه بر این، از ابزارهای متن‌کاوی برای کمک به پژوهشگران در زمینه‌های مختلف استفاده می‌شود. به‌عنوان‌مثال، یک اخترشناس که علایم اشعه ایکس را در ناحیه‌ای از فضا کشف

می‌کند ممکن است تمایل داشته باشد که در پیشینه‌های درون خطی جستجو نماید تا ببیند که آیا تاکنون هیچ نوع اشعه مادون قرمزی در همان ناحیه کشف شده است یا خیر؟ یک زیست‌شناس که دارای فهرستی از تعداد ۱۹۹ ژن است که در مقالات مختلف شناسایی شده‌اند ممکن است تمایل داشته باشد که به سراغ پژوهش‌های منتشرشده موجود برود و در جستجوی مقالاتی باشد که درباره عملکرد این ژن‌ها توضیح داده باشند (۲۵).

متن‌کاوی که به تحلیل هوشمند متن، داده‌کاوی متنی یا کشف دانش در متن، نیز مشهور است عموماً به فرایند استخراج دانش و اطلاعات موردعلاقه و مهم از مجموعه متنی غیر ساختاریافته اشاره دارد. به‌عبارت‌دیگر، متن‌کاوی فرایند تحلیل طبیعی متن به‌منظور کشف و ثبت اطلاعات معنایی برای درونداد و ذخیره‌سازی در یک ساختار سازمان دانش است. یکی از فنون مفید متن‌کاوی دسته‌بندی است که برای کشف توزیعات و انگاره‌های داده‌های مورد علاقه در داده‌های حجیم به کار می‌رود. با استفاده از این فن می‌توان بدون اتکاء بر هیچ دانش پیش‌زمینه‌ای ساختارها یا دسته‌های موردعلاقه را به‌طور مستقیم از داده‌ها شناسایی نمود (۲۶).

رابطه متن‌کاوی و داده‌کاوی

متن‌کاوی شامل طیف گسترده‌ای از وظایف است که می‌تواند اطلاعاتی را در مورد جنبه‌های مختلف متون به ارمغان آورد. وظایف معمول استخراج متن شامل (۱۹):

- طبقه‌بندی اسناد[19] - اختصاص سند به یک یا چند دسته از پیش تعریف شده (به‌عنوان‌مثال، اختصاص مقاله روزنامه به یک یا چند دسته، برچسب‌گذاری نامه‌های الکترونیکی به‌عنوان هرزنامه)؛
- خوشه‌بندی[20] - گروه‌بندی اسناد باتوجه‌به شباهت آن‌ها، به‌عنوان‌مثال، به‌منظور شناسایی اسنادی که دارای یک موضوع مشترک هستند؛
- جمع‌بندی[21] (خلاصه‌سازی)- یافتن مهم‌ترین قسمت‌ها در یک یا چند سند و ایجاد متنی که به‌طور قابل‌توجهی کوتاه‌تر از اصل باشد؛

[19] categorization of documents
[20] clustering
[21] summarization

- بازیابی اطلاعات[۲۲] - بازیابی اسنادی که با پرس‌وجو مطابقت دارند و اطلاعات موردنیاز را از مجموعه بزرگی از اسناد نشان می‌دهد؛

- استخراج معنی[۲۳] اسناد یا قسمت‌هایی از آن‌را با شناسایی موضوعات پنهان، تجزیه‌وتحلیل احساسات، عقاید یا احساسات نشان می‌دهد؛

- استخراج اطلاعات[۲۴] - استخراج اطلاعات ساخت‌یافته مانند موجودیت‌ها، رویدادها یا روابط از متون بدون ساختار؛

- استخراج انجمن[۲۵] - یافتن ارتباط بین مفاهیم یا اصطلاحات در متون؛

- تحلیل روند[۲۶] - بررسی چگونگی تغییر مفاهیم موجود در اسناد در زمان؛

- ترجمه ماشینی[۲۷] - تبدیل متن نوشته‌شده به یک زبان به متنی به زبان دیگر.

برخی از کارهای استخراج متن بسیار شبیه به وظایف داده‌کاوی است.

داده‌کاوی فرایند خودکار یا نیمه‌اتوماتیک یافتن دانش ضمنی، قبلاً ناشناخته و بالقوه مفید در مجموعه داده‌های ذخیره شده الکترونیکی است. دانش دارای شکلی از الگوهای ساختاری در داده‌ها است که می‌تواند برای پیش‌بینی یا ارائه پاسخ در آینده نیز مورداستفاده قرار گیرد (۳۳).

داده‌کاوی شامل روش‌ها، ابزارها، الگوریتم‌ها یا مدل‌های مختلف است. جهت انجام عملیات داده‌کاوی نیاز هست که داده‌ها به شکل ساختاریافته باشند. این بدان معنی است که داده‌ها را می‌توان به‌صورت جدول مانند یک پایگاه داده رابطه‌ای نشان داد. داده‌ها به شکل مجموعه‌ای از مثال‌ها (یا نمونه‌ها، نقاط داده، مشاهدات) با مقادیر خاص ویژگی‌های آن‌ها (یا ویژگی‌ها، متغیرها، زمینه‌ها) توصیف می‌شود.

ویژگی‌ها می‌توانند انواع مختلفی داشته باشند (۳۰؛ ۲۹).

- طبقه‌ای (اسمی) [۲۸] - دامنه مجموعه‌ای گسسته از مقادیر است که در آن ترتیب معنی ندارد؛

[22] information retrieval
[23] extracting the meaning
[24] information extraction
[25] association mining
[26] trend analysis
[27] machine translation
[28] categorical (nominal)

- باینری[۲۹] - نوع خاصی از ویژگی‌های طبقه‌ای فقط با دو مقدار ممکن؛

- ترتیبی[۳۰] - دامنه یک مجموعه گسسته از مقادیر است که می‌تواند مرتب شود؛

- عددی[۳۱] - مقدار یک عدد، عدد صحیح یا پیوسته است.

متن رشته‌ای است که به زبان طبیعی متشکل از قسمت‌هایی (کلمات) با معنای خاص نوشته شده است که متون همچنین می‌توانند از دامنه متفاوتی باشند. یک واحد از متن می‌تواند یک جمله، چند جمله با هم ترکیب‌شده در یک پاراگراف یا متن‌های بسیار طولانی‌تری مانند صفحات وب، ایمیل، مقاله یا کتاب باشد. گاهی اوقات، یک متن می‌تواند فقط چند کلمه باشد که جمله معتبری نیست که کاملاً معمول است، به‌عنوان‌مثال، برای پست‌های کوتاه در شبکه‌های اجتماعی، برخی قوانین (نحو) ترکیب می‌شوند؛ بنابراین برای اینکه بتوان از روش‌های داده‌کاوی در متن استفاده کرد، باید آن‌ها را به نمایشی ساختاری تبدیل کرد.

به‌طورکلی می‌توان گفت که داده‌کاوی با داده‌های ساختاریافته و نرمال‌شده سروکار دارد و در اصطلاح داده‌کاوی با پایگاه داده‌های رابطه‌ای کار می‌کند. اما در عوض متن‌کاوی با داده‌های ساختارنیافته یا نیمه‌ساختاریافته یعنی متون موجود در مقالات، اسناد و غیره سروکار دارد. علاوه بر این دسترسی کم به ساختار در متون دلیلی دیگر بر این است که متن‌کاوی کاری بسیار مشکل است. مفاهیم موجود در متون معمولاً بسیار انتزاعی هستند و به‌سختی قابل مدل کردن می‌باشند. همچنین وقوع کلمات مترادف (کلمات متفاوت ازنظر نوشتاری اما هم‌معنی) یا کلمات با تلفظ یکسان (کلمات با تلفظ یکسان اما معنی متفاوت) پیداکردن رابطه منطقی بین بخش‌های مختلف متن را مشکل می‌کند (۴۱).

فرق داده‌کاوی با متن‌کاوی این است که در متن‌کاوی الگوها از متن زبان طبیعی استخراج می‌شوند درحالی‌که در داده‌کاوی الگوها را از پایگاه‌های داده ساخت‌یافته به دست می‌آورند. متن‌کاوی، متصل کردن اطلاعات استخراج شده به یکدیگر برای تشکیل حقایق یا فرضیه‌های جدید است تا پس از آن به کمک روش‌های متعارف آزمایش، بررسی بیشتری شوند (۲۹).

[29] binary
[30] ordinal
[31] numerical

به‌عبارت‌دیگر تفاوت میان داده‌کاوی و متن‌کاوی در این است که داده‌کاوی، اطلاعات را گردآوری و فهرست‌نویسی کرده، سپس اقدام به تولید دانش از بین حجم عظیمی از داده‌ها می‌کند اما متن‌کاوی، حوزه‌ای نو و میان‌رشته‌ای است که از رشته‌های بازیابی اطلاعات، داده‌کاوی، یادگیری ماشینی، آمار و زبان‌شناسی محاسباتی مشتق شده است و عمدتاً بر مستندات متنی تکیه دارد (۱۲).

تاریخچه متن‌کاوی

آنچه امروزه به‌طور عام متن‌کاوی نامیده می‌شود، مجموعه‌ای متشکل از علوم مختلف از قبیل زبان‌شناسی، آمار، کامپیوتر، مدیریت، هوش مصنوعی و دیگر حوزه‌ها است. اما آغازگر توسعهٔ روش‌های متن‌کاوی نیاز محققان به فهرست‌گذاری متون (به‌عنوان مثال کتاب‌های یک کتابخانه) بود. اما خیلی زود جهت این توسعه به‌سمت استخراج داده‌های متنی با استفاده از فنون پردازش زبان طبیعی تغییر پیدا کرد (ماینر و همکاران، ۲۰۱۲). در اصل باید ریشه اصلی روش‌های متن‌کاوی را در تلاش‌ها و اقدامات صورت‌گرفته برای استفاده از فنون کمّی و آماری در علم زبان‌شناسی جست که به این مجموعه اقدامات، زبان‌شناسی کمی[۳۲] می‌گویند. تاریخچه استفاده از زبان‌شناسی کمی به حداقل قرن نوزدهم میلادی بازمی‌گردد.

اما فعالیت کلاسیک تئوری گونهٔ زیف ۱۹۴۹ به‌عنوان یکی از مهم‌ترین پیش‌گامان عرصهٔ تحلیل کمی زبان‌شناسی، شناخته می‌شود (۲۲). از دههٔ هفتاد میلادی تاکنون، افزایش چشمگیری در میزان علاقه به این حوزه از علوم اطلاعات مشاهده شده است. پژوهش ویلی یکی از اولین کاربردهای این فنون در مطالعه منابع و ادبیات علمی، محسوب می‌شود (۱۱).

نخستین بار لان[۳۳] در سال ۱۹۵۸ میلادی، مفهوم متن‌کاوی را در مقاله خود مطرح نمود. سپس در سال ۱۹۶۱ نیز دویله در مقاله‌ای به متن‌کاوی و روش‌های مرتبط با آن اشاره کرد و بیان داشت که «طبقه‌بندی و سازماندهی اطلاعات» می‌تواند از تجزیه‌وتحلیل تکرار و توزیع کلمات به‌کاررفته در آن صورت پذیرد. اگر این دو مورد را به‌عنوان نخستین موارد مطرح‌شدن متن‌کاوی در نظر بگیریم، نتیجه‌گیری می‌شود که متن‌کاوی ممکن است مفهومی جدید باشد، اما رویای استخراج خودکار اطلاعات از متون، هم‌زمان با پیدایش کامپیوتر

[32] Quantitative linguistics
[33] Luhn

وجود داشته است. سوانسون[۳۴] در سال ۱۹۹۱ در مقاله خود بیان کرد که متون علمی باید به‌عنوان پدیده‌های ارزشمند طبیعی، کشف، اصلاح و تجزیه‌وتحلیل شوند و به‌این‌ترتیب در واقع نظر دانشمندان را به استفاده از اطلاعات با آنالیز هوشمندانه جلب کرد.

سوانسون با ایجاد نرم‌افزاری نخستین گام را در جهت عمل سازی متن‌کاوی طی کرد. او از این نرم‌افزار در استخراج اطلاعات مفید از متون پزشکی استفاده کرد. این نرم‌افزار ارو اسمیس نام دارد. این نرم‌افزار که به‌صورت تخصصی در ارتباط با پردازش متون پزشکی تهیه شده، قادر است پس از دریافت متون مورد بررسی، کلمات کلیدی و متون مرتبط با یکدیگر را مشخص کند. سانسون هیچ‌گاه از اصطلاح متن‌کاوی برای این نر م افزار استفاده نکرد، اما به نظر می‌رسد که این نرم‌افزار نخستین نرم‌افزار متن‌کاوی بوده است. ازاین‌رو می‌توان او را پدر متن‌کاوی مدرن نامید. لیندزی[۳۵] و گاردان[۳۶] ۱۹۹۹ کارهای سوانسون را بدون آنکه نام متن‌کاوی بر آن نهند، ادامه دادند. نرم‌افزار آن‌ها دو واژه را به‌صورت هم‌زمان در میان متون جستجو می کرد و نتیجه را در لیستی قرار می‌داد تا کاربران آن‌ها را به‌عنوان مقاله‌های تکمیلی مورد مطالعه قرار دهند. لیندزی و گاردان در همان سال روش TF-IDF[۳۷] را برای رتبه‌بندی متون و واژه‌ها به این نرم‌افزار افزودند.

هیرست[۳۸] (۱۹۹۷) نیز متن‌کاوی را دسترسی به اطلاعات(بازیابی سنتی اطلاعات) متمایز ساخته است. بازیابی سنتی اطلاعات، بیشتر روی بازیابی متون مرتبط با هم تأکید می‌کند. متونی که با نیازهای اطلاعاتی کاربر مرتبط باشد. براساس تعریف هیرست داده‌کاوی-که متن‌کاوی نیز نوعی از داده‌کاوی به‌شمار می‌آید- تنها با اطلاعات و بازیابی آن سروکار ندارد، بلکه تلاش می‌کند تا اطلاعات جدیدی را از میان داده‌ها کشف کند که پیش ازاین حتی برای ایجادکننده داده‌ها (نویسنده متن هم مشخص نبود. او معتقد بود که داده‌کاوی و متن‌کاوی «توأم با شانس»[۳۹] هستند، درحالی‌که بازیابی اطلاعات «هدف‌گرا»[۴۰] است. بنابراین کارهایی نظیر جستجوی واژه‌ها برای پاسخ به پرسش‌ها، متن‌کاوی به‌شمار نمی‌روند. او بر این باور بود که بازیابی و دستیابی

[34] Swanson
[35] Lindsay
[36] Gordon
[37] Term Frequency–Inverse Document Frequency
[38] Hearst
[39] Opportunistic
[40] Goal – Driven

به اطلاعات می‌تواند به‌عنوان یک کار تکمیلی و پشتیبان برای متن‌کاوی به‌شمار رود. نخستین محصول نرم‌افزاری (Intelligent Miner for Text) در زمینه متن‌کاوی در سال ۱۹۹۹ توسط آی بی ام[۴۱] به بازار عرضه شد.

این نرم‌افزار شامل مجموعه ابزارهایی است که به استخراج اطلاعات از متن می‌پردازند و به‌این‌ترتیب متن را غنی می‌سازد. اطلاعاتی که از محتوای متن استخراج می‌شود می‌تواند ویژگی‌های متن نظیر زبان متن، اسامی افراد، تاریخ‌ها، مقادیر پول و ... باشد. استخراج این ویژگی‌ها به‌صورت خودکار بوده و براساس فرهنگ لغت از پیش تعریف شده‌ای انجام نمی‌شود. این نرم‌افزار قادر است تا روی یک متن به تنهایی و یا یک مجموعه از متون، پردازش‌های موردنیاز را انجام دهد. در این نرم‌افزار که شمارش کلمات به‌کاررفته در متن اساس پردازش‌های بعدی بر روی محتوای متن است، قسمتی برای تشخیص اصطلاح‌ها دارد و متون را خوشه‌بندی و دسته‌بندی می‌کند (۸).

با توسعه مفهوم متن‌کاوی، مفاهیم دیگری نیز پابه‌پای آن رشد کردند: بازیابی اطلاعات از مجموعه متون، بازیابی اطلاعات از یک متن، کشف دانش از بانک‌های اطلاعاتی، مدیریت دانش در سازمان‌ها و نمایش (تصویرسازی) داده‌ها و اطلاعات، این مفاهیم توسط کاتساف[۴۲]، اراد[۴۳] و لوزیویچ[۴۴] در سال ۲۰۰۰ در چند مقاله منتشر شد تا میان این مفاهیم تمایز قایل شوند. در سال ۲۰۰۱ میلادی کاتساف و دمارکو[۴۵] متن‌کاوی را با «استخراج اطلاعات از متون فنی» تعریف کردند. براساس این تعریف، متن‌کاوی شامل سه بخش می‌شود: بازیابی اطلاعات، پردازش اطلاعات و یکپارچگی اطلاعات. پردازش اطلاعات به استخراج الگوهای موجود در متون بازیابی‌شده اطلاق می‌شود و یکپارچگی اطلاعات ترکیب هم‌افزایانه خروجی کامپیوتری پردازش اطلاعات با خواندن اطلاعات بازیابی شده توسط انسان است که مفهوم سیستم انسان- ماشین را تداعی می‌کند.

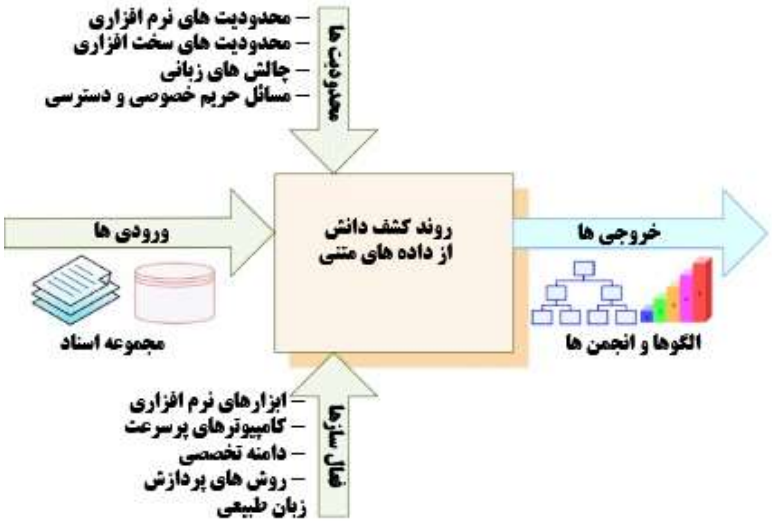
[41] IBM
[42] Kostoff
[43] Oard
[44] Losiewicz
[45] DeMarco

روش‌شناسی متن‌کاوی

متن‌کاوی فرایند نیمه‌خودکار تلقی شده که برای استخراج الگوها و کشف دانش از روی انبوهی از منابع داده‌ای غیر ساختاریافته مفید واقع می‌شود (۱۰). متن‌کاوی دارای رابطهٔ تنگاتنگی با داده‌کاوی بوده و هدف کُلی و برخی فرایندهای عملیاتی آن‌ها یکسان است. همچنین در متن‌کاوی ورودی فرایندها مجموعه‌ای از فایل‌های متنی غیر ساختاریافته بوده که از آن جمله می‌توان به فایل‌های Word، PDF، گزیده‌ای از متون و فایل‌های XML اشاره نمود. مزایای متن‌کاوی بیشتر در قلمروهایی است که انبوهی از داده‌های متنی در دست تهیه قرار می‌گیرند. از جمله مهم‌ترین این قلمروها می‌توان به بخش ادبیات/ پیشینهٔ موضوع مقالات مختلف (مورد استفادهٔ همین مقاله)، بخش مالی (گزارش‌های فصلی، تفسیرات رسانه‌ای)، خدمات بهداشتی (گزارش مرخصی بیمار، یادداشت‌های پزشک)، حقوق و مسائل قانونی (دستورات دادگاه)، زیست‌شناسی (تعاملات مولکولی)، فناوری (فایل‌های پروانه ثبت اختراع) و بازاریابی (نظرات مشتریان) اشاره نمود (۱۷).

فرایند متن‌کاوی

مطالعات متن‌کاوی جهت نائل آمدن به موفقیت مستلزم تبعیت از فرایند اصولی و کارآمدی است که با بهره‌گیری از بهترین رویه‌های کاربردی بنا نهاده می‌شود. هر پروژه‌ی متن‌کاوی به فرایند استاندارد و مورد پذیرش دست‌اندرکاران این حوزه، همچون فرایند استاندارد صنعتی Cross برای داده‌کاوی (CRISP-DM) برای اجرای وظایف محوله نیاز خواهد داشت. در سطوح بالاتر فرایند متن‌کاوی از طریق نمودار مختص به شرایط بستری نمایش داده می‌شود که یک نمونه از چنین نموداری به‌همراه اجزای تشکیل‌دهنده شامل ورودی‌ها، خروجی‌ها، کنترل‌ها (محدودیت‌ها) و سازوکارها (توانمندسازها) با پیکان‌های جهت‌دار در نمودار ۲-۲ نمایش داده شده‌اند (۱۰).



شکل ۲-فرآیند متن‌کاوی در قالب یک مدل مفهومی کلی (۱۰)

نمودار وابسته به شرایط بستری به سه مرحله/ فعالیت پی‌درپی تقسیم می‌شود. هر یک از این مراحل دارای ورودی‌های خاصی بوده که برای تولید خروجی‌های اختصاصی بکار گرفته می‌شوند. درصورتی‌که به هر دلیلی خروجی بخشی از این فرایندها با انتظارات ما سازگاری نداشته باشد، اعمال بازخورد (بازگشت در جهت رو به عقب) نسبت به وظیفه‌ی در دست اجرا ضروری خواهد بود. در شکل ۲ نمای گرافیکی از این وضعیت نمایش داده شده است.

مرحلهٔ نخست. تعیین بخش‌های منتخب از متن: هدف اصلی فعالیت نخست گردآوری تمامی اسناد مرتبط با قلمروی در دست مطالعه است. در ادامه‌ی کار اسناد گردآوری شده به فرمت خاصی تبدیل و در همان شرایط بایگانی شده تا همگی در ساختاری واحد تحت پردازش‌های رایانه‌ای متعاقب قرار بگیرند (برای مثال فایل‌های متنی در فرمت ASCII)؛

مرحلهٔ دوم. پیش‌پردازش داده‌ها (تشکیل ماتریس واژه / تفکیک برحسب سند): پس از نهایی‌سازی وضعیت متون، ماتریس واژه‌های تفکیک شده برحسب سند (TDM) براساس اطلاعات در دسترس تشکیل می‌شود. در این ماتریس سطرها نشان‌دهندهٔ اسناد و ستون‌ها بیانگر واژه‌ها (کلید واژه‌های مدنظر، هدف، جستجو شده) است. روابط برقرار شده بین واژه‌ها و اسناد به کمک شاخص‌ها نمایش داده می‌شوند (برای مثال شاخص‌های نسبی که ساده‌ترین آن‌ها تعداد تکرارپذیری هر واژه در اسناد مربوطه است)؛

مرحلهٔ سوم. استخراج دانش: هنگامی‌که به ماتریس TDM با ساختاردهی بهینه دسترسی باشد، به استخراج الگوها مبادرت ورزیده و آن‌ها به‌عنوان خوشه‌ها تعریف می‌شود. خوشه‌بندی یک فرایند غیر نظارتی بوده که طی آن اشیاء یا رخدادها در گروه‌بندی‌های نرمال (معنادار) دسته‌بندی می‌شوند. فرایند غیر نظارتی فرایندی است که از هیچ‌گونه الگو یا دانش قبلی برای هدایت فرایند خوشه‌بندی سود نخواهد بُرد. در فرایند خوشه‌بندی غیر نظارتی، مجموعه اشیاء/ مؤلفه‌های فاقد برچسب‌گذاری (اسناد، دیدگاه‌های مشتری، صفحات وب) بدون هرگونه دانش قبلی به خوشه‌های معنادار منتقل می‌شوند. فرضیهٔ بنیادی بیانگر این نکته بوده که اسناد مرتبط کاملاً شبیه یکدیگر هستند، این در حالی است که اسناد غیرمرتبط از کمترین شباهت نسبت به یکدیگر سود می‌برند. درصورتی‌که این فرضیه معتبر باشد، آنگاه خوشه‌بندی اسناد بر مبنای میزان مشابهت محتوی به ارتقای کیفیت جستجو می‌انجامد.

یافتن تعداد بهینه‌ای از خوشه‌ها به‌هیچ‌عنوان وظیفه‌ای ساده نخواهد بود. در واقع هیچ‌گونه فرمول ریاضیاتی برای انجام این عمل پیش‌بینی نشده است (یک الگوریتم با ساختار بسته). تعیین تعداد بهینه‌ای از خوشه‌ها همچنان در قلمروی فرایندهای شهودی- آزمایشی قرار گرفته و طی آن تعداد خوشه‌ها به‌تدریج از تعداد کم تا تعداد بیشتر افزایش‌یافته (یا برعکس) تا زمانی که تعداد خوشه‌های حاصله به شیوه‌ای بهینه بیانگر مجموعه داده‌ی چندبعدی در دست ارزیابی باشند (۱۹).

خوشه‌بندی

در سال‌های اخیر، به دلیل پیشرفت‌هایی که در داده‌کاوی، قدرت رایانه‌ها و بسته‌های نرم‌افزاری آماری حاوی الگوریتم‌های تحلیل خوشه، صورت‌گرفته است، تحلیل خوشه به موضوعی مهم مبدل شده است. اغلب نیاز است تا مجموعه‌ای از اسناد به گروه‌ها یا خوشه‌هایی همگن تقسیم‌بندی شوند. اگر این مجموعه حاوی تعداد اندکی مستند باشد، این کار می‌تواند به‌صورت دستی انجام گیرد. ولی اگر با حجم زیادی از اسناد سروکار داشته باشیم، انجام دستی این فرایند، زمان‌بر و غیر مؤثر، خواهد بود. داده و الگو یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. قابلیت آن در ورود به فضای داده و تشخیص ساختار آن‌ها، خوشه‌بندی را یکی از ایده‌آل‌ترین سازوکارها برای

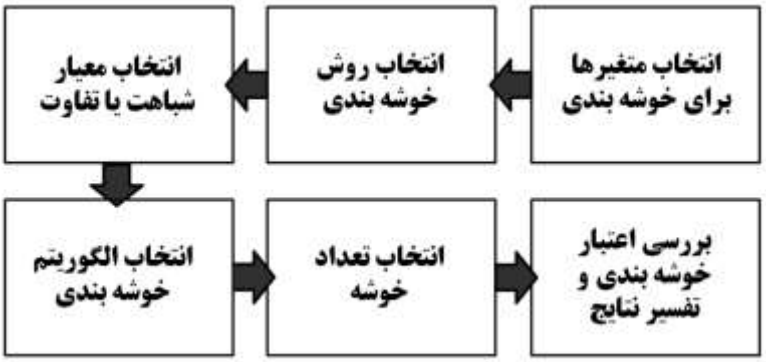
کار با دنیای عظیم داده‌ها کرده است. ایده آن نخستین‌بار در دهۀ **۱۹۳۵** ارائه شد و امروزه با پیشرفت‌ها و جهش‌های عظیمی که پدید آمده، خوشه‌بندی در کاربردها و جنبه‌های مختلفی حضور یافته است. هدف نهایی خوشه‌بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم‌بندی داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را به هم داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند. البته کیفیت نتایج خوشه‌بندی به روش اندازه‌گیری شباهت و توانایی و قدرت الگوریتم در کشف الگوهای مخفی میان داده‌ها بستگی دارد(۲۳).

خوشه‌بندی یکی از کاربردهای متن‌کاوی است. خوشه‌بندی به فرایند تقسیم مجموعه‌ای از داده‌ها (یا اشیاء) به زیر کلاس‌هایی با مفهوم خوشه اطلاق می‌شود. به‌این‌ترتیب یک خوشه شامل یک سری داده‌های مشابه است که همانند یک گروه واحد، رفتار می‌کنند (۲۳). هدف از خوشه‌بندی این است که اسناد مرتبط باید بسیار شبیه باشند از آن‌هایی که مرتبط نیستند (۲). تکنیک خوشه‌بندی تکنیکی قابل‌اعتماد است که به‌طورکلی برای تحلیل مقدار زیادی داده مورد استفاده قرار می‌گیرد. ثابت شده است که خوشه‌بندی متن یکی از مؤثرترین ابزارهای مورداستفاده برای تحلیل عنوان‌ها است (۱۱). هر مجموعه که موجودیت نامیده می‌شود، با یک خوشه مشخص می‌شود که مربوط یه یکی از عنوان‌ها در مجموعه نوشته‌هاست (۹).

خوشه‌بندی به‌عنوان یکی از روش‌های داده‌کاوی توصیفی، تکنیکی برای گروه‌بندی مشاهدات به K گروه (خوشه) مختلف است؛ به‌طوری‌که داده‌ها (که هرکدام نشان‌دهنده برداری از مقادیر کمی در یک فضای چندبعدی هستند) در هر خوشه بالاترین درجه شباهت را دارا بوده و داده‌های متعلق به خوشه‌های مختلف، دارای حداکثر درجه عدم شباهت باشند (۱۱).

خوشه‌بندی یکی از تکنیک‌هایی است که به‌طور گسترده، به‌منظور متن‌کاوی، شناسایی الگو، تحلیل صفحات وب و تحلیل بازار مورداستفاده قرار می‌گیرد. در تعریفی دیگر، خوشه‌بندی برای جداکردن یک جمعیت غیر همگن، به تعدادی از زیرگروه‌های همگن، بدون طبقه‌های از پیش تعریف شده، استفاده می‌شود. گروه‌ها با توجه به مشابهت مفاهیم و موجودیت‌های درون خوشه‌ها و براساس متغیرهایی مشخص، براساس محتوی

تحلیل، خوشه‌بندی می‌شوند (۳۳). مویی[۴۶] و سارستد[۴۷] ۲۰۱۱ مراحل شش‌گانه‌ای برای انجام فرایند خوشه‌بندی معرفی کرده‌اند که در شکل ۳. نشان داده‌شده است.



شکل۳ . مراحل شش‌گانه خوشه‌بندی (۳۳)

مراحل فرایند خوشه‌بندی

انتخاب متغیرها برای خوشه‌بندی. الگوریتم‌های خوشه‌بندی مانند بسیاری از آنالیزهای آماری برای هر نوع داده ورودی، نتایج را فارغ از اینکه داده‌ها پیش‌فرض‌های لازم را برای خوشه‌بندی دارا باشند در اختیار کاربر قرار می‌دهد؛ بنابراین انتخاب متغیرهای صحیح مرتبط برای کسب نتایج معتبر بسیار حائز اهمیت است.

انتخاب روش خوشه‌بندی. با انتخاب روش خوشه‌بندی، پژوهشگران روشی که خوشه‌ها را تشکیل می‌دهد مشخص می‌نمایند. این روش معمولاً شامل بهینه‌کردن مشخصه‌ای مانند حداقل کردن واریانس داخل خوشه‌ها یا حداکثر کردن فاصله بین خوشه‌ها است (۴۲).

انتخاب معیار شباهت یا تفاوت. گام بعد برای انجام فرایند خوشه‌بندی، انتخاب معیار تفاوت یا تشابه است تا بتوان فاصله دو مشاهده را به‌صورت یک مقدار عددی مشخص نمود (۱۸). معیارهای متفاوتی برای خوشه‌بندی مجموعه داده‌ها وجود دارد که به ماهیت داده‌ها (کمی، اسمی و غیره) وابسته است (۴۰).

[46] Mooi
[47] Sarstedt

انتخاب الگوریتم خوشه‌بندی. بسته به نوع خوشه‌بندی (سلسله مراتبی، تقسیمی و غیره) از الگوریتم‌های مختلفی همچون الگوریتم پیوند، پیوند کامل، پیوند متوسط، الگوریتم مرکز، الگوریتم K-Means و غیره استفاده می‌شود.

انتخاب تعداد خوشه. معیارهای اعتبار مختلفی برای ارزیابی خوشه‌بندی و انتخاب تعداد خوشه براساس محاسبات آماری وجود دارد. این معیارها بسته به روش خوشه‌بندی متفاوت است.

بررسی اعتبار خوشه‌بندی و تفسیر نتایج. بررسی اعتبار خوشه‌بندی برای جلوگیری از ایجاد نتایج نامعتبر و خوشه‌های غیرحقیقی لازم است. توصیه می‌شود که پژوهشگران الگوریتم‌های مختلف خوشه‌بندی را برای داده‌های خود انجام داده و نتایج حاصل را براساس معیارهای اعتبار مقایسه نمایند تا با اطمینان بیشتری خوشه‌ها را انتخاب کنند. همچنین توصیه می‌شود تا خوشه‌بندی با تعداد خوشه‌های مختلف صورت‌گرفته و نتایج مقایسه شوند. استفاده از نظرات خبرگان حوزه موردنظر نیز در بررسی اعتبار نتایج خوشه‌بندی مفید خواهد بود (۴۰).

مدل‌سازی موضوعی

در دنیای امروز، برای داشتن یک روش بهتر برای مدیریت انفجاری اسناد الکترونیکی، استفاده از روش‌ها یا ابزارهایی که به‌طور خودکار عمل سازماندهی، جستجو، ایندکس کردن و مرور مجموعه‌های عظیم را انجام می‌دهند، ضروری است. براساس پژوهش‌های کنونی در مورد یادگیری ماشین و آمارها، تکنیک‌های جدیدی برای یافتن الگوهای کلمات در مجموعه اسناد با استفاده از مدل‌های احتمالی سلسله مراتبی، توسعه‌یافته‌اند. این مدل‌ها، «مدل‌های موضوعی» نامیده می‌شوند. کشف الگوها معمولاً موضوعات اساسی که برای شکل‌دادن اسنادی همچون مدل‌های احتمالی سلسله مراتبی یکپارچه شده‌اند را منعکس می‌کند و به‌آسانی به سایر انواع داده‌ها تعمیم داده می‌شوند. مدل‌های موضوعی به‌جای تحلیل کلمات به تجزیه‌وتحلیل تصاویر، داده‌های بیولوژیکی و اطلاعات و داده‌های نظرسنجی می‌پردازند (۲۵).

اهمیت اصلی مدل‌سازی موضوعی، کشف الگوهای استفاده از کلمات و چگونگی ارتباط با اسنادی است که الگوهای مشابه را ارائه می‌دهند. به‌این‌ترتیب، ایده مدل‌های موضوعی یعنی اینکه بتواند با اسناد کار کند و این

اسناد ترکیبی از موضوعات بوده و هر موضوع احتمال توزیع کلمات است. به‌عبارت‌دیگر، مدل موضوعی یک مدل مولد اسناد است. این مدل یک‌رویه احتمالاتی ساده برای اسنادی که قرار است ایجاد شوند، فراهم می‌کند. ایجاد یک مدل جدید با انتخاب یک توزیع روی موضوعات انجام می‌گیرد. پس از آن، هر کلمه‌ای که در سند است می‌تواند یک موضوع را به‌صورت تصادفی و بسته به توزیع، انتخاب کند. سپس، یک کلمه از آن موضوع، استخراج می‌شود (۲۲). مدل‌های موضوعی علاوه بر تجزیه‌وتحلیل متن و متن‌کاوی، متکی به مجموعه فرضیات کلمات هستند که اطلاعات مربوط به ترتیب کلمات را نادیده می‌گیرند. هر سند موجود در یک مجموعه توسط هیستوگرامی که دربرگیرنده میزان وقوع کلمات است، نمایش داده می‌شود. این هیستوگرام با توزیع روی تعداد موضوعات معینی مدل‌سازی شده است که هر یک از آن‌ها یک توزیع روی کلمات موجود در واژه‌نامه اند. با درک توزیع‌ها، می‌توان از هر سند، هیستوگرامی با ابعاد بزرگ که نشان‌دهنده رتبه پایین مربوطه است را به دست آورد. مدل‌های موضوعی مختلف از جمله تحلیل معنایی پنهان[۴۸]، تحلیل معنایی پنهانی احتمالاتی[۴۹] ، تخصیص پنهان دیریکله[۵۰]، مدل موضوعی هم‌بسته[۵۱] طبقه‌بندی دقیقی در حوزه کشف مدل‌سازی موضوعی داشته‌اند (۱۳). با گذشت زمان، موضوعات موجود در مجموعه اسناد رشد داشته‌اند و باید گفت مدل‌سازی موضوعی بدون لحاظ‌کردن زمان باعث ابهام در کشف موضوع می‌شود. مدل‌سازی موضوعی با لحاظ نمودن زمان، مدل‌سازی تکاملی موضوعی نامیده می‌شود. این مدل‌سازی می‌تواند اطلاعات پنهان و مهم را در مجموعه اسناد افشا نماید و امکان شناسایی موضوعات را باتوجه‌به زمان و بررسی تکاملشان در طول زمان می‌دهد. قلمروهای مختلفی می‌توانند از مدل‌های تکاملی موضوعی استفاده کنند، یک نمونه از این مورد این است که پژوهشگر بخواهد موضوع پژوهشی را در قلمرویی خاص انتخاب نماید و از چگونگی تکامل این موضوع در طول زمان آگاهی یافته و درصدد شناسایی اسنادی باشد که آن موضوع را توضیح می‌دهند. چهار روشی که مدل‌سازی موضوعی به آن‌ها متکی است: تحلیل معنایی پنهان (LSA)، تحلیل معنایی پنهان احتمالی (PLSA)، تخصیص پنهان دیریکله (LDA) و مدل موضوعی هم‌بسته (CTM) هستند.

[48] Latent Semantic Analysis (LSA)
[49] Probabilistic Latent Semantic Analysis (PLSA)
[50] Latent Dirichlet Allocation (LDA)
[51] Correlated Topic Model (CTM)

روش‌های مدل‌سازی موضوعی

در این بخش، برخی از روش‌های مدل‌سازی موضوعی که با کلمات، اسناد و موضوعات سروکار دارند، بحث می‌شوند. به‌علاوه، ایده کلی این روش‌ها و مثال‌های از این روش‌ها، در صورت امکان ارائه می‌گردد. این روش‌ها، کاربردهای زیادی دارند و در اینجا به‌طور مختصر چگونگی کاربرد هر روش بیان می‌گردد.

- **تحلیل معنایی پنهان**

تحلیل معنایی پنهان (LSA) روش یا تکنیک مربوط به پردازش زبان طبیعی است. هدف اصلی تحلیل معنایی پنهان، ایجاد بردار بازنمایی برای متون و نشانه‌گذاری محتوای معنایی است. با استفاده از بازنمایی برداری، شباهت میان متون محاسبه می‌شود. درگذشته، تحلیل معنایی پنهان، ایندکس گذاری معنایی پنهان[۵۲] (LSI) نام داشت ولی برای بازیابی اطلاعات، ارائه شده بود؛ بنابراین، یافتن اسناد نزدیک به پرس‌وجو، نیاز به انتخاب از میان تعداد زیادی اسناد داشت. تحلیل معنایی پنهان باید ابعادی برای روشی مانند تطبیق کلمات کلیدی، تطبیق وزن کلمات کلیدی و بازنمایی بردار مربوط به رخداد کلمات در اسناد، داشته باشد (۳۶). همچنین، تحلیل معنایی پنهان از تجزیه مقدارهای تکی[۵۳] (SVD) برای مرتب‌سازی مجدد داده‌ها استفاده می‌نماید. تجزیه مقدارهای تکی روشی است که از یک ماتریس برای پیکربندی مجدد و محاسبه نقصانه‌ای فضای بردار، استفاده می‌نماید. نقصانه‌ای فضای بردار محاسبه شده و از بالاترین تا کمترین اهمیت سازماندهی می‌شوند. برای توصیف اساسی‌ترین مراحل تحلیل معنایی پنهان ابتدا، مجموعه عظیمی از متون مرتبط گردآوری‌شده و سپس توسط اسناد، تقسیم می‌شود. سپس، ماتریس وقوع همکاری برای عبارات و اسناد ایجاد شده و نام سلول مانند سند x، عبارت y و m برای میزان ابعاد عبارات و بردار n بعدی برای اسناد ذکر می‌شود. پس از آن، هر سلول انتخاب و محاسبه خواهد شد. در نهایت باید گفت، تجزیه مقدارهای تکی نقش مهمی برای محاسبه ابعاد و ایجاد ماتریس سه‌بعدی ایفا می‌کند (۳۱).

[52] Latent Semantic Indexing
[53] Singular Value Decomposition

- **تحلیل معنایی پنهان احتمالی**

تحلیل معنایی پنهان احتمالی (PLSA) روشی است که پس از روش تحلیل معنایی پنهان و برای برطرف نمودن معایب تحلیل معنایی پنهان ارائه شد. هافمن و پوزیچا[۵۴] در سال ۱۹۹۹ این روش را ارائه نمودند. تحلیل معنایی پنهان احتمالی روشی است که امکان ایندکس خودکار سند را براساس مدل کلاس پنهانی آماری برای تحلیل فاکتور شمارش می‌دهد و درصدد بهبود تحلیل معنایی پنهان به‌صورت احتمالاتی و با استفاده از مدل مولد است. هدف اصلی تحلیل معنایی پنهان احتمالی، شناسایی و تمیز میان مفاهیم مختلف کلمات به‌کاررفته بدون استفاده از واژه‌نامه است. این روش دو نتیجه مهم دارد: نخست اینکه امکان ابهام‌زدایی چندمعنایی (یعنی کلماتی که چند معنا دارند) را می‌دهد و دوم شباهت‌های معمول با گروه‌بندی کلماتی که چارچوب مشترکی دارند را نشان می‌دهد (۱۱). تحلیل معنایی پنهان احتمالی مبتنی بر یک مدل آماری است که به‌عنوان یک مدل ابعادی شناخته شده است. یک مدل ابعادی مدل متغیر پنهان برای داده‌های با رخداد مشترک است که دسته متغیرهای مشاهده نشده را با هر مشاهده، مرتبط می‌کند. روش تحلیل معنایی پنهان احتمالی برای بهبود تحلیل معنایی پنهان ارائه شده و بنابراین مشکلاتی که تحلیل معنایی پنهان نمی‌تواند حل کند را آدرس‌دهی می‌نماید. تحلیل معنایی پنهان احتمالی کاربردهای موفقی در دنیای واقعی داشته است از جمله بینایی رایانه‌ای[۵۵] و سیستم‌های توصیه گر[۵۶].

کاربردهای تحلیل معنایی پنهان احتمالی حوزه‌های مختلفی از جمله بازیابی و فیلترسازی اطلاعات، پردازش زبان طبیعی و یادگیری ماشین از متن را در برگرفته است. به‌ویژه، برخی از این کاربردها، طبقه‌بندی پردازش خودکار، دسته‌بندی، ردگیری موضوعی، بازیابی تصویر و توصیه خودکار پرسش هستند. در اینجا دو مورد از این کاربردها به توضیح داده می‌شود.

- **بازیابی تصویر**. مدل تحلیل معنایی پنهان احتمالی ویژگی‌های بصری دارد که باعث می‌شود برای نمایش هر تصویر به‌عنوان مجموعه کلمات بصری از یک واژه‌نامه بصری متناهی و گسسته به کار رود. وقوع کلمات

[54] Puzicha
[55] Computer vision
[56] recommender systems

بصری در یک تصویر در بردار رخداد مشترک محاسبه شده است. هر تصویر، بردارهای رخداد مشترکی دارد که به ساخت جدول رخداد مشترک برای به‌دست‌آوردن مدل تحلیل معنایی پنهان احتمالی کمک می‌کند. پس از شناخت مدل تحلیل معنایی پنهان احتمالی می‌توان آن‌را به همه تصاویر پایگاه داده اعمال نمود. سپس آرایش برداری برای نمایش هر تصویر به کار می‌رود و هر عنصر بردار درجه‌ای که هر تصویر، موضوع معینی را ترسیم می‌کند را نشان می‌دهد (۴۰)

- **توصیه خودکار پرسش**. یکی از کاربردهای مهمی که تحلیل معنایی پنهان احتمالی با آن سروکار دارد، وظیفه توصیه پرسش است. در این کاربرد، کلمه از کاربر مستقل است. اگر کاربر یک معنای خاصی را مدنظر داشته باشد، و نیز زمانی که کاربر، پاسخ‌ها و معناهای نهفته مربوط به پرسش را دریافت می‌کند، می‌تواند توصیه‌هایی براساس شباهت‌های معناهای نهفته ارائه دهد. وو[۵۷] و همکاران (۲۰۰۸) اظهار داشتند که تحلیل معنایی پنهان احتمالی برای مدل‌سازی پروفایل کاربر (نمایش داده شده توسط پرسش‌هایی که کاربر می‌پرسد یا پاسخ می‌دهد) و پرسش‌هایی نیز با حذف احتمالات موضوعات پنهان پشت کلمات، مدل‌سازی می‌شوند(۱۶). به دلیل اینکه پروفایل کاربر توسط همه پرسش‌هایی که او می‌پرسد یا پاسخ می‌دهد نمایش داده شده است، از این‌رو فقط باید چگونگی مدل‌سازی درست پرسش مدنظر قرار گیرد.

- **تخصیص پنهانی دیریکله**

علت بروز مدل تخصیص پنهانی دیریکله (LDA) بهبود روش مدل‌های ترکیبی است که قابلیت تبادل کلمات و اسناد را از روش قدیمی و توسط تحلیل معنایی پنهان احتمالی و تحلیل معنایی پنهان ثبت می‌کند. این امر در سال ۱۹۹۰ اتفاق افتاد، یعنی زمانی که تئوری بازنمود کلاسیک[۵۸]، عنوان کرد که هر مجموعه متغیر تصادفی قابل تعویض، بازنموده به‌صورت توزیع ترکیبی و به‌طورکلی یک ترکیب نامتناهی است (۳۰).

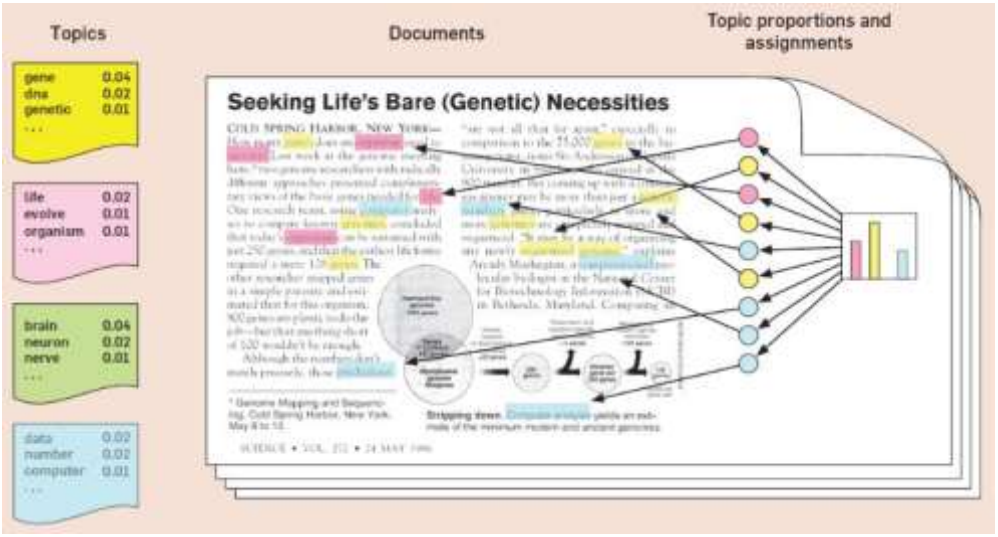
مجموعه اسناد الکترونیکی متعدد از جمله وب، وبلاگ‌های جذاب علمی، مقالات اخباری و پیشینه گذشته، چالش‌های جدیدی برای پژوهشگران در جامعه داده‌کاوی ایجاد کرده‌اند. به‌ویژه نیاز به روش‌های خودکار بصری‌سازی، تحلیل و خلاصه‌سازی این مجموعه اسناد افزایش‌یافته است. اخیراً، مدل‌سازی موضوعی پنهان

[57] Wu
[58] Classic representation theorem

به‌عنوان یک تکنیک کاملاً بدون نظارت، محبوبیت زیادی برای کشف موضوع در مجموعه اسناد عظیم، داشته است. این مدل، الگوریتمی برای متن‌کاوی است که براساس مدل‌های موضوعی آماری (بیزین) بوده و کاربرد بسیار گسترده‌ای داشته است. مدل تخصیص پنهانی دیریکله یک مدل مولد است که در صدد تقلید پردازش نوشتن است؛ بنابراین سعی می‌کند یک سند را براساس موضوع موردنظر، تولید کند. این مدل همچنین می‌تواند برای سایر انواع داده‌ها به کار رود. ده‌ها مدل مبتنی بر مدل تخصیص پنهانی دیریکله وجود دارد از جمله متن‌کاوی موقتی، تحلیل موضوع - نویسنده، مدل‌های موضوعی نظارتشده، خوشه‌بندی مشترک دیریکله و مدل تخصیص پنهانی دیریکله مبتنی بر بیوانفورماتیک (۵؛ ۹).

شکل ۴ ایده نهفته در منطق الگوریتم تخصیص پنهان دیریکله را نشان داده است. این روش فرض می‌کند، تعدادی موضوع که هرکدامشان توزیعی از کلمات است، در کل مجموعه مستندات موجود است (سمت چپ شکل). حال می‌توان تصور نمود که هر مستند این‌گونه شکل می‌گیرد:

ابتدا توزیعی از موضوعات انتخاب می‌شود (هیستوگرام قابل‌مشاهده در شکل)، سپس برای هر واژه یک تخصیص موضوعی (دایره‌های زنگی در شکل) گزینش شده و در نهایت نیز واژه موردنظر از موضوع مربوطه انتخاب می‌گردد. شایان‌ذکر است که موضوعات و تخصیص‌های موضوعی به نمایش درآمده در این تصویر تنها به‌عنوان‌مثالی توضیح‌دهنده ارائه‌شده و از اعمال بر داده‌های واقعی به دست نیامده‌اند (۱۸).



شکل ۴. ایده نهفته در الگوریتم تخصیص پنهان دیریکله (۱۸)

ایده اصلی این فرایند به‌طور ساده، این است که هر سند به‌صورت ترکیبی از موضوعات مدل شده و هر موضوع یک توزیع احتمالی گسسته است که تعیین می‌کند چگونه احتمال هر کلمه در موضوع موردنظر ظاهر شده است. این احتمالات موضوعی، بازنمود دقیقی از سند را ارائه می‌کنند. در اینجا، یک «سند» در واقع «کیسه‌ای از کلمات»[۵۹] است که ساختاری فراتر از آماره های کلمه و موضوع ندارد. از کاربردهای مدل مبتنی بر روش مدل تخصیص پنهانی دیریکله می‌توان موارد زیر را نام برد:

- **کشف نقش**[۶۰]: تحلیل شبکه اجتماعی[۶۱] (SNA) بررسی مدل‌های ریاضی برای تعامل میان افراد، سازمان‌ها و گروه‌ها است. به دلیل بروز ارتباطاتی میان مهاجمان امنیتی ۱۱ سپتامبر و مجموعه داده‌های عظیم انسانی در سرویس‌های وب محبوب از جمله فیسبوک و مای اسپیس، علاقه به تحلیل شبکه‌های اجتماعی، افزایش‌یافته است. این امر منجر به مدل موضوع - گیرنده - نویسنده[۶۲] (ART) برای تحلیل شبکه‌های اجتماعی شد. مدل تخصیص پنهانی دیریکله و مدل موضوع- نویسنده را ترکیب می‌کند. ایده مدل موضوع - گیرنده - نویسنده، درک توزیع‌های موضوعی براساس پیام‌های حساس به جهت که بین دریافت‌کنندگان و فرستندگان ارسال شده است، می‌باشد (۲۶).
- **موضوع احساسی**[۶۳]: مدل جفت- جفت- لینک[۶۴]– مدل تخصیص پنهانی دیریکله که بر مسئله مدل‌سازی اتصال متن و ارجاعات در حوزه مدل‌سازی موضوعی متمرکز است. این براساس ایده مدل تخصیص پنهانی دیریکله و مدل‌های بلوک احتمالی عضویت ترکیبی[۶۵] (MMSB) است و امکان مدل‌سازی دلخواه لینک را می‌دهد (۲۶).
- **طبقه‌بندی خودکار نوشته**[۶۶]: مسئله طبقه‌بندی خودکار نوشته همبستگی نزدیکی با طبقه‌بندی متن دارد و از دهه ۱۹۶۰ مورد بررسی قرار گرفته است. مدل تخصیص پنهانی دیریکله در مقایسه با روش‌های

[59] bag of words
[60] Role discovery
[61] Social Network Analysis
[62] Author-Recipient-Topic
[63] Emotion topic
[64] Pairwise-Link-LDA
[65] Mixed Membership Stochastic Block Models
[66] Automatic essay grading

کاهش ابعاد[۶۷] برای طبقه‌بندی خودکار نوشته، نشان داده که رویکردهای قابل اطمینانی برای وظایف مربوط به بازیابی اطلاعات از فیلترسازی و دسته‌بندی گرفته تا بازیابی و دسته‌بندی سند دارد (۲۶).

- **آنتی فیشینگ**[۶۸]: ایمیل‌های فیشینگ، روش‌هایی برای نمایش اطلاعات حساس مانند اطلاعات برای جلوگیری از ایمیل‌های فیشینگ نیست. به دلیل اینکه مدل‌های موضوعی پنهانی، خوشه‌هایی از کلمات هستند که با هم در ایمیل ظاهر می‌شوند، کاربر می‌تواند پیش‌بینی کند که در یک ایمیل فیشینگ کلمات «کلیک» و «حساب» معمولاً با هم ظاهر می‌شوند. مدل‌های موضوعی پنهان معمول در دسته‌های مختلف اسناد از جمله فیشینگ و غیر فیشینگ لحاظ نشده‌اند. به همین دلیل، پژوهشگران، مدل آماری جدیدی به نام مدل موضوع- کلاس پنهانی[۶۹] توسعه داده‌اند که تعمیم‌یافته مدل تخصیص پنهانی دیریکله است (۲۷)

- **مدل موضوعی هم‌بسته**

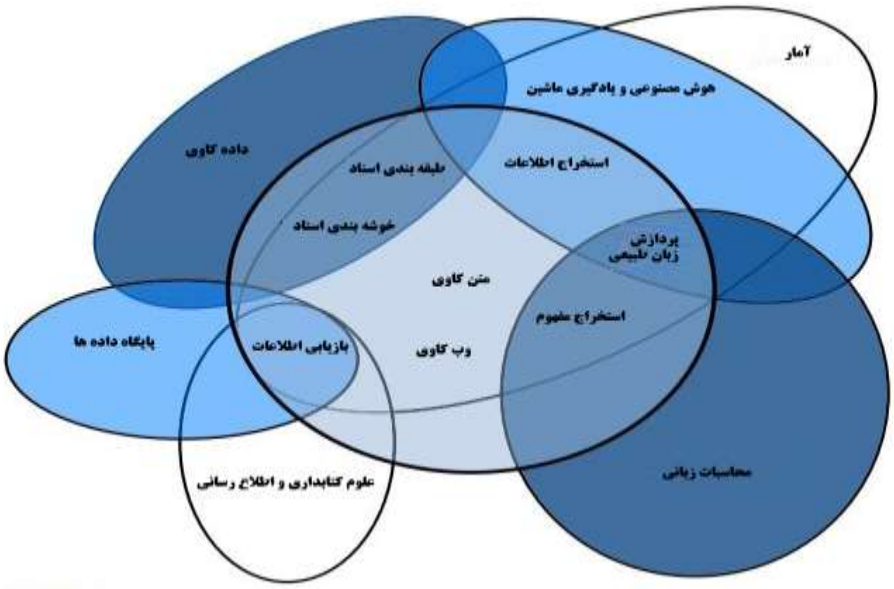
مدل موضوعی هم‌بسته[۷۰] یک نوع مدل آماری به‌کاررفته در پردازش زبان طبیعی و یادگیری ماشین است. مدل موضوعی هم‌بسته برای کشف موضوعاتی که در گروه اسناد نشان داده‌اند، به‌کاررفته است. کلید مدل موضوعی هم‌بسته توزیع نرمال لجستیک است. مدل‌های موضوعی هم‌بسته به مدل تخصیص پنهانی دیریکله متکی هستند (۱۳).

ماهیت بین‌رشته‌ای متن‌کاوی

بین‌رشته‌ای با به‌کارگیری اصول و قواعد دانش‌های موجود (دانشی بیشتر از دو قلمرو علمی) و ترکیب آن‌ها با هم، دانش جدید را ایجاد می‌نماید (۱۵).

متن‌کاوی زمینه‌ی بین‌رشته‌ای است که به‌طورکلی از شش قلمرو دانشی و اشتراک علوم بهره می‌گیرد. آمار، هوش مصنوعی و یادگیری ماشین، داده‌کاوی، محاسبات زبانی، پایگاه داده و علوم کتابداری و اطلاع‌رسانی قلمروهای علمی هستند که متن‌کاوی از آن‌ها استفاده می‌کند. شکل ۵-۲ نمودار ون اشتراکات متن‌کاوی و سایر حوزه‌های علمی را نشان می‌دهد (۱۵).

[67] Dimension Reduction Methods
[68] Anti-Phishing
[69] latent Class-Topic Model (CLTOM)
[70] Correlated Topic Model (CTM)



شکل ۵. نمودار ون اشتراکات متن‌کاوی و سایر قلمروهای علمی (۱۵)

همان‌طور که در شکل ۵ مشاهده می‌شود از اشتراک این قلمروهای علمی، هفت روش عملی برای رسیدن به اهداف متن‌کاوی ایجاد می‌شود که عبارت از پردازش زبان طبیعی، طبقه‌بندی اسناد، خوشه‌بندی اسناد، بازیابی اطلاعات، وب کاوی و استخراج مفهوم هستند.

- طبقه‌بندی اسناد[۷۱]: فرایند انتخاب بهترین برچسب (موضوع یا نمایه، نوع حس، ...) برای متون بدون برچسب[۷۲] از مجموعه برچسب‌های از قبل مشخص شده، با استفاده از مدلی که از روی متون برچسب‌گذاری شده (داده‌های آموزشی) یادگرفته و ساخته شده است.

- خوشه‌بندی اسناد[۷۳]: به فرایند گروه‌بندی مستندات مشابه درون خوشه‌های مختلف خوشه‌بندی می‌گویند.

- بازیابی اطلاعات[۷۴]: فرایند شاخص‌گذاری[۷۵]، جستجو و بازیابی و بازیابی مستندات از بین مجموعه داده‌های عظیم متنی باتوجه‌به کلیدواژه‌های پرس‌وجو را بازیابی اطلاعات می‌گویند. مهم‌ترین کاربرد بازیابی اطلاعات در موتورهای جستجوی وب از قبیل گوگل، بینگ، یاهو و ... دیده می‌شود.

---

71 Document Classification
72 untagged documents
73 Document Clustering
74 Information Retrieval-IR
75 indexing

- وب کاوی[۷۶]: فرایند داده و متن‌کاوی روی محتوای صفحات و ارتباطات (لینک‌های) بین صفحات وب را وب کاوی می‌گویند. صفحات وب حالتی نیمه‌ساختاریافته از متون و لینک (ارتباط) به صفحات دیگر تشکیل شده‌اند؛ لذا یک رویکرد متداول برای وب‌کاوی، بازنمایی صفحات وب در قالب گراف و تحلیل گراف وب هست. امروزه زیرشاخه‌ای از وب‌کاوی برای تحلیل شبکه‌های اجتماعی مانند فیس‌بوک، توییتر و ... بسیار موردتوجه پژوهشگران و کسب‌وکارهای مختلف قرار گرفته است.

- استخراج اطلاعات[۷۷]: فرایند شناسایی و استخراج موجودیت‌های مناسب و همچنین روابط بین آن‌ها از درون متن (غیر ساخت‌یافته) را استخراج اطلاعات می‌گویند. به‌عبارت‌دیگر استخراج اطلاعات فرایندی برای پردازش داده‌های غیر ساخت‌یافته (مثل متن، تصویر، صوت و ...) یا نیمه‌ساختاریافته (مثل صفحات وب، XML، ...) و ساخت (تبدیل کردن آن‌ها به) مجموعه داده ساخت‌یافته (از قبیل جداول پایگاه‌داده) است. استخراج اطلاعات به دو نوع باز (عمومی) و بسته (خاص و در حوزه مشخص) تقسیم می‌شود.

- پردازش زبان طبیعی[۷۸]: هدف آن، پردازش سطح پایین و فهم (درک) زبان و بخصوص متن توسط کامپیوترها است. معمولاً معادل با اصلاح زبان‌شناسی محاسباتی[۷۹] بکار گرفته می‌شود، هرچند که اغلب زبان‌شناسان زبان‌شناسی محاسباتی را کلی‌تر از پردازش زبان طبیعی می‌دانند.

- استخراج مفاهیم[۸۰]: فرایند گروه‌بندی کلمات و عبارات متن درون گروه‌های مشابه معنایی را استخراج مفاهیم می‌گویند. معمولاً از تکنیک‌های آماری (مانند n-grams یا هم‌رخدادی)، تعبیه کلمات (word embedding)، مدل‌سازی موضوعات[۸۱] و خوشه‌بندی متون و کلمات برای استخراج مفاهیم استفاده می‌شود.

[76] Web Mining
[77] Information Extraction-IE
[78] Natural Language Processing-NLP
[79] Computational linguistics
[80] Concept Extraction
[81] topics modeling

مروری بر پیشینه‌های پژوهش

بررسی پژوهش‌ها نشان می‌دهد که فنون متن‌کاوی و استخراج دانش از متن به طور عملیاتی بر روی انواع مختلف متون کاربرد دارد و در پژوهش های مختلف جهت استخراج دانش از متون بکار رفته است، که در این زمینه می‌توان به موارد زیر اشاره نمود:

بررسی نیاز اطلاعاتی و رضایت کاربران از وب‌سایت مرتبط با سرطان (۲۵)، مدلی جهت استخراج اطلاعات برای بهره‌برداری از روش‌های متن‌کاوی در یادگیری الکترونیکی (۲۶). با توجه به اینکه تمرکز این پژوهش بر روی متون انتشارات علمی پژوهشی منتشرشده جهانی در پایگاه اطلاعاتی WOS هست، در ادامه پیشینه مقالات مرتبط در این خصوص ذکر می‌گردد.

لام[۸۲] و همکاران ۲۰۱۶ نیز در پژوهش خود به شناسایی روند پژوهش‌ها منتشرشده در مجله اختلالات خواب پرداختند، در این پژوهش مقالات منتشرشده بین سال‌های ۲۰۰۰ تا ۲۰۱۳ با استفاده از فنون متن‌کاوی مورد تجزیه و تحلیل قرار گرفت، نتایج این پژوهش روند انتشار مقالات در مجلات علمی و همچنین مهم‌ترین موضوعات مرتبط با بی‌خوابی و اختلالات آن را نشان داده است(۲۷).

وانگ و همکاران ۲۰۱۶ در پژوهشی به شناسایی مباحث موجود در مطالعات مصرف مواد و افسردگی در بزرگسالان پرداختند، به همین منظور تعداد ۱۷۷۲۳ چکیده مقاله بین سال های ۲۰۰۰ تا ۲۰۱۴ از پایگاه اطلاعاتی پاب مد استخراج و مود تجزیه و تحلیل قرار گرفت. پژوهشگران جهت شناسایی موضوعات و خوشه های این پژوهش از الگوریتم مدلسازی موضوعی تخصیص پنهان دیریکله استفاده نمودند. پژوهشگران به این نتیجه دست یافتند که مدل سازی موضوعی این امکان را دارد که مجموعه زیادی از مقالات را در قالب های جداگانه تفکیک کند و همچنین می تواند به عنوان ابزاری برای درک روند پژوهش ها، نه تنها با بازیابی حقایق شناخته شده بلکه با کشف موضوعات مرتبط استفاده شود (۲۸).

سلواراج[۸۳] و پریاسامی[۸۴] ۲۰۱۶ در پژوهشی با استفاده از روش متن‌کاوی، گیاهان دارویی هند در درمان دیابت را در مقالات زیست پزشکی مورد بررسی قراردادند، متون مورد بررسی این پژوهش با جستجو در پایگاه‌های

[82] Lam
[83] Selvaraj
[84] Periyasamy

اطلاعاتی PubMed، Science Direct و ... به دست آمدند. در این پژوهش ۲۰۳ گیاه دارویی هند برای دیابت از ۳۵۵ مقاله از مجموع ۱۵۶۵۱ مقاله کشف شد. داده های جمع آوری شده با استفاده از زبان برنامه نویسی پایتون[۸۵] و کتابخانه NLTK[۸۶] پیش پردازش (کلمه کلمه کردن[۸۷]، حذف کلمات زائد[۸۸] و ریشه یابی[۸۹]) و تحلیل شدند. نتایج این پژوهش مهم‌ترین گیاه دارویی هند برای درمان دیابت که در مطالعات بکار گرفته شده است، را نشان داده است(۲۹).

اوزایدین[۹۰] و همکاران ۲۰۱۷ در پژوهش خود به تجزیه و تحلیل تکامل تحقیقات سلامت همراه[۹۱] با بکارگیری متن کاوی و پردازش زبان طبیعی پرداختند. نمونه مورد مطالعه این پژوهش شامل خلاصه ۵۶۴۴ مقاله پژوهشی است که از پنج موتور جستجوی دانشگاهی با استفاده از اصطلاحات جستجو مانند سلامت تلفن همراه جمع آوری شده است. جهت متن کاوی در این پژوهش از نرم افزار JMP Pro ماژول اکسپلورر متن[۹۲] استفاده شده است. در پایان نتایج به شکل ابرهای کلمه ای و تحلیل روند ارائه شده است. پژوهشگران یافته های این مطالعه را در شناسایی زمینه هایی برای مطالعات آینده مفید دانسته اند(۳۰).

پایتون[۹۳] و همکاران ۲۰۱۸ در مطالعه‌ای به بررسی مسائل بهداشت روانی تأثیر گذار بر دانشجویان، در متون مقالات خبری، گزارش‌ها، جهت کشف موضوعات برجسته مسائل بهداشت روانی دانشجویان پرداختند. در این پژوهش ۱۶۵ منبع که در سال‌های ۲۰۱۰ تا ۲۰۱۵ منتشرشده بود با استفاده فنون متن‌کاوی در نرم‌افزار SAS مورد تجزیه و تحلیل قرارگرفته است. نتایج این پژوهش خوشه‌های اصلی موضوعی در تجربیات روان‌شناختی دانشجویان در آموزش عالی را نشان داده است. و همچنین مهم‌ترین موضوعات مرتبط با دانشجویان را نیز شناسایی نموده است(۳۱).

---

85 Python
86 Natural Language Tool Kit (NLTK)
87 Tokenization
88 Stop word removal
89 Stemming
90 Ozaydin
91 Mobile Health(Mhealth)
92 Text Explorer
93 Payton

کوان[۹۴] و همکاران ۲۰۱۸ در مطالعه‌ای به بررسی روند مقالات پژوهش‌ها deqi (حوزه‌ای در خصوص طب سوزنی) با استفاده از فنون متن‌کاوی پرداخته که بدین منظور ۱۴۸ مقاله از پایگاه اطلاعاتی استخراج و مورد تجزیه و تحلیل قرارگرفته است، در این پژوهش برای نشان دادن اهمیت هر کلیدواژه از الگوریتم Tf-idf استفاده شده است، همچنین برای خوشه بندی موضوعات مقالات از الگوریتم های خوشه بندی در نرم افزار وکا[۹۵] استفاده شده است. نتایج این پژوهش به‌طورکلی روند پژوهش‌ها deqi و برنامه‌های آینده این حوزه را مشخص نموده است(۳۲).

دنسی اسکات[۹۶] و همکاران ۲۰۱۸ در پژوهشی به ارزیابی خلاصه مقالات ارائه شده در کنفرانس‌های بین‌المللی ایدز در بیش از ۲۵ سال به‌منظور شناسایی روند اصطلاحات اچ ای وی پرداخته است. در این پژوهش بیش از ۸۰،۰۰۰ خلاصه از انجمن بین‌المللی ایدز به‌دست‌آمده که جهت متن‌کاوی آن از نرم‌افزار نایم[۹۷] استفاده‌شده است. همچنین از نرم‌افزار تابلیو[۹۸] جهت رسم تصاویر و ورد کلودها استفاده‌شده است. یافته‌های اصلی روند نتایج این پژوهش اصطلاحات مربوط به اچ آی وی را در طول ۲۵ سال مشخص نموده است و نشان داد که اصطلاح "اپیدمی ایدز" از سال ۱۹۸۹ تا ۱۹۹۱ به‌شدت مورد استفاده قرار گرفت و پس از آن کاهش یافت. در مقابل، استفاده از واژه "اپیدمی اچ آی وی" از سال ۲۰۱۴ افزایش‌یافته است. از اواسط دهه ۱۹۹۰، اصطلاح "درمان باتجربه" با فراوانی بیشتر در خلاصه‌ها ظاهر شد. به‌طورکلی نتایج این مطالعه تغییراتی را در استفاده از اصطلاحات اچ آی وی در طول ۲۵ سال، ازجمله افزودن، ناپدید شدن و تغییر شرایط استفاده از اصطلاحات نشان می‌دهد که پیشرفت‌های پژوهش و اقدامات پزشکی و ناتوانی زایی بیماری را نشان می‌دهد(۳۳).

کیم و همکاران ۲۰۱۸ به شناسایی روند پژوهش‌های حوزه انفورماتیک پزشکی به‌منظور درک موقعیت فعلی حوزه‌ی علمی انفورماتیک پزشکی، مسیر پیشرو و شناسایی محدودیت‌های حاکم بر این حوزه پرداخته‌اند، به همین منظور پژوهشگران تعداد ۲۳ مجله از مجلات مهم این حوزه با تعداد ۲۶۳۰۷ مقاله را در روند ۱۲ ساله با استفاده از فنون متن‌کاوی مورد بررسی قرار داده‌اند، در این پژوهش از الگوریتم Tf-idf جهت شناسایی

[94] Kwon
[95] Weka (Waikato Environment for Knowledge Analysis)
[96] Dancy-Scott
[97] KNIME
[98] Tableau

کلیدواژه های مهم استفاده شده است. نتایج نشان داده است که برخی زمینه‌های موضوعی، مانند زیست پزشکی، در حال کاهش است، درحالی‌که سایر حوزه‌های پژوهش‌های مانند فناوری اطلاعات سلامت، پژوهش مبتنی بر اینترنت و پرونده‌های پزشکی / بهداشتی الکترونیکی در حال رشد هستند و همچنین بیشترین پژوهش‌ها در خوشه‌های "پژوهش‌ها مبتنی بر اینترنت و ارائه‌ی دانش" بوده است(۳۴).

روسناو [۹۹]و همکارن ۲۰۱۸ در مطالعه خود به شناسایی روند پژوهش‌های منتشرشده حوزه بیهوشی پرداختند به همین منظور ۲۲۲۶۲ خلاصه مقاله ارائه شده در نشست‌های معتبر بین سال‌های ۲۰۰۰ تا ۲۰۱۳ با استفاده از کتابخانه NLTK در زبان برنامه‌نویسی پایتون مورد تجزیه و تحلیل قرار گرفت، در این پژوهش جهت شناسایی موضوعات از الگوریتم مدل سازی موضوعی LDA استفاده شده است. نتایج چگونگی استفاده از یک روش منحصر به فرد را برای شناسایی موضوعات و روندهای رایج در پژوهش‌ها حوزه بیهوشی را نشان داد، همچنین موضوعات جدید در جهت تولیدِ ایده برای کارهای آتی شناسایی شدند(۳۵).

جلیلی و همکاران ۲۰۱۹ به تجزیه و تحلیل مقالات حوزه مراقبت بهداشتی و سایبری پرداختند، پژوهشگران تعداد ۴۷۲ مقاله مرتبط را در پایگاه اطلاعاتی پاب مد و وب آو ساینس استخراج و با استفاده از فنون متن‌کاوی در نرم افزار لکسی منسر [۱۰۰] مورد تجزیه و تحلیل قرارگرفته است. نتایج نشان داد که علیرغم افزایش پژوهش‌ها و توجه به امنیت سایبری، در این پژوهش‌ها نقاط ضعفی نیز وجود دارد. به‌عنوان مثال، یافته‌ها نشان می‌دهد که اکثریت مقالات بر فنّاوری متمرکزشده‌اند: مقالات متمرکز بر فناوری، بیش از نیمی از خوشه‌ها را تشکیل می‌دهند، درحالی‌که مقالات مدیریتی تنها ۳۲درصد از خوشه‌ها را تشکیل داده است. به‌طور مشابه، در تجزیه و تحلیل مجله‌ها ، ۵۸ مقاله در ۱۵ مجله منتشرشده از مجلات علمی کامپیوتر بود و ۱۲ مقاله در مجله‌های متمرکز بر سلامت بوده است(۳۶).

صاحب و صاحب ۲۰۱۹ در پژوهشی به تجزیه و تحلیل مقالات حوزه اطلاعات سلامت با روش متن کاوی پرداختند، بدین منظور تعداد ۳۰۱۱۵ مقاله بین سالهای ۱۹۷۴ تا ۲۰۱۸ را از پایگاه اطلاعاتی ساینس دایرکت استخراج نمودند، سپس با استفاده از روش های متن کاوی کلمات کلیدی پژوهش ها را بدست آوردند،

[99] Rusanov
[100] Leximancer

پژوهشگران از نرم افزارهای سایت اسپیس[۱۰۱] و ووس ویوور[۱۰۲] برای تجزیه و تحلیل داده ها استفاده نموده اند همچنین برای برای ترسیم شبکه ها و گراف ها از نرم افزار گفی[۱۰۳] و ووس ویوور بهره برده اند. نتایج این پژوهش نشان داد که سه موضوع عمده پژوهش های انجام شده عبارت بودند از: استفاده از علوم رایانه در مراقبت های بهداشتی، تاثیر فناوری اطلاعات سلامت بر سلامت بیماران و کیفیت مراقبت های بهداشتی و سیستم های پشتیبانی تصمیم. نتایج این پژوهش نشان داده است که آینده این پژوهش ها به سمت داده های سلامتی تولید شده توسط بیمار، الگوریتم های یادگیری عمیق، ابزار سنجش از خود و خود سنجی و سیستم های پشتیبانی تصمیم گیری مبتنی بر اینترنت حرکت می کند(۳۷).

بینگتون[۱۰۴] و همکاران ۲۰۱۹ در پژوهش خود به نگاشت نقشه علمی مجله رفتار حرفه ای[۱۰۵] در بازه زمانی ۲۳ ساله از سال ۱۹۹۴ تا ۲۰۰۶ پرداختند به همین منظور ۱۴۹۰ مقاله از مجله مذکور را استخراج و با استفاده از نرم افزار ووس ویور مورد تجزیه و تحلیل قرار دادند. نتایج این پژوهش یک طبقه بندی تجربی مبتنی بر زمینه های اصلی محتوا در مجله را براساس میزان تمایل به اصطلاحات ایجاد شده، ارائه نموده است. همچنین مهم ترین خوشه های موضوعی را مشخص نموده است(۳۸).

خاصه و همکاران ۱۳۹۵ به تحلیل خوشه های موضوعی و ترسیم نقشه های علمی پژوهشگران ایرانی حوزه انگل شناسی با تأکید بر شاخصهای هم تألیفی و شاخص اچ پرداختند. بدین منظور تعداد ۱۲۷۱ مقاله بین سالهای ۱۲۷۹ تا ۲۰۱۵ از پایگاه اطلاعاتی استخراج و با فنون علم سنجی مورد تجزیه و تحلیل قرار گرفت. در این پژوهش از نرم افزار های ووس ویوور، یو سی نت[۱۰۶] و نت دارو[۱۰۷] استفاده شده است. نتایج این پژوهش فراوانی رخداد کلمات در مقالات، خوشه بندی موضوعات و تاثیر گذارترین پژوهشگران شناسایی شده است(۳۹).

[101] CiteSpace
[102] VOSviewer
[103] Gephi
[104] Byington
[105] Vocational Behavior
[106] UciNet
[107] NetDraw

مرور پیشینه پژوهش نشان می دهد که فنون متن‌کاوی در تجزیه و تحلیل متون و کشف و استخراج دانش در حجم عظیمی از متون کاربرد دارد و تحزیه و تحلیل متون منتشر شده در پایگاه های استنادی و شناسایی روند پژوهش های انجام شده در قلمروهای مختلف علمی یکی از مهمترین کاربردهای متن کاوی است. در هرکدام از پژوهش‌های انجام‌شده با توجه به اهداف و جامعه پژوهش فنون خاصی از متن‌کاوی استفاده شده است. و از مهمترین فنون استفاده شده در تجزیه و تحلیل متون علمی می توان شناسایی پر تکرارترین کلیدواژه‌ها و مهمترین موضوعات را با استفاده از الگوریتم های خاص متن کاوی نام برد. الگوریتم های خوشه بندی K-Means و الگوریتم مدلسازی موضوعیLDA از مهم ترین الگوریتم های بکاربرده شده در پژوهش های پیشین جهت تعیین خوشه ها و موضوعات بوده است، که با توجه به دقت بالایی که دارد در پژوهش ها مورد استفاده قرار گرفته و پیشنهاد شده است.

نتیجه جستجوهای انجام شده در پایگاه های اطلاعاتی و استنادی ملی و بین المللی و مرور پیشینه‌های خارجی و داخلی حاکی از آن است که در زمینه موضوعی کرونا ویروس، پژوهش مشابهی مشاهده نگردید و پژوهش حاضر از نظر موضوعی و یافته‌ها و نتایجی که ارائه نموده، تکراری نیست.

فصل سوم

روش‌شناسی پژوهش

روش‌شناسی اجرای طرح

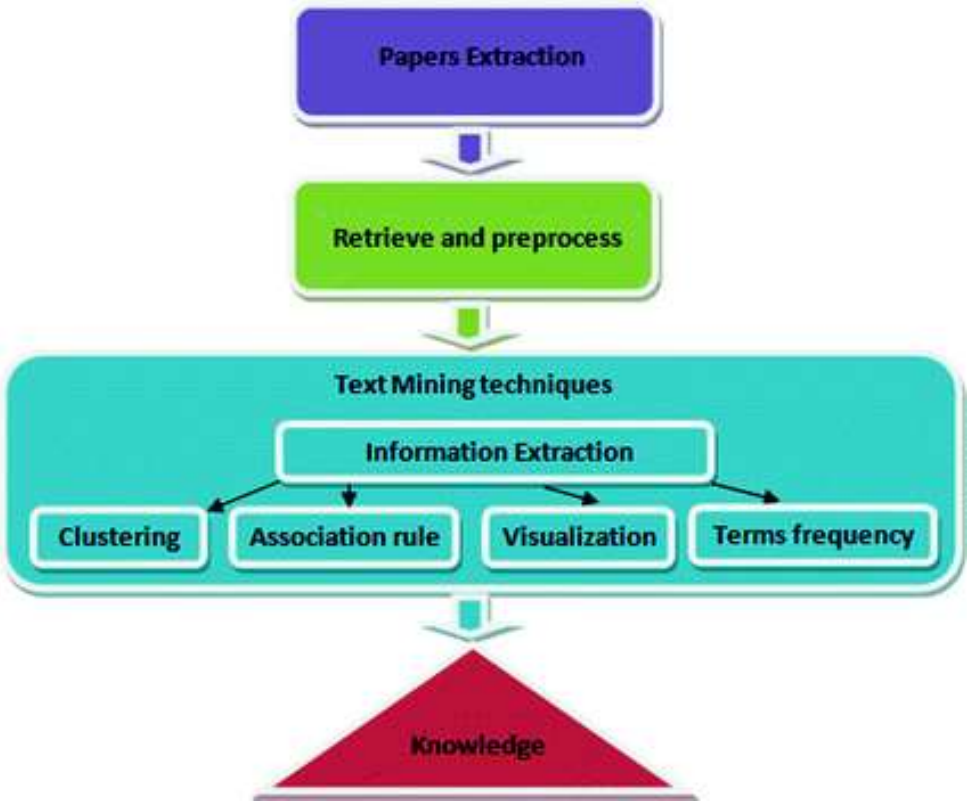
این پژوهش از نوع کاربردی است که به روش متن کاوی و با رویکرد تحلیلی انجام شده است. جامعه آماری، کلیه انتشارات علمی قلمرو موضوعی کروناویروس در بازه زمانی ۱۹۷۰ تا ۲۰۲۰ به زبان انگلیسی است. پس از مشورت با متخصصان بیماریهای تنفسی و عفونی، راهبرد جستجو طراحی شد. جهت مشخص نمودن راهبرد جستجو از مرورگر سرعنوانهای موضوعی پزشکی (MESH) استفاده گردید. در مرحله بعدی به منظور جستجو و بازیابی مدارک کروناویروس از جستجوی پیشرفته Web of Science Core Collectionکه معتبرترین، پرکاربردترین و قدیمی ترین پایگاه استنادی جهان است، استفاده شد(۴۷).

راهبرد جستجوی بکار رفته دراین پژوهش در تاریخ ۱۵-۳-۲۰۲۰ به شرح زیر بوده است:

(TI=(COVID-19 OR 2019 novel coronavirus disease OR 2019 novel coronavirus infection OR 2019-nCoV disease OR 2019-nCoV infection OR COVID19 OR coronavirus disease 2019 OR coronavirus disease-19 OR coronavirus OR ALPHACORONAVIRUS OR BETACORONAVIRUS OR DELTACORONAVIRUS OR GAMMACORONAVIRUS OR Bulbul coronavirus HKU11 OR Coronavirus HKU15 OR Munia coronavirus HKU13 OR Rabbit Coronavirus OR Thrush coronavirus HKU12 OR Alphacoronavirus 1 OR Coronavirus 229E, Human OR Coronavirus NL63, Human OR Coronavirus, Canine OR Coronavirus, Feline OR Betacoronavirus 1 OR Coronavirus, Rat OR Middle East Respiratory Syndrome Coronavirus OR Coronavirus, Turkey OR Coronavirus OC43, Human OR Coronavirus, Bovine))

LANGUAGE: (English)

پس از جستجو، ۶۵۶۵ مدرک بازیابی شد. روش متن‌کاوی به‌کار رفته در این پژوهش برگرفته از چارچوب طراحی‌شده Zhang & Chen (۴۸) است که توسط توسط Salloum و همکاران توسعه یافته است(۱۷) (شکل ۶).

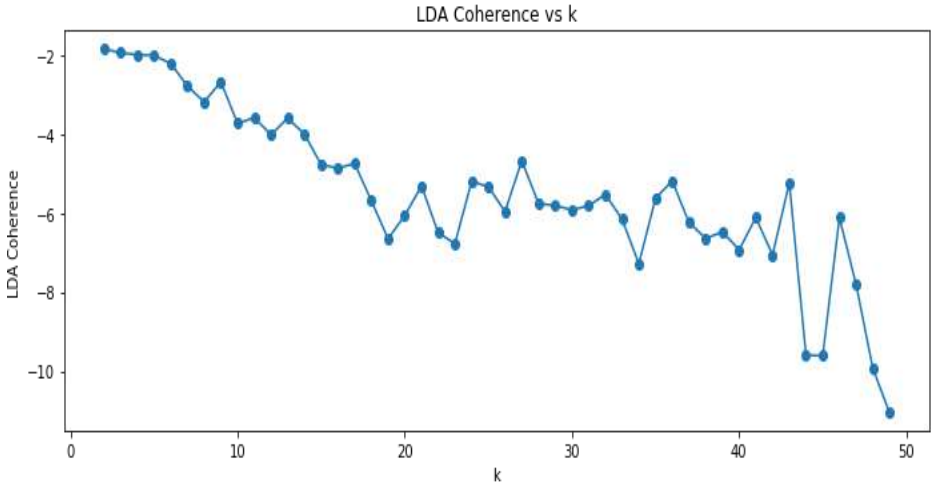


شکل ۶. چارچوب متن‌کاوی مورد استفاده در این پژوهش

در این پژوهش جهت انجام متن‌کاوی، مراحل زیر روی داده‌ها انجام می گیرد. پیش‌پردازش متون که شامل انتخاب اسناد، استخراج واژه های بکار رفته در متون، یکدست سازی متون با استفاده از بررسی دستی واژه‌های مقالات و یکی کردن واژه های مترادف، حذف واژه های بی معنی و استاپ وردها، سپس در این مرحله واژه‌های بکار رفته در متون مقالات با استفاده از الگوریتم ریشه یابی پورتر (Porter stemmer) (۴۹، ۵۰) ریشه یابی شدند. در مرحله دوم انجام فراوانی و وزن دهی واژگان، اجرای الگوریتم مدل سازی موضوعی و بصری سازی با فنون مختلف متن کاوی انجام می شود و در نهایت با استخراج دانش از متون و تفسیر آن به پایان می رسد. در این پژوهش مهم‌ترین واژگان بر اساس وزن TF-IDF نیز شناسایی و گزارش شده‌اند. TF-IDF یک آمار عددی است که میزان اهمیت یک کلمه نسبت به یک سند در مجموعه‌ای از اسناد را نشان می‌دهد. در واقع هدف آن، نشان دادن اهمیت کلمه در متن است. مقدار TF-IDF به‌تناسب تعداد تکرار کلمه در سند افزایش می‌یابد و توسط تعداد اسنادی که در مجموعه هستند و شامل کلمه نیز می‌باشند متعادل می‌شود. به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی

متن ندارد (۳۴، ۵۱، ۵۲). جهت انجام عملیات مدل‌سازی موضوعی در این پژوهش از الگوریتم مدل‌سازی موضوعی LDA استفاده شده است (۵۳). مدل سازی موضوعی یک رویکرد یادگیری ماشین برای کشف الگوها یا موضوعات درون یک مجموعه اسناد است. در این پژوهش، روش تخصیص پنهان دیریکله (LDA) ، را که یکی از روش های پیاده سازی مدل سازی موضوعی است، انتخاب شده است(۵۴). LDA هم به دلیل اینکه یکی از بهترین الگوریتم هایی است که به طور گسترده استفاده شده است و هم به دلیل اینکه در شناسایی موضوعات معنایی مرتبط در متون علمی بسیار اثر بخش بوده (۵۵) و بهتر از خیلی از الگوریتم های جدید تر در این رابطه عمل می کند، انتخاب شد(۵۶). یکی از محدودیت‌های استفاده از مدل‌سازی موضوعی پیش‌بینی تعداد موضوعات است که در این پژوهش با استفاده از معیار logarithmic (log) UMass Coherence تعداد موضوعات پیش بینی شده است(۵۷). در گام بعدی از elbow criterion جهت شناسایی تعداد مطلوب موضوعات استفاده‌شده است. که روشی برای تخمین تعداد موضوعات مطلوب می‌باشد(۵۸). با استفاده از logarithmic (log) UMass Coherence و ترسیم نمودار elbow criterion (شکل ۷) بین ۸ تا ۵۰ موضوع را برای مقالات مستخرج این پژوهش در حوزه کرونا ویروس می‌توان انتخاب نمود که با توجه به بررسی و تفسیر موضوعات با مقادیر مختلف، تعداد ۸ موضوع انتخاب‌شده است. با انتخاب تعداد موضوعات، مباحث بسیار گسترده ای ایجاد می شود، در حالی که انتخاب تعداد بیش از حد نیز منجر به تعداد زیادی مباحث کوچک و بسیار مشابه خواهد شد (۵۹، ۶۰). تعداد موضوعات بالاتر همچنین باعث می‌شود که هیچ اطلاعات اضافی موضوعی به دست نیاید. همچنین با توجه به پراکندگی کلمات کلیدی بین موضوعات، تفسیر موضوعات سخت‌تر می‌شود (۶۱). موضوعات با استفاده از مهم‌ترین واژگان و مقالات هر موضوع به‌دست‌آمده از اجرای الگوریتم LDA تفسیر شدند. جهت اجرای الگوربتم مدلسازی موضوعی از زبان برنامه‌نویسی پایتون و کتابخانه Grnsim استفاده شده است (۶۲). زیرا یک ابزار مدلسازی موضوعی متن باز است، دارای نحو (syntax) ساده ، کم حجم ، ماهیت چند منظوره و سهولت توسعه است و کتابخانه های متنوعی را برای کار با متون در اختیار کاربر قرار می دهد (۶۲). همچنین پژوهش های متعددی برای اجرای LDA این ابزار را به کار برده اند (۶۳-۶۵). شکل ۷ از اجرای معیار logarithmic (log)

UMass Coherence و براساس داده های بدست آمده از آن بدست آمده است. نقطه انتخاب Elbow که قسمت با شیب تندتر و قسمت مسطح را متمایز می کند با انتخاب نقطه ای که تشخیص داده می شود(۶۶). که این امر به اصطلاح Elbow در نمودار ایجاد می کند.



شکل ۷. نمودار elbow حاصل از اجری الگوریتم UMass Coherence جهت انتخاب تعداد موضوعات

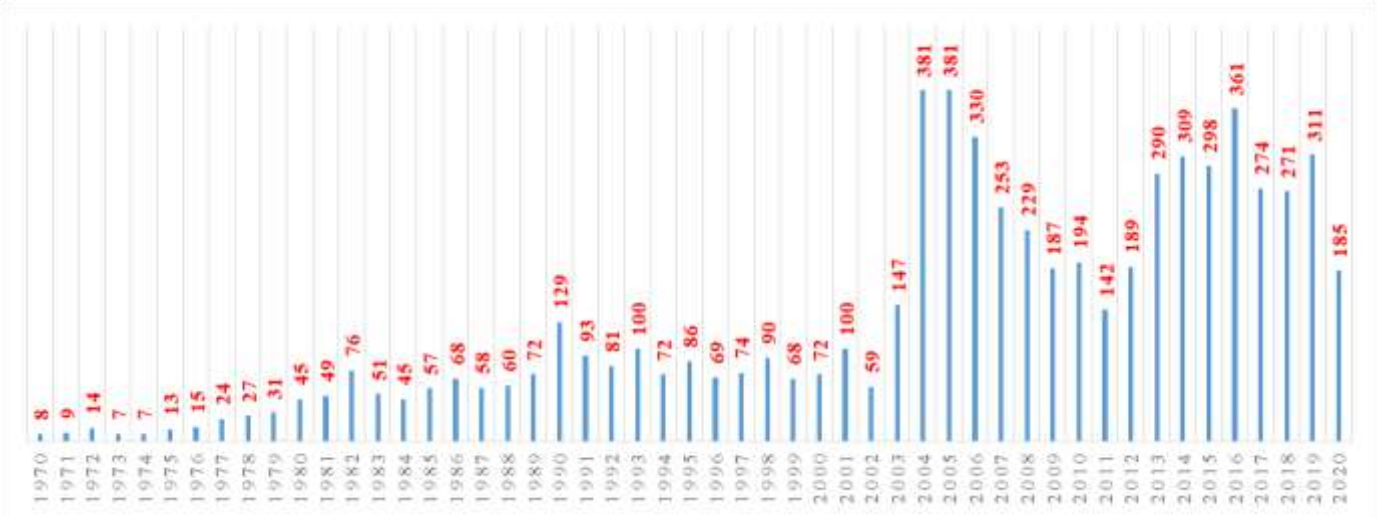
با توجه به اینکه الگوریتم LDA مورد استفاده در این پژوهش تعداد بهینه موضوعات ، توزیع هر سند در این موضوعات و لیست اصطلاحات مرتبط با هر موضوع را مشخص می کند ولی برچسب برای موضوعات ارائه نمی دهد ، که این کار باید به صورت دستی تعریف شود(۵۴). در این پژوهش بر چسب هر یک از موضوعات با مشورت نویسندگان این پژوهش و کمک متخصصان موضوعی تعیین شدند. در پایان نیز با توجه به اطلاعات به‌دست‌آمده از تجزیه‌وتحلیل داده‌ها و مقالات، به تفسیر نتایج پرداخته‌شده است.

فصل چهارم

تحلیل داده‌ها و ارائه یافته‌ها

روند انتشارات جهانی کروناویروس در نیم قرن اخیر

با توجه به حذف موارد مشابه در داده های بازیابی شده، تعداد ۶۵۶۱ مدرک جهت انجام فنون متن کاوی و مدلسازی موضوعی انتخاب شدند. که نمودار ۱ روند انتشارات کرونا ویروس در پنجاه سال گذشته را نشان می دهد.



نمودار۱. نمودار روند انتشارات کرونا ویروس درنیم قرن اخیر (۱۹۷۰-۲۰۲۰)

واژگان مهم انتشارات جهانی کروناویروس در نیم قرن اخیر

جدول ۱: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	sar	87.60119	16	infecti	48.85843
2	scienc	71.6268	17	microbiolog	48.43426
3	protein	69.84372	18	intern	47.09281
4	mer	64.08395	19	felin	46.8567
5	veterinari	62.90938	20	bovin	45.96076
6	cell	62.77219	21	sequenc	45.49739
7	human	62.62166	22	syndrom	44.72848
8	rna	59.48379	23	strain	44.27336
9	medicin	58.01514	24	hcov	42.71015
10	virolog	57.67949	25	detect	42.44752
11	mice	55.21389	26	immunolog	40.85823
12	antibodi	52.53795	27	middl	40.42126
13	cov	51.4668	28	infect	39.88012
14	diseas	51.43914	29	gener	39.86575
15	respiratori	51.09082	30	structur	39.69041

جدول یک ۳۰ واژگان مهم انتشارات کرونا ویروس را نشان می دهد که لازم به ذکر است که این واژگان قبل از استخراج با استفاده از الگوریتم ریشه یابی پورتر ریشه یابی شده اند.

روند تغییرات واژگانی در نیم قرن اخیر

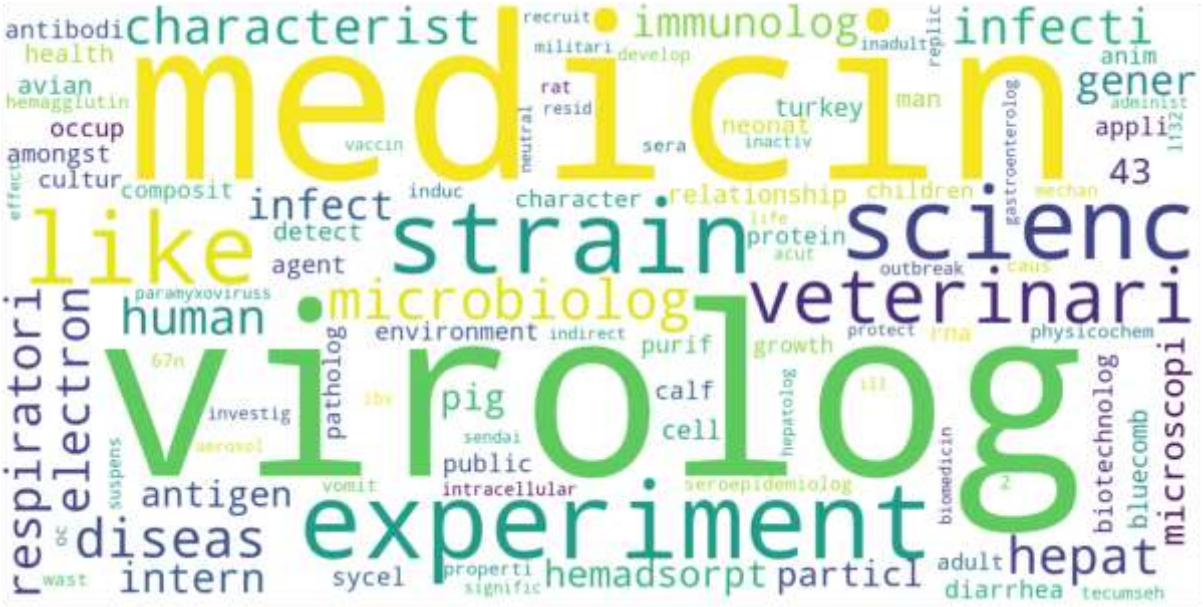
جدول دو تا ۱۲ ۳۰ واژگان مهم به کار رفته در انتشارات جهانی کرونا ویروس را در دوره های زمانی معین نشان داده است، بر همین اساس اشکال شماره ۸ تا ۱۸ نیز ابرواژگان انتشارات جهانی کرونا ویروس را در آن دوره زمانی معین نشان داده است.

در شکل ابر واژگان کلمات و واژه های بزرگتر از اهمیت بالاتری نسبت به سایر واژه ها در اسنااد مورد بررسی برخوردار هستند. این جداول و اشکال مهمترین واژگان مرتبط با کرونا ویروس را در دوره های زمانی معین به خوبی نشان داده است.

دوره زمانی ۱۹۷۴-۱۹۷۰

جدول ۲: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۷۴-۱۹۷۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	virolog	2.234692	16	immunolog	1.095281
2	medicin	2.070715	17	gener	1.043682
3	experiment	1.851236	18	intern	1.043682
4	strain	1.685368	19	infect	1.037783
5	scienc	1.628741	20	hemadsorpt	1.011892
6	like	1.587648	21	antigen	0.934055
7	veterinari	1.58687	22	pig	0.924176
8	microbiolog	1.525408	23	particl	0.885587
9	diseas	1.396722	24	microscopi	0.883801
10	hepat	1.297586	25	relationship	0.862204
11	characterist	1.227022	26	bluecomb	0.858722
12	infecti	1.180645	27	turkey	0.858722
13	human	1.170892	28	cell	0.851068
14	respiratori	1.134459	29	antibodi	0.8101
15	electron	1.116209	30	environment	0.802652

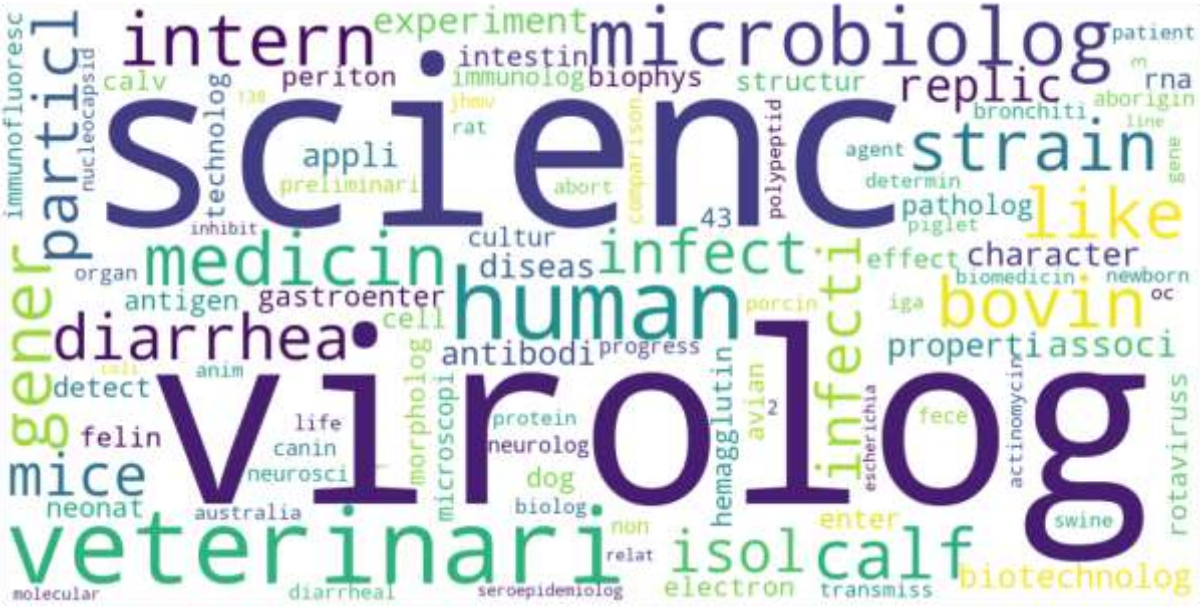


شکل ۸. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۷۴-۱۹۷۰

دوره زمانی ۱۹۷۹-۱۹۷۵

جدول ۳: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۷۹-۱۹۷۵

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	virolog	4.940521	16	infect	2.543093
2	scienc	4.567458	17	isol	2.405314
3	veterinari	4.523222	18	mice	2.401628
4	human	4.084453	19	replic	2.332571
5	microbiolog	3.846731	20	experiment	2.295584
6	medicin	3.206674	21	associ	2.240821
7	intern	3.182849	22	biotechnolog	2.19148
8	gener	3.020025	23	properti	2.168844
9	bovin	2.941326	24	character	2.144349
10	strain	2.875573	25	antibodi	2.101399
11	calf	2.831411	26	appli	2.093
12	like	2.801798	27	diseas	2.049649
13	diarrhea	2.799995	28	gastroenter	2.013509
14	particl	2.689199	29	intestin	1.913372
15	infecti	2.568147	30	rna	1.890669

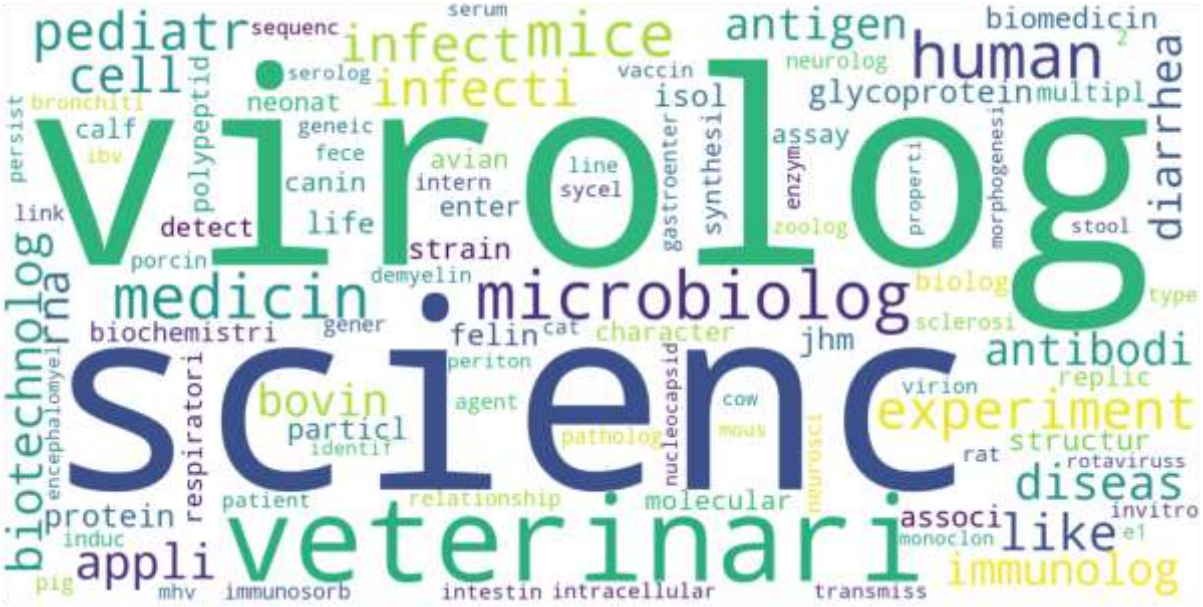


شکل ۹. ابرواژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۷۹-۱۹۷۵

دوره زمانی ۱۹۸۴-۱۹۸۰

جدول ۴: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۸۴-۱۹۸۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	virolog	10.53659	16	diseas	5.921216
2	scienc	9.655163	17	immunolog	5.885889
3	veterinari	9.066526	18	bovin	5.822581
4	microbiolog	8.766856	19	rna	5.814931
5	human	8.191571	20	antibodi	5.749395
6	mice	7.688369	21	antigen	5.61262
7	medicin	7.442954	22	diarrhea	5.527209
8	experiment	7.383938	23	glycoprotein	5.492262
9	pediatr	6.947632	24	protein	5.458563
10	infect	6.638935	25	jhm	5.255443
11	like	6.535287	26	isol	5.187459
12	infecti	6.446291	27	structur	4.99218
13	cell	6.117629	28	felin	4.898544
14	biotechnolog	5.941455	29	strain	4.635243
15	appli	5.941455	30	biomedicin	4.628072

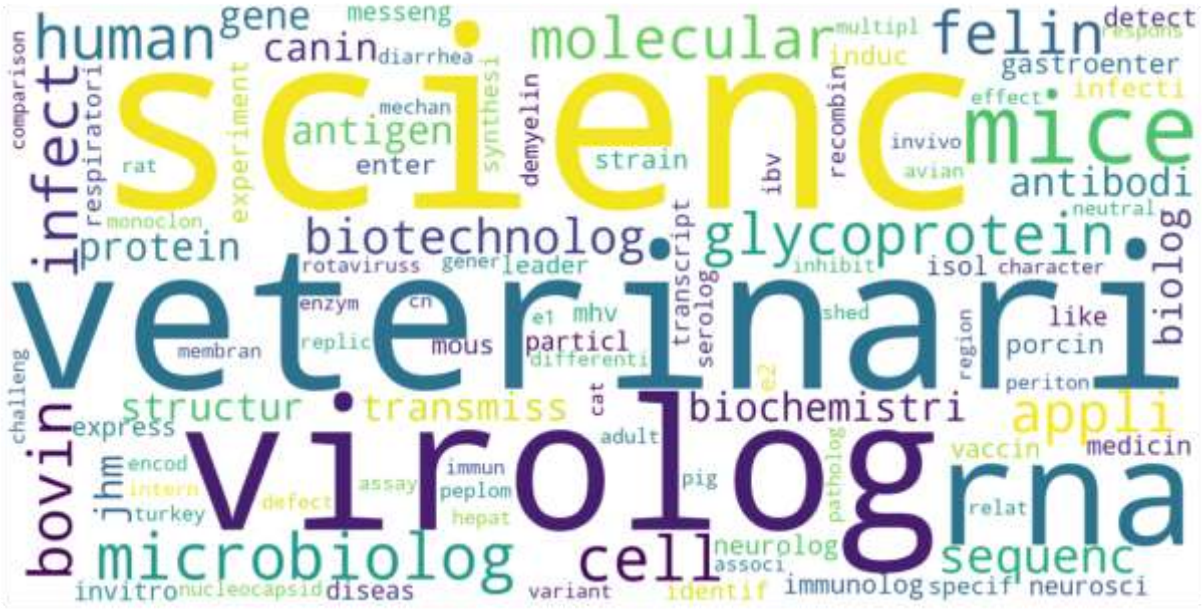


شکل ۱۰. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۸۴-۱۹۸۰

دوره زمانی ۱۹۸۹-۱۹۸۵

جدول ۵: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۸۹-۱۹۸۵

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	scienc	11.22675	16	sequenc	6.137285
2	veterinari	10.92477	17	biochemistri	5.918379
3	virolog	10.88295	18	gene	5.828028
4	rna	9.765854	19	protein	5.723071
5	mice	9.004116	20	antigen	5.690168
6	microbiolog	8.129129	21	antibodi	5.683462
7	cell	7.655577	22	transmiss	5.390346
8	human	7.539678	23	canin	5.30111
9	infect	7.511839	24	structur	5.26307
10	bovin	7.115721	25	biolog	5.233675
11	felin	6.672372	26	jhm	5.191457
12	glycoprotein	6.278656	27	gastroenter	5.030894
13	molecular	6.263146	28	porcin	4.780367
14	biotechnolog	6.16694	29	enter	4.729001
15	appli	6.16694	30	induc	4.635416

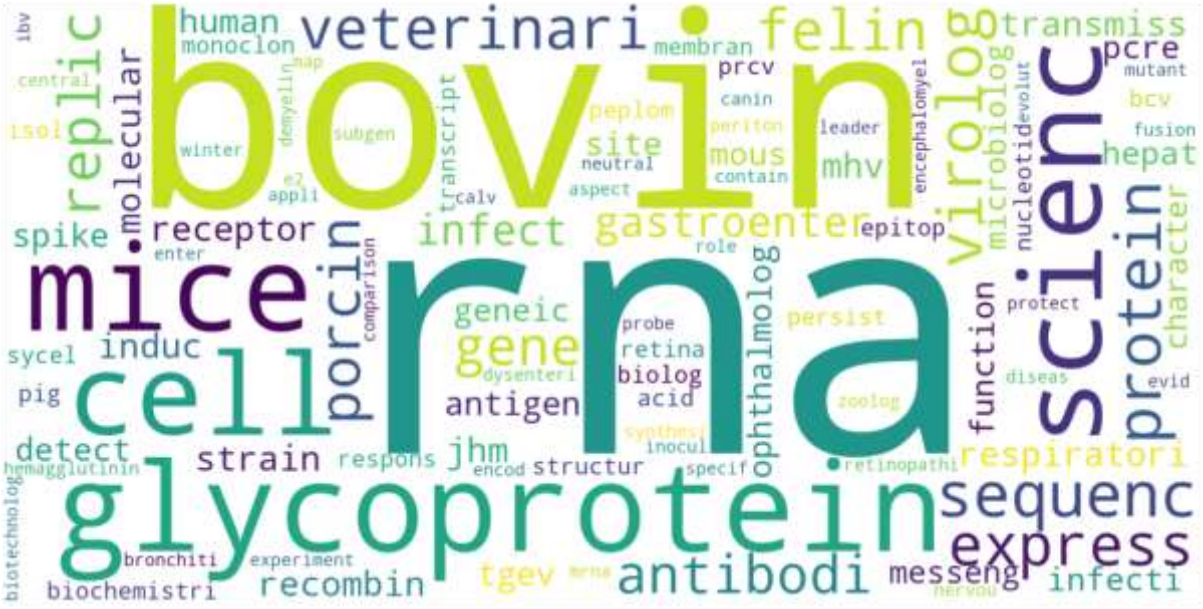


شکل ۱۱. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۸۹-۱۹۸۵

دوره زمانی ۱۹۹۴-۱۹۹۰

جدول ۶: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۹۴-۱۹۹۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	rna	9.199659	16	gene	4.919353
2	bovin	6.891564	17	gastroenter	4.882954
3	glycoprotein	6.865815	18	infect	4.824446
4	mice	6.442043	19	respiratori	4.785099
5	cell	5.770077	20	molecular	4.755073
6	scienc	5.756795	21	strain	4.737434
7	sequenc	5.753772	22	transmiss	4.727857
8	protein	5.728351	23	receptor	4.681765
9	virolog	5.586933	24	induc	4.53969
10	veterinari	5.323153	25	mhv	4.526854
11	express	5.276718	26	recombin	4.448481
12	felin	5.05487	27	jhm	4.389839
13	replic	5.014319	28	human	4.280516
14	porcin	4.956669	29	tgev	4.17395
15	antibodi	4.945555	30	site	4.104986

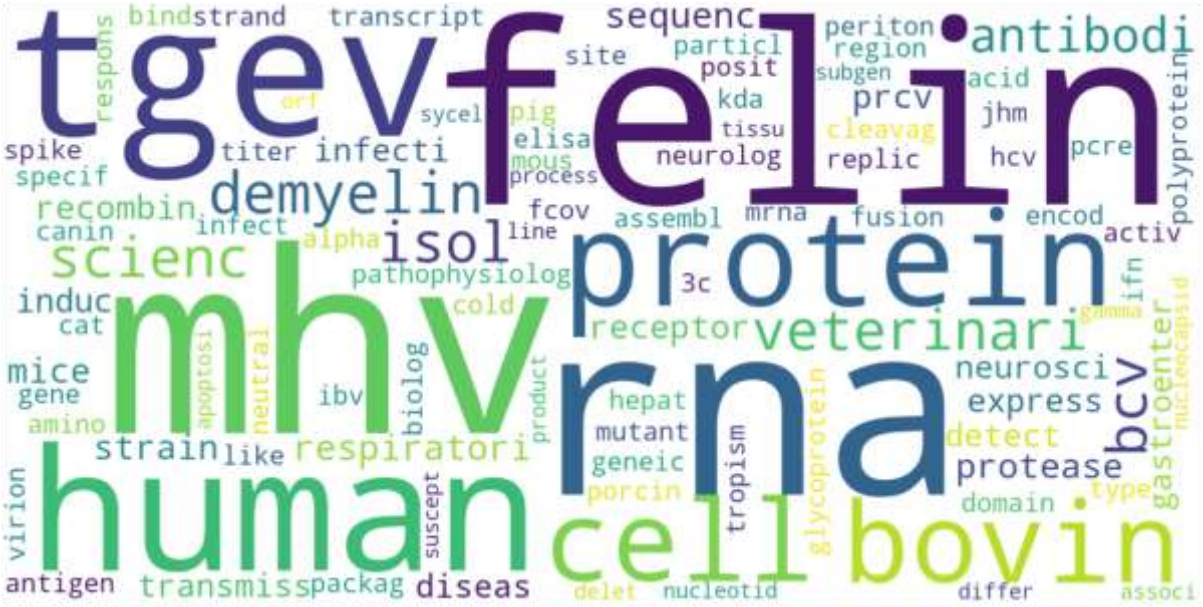


شکل ۱۲. ابرواژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۹۴-۱۹۹۰

دوره زمانی ۱۹۹۹-۱۹۹۵

جدول ۷: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۹۹-۱۹۹۵

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	mhv	4.457778	16	sequenc	2.624647
2	felin	4.364362	17	protease	2.541933
3	rna	4.128483	18	strain	2.417498
4	tgev	3.647472	19	mice	2.383494
5	human	3.096765	20	recombin	2.263649
6	protein	3.083766	21	prcv	2.235314
7	cell	3.016254	22	receptor	2.22291
8	bovin	2.987302	23	induc	2.146945
9	demyelin	2.920744	24	express	2.134462
10	bcv	2.762471	25	neurosci	2.038733
11	veterinari	2.729265	26	diseas	2.029748
12	isol	2.72609	27	transmiss	2.0291
13	scienc	2.717611	28	detect	2.0067
14	antibodi	2.684135	29	gastroenter	2.000823
15	respiratori	2.657613	30	infecti	1.986726

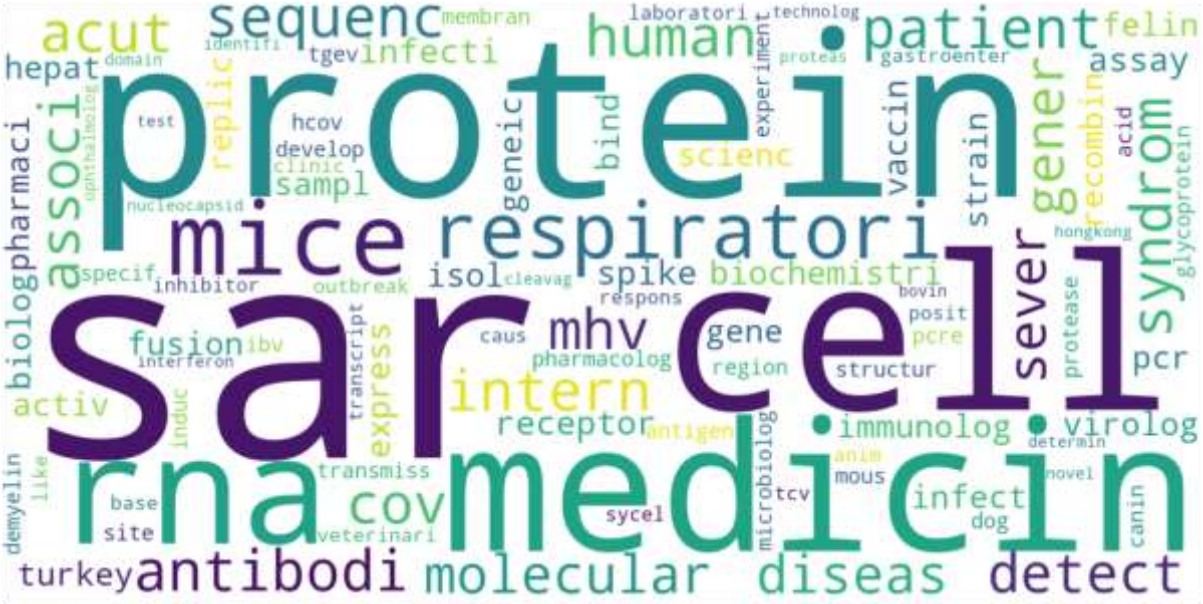


شکل ۱۳. ابرواژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۱۹۹۹-۱۹۹۵

دوره زمانی ۲۰۰۴-۲۰۰۰

جدول ۸: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۰۴-۲۰۰۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	sar	14.89556	16	mhv	5.556315
2	protein	7.607677	17	associ	5.548664
3	cell	6.896489	18	gener	5.31152
4	medicin	6.127781	19	detect	5.251809
5	rna	6.082367	20	molecular	4.918983
6	mice	6.042668	21	diseas	4.918893
7	respiratori	5.886149	22	sever	4.889543
8	sequenc	5.8573	23	receptor	4.828367
9	human	5.848199	24	infecti	4.784978
10	patient	5.83341	25	assay	4.703011
11	syndrom	5.782379	26	spike	4.679589
12	intern	5.738473	27	express	4.489658
13	antibodi	5.735749	28	isol	4.424045
14	acut	5.626481	29	vaccin	4.418674
15	cov	5.575543	30	biochemistri	4.408628



شکل ۱۴. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۰۴-۲۰۰۰

دوره زمانی ۲۰۰۹-۲۰۰۵

جدول ۹: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۰۹-۲۰۰۵

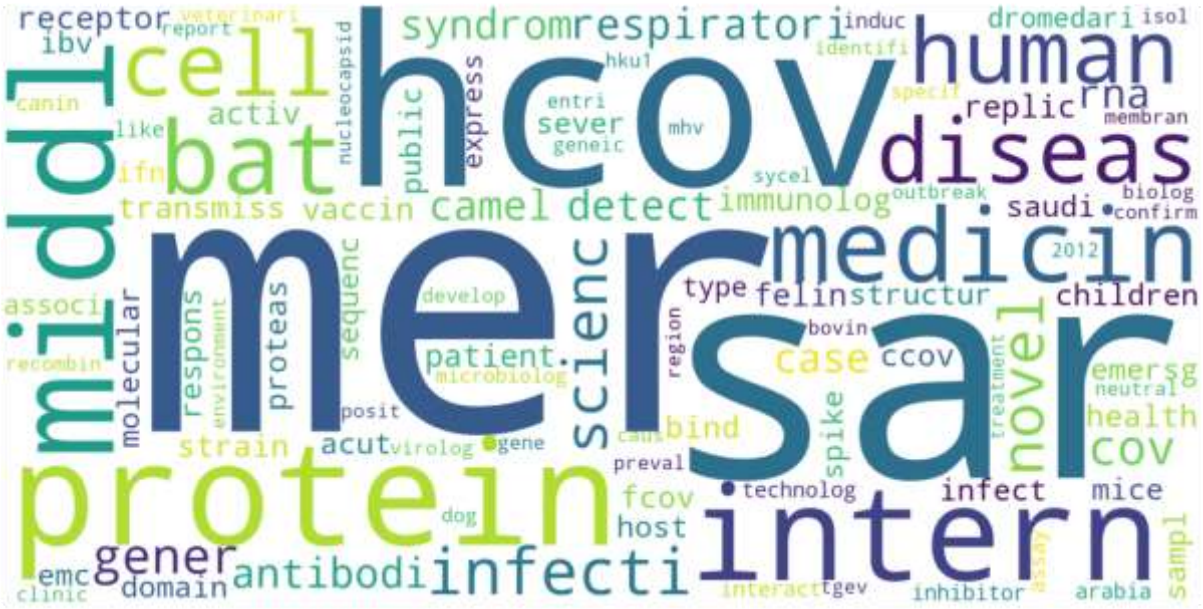
No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	sar	14.61297	16	fusion	7.700817
2	protein	13.76919	17	vaccin	7.639387
3	human	11.79959	18	associ	7.538292
4	hcov	11.70283	19	bind	7.526169
5	cell	11.531	20	replic	7.490637
6	antibodi	10.33576	21	detect	7.461764
7	diseas	10.20709	22	nucleocapsid	7.350627
8	rna	9.579879	23	express	7.243263
9	structur	9.10391	24	domain	7.229944
10	mice	8.890629	25	sequenc	7.209071
11	infecti	8.365669	26	respons	7.184
12	cov	8.225498	27	felin	6.987089
13	receptor	8.112245	28	activ	6.969257
14	proteas	7.870577	29	children	6.754565
15	spike	7.838312	30	type	6.728563

شکل ۱۵. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۰۹-۲۰۰۵

دوره زمانی ۲۰۱۴-۲۰۱۰

جدول ۱۰: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۱۴-۲۰۱۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	mer	12.71098	16	cov	7.286838
2	sar	12.48556	17	respiratori	7.224465
3	hcov	10.8714	18	camel	7.035881
4	protein	10.35923	19	rna	6.973949
5	middl	9.78609	20	antibodi	6.89631
6	intern	9.655664	21	syndrom	6.805376
7	medicin	9.556601	22	detect	6.740191
8	bat	9.452478	23	case	6.60006
9	diseas	8.914911	24	immunolog	6.546548
10	human	8.895216	25	receptor	6.527402
11	cell	8.892759	26	felin	6.302513
12	infecti	8.829528	27	activ	6.218115
13	novel	8.01145	28	patient	6.174271
14	scienc	7.930456	29	replic	6.158379
15	gener	7.567022	30	strain	5.94959

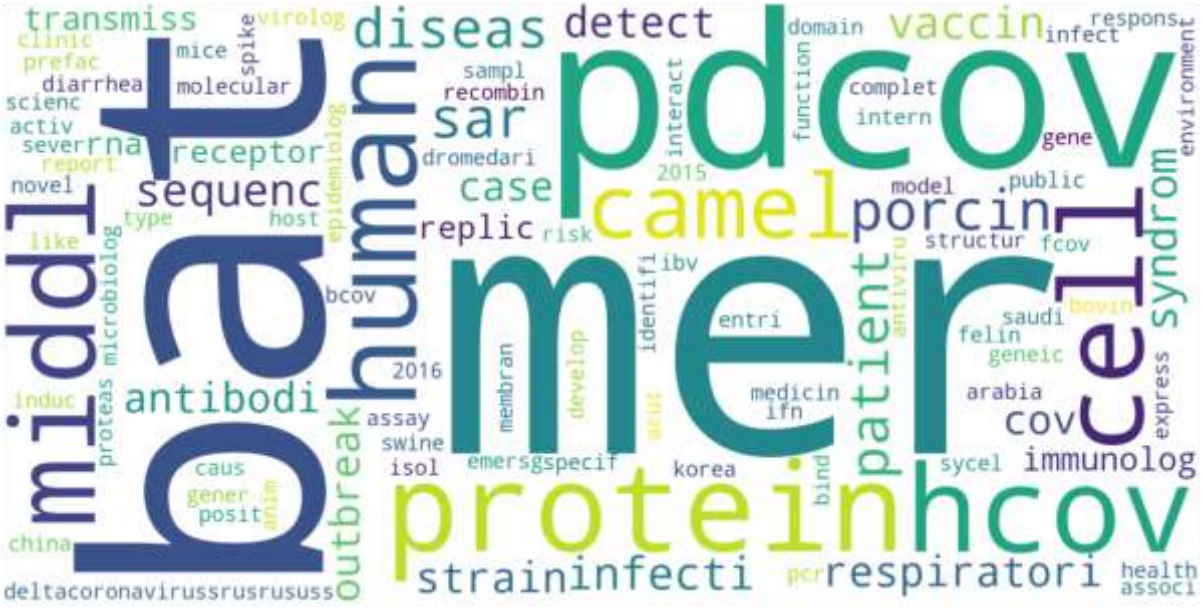


شکل ۱۶. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۱۴-۲۰۱۰

دوره زمانی ۲۰۱۹-۲۰۱۵

جدول ۱۱: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۱۹-۲۰۱۵

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	mer	16.33597	16	syndrom	8.938658
2	bat	13.43484	17	vaccin	8.904365
3	pdcov	12.48145	18	respiratori	8.866603
4	protein	12.29108	19	detect	8.750852
5	hcov	11.50728	20	strain	8.626559
6	human	11.23645	21	antibodi	8.428922
7	middl	10.97033	22	outbreak	8.353423
8	cell	10.71429	23	cov	8.215657
9	camel	10.34737	24	case	7.975594
10	porcin	10.16952	25	immunolog	7.968952
11	sar	10.12387	26	transmiss	7.823613
12	patient	9.591707	27	rna	7.375845
13	diseas	9.560869	28	receptor	7.344309
14	sequenc	9.186165	29	replic	7.328039
15	infecti	9.111715	30	spike	7.280705



شکل ۱۷. ابر واژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۱۹-۲۰۱۵

دوره زمانی سال ۲۰۲۰

جدول ۱۲: مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۲۰

No	Keywords	TF-IDF Weight	No	Keywords	TF-IDF Weight
1	intern	4.667363	16	emersg	2.030008
2	medicin	4.660293	17	wuhan	2.003299
3	gener	4.392905	18	pdcov	1.964452
4	chemistri	3.543201	19	case	1.921509
5	engin	3.522188	20	diseas	1.908429
6	covid	3.369206	21	human	1.879251
7	19	3.239838	22	global	1.786458
8	ncov	3.086084	23	respons	1.766238
9	scienc	3.000812	24	public	1.714515
10	novel	2.958239	25	sar	1.684514
11	technolog	2.778203	26	patient	1.684043
12	outbreak	2.605343	27	pneumonia	1.632082
13	china	2.433129	28	respiratori	1.614598
14	health	2.268654	29	infecti	1.5592
15	vaccin	2.148692	30	protein	1.428665

شکل ۱۸. ابرواژگان مهمترین واژگان به کار رفته در انتشارات جهانی کرونا ویروس در دوره زمانی ۲۰۲۰

موضوع انتشارات جهانی کروناویروس در نیم قرن اخیر

نتایج حاصل از اجرای الگوریتم مدل‌سازی موضوعی LDA در جدول ۱۳ نشان داده‌شده است. در این جدول ۸ موضوع بدست آمده به همراه مهمترین واژگان به همراه مرتبط ترین مقالات هر موضوع نشان داده شده است.

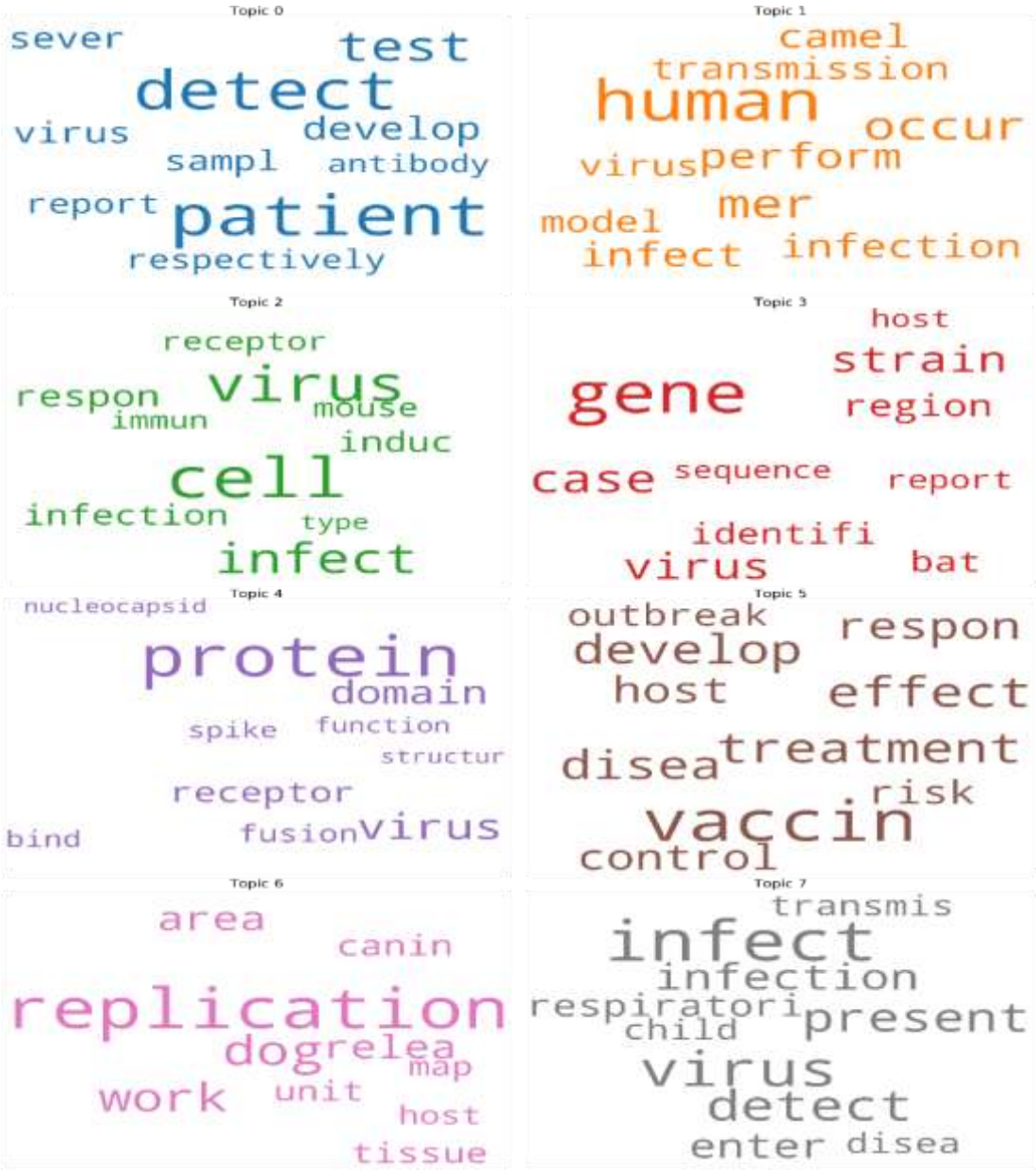
جدول ۱۳: موضوعات حاصل از الگوریتم مدل‌سازی موضوعی مقالات

Topic \Num Doc		10 Top Keywords	Top Related Articles
0	Coronavirus: diagnosis tests	patient, detect, test, develop, sever, virus, sampl, report, respectively, antibody	Incidence Of Bovine Enterovirus, Coronavirus, And Group A Rotavirus, And Concentration Of Total Coliforms In Midwestern Pasture Streams
			An Immunofluorescence Test For Detection Of Serum Antibody To Rodent Coronaviruses
			Preliminary-Observations On Enteritis Associated With A Coronavirus-Like Agent In Rabbits
			Detection And Monitoring Of SARS Coronavirus In The Plasma And Peripheral Blood Lymphocytes Of Patients With Severe Acute Respiratory Syndrome
			Comparison Of Coronavirus Sk Specific Serum Antibody-Levels In Multiple-Sclerosis Patients And Controls
1	Coronavirus: Epidemiology and Transmission	human, mer, occur, perform, infect, infection, camel, transmission, model, virus	Use Of Clinical Criteria And Molecular Diagnosis To More Effectively Monitor Patients Recovering After Severe Acute Respiratory Syndrome Coronavirus Infection
			Seroepidemiology For MERS Coronavirus Using Microneutralisation And Pseudoparticle Virus Neutralisation Assays Reveal A High Prevalence Of Antibody In Dromedary Camels In Egypt, June 2013
			Neonatal Calf Diarrhea - Propagation, Attenuation, And Characteristics Of A Coronavirus-Like Agent
			Camels Could Be The Source Of MERS Coronavirus, Research Finds
			First Confirmed Cases Of Middle East Respiratory Syndrome Coronavirus (MERS-Cov) Infection In The United States, Updated Information On The Epidemiology Of MERS-Cov Infection, And Guidance For The Public, Clinicians,

Topic \Num Doc		10 Top Keywords	Top Related Articles
			And Public Health Authorities - May 2014 (Vol 14, Pg 1693, 2014)
2	Cell signaling and immune respone	cell, virus, infect, respon, induc, infection, mouse, receptor, immun, type	Priming Of CD8(+) T Cells During Central Nervous System Infection With A Murine Coronavirus Is Strain Dependent
			Important Roles For Gamma Interferon And NKG2D In Gamma Delta T-Cell-Induced Demyelination In T-Cell Receptor Beta-Deficient Mice Infected With A Coronavirus
			Maintenance Of Pluripotency In Mouse Embryonic Stem Cells Persistently Infected With Murine Coronavirus
			Rat Coronaviruses Infect Rat Alveolar Type I Epithelial Cells And Induce Expression Of CXC Chemokines
			Cutting Edge: CD8 T Cell-Mediated Demyelination Is IFN-Gamma Dependent In Mice Infected With A Neurotropic Coronavirus
3	Coronavirus: Gene sequence and genomics	gene, case, virus, strain, region, bat, identifi, report, host, sequence	Complete Genome Sequence Of A Novel Swine Acute Diarrhea Syndrome Coronavirus, CH/FJWT/2018, Isolated In Fujian, China, In 2018
			Common RNA Replication Signals Exist Among Group 2 Coronaviruses: Evidence For In Vivo Recombination Between Animal And Human Coronavius Molecules
			Complete Genome Sequence Of Porcine Deltacoronavirus Strain CH/Sichuan/S27/2012 From Mainland China
			3 Intergenic Regions Of Coronavirus Mouse Hepatitis-Virus Strain-A59 Genome Rna Contain A Common Nucleotide-Sequence That Is Homologous To The 3' End Of The Viral Messenger-Rna Leader Sequence
			Effect Of Intergenic Consensus Sequence Flanking Sequences On Coronavirus Transcription
4	Coronavirus : structure and proteomics	protein, virus, domain, receptor, fusion, bind, spike, function,	Palmitoylation Of The Alphacoronavirus TGEV Spike Protein S Is Essential For Incorporation Into Virus-Like Particles But Dispensable For S-M Interaction

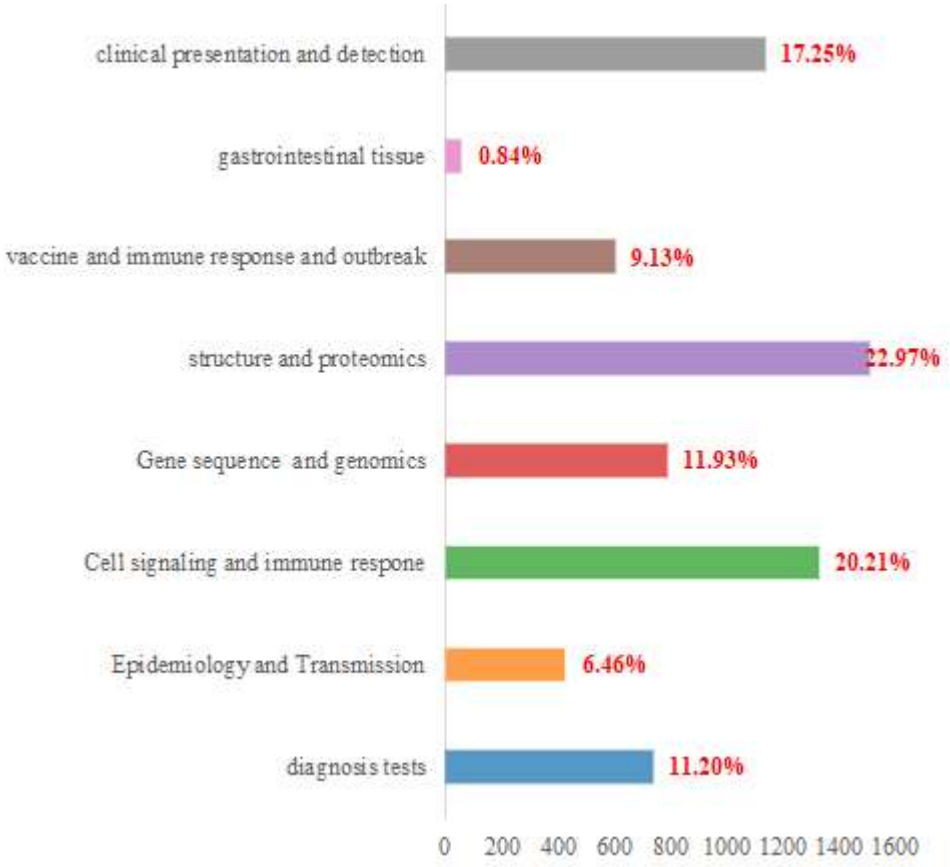
Topic \Num Doc		10 Top Keywords	Top Related Articles
		structur, nucleocapsid	Incorporation Of Spike And Membrane Glycoproteins Into Coronavirus Virions
			Structure Of A Conserved Golgi Complex-Targeting Signal In Coronavirus Envelope Proteins
			Coronavirus N Protein N-Terminal Domain (NTD) Specifically Binds The Transcriptional Regulatory Sequence (TRS) And Melts TRS-Ctrs RNA Duplexes
			Characterisation Of A Papain-Like Proteinase Domain Encoded By ORF1a Of The Coronavirus IBV And Determination Of The C-Terminal Cleavage Site Of An 87 Kda Protein
5	Coronavirus and vaccine and immune response and outbreak	vaccin, effect, treatment, develop, disea, respon, host, control, risk, outbreak	Middle East Respiratory Syndrome Coronavirus Seropositivity In Camel Handlers And Their Families, Pakistan
			Age-Dependent Resistance To Transmissible Gastroenteritis Of Swine .3. Effects Of Epithelial-Cell Kinetics On Coronavirus Production And On Atrophy Of Intestinal Villi
			The Novel Chinese Coronavirus (2019-Ncov) Infections: Challenges For Fighting The Storm
			Effect Of Recent Vaccination On Feline Coronavirus Antibody-Test Results
			Effect Of Specific Humoral Immunity And Some Non-Specific Factors On Resistance Of Volunteers To Respiratory Coronavirus Infection
6	Coronavirus and gastrointestinal tissue	replication, dog, work, area, relea, canin, tissue, unit, map, host	Quantitation, Phenotypic Characterization And Insitu Localization Of Lymphoid-Cells In The Brain Parenchyma Of Rats With Differing Susceptibility To Coronavirus Jhm-Induced Encephalomyelitis
			Penetration, Replication, And Release Of Coronavirus Particles In Infected Neonatal Small Intestinal Epithelium
			Coronavirus-Like Particles In Stools From Dogs, From Some Country Areas Of Australia
			Are Coronavirus-Like Particles Seen In Diarrhea Stools Really Viruses

Topic \Num Doc		10 Top Keywords	Top Related Articles
			Dissecting The Mechanism Of Host Shutoff By SARS Coronavirus
7	Coronavirus: clinical presentation and detection	infect, virus, detect, present, infection, enter, respiratori, disea, transmis, child	Mechanical Transmission Of Turkey Coronavirus By Domestic Houseflies (Musca Domestica Linnaeaus)
			Detection Of Respiratory And Enteric Shedding Of Bovine Coronaviruses In Cattle In Northwestern Turkey
			The Detection Of Feline Coronaviruses In Blood Samples From Cats By Mrna RT-PCR
			Proving Etiologic Relationships To Disease The Particular Problem Of Human Coronaviruses
			Isolation Of Respiratory Bovine Coronavirus, Other Cytocidal Viruses, And Pasteurella Spp From Cattle Involved In Two Natural Outbreaks Of Shipping Fever



شکل ۱۹. ابر واژگان موضوعات بدست آمده از الگوریتم مدل سازی انتشارات کروناویروس (۱۹۷۰-۲۰۲۰)

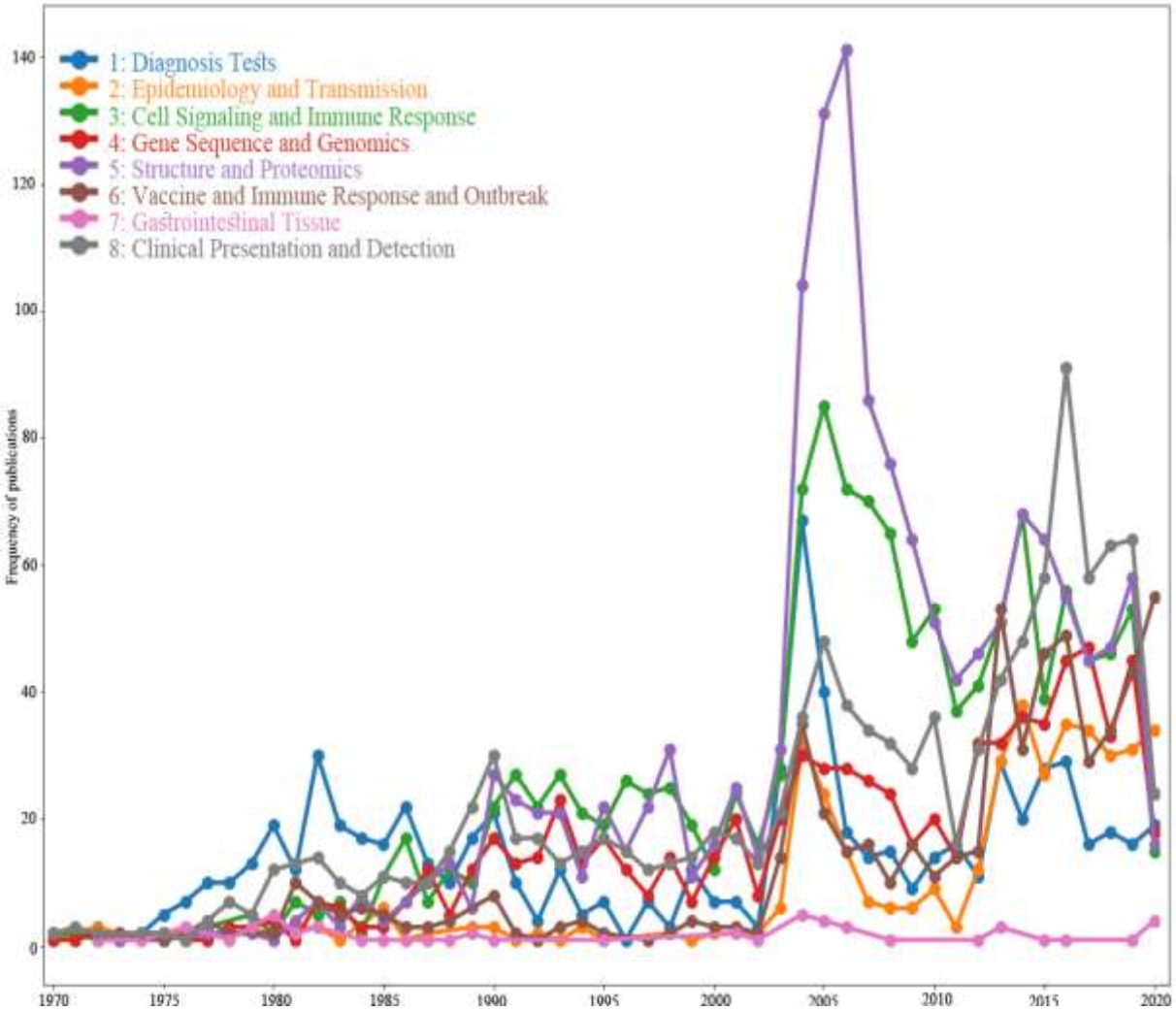
شکل ۱۹ نیز به صورت مصور، ده واژگان مهم هر کدام از موضوعات در قالب ابر واژگان نشان داده است. در ابر واژگان، واژگان دارای فونت بزرگتر، دارای اهمیت و کاربرد بیشتری در آن موضوع می باشند



نمودار ۲. سهم مقالات منتشرشده در هرکدام از موضوعات بدست آمده از الگوریتم مدل سازی انتشارات کروناویروس (۱۹۷۰-۲۰۲۰)

نمودار ۲ میزان انتشار مقالات در هر کدام از موضوعات را نشان داده است. موضوع شمار ۴ بیشترین میزان انتشار را داشته است و کمترین میزان انتشار نیز مربوط به موضوع شماره ۶ بوده است.

روند انتشار موضوعات انتشارات جهانی کروناویروس در نیم قرن اخیر



نمودار ۳. روند انتشار مقالات کروناویروس در موضوع‌های به‌دست‌آمده از الگوریتم مدل سازی (۱۹۷۰-۲۰۲۰)

نمودار ۳ روند انتشار موضوعات مختلف کرونا ویروس در جهان را نشان می دهد. اوج انتشار موضوعات مختلف کرونا ویروس از سال ۲۰۰۳ به بعد شروع شده و در سال‌های ۲۰۰۴ و ۲۰۰۵به اوج خود رسیده است. بعد از آن انتشار در موضوعات مختلف کاهش داشته است سپس از سال ۲۰۱۲ روند انتشار افزایشی موضوعات شروع شده است. نگاهی به نمودار بالا نشان می دهد که موضوع structure and proteomics بالاترین سهم انتشارات کروناویروس در نیم قرن اخیر را به خود اختصاص داده است. این موضوع در سال‌ها ۲۰۰۵ و ۲۰۱۵ بیشترین سهم انتشارات را داشته است.

موضوع Cell signaling and immune response نیز از موضوعات پر انتشار بوده است و بیشترین تعداد مقالات منتشر شده این موضوع در سال ۲۰۰۴ بوده است و سپس روند آرامی داشته است، قله دوم این موضوع در سال ۲۰۱۵ بوده است. در رده بعدی بیشترین انتشار موضوع clinical presentation and detection بوده است که بیشترین میزان انتشار این موضوع در سال ۲۰۰۳ بوده است، سپس در سال ۲۰۱۶ بیشترین انتشار این موضوع بوده است. این موضوع در ۵ سال اخیر بیشترین انتشار را در بین سایر موضوعات

داشته است و مورد توجه پژوهشگران بوده است. موضوع Gene sequence and genomics نیز از موضوعاتی هست که روند انتشار آرامی را در طول زمان داشته است بیشترین میزان انتشار این موضوع در سال ۲۰۰۳ بوده است سپس بیشترین میزان انتشار بعدی این موضوع در سال های ۲۰۱۶ و ۲۰۱۷ بوده است. موضوع diagnosis tests نیز بیشترین انتشار را در سال ۲۰۰۳ داشته است و بعد از آن نیز روند انتشار ملایمی را داشته است. موضوع vaccine and immune response and outbreak نیز بیشترین رشد را در سال ۲۰۱۳ داشته است. موضوع Epidemiology and Transmission نیز بیشترین انتشار را در سال ۲۰۰۳ و ۲۰۱۴ داشته است. موضوع gastrointestinal tissue نیز کمترین میزان انتشار را در بین سایر موضوعات داشته است و به طور کلی روند انتشاراتی آرامی را در طول زمان داشته است.

فصل پنجم

نتیجه‌گیری و پیشنهادها

نتیجه‌گیری

در سال های اخیر به دلیل رویدادها و کوشش های پژوهشگران در جهت اشراف همه جانبه بر شرایط اپیدمیک و پاندمیک قلمرو موضوعی کرونا، تولید علم در این حوزه افزایش یافته است(۶۷). nCoV-2019 که به کووید۱۹ معروف است، شکل جدید ویروسی است که اولین گزارش آن در ژانویه ۲۰۲۰ ارائه شده است. با توجه به اهمیت این موضوع تولیدات و انتشارت علمی این حوزه نیز در حال افزایش است و بررسی موضوعات منتشر شده کروناویروس دارای اهمیت ویژه ای برای متخصصین و پژوهشگران است. مدل سازی موضوعی به عنوان ابزار متن کاوی برای پردازش ، سازماندهی ، مدیریت و استخراج دانش عمل می کند و معمولاً برای تعیین "موضوعات" اساسی در متون مورد استفاده قرار می گیرد(۶۸) و می تواند یک نمای مفید از یک مجموعه بزرگ مجموعه به عنوان یک کل، اسناد مستقل و روابط بین اسناد را ارائه دهد(۶۹). متن کاوی و مدل‌سازی موضوعی انتشارات کروناویروس، که موضوع اصلی این پژوهش است، تصویر روشنی از موضوعات منتشر شده کروناویروس در نیم قرن اخیر را نشان می‌دهد. نتایج این پژوهش حاکی از آن است که روند رشد انتشار در موضوعات مختلف کرونا ویروس بین سال‌های ۲۰۰۳ تا ۲۰۰۶ بوده و پس از آن تا سال ۲۰۱۲ در حال کاهش بوده است. از سال ۲۰۱۲ تا کنون نیز روند روند ثابتی را نداشته است و در مواقع شیوع بیماری و ویروس مانند mers و sars رشد مشهودی داشته است. پژوهش های گذشته نیز نشان دادند که شیوع نوع جدیدی از کرونا ویروس یا نام Sars-cov از سال ۲۰۰۳ به بعد دلیل اصلی رشد انتشارات این حوزه بوده است (۷۰-۷۲) ، همچنین در سپتامبر ۲۰۱۲ سازمان بهداشت جهانی (who) مواردی از پنومونی شدید ناشی از کرونا ویروس جدید انسانی (novel human β-coronavirus) که سندرم تنفسی خاور میانه (Middle East respiratory syndrome coronavirus (MERS-CoV)) نام گذاری شده را گزارش داده است(۷۳). Ram نیز نشان داد که پژوهش های Cov تا سال ۲۰۰۲ روند آهسته ای را داشته است و بیشترین میزان انتشار این موضوع در سالهای ۲۰۰۳ و ۲۰۱۳ بوده است(۲۲). همچنین در سپتامبر ۲۰۱۲ سازمان بهداشت جهانی (World Health Organization) مواردی از پنومونی شدید ناشی از کرونا ویروس جدید انسانی (novel human β-coronavirus) که سندرم تنفسی خاورمیانه (Middle East

respiratory syndrome coronavirus (MERS-CoV) نام گذاری شده را گزارش داده است(۷۴) که دلیلی بر رشد انتشارات این حوزه در سال ۲۰۱۲ به بعد بوده است.

Kagan و همکاران نیز نشان دادند که علاقه جامعه پژوهش به یک ویروس و بیماری در حال ظهور، به طور موقت با پویایی این وضعیت همراه است و یک افت شدید علاقه بعد از فروکش شدن اپیدمی اولیه به وجود می آید(۷۵).

نتایج پژوهش حاضر ده واژگان پر کاربرد در انتشارات حوزه کرونا ویروس با استفاده از روش TF-IDF (۷۶) را واژگان

SARS, science, protein, MERS, veterinary, cell, human, RNA, medicine, virology

نشان داده است. که این ها واژگان اصلی و پرکاربرد در انتشارات حوزه کرونا ویروس هستند که نشان دهنده تم و قالب اصلی تحقیقات این حوزه علمی است. و به طور کلی حوزه علمی کرونا ویروس را نشان می دهد، به هر حال SARS, MERS جزئی از خانواده کرونا ویروس ها هستند و بیشترین میزان کاربرد در انتشارات این حوزه را تا حال حاضر داشته است. گارفیلد بیان نموده است که واژگان اصلی، مجموعه ای از اصطلاحات هستند موضوعات مورد بحث در یک سند را نشان می دهد (۷۷). Ram نیز نشان داده است که تحقیقات در مورد CoV حول کلمات کلیدی " "Coronavirus، SARS، MERS و آنفلوانزا ، که متداول ترین کلمات کلیدی در آن هست، می چرخد(۷۴). Haghani نیز نشان داد که واژگان متداول در ادبیات حوزه کرونا ویروس

human* and Coronavirus infection* and viral pneumonia (including their variations, humans, Coronavirus infections and pneumonia virus)

بوده است(۷۸).

Mahbub Hossain نشان داد که کلمات کلیدی بکار رفته در انتشارات حوزه کوید۱۹ بیانگر پیچدگی و گستردگی این حوزه علمی است و رشته های مختلفی از قبیل

virology, microbiology, infectious diseases, clinical medicine, public health, allied health sciences, social sciences, and other branches of knowledge.

را شامل می شود(۷۹).

نتایج حاصل از مدل‌سازی موضوعی ۸ موضوع را برای انتشارات جهانی کرونا ویروس نشان داده است. عناوین هر کدام از موضوعات به ترتیب بیشترین انتشارات هر موضوع عبارتند از:

“structure and proteomics”, “Cell signaling and immune response”, “clinical presentation and detection”, “Gene sequence and genomics”, “Diagnosis tests”, ‘‘vaccine and immune response and outbreak”, “Epidemiology and Transmission” and “gastrointestinal tissue” .

Colavizza در این خصوص نشان داد در مجموعه پژوهش‌های Cord19 (COVID-19 Open Research Dataset) موضوعات این مجموعه بر مباحث خاصی مانند: کرونا ویروس ها(SARS، MERS و Covid19)، بهداشت عمومی و همه گیری ویروس، زیست شناسی مولکولی ویروس، آنفولانزا و خانواده های ویروس، ایمونولوژی و آنتی ویروس ها و روش شناسی (آزمایش، تشخیص، آزمایشات بالینی) متمرکز است(۸۰).

در سال ۲۰۲۰ با شیوع COVID-19 بیشترین انتشار در موضوعات

‘‘vaccine and immune response and outbreak’’, “Clinical presentation and detection”, “structure and proteomics”

است. موضوع ’’vaccine and immune response and outbreak‘‘ در سال ۲۰۲۰ بیشترین میزان انتشار را در بین سایر موضوعات داشته است. اهمیت این موضوع از چندین جنبه قابل بررسی است پژوهش بر روی واکسن کروناویروس و ایمنی انسان‌ها در مقابله با ابتلا به کروناویروس مهمترین و پر استنادترین موضوع مورد پژوهش است. چرا که با کشف واکسن علاوه بر ایجاد ایمنی و حفاظت افراد در برابر ابتلا به این ویروس در کاهش نگرانی و اضطراب افراد جامعه هم موثر است. در خصوص immune response نیز از دو جنبه قابل بررسی است نخست اینکه با ایجاد پاسخ ایمنی هومورال Humoral immunity در بدن بیماران و تولید آنتی بادی ، می‌توان از پلاسمای این افراد پس از بهبودی در درمان سایر بیماران استفاده نمود(۸۱) و دوم اینکه با تولید انتی بادی می‌توان در طراحی تست های ازمایشگاهی تشخیصی و همچنین روش تشخیصی این بیماری متمرکز شد(۸۲). بحث outbreak در این موضوع هم در مدیریت و کنترل بیماری بسیار مهم است. با توجه

به اینکه COVID-19 ویروس جدیدی است، کلید واژه های حفاظت و پیشگیری(واکسیناسیون) ، تشخیص و درمان(پاسخ ایمنی) و میزان درگیری و morbidity و mortality (در قالب شیوع) گنجانده می شود(۸۳) و طبیعی است که در وضعیت فعلی بیشترین مطالعات و تحقیقات در رابطه با این موضوع انجام شود. Colavizzaنشان داده است که شیوع SARS در سال ۲۰۰۳ با افزایش انتشارات مربوط به کروناویروس و مدیریت اپیدمی همراه بود. از این‌رو، در گذر زمان آثار و انتشارات بیولوژی مولکولی ویروس‌ها همواره غالب بوده است، همچنین با اپیدمی COVID-19، همانطور که انتظار می رود ، تعداد زیادی از مدارک در درجه نخست در مورد موضوعات مربوط به کروناویروس و مدیریت اپیدمی آن منتشر می شوند(۸۰).

Mahbub Hossain نیز انتشارات علمی کووید ۱۹ را در دو خوشه دسته بندی نموده است، خوشه اول مربوط به اصطلاحات پژوهشی

research terms including diversity , multiple sequence alignment, and sars-like coronaviruses.

و خوشه دوم شامل اصطلاحات مرتبط با

pathogenicity of coronavirus outbreak, earlier outbreaks with other typologies, epidemiology, and diagnostic approaches.

بوده است(۲۴). Haghani نشان داد که در جنبه های

biological and immunological aspects and of the research on vaccines and medical treatment

مورد کووید ۱۹ از اولویت های پژوهشی است، با این وجود حجم زیادی از دانش منتشر شده را مباحث ایمنی از جمله ایمنی جسمی و روانی در متخصصان بهداشتی و بیماران تشکیل داده است(۷۹).

Dehghanbanadaki نیز تمرکز پژوهشگران در خصوص کوید ۱۹ را سابجکت های

various aspects of this infection such as pathogenesis, epidemiology, transmission, diagnosis, treatment, prevention, and its complications

نشان داده است(۸۴).

نتایج پژوهش حاضر همچنین نشان داده است که کمترین میزان انتشارت نیز مربوط به موضوع gastrointestinal tissue بوده که این موضوع در کرونا ویروس جزء عوارض ناشایع و غیر معمول می باشد که در برخی از بیماران مشاهده گردیده است(۸۵، ۸۶).

بحث

این پژوهش به طور شفاف قلمروهای موضوعی انتشارات جهانی کرونا ویروس را نشان داده است. نتایج حاکی از آن است که روند انتشارات علمی حوزه کرونا ویروس، ثابت نبوده و با شیوع این ویروس در هر دوره زمانی انتشارات علمی این حوزه نیز رشد داشته است. همچنین این انتشارات از گذشته تا کنون در موضوعات مختلف از ویروس شناسی تا درمان در پاسخ به شرایط موجود منتشر می شوند و انتشارات این حوزه علمی همچنان در حال رشد و گسترش است، روند انتشار در موضوعات مختلف این حوزه نیزثابت نبوده است. همچنین بعد از شیوع covid19 بیشترین انتشار مربوط به موضوعات ویروس شناسی و ساختار ویروس، تشخیص و مدیریت اپیدمی بوده است.

است.

پیشنهاد برای پژوهش‌های آینده

- با توجه به اینکه نتایج این پژوهش موضوعات کلی انتشارات در حوزه کرونا ویروس را شناسایی نموده است، بنابراین پژوهشگران می توانند به طور خاص انتشارات هر کدام از موضوعات بدست آمده این پژوهش را با استفاده از فنون علم سنجی و متن کاوی را بررسی و تحلیل نمایند.

- نتایج این پژوهش به طور کلی انتشارات حوزه کرونا ویروس را بررسی و تحلیل نموده است، بنابراین پژوهشگران می توانند به طور خاص به تحلیل و مقایسه موضوعی انتشارات هر کدام از کرونا ویروس ها از قبیل Sars, Mers, … با استفاده از روش ها و فنون متن کاوی بپردازند.

- بررسی‌های اولیه حاکی از آن است که انتشارات مرتبط با COVID-19 از ابتدای سال ۲۰۲۰ تاکنون رشد چشمگیری داشته و تا پایان سال نیز پیش‌بینی می‌شود، انتشارات با شیب تندی به رشد خود ادامه دهند بنابراین پیشنهاد می‌گردد پژوهشگران علم‌سنجی به طور خاص COVID-19 را بررسی و روند انتشارات آن را

مشخص نمایند. همچنین متن کاوی و تحلیل متن می تواند نتایج کاربردی و ارزشمند قابل توجهی را به پژوهشگران و متخصصان نظام سلامت ارائه نماید.

منابع

1. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, Evaluation and Treatment Coronavirus (COVID-19). StatPearls. Treasure Island FL: StatPearls Publishing LLC.; 2020.
2. Wang MY, Fang SC, Chang YH. Exploring technological opportunities by mining the gaps between science and technology: Microalgal biofuels. Technological Forecasting and Social Change. 2015 Mar 1;92:182-95.
3. Behkami NA, Daim TU. Research forecasting for health information technology (HIT), using technology intelligence. Technological Forecasting and Social Change. 2012;79(3):498-508.
4. Soleimani Nezhad A, Salajegheh M, Tayyebi Nia E. Clustering scientific articles based on the k_means algorithmCase Study: Iranian Research Institute for information Science and Technology . Iranian Journal of Information Processing and Management. 2019;34(2):871-96.
5. Le MH, Ho TB, Nakamori Y. Detecting emerging trends from scientific corpora. International Journal of Knowledge and Systems Sciences. 2005 Jun;2(2):53-9.
6. Kontostathis A, De I, Holzman L, Pottenger WJP. Use of term clusters for emerging trend detection. 2004.
7. Kao A, Poteet SR. Natural language processing and text mining: Springer Science & Business Media; 2007.
8. Hashimi H, Hafez A, Mathkour H. Selection criteria for text mining approaches. Computers in Human Behavior. 2015 Oct 1;51:729-33.
9. Salloum SA, Al-Emran M, Monem AA, Shaalan K. A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. 2017 Jan;2(1):127-33.
10. Choudhary AK, Oluikpe PI, Harding JA, Carrillo PM. The needs and benefits of Text Mining applications on Post-Project Reviews. Computers in Industry. 2009 Dec 1;60(9):728-40.
11. Srivastava AN, Sahami M. Text mining: Classification, clustering, and applications. Chapman and Hall/CRC; 2009 Jun 15.
12. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nature reviews. Genetics. 2012 Dec;13(12):829-839.
13. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012 Jun;13(6):395-405.
14. Rodriguez-Esteban R, Bundschus M. Text mining patents for biomedical knowledge. Drug discovery today. 2016 Jun 1;21(6):997-1002.
15. Hung JL, Zhang K. Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. Journal of Computing in Higher education. 2012 Apr 1;24(1):1-7.
16. Lee S, Lee S, Seol H, Park Y. Using patent information for designing new product and technology: keyword based technology roadmapping. R&d Management. 2008 Mar;38(2):169-88.
17. Salloum SA, Al-Emran M, Monem AA, Shaalan K. Using text mining techniques for extracting information from research articles. Intelligent Natural Language Processing: Trends and Applications: Springer; 2018. p. 373-97.
18. Blei DM. Probabilistic topic models, Commun. ACM. 2012;55(4):77-84.

19. Abramson D, Lees M, Krzhizhanovskaya VV, Dongarra JJ, Sloot PM, editors. Big Data Meets Computational Science, Preface for ICCS 2014. ICCS; 2014.
20. Wang C, Blei D, Heckerman D. Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298. 2012 Jun 13.
21. Steyvers M, Griffiths T. Probabilistic topic models. Handbook of latent semantic analysis. 2007;427(7):424-40.
22. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019 Jun 15;78(11):15169-211.
23. O'callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications. 2015 Aug 1;42(13):5645-57.
24. Cheng X, Shuai C, Liu J, Wang J, Liu Y, Li W, Shuai J. Topic modelling of ecology, environment and poverty nexus: An integrated framework. Agriculture, Ecosystems & Environment. 2018 Nov 15;267:1-4.
25. Okuhara T, Ishikawa H, Urakubo A, Hayakawa M, Yamaki C, Takayama T, et al. Cancer information needs according to cancer type: A content analysis of data from Japan's largest cancer information website. reventive medicine reports. 2018;12:245-52.
26. Aghakardan A, KeyhabiNejad M. Proposing a Model for Extracting Information from Textual Documents, Based on Text Mining in E-learning. Iranian Communication and Information Technology. Journal Of Information and Communication Technology. 2012;4(11):47-54.
27. Lam C, Lai FC, Wang CH, Lai MH, Hsu N, Chung MH. Text mining of journal articles for sleep disorder terminologies. PloS one. 2016;11(5).
28. Wang SH, Ding Y, Zhao W, Huang YH, Perkins R, Zou W, et al. Text mining for identifying topics in the literatures about adolescent substance use and depression. BMC public health. 2016;16:279.
29. Selvaraj B, Periyasamy S. Indian medicinal plants for diabetes: text data mining the literature of different electronic databases for future therapeutics, Biomedical Research . 2016.
30. Ozaydin B, Zengul F, Oner N, Delen D. Text-mining analysis of mHealth research. mHealth. 2017;3:53.
31. Payton FC, Yarger LK, Pinter AT. Text Mining Mental Health Reports for Issues Impacting Today's College Students: Qualitative Study. JMIR mental health. 2018;5(4):e10032.
32. Kwon OS, Kim J, Choi KH, Ryu Y, Park JE. Trends in deqi research: a text mining and network analysis. Integrative medicine research. 2018 Sep 1;7(3):231-7.
33. Dancy-Scott N, Dutcher GA, Keselman A, Hochstein C, Copty C, Ben-Senia D, et al. Trends in HIV Terminology: Text Mining and Data Visualization Assessment of International AIDS Conference Abstracts Over 25 Years. 2018;4(2).
34. Kim YM, Delen D. Medical informatics research trend analysis: A text mining approach. Health informatics journal. 2018; 24(4):432-52.
35. Rusanov A, Miotto R, Weng C. Trends in anesthesiology research: a machine learning approach to theme discovery and summarization. JAMIA open. 2018;1(2):283-93.
36. Jalali MS, Razak S, Gordon W, Perakslis E, Madnick S. Health care and cybersecurity: bibliometric analysis of the literature. Journal of medical Internet research. 2019;21(2):e12644.

37. Saheb T, Saheb M. Analyzing and visualizing knowledge structures of health informatics from 1974 to 2018: A bibliometric and social network analysis. Healthcare informatics research. 2019 Apr 1;25(2):61-72.
38. Byington EK, Felps W, Baruch Y. Mapping the Journal of Vocational Behavior: A 23-year review. Journal of Vocational Behavior. 2019;110:229-44.
39. Khasseh A A, Soosaraei M, Fakhar M. Cluster Analysis and Mapping of Iranian Researchers in the Field of Parasitology:With an Emphasis on the Co-authoreship Indicators and H Index. Iran J Med Microbiol. 2016; 10 (2) :63-74.
40. Truyens M, Van Eecke P. Legal aspects of text mining. Computer law & security review. 2014 Apr 1;30(2):153-70.
41. Tseng YH, Lin CJ, Lin YI. Text mining techniques for patent analysis. Information processing & management. 2007 Sep 1;43(5):1216-47.
42. Weiss SM, Indurkhya N, Zhang T. Fundamentals of predictive text mining: Springer; 2015.
43. Lamba M, Madhusudhan M. Application of Topic Mining and Prediction Modeling Tools for Library and Information Science Journals. Library Practices in Digital Era. Eds. MR Murali Prasad et al. Hyderabad: BS Publications. 2018:395-401.
44. Blei DM. Probabilistic topic models, Commun. ACM. 2012;55(4):77-84.
45. Maskeri G, Sarkar S, Heafield K, editors. Mining business topics in source code using latent dirichlet allocation. Proceedings of the 1st India software engineering conference; 2008: ACM.
46. Kontostathis A, Galitsky LM, Pottenger WM, Roy S, Phelps DJ. A survey of emerging trend detection in textual data mining. Survey of text mining: Springer; 2004. p. 185-224.
47. Birkle C, Pendlebury DA, Schnell J, Adams J. Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies. 2020;1(1):363-76
48. Zhang Y, Chen M, Liu L, editors. A review on text mining. 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS); 2015: IEEE.
49. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval: Cambridge university press; 2008.
50. Abuhay TM, Kovalchuk SV, Bochenina KO, Kampis G, Krzhizhanovskaya VV, Lees MH. Analysis of computational science papers from ICCS 2001-2016 using topic modeling and graph theory. arXiv preprint arXiv:170502203. 2017.
51. Frakes WB, Baeza-Yates R. Information retrieval: data structures and algorithms. Prentice-Hall, Inc.; 1992.
52. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inf Syst. 2008;26(3):Article 13.
53. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research (3). 2003.
54. Blei DM. Probabilistic topic models . Commun. ACM. 2012;55(4):77-84.
55. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences. 2004;101(suppl 1):5228-35.
56. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM, editors. Reading tea leaves: How humans interpret topic models .Advances in neural information processing systems; 2009.
57. Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Shanghai, China: Association for Computing Machinery; 2015. p. 399–408.

58. Sbalchiero S, Eder M. Topic modeling, long texts and the best number of topics. Some Problems and solutions. Quality & Quantity. 2020:1-14.
59. Greene D, O'Callaghan D, Cunningham P, editors. How many topics? stability analysis for topic models. Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2014: Springer.
60. Greene D, Cross JP. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Political Analysis. 2017;25(1):77-94.
61. Wieczorek O, Schubert D. The Symbolic Power of the Research Excellence Framework. Evidence from a case study on the individual and collective adaptation of British Sociologists.
62. Rehurek R, Sojka P, editors. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010: Citeseer.
63. Vorontsov K, Frei O, Apishev M, Romov P, Dudarenko M, editors. Bigartm: Open source library for regularized multimodal topic modeling of large collections. International Conference on Analysis of Images, Social Networks and Texts; 2015: Springer.
64. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus. 2016;5(1):1608.
65. Syed S, Weber CT. Using machine learning to uncover latent research topics in fishery models. Reviews in Fisheries Science & Aquaculture. 2018;26(3):319-36.
66. Chen K, Kou G, Shang J, Chen Y. Visualizing market structure through online product reviews: Integrate topic modeling, TOPSIS, and multi-dimensional scaling approaches. Electronic Commerce Research and Applications. 2015;14(1):58-74.
67. Danesh F, Ghavidel S. Coronavirus: Scientometrics of 50 Years of global scientific productions. Iranian Journal of Medical Microbiology. 2020;14(1):1-16.
68. Lamba M, Madhusudhan M. Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. Scientometrics. 2019;120(2):477-505.
69. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019;78(11):15169-211.
70. Peiris J, Lai S, Poon L, Guan Y, Yam L, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. The Lancet. 2003;361(9366):1319-25.
71. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, et al. A novel coronavirus associated with severe acute respiratory syndrome. New England journal of medicine. 2003;348(20):1953-66.
72. Poutanen SM, Low DE, Henry B, Finkelstein S, Rose D, Green K, Tellier R, Draker R, Adachi D, Ayers M, Chan AK. Identification of severe acute respiratory syndrome in Canada. New England Journal of Medicine. 2003;348(20):1995-2005.
73. Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New England Journal of Medicine. 2012;367(19):1814-20.
74. Ram S. Coronavirus Research Trends: A 50–Year Bibliometric Assessment. Science & Technology Libraries. 2020:1-17.
75. Kagan D, Moran-Gilad J, Fire M. Scientometric trends for coronaviruses and other emerging viral infections. BioRxiv. 2020.
76. Li J, Zhang K. Keyword extraction based on tf/idf for Chinese news document. Wuhan University Journal of Natural Sciences. 2007;12(5):917-21.
77. Garfield E. Keywords plus-ISI's breakthrough retrieval method. 1. Expanding your searching power on current-contents on diskette. Current contents. 1990;32:5-9.

78. Haghani M, Bliemer MC, Goerlandt F, Li J. The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review. Safety Science. 2020:104806.
79. Hossain MM. Current status of global research on novel coronavirus disease (COVID-19): a bibliometric analysis and knowledge mapping [version 1; peer review: awaiting peer review]. F1000Research 2020, 9:374.
80. Colavizza G, Costas R, Traag VA, Van Eck NJ, Van Leeuwen T, Waltman L. A scientometric overview of CORD-19. BioRxiv. 2020.
81. Sullivan HC, Roback JD. Convalescent plasma: therapeutic hope or hopeless strategy in the SARS-CoV-2 pandemic. Transfusion Medicine Reviews. 2020.
82. Zhao J, Yuan Q, Wang H, Liu W, Liao X, Su Y, Wang X, Yuan J, Li T, Li J, Qian S. Antibody responses to SARS-CoV-2 in patients of novel coronavirus disease 2019. Clinical Infectious Diseases: an Official Publication of the Infectious Diseases Society of America. 2020.
83. Hafeez A, Ahmad S, Siddqui SA, Ahmad M, Mishra S. A Review of COVID-19 (Coronavirus Disease-2019) Diagnosis, Treatments and Prevention. EJMO 2020, 4, 116-125.
84. Dehghanbanadaki H, Seif F, Vahidi Y, Razi F, Hashemi E, Khoshmirsafa M, Aazami H. Bibliometric analysis of global scientific research on Coronavirus (COVID-19). Medical Journal of The Islamic Republic of Iran (MJIRI). 2020 Feb 10;34(1):354-62.
85. Xiao F, Tang M, Zheng X ,Liu Y, Li X, Shan H. Evidence for gastrointestinal infection of SARS-CoV-2. Gastroenterology. 2020.
86. Bertram S, Heurich A, Lavender H, Gierer S, Danisch S, Perin P, et al. Influenza and SARS-coronavirus activating proteases TMPRSS2 and HAT are expressed at multiple sites in human respiratory and gastrointestinal tracts. PloS one. 2012;7(4).

Ministry of Health and Medical Education
Gonabad University of Medical Sciences
Vice Chancellor of Research and Technology

Text mining of Coronavirus World Publications

Research Ethics Code**:**
IR.GMU.REC.1398.189

2020 March 17

Researchers

Meisam Dastani, Farshid Danesh & Mohammad Ghorbani

2021 Spring