وزارت علوم، تحقیقات، و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

طرح پژوهشی

تهیه مجموعه داده استاندارد فارسی برای مساله استخراج خودکار کلیدواژه از اسناد علمی

مجری

آزاده محبی

همکاران

مرضیه زرین‌بال، عمار جلالی‌منش، زهرا دهسرایی،
علی سلیمی‌صدر، حمید حسنی

گروه تحقیقاتی

امیر بادامچی، الهه محرابی

فروردین ۱۴۰۰


بسم الله الرحمن الرحیم

وزارت علوم تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
ایرانداک

شماره: ۳۳/۸۰۵
تاریخ: ۱۴۰۰/۰۲/۰۵
پیوست: ندارد

پژوهش . مدیریت دانش . آموزش

بنام خدا

ایرانداک در نیم‌قرن دوم کار خود همچنان جوان و کوشا، همراه هر پژوهش، پژوهشگر، و سیاست‌گذار

۱۳۹۹ – ۱۳۴۷

گواهی

انجام طرح پژوهشی:

- **عنوان:** تهیه مجموعه داده استاندارد فارسی برای مساله استخراج خودکار کلیدواژه از اسناد علمی
- **مجری:** دکتر آزاده مجبی
- **همکار:** دکتر مرضیه زرین‌بال، دکتر عمار جلالی‌منش، زهرا دهسرائی، علی سلیمی‌صدر، حمید حسنی
- **گروه تحقیقاتی:** امیر بادامچی، الهه محرابی
- **شمارهٔ قرارداد:** ۳۳/۵۹۶۱
- **تاریخ قرارداد:** ۱۳۹۸/۶/۱۰
- **تاریخ شروع:** ۱۳۹۸/۶/۱۰
- **تاریخ پایان:** ۱۴۰۰/۱/۱۸
- **ناظر:** دکتر نصرالله پاک‌نیت

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایهٔ داوری در نشست ۴۰۷ (تاریخ ۱۴۰۰/۱/۳۰) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری، همکاران، و ناظر این طرح تقدیم دارم.

با آرزوی بهروزی
پرویز شهریاری
معاون پژوهش و آموزش

تهران . خیابان انقلاب اسلامی . چهارراه فلسطین . شمارهٔ ۱۰۹۰ صندوق پستی ۱۳۱۸۵-۱۳۷۱
تلفن: ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۴۳۰ دورنگار: ۶۶۴۶۲۲۵۴ تلفن گویا: ۶۶۴۹۴۹۵۴
irandoc.ac.ir
info@irandoc.ac.ir

تهیه مجموعه داده استاندارد فارسی برای مساله استخراج خودکار کلیدواژه از اسناد علمی

مدیر طرح پژوهشی: آزاده محبی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه ۴، اتاق ۴۰۴
صندوق پستی: ۱۳۱۵۷۷۳۳۱۴ **تلفن:** ۶۶۴۹۴۹۸۰-۰۲۱(داخلی۴۷۷)
وبگاه: https://www.irandoc.ac.ir
رایانامه: *mohebi@irandoc.ac.ir*

همکاراصلی طرح پژوهشی: مرضیه زرین‌بال، عمار جلالی‌منش، زهرا دهسرایی، علی سلیمی صدر، حمید حسنی
گروه تحقیقاتی: امیر بادامچی، الهه محرابی

چکیده

با توجه به حجم داده‌ها و مستندات ایجاد شده در پایگاه‌های تخصصی علمی مانند گنج، نیاز است که روش‌های هوشمند و خودکار برای افزایش دقت و سرعت تخصیص کلیدواژه به این مستندات به کار گرفته شود. توسعه و بکارگیری این روش‌ها نیازمند وجود معیارهای استاندارد برای ارزیابی و بهبود آنهاست. یکی از این معیارها، وجود یک مجموعه داده استاندارد است که حاوی مجموعه‌ای از اسناد و کلیدواژه‌های تخصیص داده شده درست و کامل به آنها است. اگر چنین مجموعه داده‌ای موجود نباشد، نمی‌توان به راحتی الگوریتم‌ها و روش‌های استخراج خودکار کلیدواژه از مستندات علمی را با یک معیار مشخص با هم مقایسه کرد و بکارگرفت.

در برخی از پژوهش‌های انجام شده در حوزه استخراج خودکار کلیدواژه از مستندات فارسی، از مجموعه داده‌هایی استفاده شده لیکن این مجموعه داده‌ها یا داده‌های علمی نیستند و یا در صورت استفاده از داده‌ها علمی، تنها کلیدواژه‌های نویسنده به عنوان کلیدواژه‌های تخصیص داده شده به یک سند علمی در آنها در نظر گرفته شده است. این در حالی است که در بسیاری از پژوهش‌های انجام شده در زبان انگلیسی، مجموعه داده‌های استاندارد و متعددی برای این منظور وجود دارد که در اغلب پژوهش‌ها از آنها استفاده می‌شود.

بنابراین در این طرح پژوهشی، مجموعه داده‌ای از اسناد علمی (پایان‌نامه‌ها و رساله‌ها) گنج به همراه کلیدواژه‌های مرتبط با آنها ایجاد می‌شود. مدارک نمایه‌شده در گنج، می‌تواند منبعی غنی برای ساخت مجموعه داده برای مساله استخراج کلیدواژه از مستندات علمی فارسی باشد. لیکن کلیدواژه‌های تخصیص داده شده به هر مدرک در فرایند نمایه‌سازی اغلب محدود هستند و حتی در برخی از موارد به دلیل محدودیت‌های موجود در فرایند نمایه‌سازی، کامل نیستند. علاوه بر آن، مشخص نیست که هر کلیدواژه تخصیص داده شده به هر مدرک تا چه میزان به آن مدرک مرتبط است.

در فرایند ایجاد مجموعه داده در این طرح، ابتدا از مجموعه‌ای از اسناد پایگاه گنج که کلیدواژه‌های تخصصی نمایه‌ساز و نویسنده را دارند استفاده شده است. سپس براساس فرایندی ماشینی مجموعه کلیدواژه وسیع‌تری، به عنوان کلیدواژه‌های کاندید برای هر سند در نظر گرفته شده و در انتها نیز متخصصین نمایه‌ساز، با بررسی مجدد هر سند و کلیدواژه‌های کاندید، کلیدواژه‌های تخصصی مناسب را از بین آنها انتخاب کرده‌اند. در نهایت ۲۸۳۸ داده در چهار حوز ه موضوعی فنی و مهندسی، علوم انسانی، علوم و علوم کاربردی، و هنر در مجموعه داده قرار گرفته‌اند که هر یک به طور متوسط دارای ۱۰ کلیدواژه هستند. در انتها نیز پیشنهادهایی برای چگونگی بهره‌برداری از مجموعه داده استاندارد توسط جامعه علمی و انتشار آن ارائه شده است.

کلیدواژه‌ها: استخراج کلیدواژه؛ مجموعه داده؛روش RAKE؛ روش TEXTRANK؛ بازیابی اطلاعات؛ روش TFIDF؛ روش‌های رتبه‌بندی کلیدواژه‌ها.

فهرست مطالب

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان **شماره صفحه**

فهرست جدول‌ها

عنوان **شماره صفحه**

سپاسگزاری

ابتدا خداوند بزرگ را شاکرم که فرصت خدمت را به اینجانب عطا فرمود.

ضمن تشکر از شورای محترم پژوهش برای بررسی این طرح، از ناظر محترم جناب آقای دکتر پاک نیت سپاسگزارم که نظرات و پیشنهادهای مفیدی را هنگام داوری و نظارت طرح ارائه کردند و بدون شک نظرات ارزشمند ایشان در ارتقا کیفیت این طرح تاثیر بسزایی داشته است.

همچنین از خانم دهسرایی، مدیر واحد سازماندهی و تحلیل اطلاعات پژوهشگاه و متخصصین نمایه‌ساز خانم‌ها پیروزکیا، صادقیان، محمودیان و موسیوند نیز تشکر می‌کنم که در این طرح با صبوری و دلسوزانه همکاری داشته‌اند.

از سرکار خانم دکتر زرین‌بال، و آقای دکتر جلالی‌منش سپاسگزارم که رهنمودهای بسیار کاربردی در این طرح ارائه کردند و در بخش‌های مختلف انجام کار، همراه بودند. آقای مهندس سلیمی و آقای مهندسی حسنی با پیاده‌سازی الگوریتم‌های استخراج کلیدواژه در این طرح یاری رساندند و از ایشان متشکرم.

از آقای مهندس بادامچی که سامانه ارزیابی را طراحی و پیاده‌سازی کردند و با مهارت و حوصله نتیجه را به اتمام رساندند و خانم مهندس محرابی که الگوریتم RAKE را پیاده‌سازی کردند، سپاسگزارم.

بخشی از این نتایج این طرح در آزمایشگاه تعامل انسان و ماشین روبوداک پیاده‌سازی شده که از سایر اعضای این آزمایشگاه که به توعی در دستیابی به نتایج طرح موثر بودند، سپاسگزارم.

فصل اول:

مقدمه و مساله پژوهش

۱. مقدمه و مساله پژوهش

برای ارزیابی بسیاری از الگوریتم‌ها و روش‌های هوشمند پردازش متن در مسائل یادگیری ماشین، نیاز است که مجموعه داده‌های استانداردی موجود باشد که براساس آن مجموعه داده‌ها، در قالب یک معیار واحد، عملکرد روش‌ها و الگوریتم‌ها بررسی شوند.

هم اکنون پایگاه‌های استاندارد متعددی وجود دارد که از طریق آنها این مجموعه داده‌ها، اغلب به صورت رایگان، در اختیار محققین قرار داده می‌شوند، مانند پایگاه داده مخزن یادگیری ماشین یو.سی.آی (UCI machine learning repository) یا مجموعه داده‌های مربوط به مسابقات علمی کاگل (kaggle) که به طور تخصصی در حوزه پردازش زبان طبیعی و پردازش متن برگزار می‌شوند. اما بیشتر تمرکز این دست از داده‌ها روی داده‌های غیرفارسی است و به طور خاص برای مساله استخراج خودکار کلیدواژه‌ها از یک سند علمی فارسی، هنوز مجموعه داده استانداردی ارائه نشده است. برای اسناد علمی انگلیسی مجموعه داده‌های متعددی وجود دارد، مانند داده‌های مربوط به اسناد پابمد (PubMed) با ۵۰۰ سند نمایه شده، یا داده‌های مربوط به اسناد سازمان غذا و دارو (FAO) با ۷۸۰ سند نمایه شده.

چنین مجموعه داده‌ای می‌تواند برای سایر مسائل متن‌کاوی نظیر دسته‌بندی و خوشه‌بندی متون، تحلیل محتوی، مدل‌سازی موضوع، و خلاصه‌سازی خودکار نیز به عنوان یک مجموعه داده استاندارد به‌کار گرفته شود.

وظیفه اصلی بخش سازماندهی اطلاعات در ایرانداک، اختصاص نمایه‌های استاندارد (کلیدواژه‌ها) به اسناد علمی پس از ثبت آنها در سامانه ثبت است. بنابراین در حال حاضر مجموعه عظیمی از اسناد علمی نمایه شده در پایگاه داده گنج موجود است، که می‌توان مجموعه داده استاندارد را از بین آنها انتخاب نمود.

علی‌رغم اینکه فرایند نمایه‌سازی و تخصیص کلیدواژه در ایرانداک توسط متخصصین انجام می‌شود، ایجاد یک مجموعه داده استاندارد از پایگاه گنج همچنان چالش برانگیز خواهد بود.

یکی از این چالش‌ها برخواسته از محدودیت‌های فرایند دستی نمایه‌سازی است که معمولا در اغلب فرایندهای دستی اجتناب‌ناپذیر است. به طور خاص محبی و جلالی‌منش (۱۳۹۷) در بخشی از طرح پژوهشی خود، برای ارزیابی روش پیشنهادی استخراج کلیدواژه از اسناد علمی پایگاه گنج، از دو رویکرد مختلف استفاده کردند. در رویکرد اول، دقت و فراخوانی روش پیشنهادی با مقایسه با کلیدواژه‌های تخصیص داده شده توسط نمایه‌ساز محاسبه شد و در رویکرد دوم، با استفاده از نظر متخصصین، بدون درنظر گرفتن نمایه‌های اولیه سند، نمایه‌های پیشنهادی ارزیابی شد. با مقایسه بین نتایج این دو رویکرد ارزیابی، مشخص شد که لزوما نمایه‌های تخصیص داده شده به یک سند که توسط متخصصین انجام می‌شود کامل و دقیق نیستند.

چالش دیگر اینست که با اینکه نمایه‌های یک سند توسط نمایه‌ساز تعیین شده، لیکن مشخص نیست که هر یک از این نمایه‌ها تا چه میزان با سند مرتبط هستند و در واقع رتبه‌بندی یا امتیازدهی برای آنها در دست نیست. مشخص بودن میزان ارتباط یک نمایه (کلیدواژه) با سند می‌تواند نقش مهمی را در عملیات بازیابی اطلاعات ایفا نماید.

با توجه به حجم وسیع داده‌های نمایه شده در گنج و تنوع موضوعی آنها، چالش بعدی اینست که چه حجم از داده‌ها با چه توزیع موضوعی باید برای مجموعه داده استاندارد در نظر گرفته شود.

در نهایت نیز لازم است که عملکرد برخی از روش‌های مطرح و جدید در حوزه استخراج خودکار کلیدواژه بر روی مجموعه داده استاندارد فارسی بررسی شود.

در طرح پژوهشی حاضر، برای ایجاد چنین مجموعه داده‌ای متدولوژی پیشنهاد می‌شود که براساس آن بتوان با هدف رفع چالش‌های پیش گفته، مجموعه داده استاندارد را ایجاد نمود. پس از ایجاد مجموعه داده استاندارد براساس متدولوژی پیشنهادی، عملکرد برخی از مهمترین روش‌های موجود اسنخراج خودکار کلیدواژه بر روی آن مجموعه داده، ارزیابی می‌شود.

۱- ۱. هدف و گام‌های پژوهش

هدف اصلی از انجام این پژوهش توسعه مجموعه داده استاندارد فارسی برای مساله استخراج کلیدواژه‌ها از اسناد علمی است. علاوه بر آن عملکرد چند روش استخراج خودکار کلیدواژه بر روی مجموعه داده استاندارد نیز بررسی می‌شود و پیشنهادی برای چگونگی در دسترس قرار دادن مجموعه داده و بهره‌برداری از آن ارائه می‌گردد.

این پژوهش از دو مرحله اصلی تشکیل شده است. در مرحله نخست مجموعه داده استاندارد ایجاد می‌شود. برای این منظور ابتدا از روش کتابخانه‌ای برای بررسی انواع روش‌های موجود برای ساخت چنین مجموعه داده‌ای استفاده می‌شود. سپس با بهره‌گیری از گروهی از متخصصین حوزه نمایه‌سازی، کلیدواژه‌ها به مجموعه‌ای از اسناد گنج اختصاص می‌یابد.

در مرحله دوم، برخی از مهمترین الگوریتم‌های استخراج خودکار کلیدواژه که عمدتا در زبان انگلیسی توسعه یافته‌اند بر روی مجموعه داده ایجاد شده، پیاده‌سازی می‌شوند. در انتها نیز مدل کسب و کار برای چگونگی بهره‌برداری از این مجموعه داده ارائه می‌گردد.

مرحله اول: ساخت و آماده‌سازی مجموعه داده استاندارد

- بررسی روش‌های موجود برای ایجاد مجموعه داده استاندارد در پژوهش‌های پیشین (عمدتا پژوهش‌های مربوط به زبان انگلیسی)
- شناسایی و تعیین شاخص‌ها و ویژگی‌های استاندارد برای مجموعه داده
- انتخاب مجموعه‌ای از اسناد گنج (اسناد نمایه شده) به منظور استخراج داده‌های اصلی از بین آنها
- پاکسازی اسناد در صورت نیاز، و تخصیص کلیدواژه‌های تخصصی توسط متخصصین نمایه‌سازی و با استفاده از اصطلاحنامه‌های تخصصی

  برای این منظور، علاوه بر کلیدواژه‌های موجود برای هر سند، مجموعه‌ای از کلیدواژه‌های کاندید براساس روش‌های ماشینی برای هر سند در نظر گرفته می‌شود که نمایه‌ساز می‌تواند کلیدواژه‌های بیشتری را از بین آنها انتخاب نماید. در واقع فرایند تخصیص کلیدواژه به گونه‌ای در نظر گرفته می‌شود که با نظر متخصص، کلیدواژه‌ها مجدد بررسی شوند و اعتبار مجموعه داده ایجاد شده نیز مورد توجه قرار گیرد. نکته قابل توجه اینجاست که تنها در دو نمونه از مجموعه داده‌های انگلیسی فرایندی برای ارزیابی و اعتبارسنجی کلیدواژه‌ها در نظر گرفته شده است و در اکثر موارد به کلیدواژه‌های نویسنده، اکتفا شده است.
- بررسی مجموعه داده ایجاد شده و نهایی‌سازی آن

مرحله دوم: پیاده‌سازی چند روش اصلی استخراج کلیدواژه روی مجموعه داده و توسعه مدل کسب و کار

- بررسی برخی از روش‌های اخیر در زمینه استخراج کلیدواژه
- پیاده‌سازی الگوریتم‌ها و روش‌ها و تست آنها بر روی مجموعه داده
- ارائه پیشنهادهایی برای نحوه بهره‌برداری از مجموعه داده و چگونگی در دسترسی قرار دادن آن

فصل دوم:

انواع روش‌های استخراج کلیدواژه

۲. روش‌های استخراج کلیدواژه

برای استخراج یا تخصیص خودکار کلیدواژه‌ها به یک سند یا مجموعه‌ای از اسناد، روش‌های متنوعی تاکنون پیشنهاد شده است. اکثر این روش‌ها برای اسناد زبان انگلیسی به‌کار گرفته شده‌اند و تعداد بسیاری از آنها مستقل از زبان عمل می‌کنند.
در این فصل ابتدا مروری بر انواع روش‌های استخراج خودکار کلیدواژه ارائه می‌شود، سپس روش‌هایی که در این طرح پژوهشی از آنها استفاده شده، تشریح می‌گردند.

۲- ۱. دسته‌بندی انواع روش‌های استخراج کلیدواژه

انواع روش‌های استخراج کلیدواژه را می‌توان از جهات مختلف دسته‌بندی کرد:

- تعداد اسناد: روش‌های استخراج کلیدواژه برای یک سند و روش‌های استخراج کلیدواژه برای چند سند
- حجم سند: اسناد بلند و اسناد کوتاه
- شیوه استخراج کلیدواژه و رتبه‌بندی آنها: با ناظر، بدون ناظر، و مبتنی بر یادگیری عمیق

در ادامه توضیحاتی درباره ماهیت و نحوه عملکرد انواع روش‌های استخراج کلیدواژه براساس نوع روش یعنی با ناظر، بدون ناظر و یادگیری عمیق ارائه می‌شود.

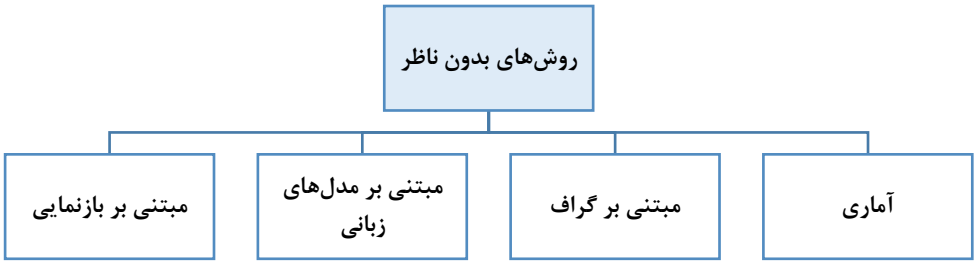
۲- ۱- ۱. روش‌های بدون ناظر

روش‌های بدون ناظر روش‌هایی هستند که در آنها از مجموعه اسناد با کلیدواژه‌های مشخص شده استفاده نمی‌شود. در این روش‌ها چالش اصلی پس از تعیین مجموعه‌ای از عبارات کاندید برای

کلیدواژه، رتبه‌بندی آنهاست (Alami Merrouni, Frikh, and Ouhbi 2020). در انواع روش‌های استخراج کلیدواژه بدون ناظر ۳ قدم اصلی انجام می‌شود (Papagiannopoulou and Tsoumakas 2019):

۱. ایجاد مجموعه‌ای از واژگان یا عبارات کاندید
۲. امتیازدهی و رتبه‌بندی واژگان یا عبارات کاندید
۳. استخراج کلیدواژه‌ها با انتخاب واژگان یا عباراتی که بالاترین رتبه یا امتیاز را دارند.

روش‌های بدون ناظر نیز به دسته‌های مختلف طبقه‌بندی شده‌اند (شکل ۲-۱). مهمترین این روش‌ها عبارتند از: روش های آماری[۱]، روش های مبتنی بر گراف[۲]، روش های مبتنی بر مدل های زبانی[۳]، و روش های مبتنی بر بازنمایی[۴] (Papagiannopoulou and Tsoumakas 2019). مرونی و همکاران (۲۰۲۰) نیز در مطالعه‌ای انواع روش‌های ا ستخراج کلیدواژه را برر سی کردند. ایشان روش‌های بدون ناظر را به دسته‌های آماری، مبتنی بر گراف، مدل‌های زبانی، احتمالاتی، یادگیری همزمان، و ترکیبی دسته‌بندی کردند. در این دسته‌بندی روش‌های احتمالاتی بیشتر در زمره روش‌های مبتنی بر بازنمایی قرار می‌گیرند که در آنها از مدل های برداری بازنمایی کلمات و متون مانند مدل کلمه-به-بردار word2vec (Mikolov et al. 2013) و مدل سند-به-بردار doc2vec (Le and Mikolov 2014) استفاده می‌شود.



شکل ۲-۱- انواع روش‌های بدون ناظر برای استخراج کلیدواژه (Papagiannopoulou and Tsoumakas 2019)

[۱] Statistical

[۲] Graph-based

[۳] Language model-based

[۴] Embedding-based

۲-۱-۱-۱. روش‌های آماری

در این روش‌ها از ویژگی‌های آماری واژگان و عبارات در متن استفاده می‌شود. در بین انواع این ویژگی‌ها، سه نوع هستند که بیشترین استفاده را داشته‌اند (Hasan and Ng 2014; Witten et al. 1999):

- ویژگی TF-IDF: این ویژگی برای اولین بار توسط سالتون و باکلی (۱۹۸۸) پیشنهاد شد که براساس فراوانی واژگان و عبارات کاندید (TF) در متن سند و نیز فراوانی آن در کلیه متون (IDF) محاسبه می‌شود (Salton and Buckley 1988).
- ویژگی فاصله عبارت کاندید: این ویژگی عبارتست از تعداد کلماتی که پیش از اولین رخداد عبارت در متن آمده است که براساس تعداد کل واژگان متن نیز نرمال می‌شود.
- ویژگی تعداد دفعات به عنوان کلیدواژه: این ویژگی عبارتست از تعداد دفعاتی که عبارت کلیدی قبلا (در یک مجموعه داده آموزشی) به عنوان کلیدواژه آمده باشد.
- سایر ویژگی‌ها نظیر طول عبارت، فاصله بین دو رخداد عبارت در متن، و توزیع آماری در متن، و طول جملاتی که عبارت در آن آمده (Lynn et al. 2017).

در برخی از مطالعات اخیر، علاوه بر شاخص‌های آماری معمول، شاخص‌های آماری دیگری تعریف شده که منعکس کننده ویژگی‌های معنایی و محتوایی سند است. یکی از این روش‌ها که اخیر پیشنهاد شده و نتایج خوبی را هم به دنبال داشته روش YAKE است (Campos et al. 2020). در این روش علاوه بر موقعیت عبارت یا کلمه در متن از ویژگی‌های آماری دیگری برای بیان کردن معنا و محتوی سند نیز استفاده شده است. ازآنجاییکه از این روش در این پژوهش استفاده شده، در بخش‌های بعدی این فصل، نحوه عملکرد آن تشریح خواهد شد.

۲-۱-۱-۲. روش‌های مبتنی بر گراف

در روش‌های مبتنی بر گراف، برای یک سند گرافی در نظر گرفته می‌شود که هر گره از آن یک عبارت یا کلمه کلیدی کاندید و یال‌های گراف بیانگر ارتباط بین عبارات است (Hasan and Ng 2014).
یال‌ها یا همان ارتباط بین واژگان می‌تواند براساس رابطه هم‌رخدادی در یک پنجره از کلمات در متن، رابطه ساختاری از نظر ساختار جمله‌ای، رابطه معنایی بین واژگان، و یا سایر رابطه‌ها نظیر رخداد در یک جمله یا در یک پاراگراف تعریف کرد (Beliga, Mestrovic, and Martineie-Ipsie 2015).
انواع روش‌های مبتنی بر گراف را می‌توان در دسته‌های زیر قرار داد (Papagiannopoulou and Tsoumakas 2019):

- روش‌های مبتنی بر آمار که در آنها تنها از اطلاعات آماری واژگان و عبارات در متن برای تعریف ارتباط بین واژگان استفاده می‌شود.
- روش‌های مبتنی بر همسایگی و شبکه ارجاعات
- روش‌های مبتنی بر موضوع
- روش‌های مبتنی بر معنا

یکی از مهمترین روش‌های مبتنی بر گراف روش TextRank است که توسط میهالسیا و تارو (Mihalcea and Tarau 2004) پیشنهاد شد. این روش براساس الگوریتم PageRank که در بازیابی اطلاعات استفاده می‌شود، عمل می‌کند. بعدها تحقیقات زیادی برای بهبود الگوریتم TextRank نیز پیشنهاد شد. از آنجاییکه در این پژوهش از این الگوریتم به عنوان یکی از روش‌های پایه استفاده می‌شود، در بخش‌های بعدی، نحوه عملکرد آن تشریح می‌شود.

لازم به ذکر است که اغلب روش‌های مبتنی بر گراف، دو چالش اساسی دارند (Liu et al. 2010):

- در این روش‌ها کلماتی که پیوندهای بسیار با سایر کلمات دارند، اهمیت بیشتری پیدا می‌کنند، که لزوما ممکن است این کلمات بیانگر موضوع اصلی متن نباشند.
- عبارت یا کلمات کلیدی باید پوشش مناسبی از موضوعات اصلی مطرح در یک متن را داشته باشند. در حالیکه در روش‌های مبتنی بر گراف، کلمات استخراج شده ممکن است تنها بیانگر یک موضوع خاص در یک متن باشند و نتوانند موضوع اصلی متن را بیان کنند.

برای رفع این چالش‌ها نیز تحقیقات متنوعی انجام شده از جمله مطالعه لیو و همکاران (Liu et al. 2010). آنها روشی را پیشنهاد دادند که در آن اهمیت کلمات در روش‌های مبتنی بر گراف در موضوعات مختلف نیز در نظر گرفته می‌شود. در این روش، موضوعات برای هر کلمه براساس پیکره ویکی پدیا یادگیری می‌شود.

۲-۱-۱-۳. روش‌های مبتنی بر مدل‌های زبانی

در روش‌های مبتنی بر مدل‌های زبانی، از مدل‌های زبانی n-gram استفاده می‌شود. در این مدل‌ها معمولا مجموعه‌ای از متون برای یادگیری مدل به‌کار گرفته می‌شود. برای هر عبارت کاندید، احتمال آنکه این عبارت یک عبارت کلیدی باشد محاسبه می‌شود. در پژوهش (Tomokiyo and Hurst 2003) دو ویژگی برای یک عبارت کاندید محاسبه می‌شود: قابلیت عبارت بودن آن (Phraseness) و قابلیت در برداشتن اطلاعات (Informativeness). این ویژگی‌ها براساس مدل‌های زبانی محاسبه

می‌شود و عبارات کاندید براساس این ویژگی‌ها با هم مقایسه می‌شوند. برخی از مدل‌های زبانی، از شبکه‌های عصبی برای یادگیری استفاده می‌کنند که در مطالعات اخیر از مدل‌های زبانی نظیر ELMo (Peters et al. 2018) برای بازنمایی عبارات کاندید استفاده شده است.

۲-۱-۱-٤. روش‌های مبتنی بر بازنمایی

در روش‌های مبتنی بر بازنمایی از بازنمایی برداری کلمات و عبارات به عنوان مدل‌های پایه جهت رتبه‌بندی کلیدواژه‌ها و عبارات کاندید استفاده می‌شود. برخی از این مدل‌ها براساس هم‌رخدادی هستند مانند مدل‌های LDA و LSA (Krestel, Fankhauser, and Nejdl 2009). برخی دیگر از این مدل‌ها براساس شبکه‌های عصبی آموزش دیده‌اند مانند مدل‌های Word2Vec (Mikolov et al. 2013) و Doc2Vec (Le and Mikolov 2014). مدل‌های اخیر بازنمایی کلمات مانند مدل GLoVe (Pennington, Socher, and Manning 2014) نیز در پژوهش‌های استخراج کلیدواژه استفاده شده است.

در پژوهش (Bennani-Smires et al. 2018) کلیدواژه‌های کاندید براساس برچسب‌گذاری اجزای کلام (POS) انجام می‌شود و از مدل‌های بازنمایی Sent2Vec یا Doc2Vec برای بازنمایی عبارات استفاده می شود. پس از آن (Papagiannopoulou and Tsoumakas 2018) در پژوهشی الگوریتم RVA را پیشنهاد کردند که در آن از بازنمایی کلمات براساس مدل GLoVe استفاده شده و مهمترین نوآوری ایشان بهره‌مندی از مفاهیم محلی کلمات در جملات بوده است.

۲- ۱- ۲. روش‌های با ناظر

روش‌های با ناظر برای استخراج کلیدواژه، براساس الگوریتم‌های دسته‌بندی عمل می‌کنند، که هر عبارت یا لغت می‌تواند در کلاس کلیدواژه‌ها یا غیرکلیدواژه‌ها قرار گیرد. در این دست از روش‌ها نیاز است که مجموعه داده‌های آموزشی موجود باشد که در آنها اسنادی با کلیدواژه‌های مشخص موجودند. دو موضوع اصلی در روش‌های با ناظر مورد توجه قرار می‌گیرد: چگونگی بازتعریف مساله در قالب یک مساله یادگیری با ناظر، و انتخاب ویژگی‌ها (Hasan and Ng 2014).

یک شیوه معمول برای تعریف مساله استخراج کلیدواژه به عنوان یک مساله یادگیری با ناظر اینست که آن را در قالب یک مساله دسته‌بندی در نظر بگیریم. در این شیوه، مجموعه داده آموزشی در دست است که در آن مستندات دارای چندین کلیدواژه هستند. با استفاده از این داده یک دسته‌بند دو دسته‌ای یادگیری می شود که با به‌کارگیری آن می‌توان برای هر واژه کاندید مشخص کرد که متعلق به دسته کلیدواژه‌هاست یا غیرکلیدواژه‌ها. برای یادگیری این دسته‌بند از الگوریتم‌های مختلفی نظیر

دسته‌بند بیز ضعیف[۵]، درخت تصمیم[۶]، بگینگ و بوستینگ[۷]، ماکزیمم آنتروپی[۸]، پرسپترون چندلایه[۹]، و ماشین بردار پشتیبان[۱۰] استفاده شده است (Hasan and Ng 2014).

انتخاب ویژگی‌های مناسب یکی دیگر از موضوعاتی است که در مساله استخراج کلیدواژه با ناظر مورد توجه قرار می‌گیرد. ویژگی‌هایی که برای این منظور استفاده می‌شوند را می‌توان در دو دسته اصلی قرار داد: ویژگی های برگرفته از خود متن، ویژگی های مستخرج از منابعی به غیر از خود متن(Hasan and Ng 2014; Papagiannopoulou and Tsoumakas 2019).

ویژگی‌های برگرفته از خود متن می‌توانند ویژگی‌های آماری باشند مانند ویژگی‌های مبتنی بر TF-IDF، یا ویژگی‌های ساختاری باشند، مانند اینکه عبارت کاندید در کجای متن آمده است، و یا ویژگی‌های نحوی باشند. ترکیب مجموعه‌ای از این ویژگی‌ها در پژوهش‌های مختلف مورد توجه قرار گرفته است.

ویژگی‌های برگرفته از متون و منابع خارجی، به غیر از متن سند، براساس اطلاعاتی به‌دست می‌آیند که از متون و مجموعه دادگانی مانند ویکی‌پدیا، اصطلاحنامه‌های تخصصی، و لاگ جستجوها به‌دست می‌آیند (Hasan and Ng 2014).

در بین الگوریتم‌های با ناظر، الگوریتمKEA به عنوان یک الگوریتم پایه محسوب می‌شود که در عملکرد آن با سایر الگوریتم‌های جدید نیز مقایسه می شود (Witten et al. 1999). در این الگوریتم دو مرحله اصلی وجود دارد: آموزش، و استخراج. در مرحله آموزش، همانند سایر روش‌های یادگیری با ناظر، مدل براساس مجموعه داده‌های آموزش، ایجاد می‌شود. در مرحله استخراج، از یک سند جدید، با استفاده از مدل آموزش داده شده، کلیدواژه‌ها استخراج می شود. در این روش، نیز در ابتدا پس از پیش‌پردازش متن، مجموعه‌ای از عبارات به عنوان عبارات کاندید در نظر گرفته می‌شوند. سپس دو ویژگی زیربرای هر عبارت کاندیدمحاسبه می‌شود:

- ویژگی TF-IDF که عبارتست از حاصلضرب فراوانی عبارت کاندید
- ویژگی اولین رخداد که نسبت تعداد کلماتی که قبل از اولین رخداد یک عبارت کاندید در متن آمده به کل کلمات متن.

---

[۵] Naïve Bayes Classifier

[۶] Decision Tree

[۷] Bagging and Boosting

[۸] Mazimum Antropy

[۹] Multi-layered Perceptron

[۱۰] Support vector Machine

در نهایت براساس مدل آموزش دیده شده، از بین عبارات کاندید، کلیدواژه‌ها شناسایی می‌شوند. مدل، براساس مجموعه داده آموزشی و ویژگی‌های فوق آموزش داده می‌شود. مجموعه داده آموزشی شامل اسنادی است که هریک کلیدواژه‌هایی دارند. برای هر سند، مجموعه‌ای از عبارات کاندید شناسایی می‌شود و براساس کلیدواژه‌های موجود، برای هر عبارت کاندید، برچسب «کلیدواژه» یا «غیر کلیدواژه» در نظر گرفته می‌شود. سپس با استفاده از «دسته‌بند بیز ضعیف»، مدل ساخته می‌شود.
مدل بیز ضعیف، برای هر عبارت کاندید از یک سند جدید، احتمال «کلیدواژه» بودن ($P[yes]$) و به طور مشابه احتمال «غیر کلیدواژه» ($P[no]$) را به صورت زیر محاسبه می‌کند:

$$P[yes] = \frac{Y}{Y+N} P_{TFIDF}[t|yes]\, P_{distance}[d|yes] \quad (۱-۲)$$

که در آن Y تعداد نمونه های مثبت در مجموعه آموزش، و N تعداد نمونه های منفی، یعنی تعداد دفعاتی است که عبارت کاندید به عنوان کلیدواژه نبوده است. در نهایت احتمال کل به صورت زیر محاسبه می‌شود:

$$P(overall) = \frac{p[yes]}{p[yes] + p[no]} \quad (۲-۲)$$

سپس عبارات کاندید براساس احتمال کل رتبه‌بندی می شوند و عباراتی که بیشترین احتمال را دارند، انتخاب می‌شوند.

۲-۱-۲-۱. یادگیری عمیق

علاوه بر روش‌های با ناظر که براساس دسته‌بندی‌های مختلف توسعه پیدا کرده‌اند، روش‌هایی نیز براساس الگوریتم‌های شبکه‌های یادگیری عمیق پیشنهاد شده‌اند که در آنها مجموعه وسیع‌تری از انواع ویژگی‌ها استفاده می شود. از بین مدل‌های یادگیری عمیق، شبکه‌های RNN دوجهته بیشترین استفاده را در پردازش زبان طبیعی دارند. در این شبکه‌ها، یک عبارت کلیدی به عنوان دنباله‌ای از واژگان در نظر گرفته می‌شود که اطلاعاتی را نیز از واژگان قبلی در بردارد. در پژوهش (Zhang et al. 2016) شبکه عمیق از نوع RNN برای استخراج کلیدواژه از توییتر پیشنهاد شد. پس از آن در (Meng et al. 2017) مدلی براساس RNN پیشنهاد شد که ساختار آن به صورت decoder-encoder طراحی شده بود و روی مجموعه‌ای از ۲۰۰۰۰ متن علمی آموزش داده شد. برای رفع مشکل افزونگی اطلاعات در کلیدواژه‌های استخراج شده، (Chen et al. 2018) با توجه به وجود این چالش در مطالعه (Meng et al. 2017) معماری جدیدی را با عنوان CorrRNN پیشنهاد کردند که همبستگی بین کلیدواژه‌ها را

هم در نظر می‌گرفت تا از افرونگی اطلاعات در کلیدواژه‌های استخراج شده، جلوگیری کند. یکی از مهمترین چالش‌ها در روش‌های مبتنی بر یادگیری عمیق نیاز به حجم بالایی از داده‌ها برای آموزش شبکه است که در برخی از پژوهش‌های اخیر (Ye and Wang 2018) نیز سعی شده تا حدی این چالش مرتفع گردد.

۲- ۲. روش‌های استفاده شده در این پژوهش

ازآنجاییکه یکی از مهمترین گام‌های این پژوهش انتخاب روش استخراج کلیدواژه مناسب برای تولید مجموعه اولیه از کلیدواژه‌ها برای یک سند است، در این بخش توضیحی درباره هر یک از روش‌های به‌کارگرفته شده در این پژوهش، ارائه می‌گردد.

۲- ۲- ۱. روش TextRank

این روش یکی از روش‌های بدون ناظر برای استخراج کلیدواژه است که به عنوان یکی از اولین روش‌های مبتنی بر گراف توسط میهالسیا و همکاران (Mihalcea and Tarau 2004) مطرح شده است. این روش بر مبنای روش بازیابی اطلاعات PageRank عمل می‌کند. در الگوریتم PageRank که در بازیابی اطلاعات در وب استفاده می‌شود، یک گراف جهت‌دار در نظر گرفته می‌شود که هر گره، یک صفحه است. هر گره به گره دیگر با یک یال جهت‌دار متصل می‌شود، اگر از آن صفحه به صفحه دیگر لینکی باشد. پس از ساختن کل گراف، برای هر صفحه وب امتیازی در نظر گرفته می‌شود و صفحات با بالاترین امتیاز بازیابی می‌شوند.

ارتباطات در الگوریتم TextRank، براساس هم‌رخدادی واژگان است. برای ایجاد گراف از یک متن، هر گره معادل یک کلمه است. با در نظر گرفتن پنجره‌ای با اندازه مشخص (بین ۲ تا ۱۰ کلمه)، وجود یال بین دو گره به این معنی است که کلمات دو گره در یک پنجره قرار گرفته‌اند. به این ترتیب گراف بدون جهت با یال‌ها و گره‌های مشخص برای هر متن ایجاد می‌شود. فهرست کلماتی که به عنوان گره در نظر گرفته می‌شوند، معمولا پس از حذف ایست‌واژه‌ها و کلمات با برچسب نحوی نامرتبط مانند حروف، افعال، قیدها، و... ایجاد می‌شوند.

این الگوریتم در قالب یک فرایند تکرارشونده عمل می‌کند که در هر مرحله، وزنی برای گره‌ها محاسبه می‌شود پس از انجام چندین تکرار، همگرایی حاصل می‌شود. در نهایت گره‌هایی که بیشترین وزن را دارند به عنوان کلیدواژه در نظر گرفته می‌شوند. در هر مرحله وزن گره‌ها به صورت زیر محاسبه می‌شود:

$$WS(V_i) = (1-d) + d * \sum_{V_j \epsilon In(V_i)} \frac{w_{ij}}{\sum_{V_k \epsilon Out(V_j)} w_{jk}} \quad (۳-۲)$$

که در آن d پارامتر تنظیم کننده بین ۰ و ۱ است، w_{ij} وزن بین گره V_i و V_j و منظور از $In(V_i)$ و $out(V_i)$ به ترتیب مجموعه گره‌های ورودی‌ها و خروجی از V_i است.

الگوریتم TextRank یکی از روش های پا یه مبتنی بر گراف برای اســتخراج کلیدواژه اســت که پژوهش‌های متعددی نیز تا کنون برای بهبود آن صورت گرفته ا ست. به عنوان مثال در پژوهشی، وزن اولیه گره‌ها و انتقال از یک مرحله به مرحله بعدی، با استفاده وزن جامع کلمات بهبود یافته است (Pan, Li, and Dai 2019). در پژوهشی دیگر نیز، وزن‌دهی اولیه گره‌ها به صورت تصادفی در نظر گرفته نشــده، و براســاس بازنمایی برداری کلمات در اولین مرحله از الگوریتم اســت (Kazemi, Pérez-Rosas, and Mihalcea 2021). الگوریتمTextRank برای اســتخراج جملات و کلیدواژه ها در خلاصه‌سازی متون نیز بکار گرفته شده است و نمونه‌های بهبودیافته آن نیز برای این منظور توسعه یافته است(Mallick et al. 2019).

۲- ۲- ۲. روش YAKE

الگوریتم YAKE برای استخراج کلیدواژه پنج قدم اصلی دارد (Campos et al. 2020):

۱. پیش‌پردازش متن و شناسایی واژه‌های کاندید
۲. استخراج ویژگی
۳. محاسبه امتیاز واژه‌های کاندید
۴. تولید n-gramها و محاسبه امتیاز کلیدواژه‌های کاندید
۵. حذف افزونگی در داده‌ها و رتبه‌بندی

کارهایی که در مرحله پیش‌پردازش روی متن انجام می‌شــود عموما عبارتند از جداســای جملات، قطعه‌بندی کردن متن، ریشــه یابی و اســتفاده از برچســب های نحوی مانند POS Tag، و حذف ایست‌واژه‌ها. خروجی این مرحله مجموعه‌ای از واژگان یا عبارات هستند که در مراحل بعدی براساس مجموعه‌ای از ویژگی‌ها امتیازدهی می‌شوند.

در مرحله استخراج ویژگی، برای واژگان و عبارات کاندید، پنج ویژگی زیر محاسبه می‌شود:

- ویژگی T_{Case}
- ویژگی $T_{Position}$
- ویژگی TF_{Norm}

- ویژگی T_{Rel}
- ویژگی T_{Sent}

ویژگی T_{Case} بیانگر این است که آیا آن واژه یا حرف اول آن با حرف بزرگ نوشته شده است. در زبان انگلیسی واژگانی که با حروف بزرگ نوشته می‌شوند یا حرف اول آنها بزرگ است، معمولا اهمیت بیشتری در متن دارند و می‌توانند بیانگر بخشی از مفهوم متن باشند. برای محاسبه این ویژگی از فرمول زیر استفاده می‌شود:

$$T_{case} = \frac{\max(TF(U(t)),\ TF(A(t))}{\ln(TF(t))} \quad (۴-۲)$$

که در آن $TF(U(t))$ تعداد رخدادهای واژه t است که با حروف بزرگ نوشته شده یا شروع شده، $TF(A(t))$ تعداد دفعاتی است که همه حروف واژه t به صورت حروف بزرگ نوشته شده، و $TF(t)$ فراوانی واژه است.

ویژگی $T_{Position}$ یک ویژگی مکانی است که به محل رخداد واژه در جمله ارتباط دارد و به صورت زیر تعریف می‌شود:

$$T_{position} = \ln(\ln(3 + Median(Sent_t))) \quad (۵-۲)$$

که در آن $Sent_t$ مکان جملاتی در متن است که واژه t را دارند و تابع $Median$ همان میانه مجموعه در بردارنده همه $Sent_t$ هاست.

ویژگی TF_{Norm} بیانگر فراوانی واژه است و به صورت زیر تعریف می‌شود:

$$TF_{Norm} = \frac{\mathrm{TF(t)}}{Mean\,TF + 1 * \sigma} \quad (۶-۲)$$

که در آن $Mean\,TF$ میانگین فراوانی‌ها و σ انحراف معیار آنهاست.

ویژگی T_{Rel} یک ویژگی آماری است که میزان پراکندگی واژه را براساس بافتی که در آن قرار دارد، نشان می‌دهد. در واقع هر چه کلماتی که با یک واژه هم‌رخداد هستند متنوع‌تر و مختلف باشد، اهمیت آن واژه به عنوان کلیدواژه، کمتر خواهد بود. برای تعریف این ویژگی، ابتدا پراکندگی چپ (و راست) به صورت زیر تعریف می‌شود:

$$DL[DR] = \frac{|\mathrm{A}_{t,w}|}{\sum_{k\epsilon \mathrm{A}_{t,w}} CoOccur_{t,l}} \quad (۷-۲)$$

بطوریکه $|A_{t,w}|$ تعداد کلمات متمایزی است که با واژه t در طرف چپ (یا راست) هم‌رخداد بوده‌اند و در مخرج هم تعداد کل کلماتی است که با واژه هم‌رخداد بوده‌اند. منظور از DR پراکندگی راست است که همانند پراکندگی چپ برای کلمات سمت راست تعریف می‌شود. بر همین اساس، ویژگی T_{Rel} به صورت زیر تعریف می‌شود:

$$T_{Rel} = 1 + (DL + DR) * \frac{TF(t)}{Max\ TF} \quad (۸-۲)$$

ویژگی T_{Sent} بیانگر تعداد دفعاتی است که واژه در جملات مختلف آمده باشد. در واقع واژگانی که در جملات متمایز زیادی آمده باشند، اهمیت بیشتری دارند. این ویژگی به صورت زیر محاسبه می‌شود:

$$T_{Sent} = \frac{SF(t)}{\#Sentences} \quad (۹-۲)$$

بطوریکه $SF(t)$ فراوانی جمله‌ای واژه و مخرج کسر کل جملات متن است.

در نهایت براساس ویژگی‌های فوق، امتیاز هر واژه به صورت زیر حساب می‌شود:

$$S(t) = \frac{T_{Rel} * T_{Position}}{T_{Case} + \frac{TF_{Norm}}{T_{Rel}} + \frac{T_{Sentence}}{T_{Rel}}} \quad (۱۰-۲)$$

برای تولید و محاسبه امتیاز هر عبارت، ابتدا یک پنجره با اندازه مشخص در هر جمله متن در نظر گرفته می‌شود و عبارت‌های کاندید تولید می‌شوند. سپس برای هر عبارت کاندید، براساس فرمول زیر امتیاز محاسبه می‌شود:

$$S(kw) = \frac{\prod_{t \epsilon kw} S(t)}{KF(kw) * (1 + \sum_{t \epsilon kw} S(t))} \quad (۱۱-۲)$$

که در آن $KF(kw)$ فراوانی عبارت کلیدی است. عباراتی که $S(kw)$ کمتری دارند مرتبط‌تر هستند. بنابراین در نهایت کلیدواژه‌ها کاندید براساس امتیازشان از کمترین امتیاز تا بیشترین امتیاز مرتب می‌شوند.

پس از محاسبه امتیاز هر کلیدواژه یا عبارت کلیدی کاندید در مرحله آخر، افزونگی اطلاعات حذف می‌شود. برای این منظور شباهت کلیدواژه‌ها با هم سنجیده می‌شود و فهرست نهایی از کلیدواژه‌ها تولید می‌شود. برای تولید این فهرست ابتدا کلیدواژه کاندیدی که کمترین مقدار $S(kw)$ را کسب

کرده، در فهرست قرار می‌گیرد و بعد کلیدواژه‌های دیگر کاندید به ترتیب با کلیدواژه موجود در فهرست از نظر میزان شباهت مقایسه می شوند. اگر شباهت آنها بیش از یک مقدار آستانه بود، آنگاه کلیدواژه‌ای در نظر گرفته می‌شود که $S(kw)$ کمتری دارد. اگر شباهت از مقدار آستانه کمتر بود، کلیدواژه کاندید به فهرست اضافه می‌شود.

۲- ۲- ۳. روش RAKE

ابتدا الگوریتم RAKE [۱۱] برای استخراج خودکار کلیدواژه در زبان انگلیسی معرفی شد. این الگوریتم تنها روی یک سند اجرا می‌شود و معمولا برای اسناد کوتاه نیز نتایج خوبی را ارائه می‌کند. همچنین برای استفاده از آن نیازی به در دسترس بودن مجموعه بزرگ متون (corpus) نیست (Rose et al. 2010). این الگوریتم از یک روش بدون ناظر بهره می‌گیرد و مستقل از زبان اسناد و مدل زبانی عمل می‌کند و می‌تواند برای متون فارسی نیز به‌کار گرفته شود. همچنین برخلاف برخی از الگوریتم‌های پیشین، خروجی‌های این الگوریتم اغلب عبارات کلیدی (در برگیرنده شمار دو یا بیشتر کلمه) هستند. اگرچه این الگوریتم، همانند دیگر الگوریتم‌ها در زبان انگلیسی، به ظاهر وابستگی زبانی ندارد، لیکن پیش‌پردازش‌هایی که باید روی متن در زمان به‌کارگیری الگوریتم انجام شود، روی نتیجه تاثیر زیادی خواهد گذاشت. الگوریتم RAKE از دو مرحله اصلی پیش‌پردازش متن و پردازش تشکیل شده است.

۲-۲-۳-۱. پیش‌پردازش متن

یکی از مهمترین کارها در مرحله پیش‌پردازش، حذف ایست‌واژه‌ها[۱۲] و علایم نگارشی است. برچسب‌گذاری دستوری کلمات[۱۳] نیز برای تعیین ضریب اهمیت کلمات و تعیین نقش کلمات در جمله است، که می‌تواند برای حذف ایست‌واژه‌ها نیز به‌کارگرفته شود. منظور از برچسب‌گذاری کلمات نسبت دادن برچسب‌های دستوری همچون اسم، صفت، حرف اضافه و ... به آنها است. از آنجایی که معمولاً فعل ها و حروف اضافه، نقش تعیین کننده‌ای در کلیدواژه ها ندارند، در الگوریتم‌های استخراج کلیدواژه عموماً واژگانی که این برچسب‌ها را در متن دارند از فهرست اولیه کلمات کاندید برای کلیدواژه، حذف می‌شوند.

[۱۱] Rapid Automatic Keyphrase Extraction

[۱۲] Stop words

[۱۳] Part of Speech (POS) Tagging

برای فهرست ایست‌واژه‌ها از یکی از پروژه‌های گیت هاب [۱۴] استفاده شده و برای غنی‌سازی آن نتایج پژوهش (سمائی و رسولی، ۱۳۹۹) نیز در آن لحاظ شده است. در این پژوهش از مدل برچسب‌گذار موجود در کتابخانه هضم (Hazm[۱۵]) استفاده شده است. هضم یکی از بسته‌های زبان پایتون است.

۲-۲-۳-۲. پردازش متن و رتبه‌بندی عبارات کاندید

در این مرحله، با دسته‌بندی کلمات، عبارات کلیدی کاندید تشکیل می‌شوند. سپس معیارهای وزن‌دهی شامل درجه (degree)، و فراوانی (frequency)، امتیاز (score) عبارات کلیدی کاندید محاسبه می شود. در پایان پس از مرتب سازی عبارات بر پایه امتیازهایشان، عبارات برگزیده استخراج می‌گردد. در ادامه گام‌های مرحله‌ی پردازش متن تشریح می‌شوند.

در مرحله پردازش متن، ابتدا عبارات کاندید برای کلیدواژه مشخص می شوند. برای این منظور، در هر جمله، کلماتی که قبل و بعد از آن‌ها ایست‌واژه وجود داشته باشد، در یک دسته به عنوان عبارت کاندید قرار می‌گیرد. به عنوان مثال در جمله زیر، ایست‌واژه‌ها، مشخص شده‌اند:

روش‌های با ناظر برای استخراج کلیدواژه، براساس الگوریتم‌های دسته‌بندی عمل می‌کنند.

با حذف ایست‌واژه‌ها، عبارات کلیدی کاندید زیر حاصل می‌شوند:

روش‌های با ناظر، استخراج کلیدواژه، الگوریتم‌های دسته‌بندی

سپس برای هر عبارت کاندید، یک امتیاز براساس فراوانی آن و درجه آن به صورت زیر محاسبه می‌شود.

برای هر کلمه w_j که در عبارات مختلفی دیده شده، درجه (degree) کلمه به صورت زیر تعریف می‌شود:

$$deg_w(w_j) = \sum_{p_i \in P_{w_j}} length(p_i.w_j) \qquad (۱۲-۲)$$

که در آن P_{w_j} مجموعه تمام عباراتی است که شامل کلمه w_j است:

[۱۴] https://github.com/kharazi/persian-stopwords

[۱۵] https://www.sobhe.ir/hazm/

$$\mathrm{P}_{w_j} = \{p_i | w_j \in p_i\} \quad (۱۳-۲)$$

و p_i ها عبارت‌هایی هستند که کلمه w_j در آنها وجود دارد و $length(p_i.w_j)$ تعداد کلمات هر عبارت p_i است. اگر $freq(w_j)$ بیانگر فراوانی کلمه w_j باشد، آنگاه امتیاز هر کلمه این‌گونه تعریف می‌شود:

$$score_w(w_j) = \frac{deg_w(w_j)}{freq(w_j)} \quad (۱۴-۲)$$

در پایان برای هر عبارت کاندید p_i، امتیاز آن برابر است با مجموع امتیازهای کلمات تشکیل‌دهنده آن. بنابراین:

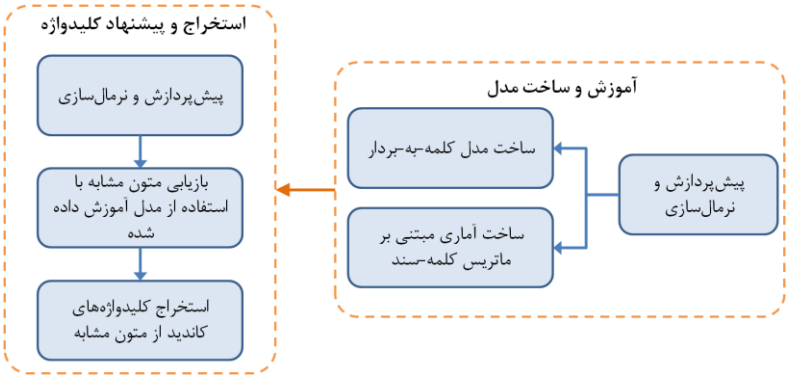
$$score_{RAKE}(p_i) = \sum_{j=1}^{n} score_w(w_j) \quad (۱۵-۲)$$

که در آن فرض بر این است که عبارت p_i دارای n کلمه $w_1, w_2, \dots, w_n$ است.

در نهایت عباراتی که بیشترین امتیاز را داشته باشند، به عنوان عبارات کلیدی انتخاب می‌شوند.

۲- ۲- ٤. روش تلفیقی

محبی و جلالی‌منش (۱۳۹۷) روشی تلفیقی را برای استخراج کلیدواژه از مستندات علمی ارائه کردند که براساس سیستم‌های پیشنهاددهنده و استدلال نمونه-محور عمل می‌کند. مبنای استدلال در این روش اینست که اسناد مشابه می‌توانند کلیدواژه‌های مشابه داشته باشند. بر همین اساس، ابتدا اسناد مشابه با یک سند جدید براساس روش‌های TFIDF و روش‌های بازنمایی کلمه-به-بردار، بازیابی می‌شوند. سپس کلیدواژه‌های کاندید از بین اسناد مشابه در نظر گرفته می‌شوند و در نهایت براساس یک تابع رتبه‌بندی، کلیدواژه‌های مناسب از بین آنها انتخاب می‌شوند. روش پیشنهادی بر روی مجموعه‌ای از اسناد پایگاه گنج در سه حوزه فنی و مهندسی، هنر و ادبیات، و علوم انسانی، پیاده‌سازی شده و نتایج آن با معیارهایی نظیر دقت، فراخوانی و نظرات متخصصین ارزیابی شده است. ساختار روش پیشنهاد شده در محبی و جلالی‌منش (۱۳۹۷) در شکل ۲-۲ نمایش داده شده است.



شکل ۲-۲- ساختار روش محبی و جلالی‌منش (۱۳۹۷)

کلیدواژه‌های کاندید، همان کلیدواژه‌های اسناد مشابه هستند. برای امتیازدهی و رتبه‌بندی کلیدواژه‌های کاندید، امتیاز هر کلیدواژه براساس دو روش محاسبه می‌شود و امتیاز نهایی حاصل ضرب امتیاز حاصلی از این دو روش خواهد بود (محبی و جلالی منش، ۱۳۹۷).

روش اول: بر مبنای رتبه سند در بازیابی متون مشابه

برای هر کلیدواژه کاندید $\boldsymbol{p_l}$ یک بردار، $V_{key}(p_l)$، M تایی (M تعداد اسناد بازیابی شده است) بر مبنای مدل تک-مولفه‌ای باینری در نظر گرفته می‌شود که هر مولفه از آن می‌تواند مقدار ۰ یا ۱ را اتخاذ کند که مقدار صفر یا یک برای آن بیانگر اینست که کلیدواژه متعلق به کدام سند بازیابی شده از بین M سند است:

(۱۶-۲)

$$V_{key}(p_l) = (v_{1l}, \dots, v_{Ml})$$

$$v_{mk} = \begin{cases} 1 & if\ k_l\ is\ a\ keyword\ for\ the\ m^{th}\ document \\ 0 & otherwise \end{cases}$$

از طرفی هر سند $\boldsymbol{d_m}$ دارای یک امتیاز بازیابی $\boldsymbol{score(d_m)}$ است. بنابراین می‌توان یک بردار $\boldsymbol{M}$ تایی از امتیازها نیز مانند عبارت زیر در نظر گرفت:

$$V_{ret} = (score(d_1), \dots, score(d_M)) \qquad (۱۷-۲)$$

حال می‌توان یک تابع امتیاز براساس ضرب نقطه‌ای بین دو بردار $\boldsymbol{V_{ret}}$ و $\boldsymbol{V_{key}(k_l)}$ در نظر گرفت که بیانگر امتیاز هر کلیدواژه باشد. کلیدواژه‌هایی که در تعداد اسناد بیشتری از بین اسناد بازیابی شده هستند،

نیز نیاز است که وزن بیشتری را نسبت به سایر کلیدواژه‌ها داشته باشند. بنابراین در محاسبه امتیاز نهایی می‌توان ضریبی را برای این منظور لحاظ نمود. بنابراین امتیاز کلیدواژه $\boldsymbol{k_l}$ براساس امتیاز عبارتست از:

$$score_1(p_l) = \frac{V_{ret} \cdot V_{key}(p_l)}{|V_{ret}|} \; log(f(p_l) + \frac{1}{4}M) \qquad (۱۸-۲)$$

که $f(p_l)$ بیانگر تعداد اسناد از بین M سند بازیابی شده است، که p_l کلیدواژه آنها بوده است. در واقع با ضرب عبارت فوق در لگاریتم، می‌خواهیم وزن کلیدواژه‌هایی را که در بیش از یک چهارم اسناد بازیابی شده آمده‌اند، بیشتر در نظر بگیریم. همانطور که از شیوه محاسبه مشخص است، در این روش، امتیاز یک کلیدواژه تنها براساس امتیاز شباهت اسناد دربردارنده آن کلیدواژه محاسبه می‌شود و ماهیت معنایی آن در نظر گرفته نمی‌شود.

روش دوم: امتیازدهی براساس ارتباط معنایی با سند

در این روش ارتباط معنایی کلیدواژه کاندید با سند جدید به عنوان معیاری برای محاسبه امتیاز کلمه در نظر گرفته می شود. برای این منظور برای هر کلیدواژه یک بردار معنایی براساس مدل آموزش داده شده کلمه–به–بردار، در نظر گرفته می‌شود.

اگر واژگان سند که مدل کلمه–به–بردار (Word2Vec) هم برای آنها موجود است، داخل مجموعه $\mathcal{R}_{\text{new}}$ قرار گیرد، آنگاه می‌توان شباهت معنایی هر واژه با کلیدواژه k_l را براساس مدل برداری آنها محاسبه کرد و در نهایت امتیاز هر کلیدواژه براساس میانگین این شباهت‌ها حساب شود. بنابراین:

$$score_2(p_l) = \frac{1}{|\mathcal{R}_{new}|} \sum_{\text{r}\in\mathcal{R}_{\text{new}}} \cos(\, V_{sem}(p_l),\, V_{W2V}(r)\,) \qquad (۱۹-۲)$$

که در آن $V_{sem}(p_l)$ نمایش برداری کلیدواژه کاندید k_l براساس مدل کلمه–به–بردار است و مقصود از $V_{W2V}(r)$ نمایش برداری واژه r از سند است. در نهایت امتیاز نهایی برای هر کلیدواژه کاندید به صورت زیر محاسبه می‌شود:

$$score_{final\_M}(p_l) = \; score_1(p_l) \times \; score_2(p_l) \qquad (۲۰-۲)$$

فصل سوم:

مجموعه داده‌های موجود

۳. مجموعه داده‌های موجود

در این فصل مهمترین مجموعه داده های موجود برای ارزیابی الگوریتم های استخراج خود کار کلیدواژه‌ها از متن، معرفی می‌شوند و در انتها ویژگی‌های لازم برای ایجاد یک مجموعه داده استاندارد فارسی برای استخراج کلیدواژه براساس ویژگی‌های داده‌های موجود معرفی می‌شود.

۳- ۱. مجموعه داده‌های انگلیسی

در زبان انگلیسی مجموعه داده‌های متنوعی برای مساله استخراج کلیدواژه وجود دارد. که نوع و محتوی آنها به صورت متون خبری، متن کامل مقالات علمی، چکیده متون علمی، پاراگراف، صفحات وب، و یا متون وبلاگ‌ها، است. این مجموعه داده‌ها با مطالعه پژوهش‌های مختلف در حوزه استخراج کلیدواژه و مقالات مروری در این حوزه شناسایی شده‌اند. (Campos et al. 2020; Papagiannopoulou and Tsoumakas 2019; Siddiqi and Sharan 2015).

از نظر تعداد اسناد موجود در مجموعه داده‌ها، در بین مجموعه‌های زبان انگلیسی، کمترین تعداد مربوط به مجموعه داده wiki20 است که شامل ۲۰ متن علمی است، و بیشترین مربوط به مجموعه داده KP20k است که شامل ۲۱۵۷۸ چکیده مقاله علمی است. اطلاعات مربوط به مهمترین مجموعه داده‌های استاندارد در جدول زیر آمده است.

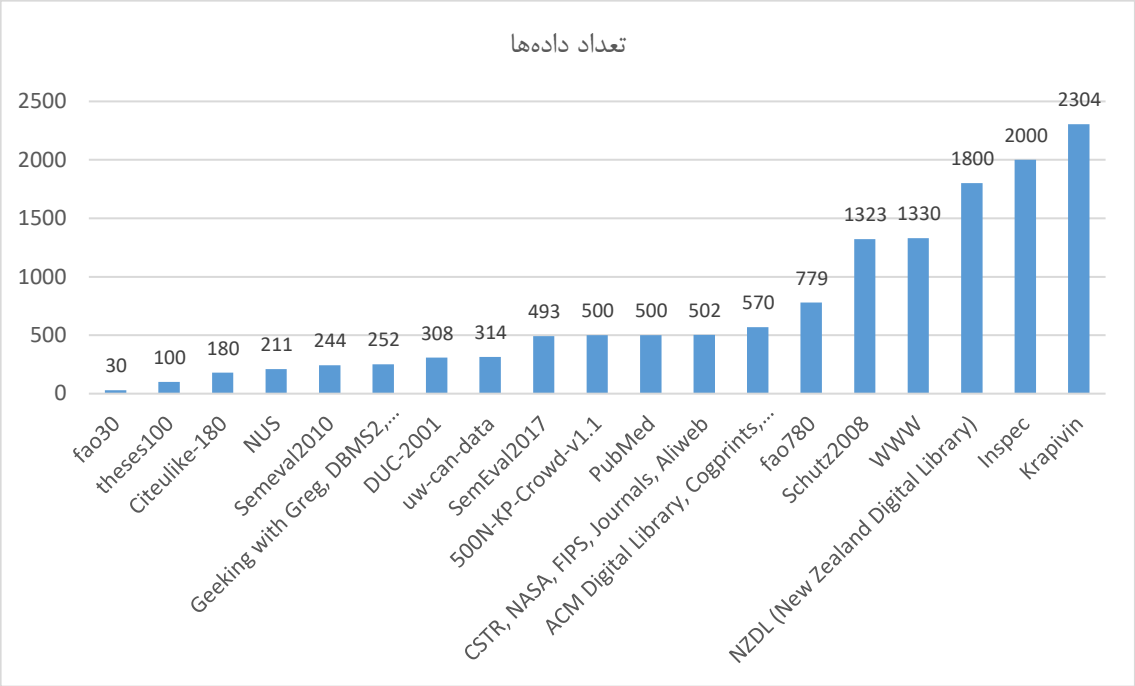
جدول ۳-۱- مجموعه داده‌های انگلیسی موجود برای استخراج کلیدواژه

ردیف	نام مجموعه داده	تهیه کننده	تعداد	نوع داده‌ها	نحوه اختصاص کلیدواژه	متوسط تعداد کلیدواژه‌های هر سند
۱	NZDL (New Zealand Digital Library)	(Witten et al. 1999)	۱۸۰۰	گزارش‌های فنی	نویسندگان مقالات	۵/۴

ردیف	نام مجموعه داده	تهیه کننده	تعداد	نوع داده‌ها	نحوه اختصاص کلیدواژه	متوسط تعداد کلیدواژه‌های هر سند
۲	uw-can-data	(Hammouda, Matute, and Kamel 2005)	۳۱۴	صفحات وب	موضوعات صفحات وب	-
۳	CSTR, NASA, FIPS, Journals, Aliweb	(El-Beltagy and Rafea 2009)	۵۰۲	مقالات و کزارش های فنی	نویسندگان مقالات	۵ تا ۱۲
۴	ACM Digital Library, Cogprints, Springerlink, Wikipedia	(You, Fontaine, and Barthès 2009)	۵۷۰	مقاله‌های علمی	نویسندگان مقالات	–
۵	Geeking with Greg, DBMS2, Stanford Infoblog	(Grineva, Grinev, and Lizorkin 2009)	۲۵۲	وب لاگ	نویسندگان	–
۶	ACL Anthology Network (AAN) corpus	(Radev et al. 2013)	۱۳۷۳۹	مقاله‌های علمی	نویسندگان مقالات	–
۷	Reuters collection	(Apté, Damerau, and Weiss 1994)	۲۱۵۷۸	متون خبری	نوسندگان مقالات	–
۸	NUS	(Nguyen and Kan 2007)	۲۱۱	متن کامل مقالات علمی	نویسندگان مقالات و متخصصین موضوعی	۱۱/۳۳
۹	Krapivin	(Krapivin et al. 2010)	۲۳۰۴	متن کامل مقالات علمی	نویسندگان مقالات و سردبیر ACM	۶/۳۴
۱۰	Semeval2010	(Kim et al. 2010)	۲۴۴	متن کامل مقالات علمی	نویسندگان مقالات و متخصصین موضوعی	۱۶/۴۷
۱۱	Citeulike-180	(Wang, Chen, and Li 2013)	۱۸۰	متن کامل مقالات علمی	متخصصین موضوعی	–
۱۲	Inspec	(Hulth 2003)	۲۰۰۰	چکیده مقالات علمی	نمایه‌سازان	۱۴/۶۲
۱۳	KP20k	(Meng et al. 2017)	۲۰۰۰۰	چکیده مقالات علمی	نویسندگان مقالات	–
۱۴	WWW	(Gollapalli and Caragea 2014)	۱۳۳۰	چکیده مقالات علمی	نویسندگان مقالات	۵/۸

ردیف	نام مجموعه داده	تهیه کننده	تعداد	نوع داده‌ها	نحوه اختصاص کلیدواژه	متوسط تعداد کلیدواژه‌های هر سند
۱۵	DUC-2001	(Wan and Xiao 2008)	۳۰۸	متون خبری	متخصصین موضوعی	-
۱۶	500N-KP-Crowd-v1.1	(Marujo, Viveiros, and da Silva Neto 2013)	۵۰۰	متون خبری	سرویس آمازون Mechanical Turker	۴۸/۹۲
۱۷	PubMed	(Aronson et al. 2000)	۵۰۰	متن کامل مقالات علمی	نویسندگان مقالات و اصطلاح‌نامه MeSH	۱۵/۲۴
۱۸	Schutz2008	(Schutz 2008)	۱۳۲۳	مقالات علمی	نویسندگان مقالات	۵
۱۹	SemEval2017	(Augenstein et al. 2017)	۴۹۳	Paragraph	متخصصین موضوعی	۱۸/۱۹
۲۰	fao30	(Medelyan and Witten 2008)	۳۰	مقالات علمی	متخصصین موضوعی	۲۳/۳۳
۲۱	fao780	(Medelyan and Witten 2008)	۷۷۹	مقالات علمی	متخصصین موضوعی	۸/۹۷
۲۲	theses100	(Theses100 2019)	۱۰۰	پایان نامه‌ها و رساله ها	نویسندگان مقالات	۷/۶۷

تعداد داده‌های موجود در مجموعه داده‌های این جدول در شکل ۳-۱ آمده است. مجموعه داده ردیف‌های ۱۳، ۷ و ۶ در جدول فوق در این شکل نمایش داده نشده است زیاد تعداد آنها با سایر مجموعه داده‌های بسیار متفاوت است.



شکل ۳-۱- تعداد داده‌های مجموعه داده‌های انگلیسی موجود

در بین ۲۲ مجموعه داده‌ها، ۱۷ مجموعه متون علمی، ۳ مجموعه متون خبری، و ۲ مجموعه دارای متون عمومی هستند.

۳- ۲. مجموعه داده‌های فارسی

مجموعه داده‌های فارسی که در پژوهش‌های پیشین برای مساله استخراج کلیدواژه استفاده شده است، به اندازه مجموعه‌های انگلیسی متنوع نیستند. در بیشتر پژوهش‌ها از مجموعه داده‌های متون خبری استفاده شده است. در داده‌هایی که از متون علمی استفاده شده‌اند، عموما نویسندگان، مجموعه داده‌ای را متشکل از متون علمی یا مقالات ایجاد کرده‌اند که هر یک دارای کلیدواژه‌های نویسنده بوده‌اند. در جدول ۳-۲ اطلاعات مربوط به این مجموعه داده‌ها و پژوهش‌های مرتبط با آنها آمده است.

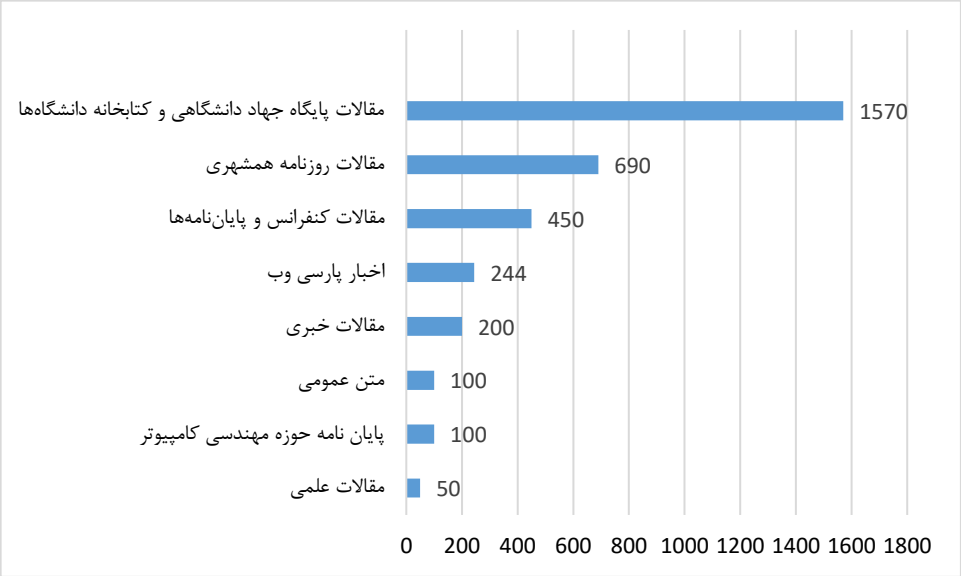
جدول ۳-۲-مجموعه داده‌های فارسی

ردیف	عنوان داده	تهیه‌کنندگان	تعداد	نوع داده
۱	PerKey	دوستمحمدی و همکاران (Doostmohammadi, Bokaei, and Sameti 2019)	۵۵۳،۱۱۱	متون خبری از منابع خبری متعدد
۲	مقالات پایگاه جهاد دانشگاهی و کتابخانه دانشگاه‌ها	لازمی و همکاران (Lazemi, Ebrahimpour-Komleh, and Noroozi 2019)	۱۵۷۰	متن کامل مقالات علمی

۳	پایان نامه حوزه مهندسی کامپیوتر	احمدی و حبیبی (۱۳۹۵)	۱۰۰	چکیده پایان نامه
۴	مقالات روزنامه همشهری	احمدی و حسینی‌خواه (۱۳۹۲) راد و پروین (۱۳۹۵) رضایی و همکاران (۱۳۹۶)	۶۹۰	متون خبری در ۵ حوزه موضوعی
۵	مقالات خبری	ویسی و افلاکی (۱۳۹۴)	۲۰۰	متن خبری با موضوعات مختلف
۶	اخبار پارسی وب	باسره و همکاران (۱۳۹۴)	۲۴۴	متون خبری از اخبار ۳۲ خبرگزاری
۷	متن عمومی	عرب نرئی و همکاران (۱۳۸۶)	۱۰۰	متن عمومی در موضوعات مختلف
۸	مقالات علمی	گزنی (۱۳۸۵)	۵۰	متن کامل مقالات علمی
۹	مقالات کنفرانس و پایان‌نامه‌ها	بشیری و همکاران (۱۳۸۴) تشکری و میبدی (۱۳۸۳)	۴۵۰	متن چکیده مقالات

تنها پژوهشی که به صورت تخصصی روی ایجاد مجموعه داده برای استخراج کلیدواژه تمرکز داشته، پژوهش(Doostmohammadi, Bokaei, and Sameti 2019) است که در آن مجموعه داده متون خبری ایجاد شده است. یکی از مجموعه داده‌هایی که در تعدادی از پژوهش‌ها به صور مشترک استفاده شده، مجموعه داده همشهری است که مشتمل بر متون خبری در حوزه‌های موضوعی مختلف است.

تعداد مقالات در هر مجموعه داده در شکل ۳-۲ مقایسه شده است. مجموعه داده PerKey به دلیل تفاوت زیاد با سایر مجموعه داده‌ها در شکل نیامده است.



شکل ۳-۲-تعداد مقالات در هر مجموعه داده فارسی

از بین مجموعه داده‌های علمی استفاده شده، مجموعه ۴۵۰ تایی مقالات کنفرانس‌ها و پایان‌نامه‌ها و مجموعه ۱۵۷۰ تایی مقالات جهاد دانشگاهی و کتابخانه دانشگاه‌ها بیشتر مورد توجه قرار گرفته‌اند که هر دو دربردارنده چکیده متون هستند. علاوه بر آن، در بین ۹ مجموعه داده معرفی شده، ۴ مجموعه داده متون علمی، ۴ مجموعه متون خبری، و ۱ مجموعه دارای متون عمومی است.

۳- ۳. ویژگی‌های مشترک مجموعه داده‌ها

با مقایسه نوع متون مجموعه داده‌ها در داده‌های فار سی و انگلیسی، می‌توان دریافت که پژوهش‌ها در حوزه ایجاد مجموعه داده‌های علمی فارسی نسبت به پژوهش‌های انگلیسی، محدود است.
درشکل ۳-۳ این مقایسه نمایش داده شده است.

شکل ۳-۳-مقایسه درصد نوع متون در مجموعه داده‌های فارسی و انگلیسی

در بین مجموعه داده‌های علمی فارسی و انگلیسی موجود، ویژگی‌های مشترکی وجود دارد که عبارتند از:

- وجود عنوان برای هر متن
- وجود کلیدواژه برای هر سند (حداقل ۳ و حداکثر ۳۰)
- استفاده از متون علمی منتشر شده (مقالات به چاپ رسیده، پایان‌نامه‌ها و رساله‌ها، یا گزارش‌های فنی منتشر شده)
- استفاده از چکیده یا تمام‌متن سند

کلیدواژه‌های در نظر گرفته شده در مجموعه داده‌ها به چند طریق اختصاص داده شده‌اند. در هر مجموعه داده کلیدواژه‌های متون علمی براساس یکی از حالت‌های زیر برای هر سند تعیین شده است:

- استفاده از کلیدواژه‌های نویسنده سند
- استفاده از متخصصین موضوعی (دانشجویان رشته تخصصی، کارشناسان متخصص) برای تخصیص کلیدواژه
- استفاده از کلیدواژه‌های نویسنده و متخصص موضوعی با هم

با اینکه استفاده از متخصص موضوعی و کلیدواژه‌های نویسنده به طور همزمان، هزینه زیادی را به همراه دارد، اما می‌تواند بر غنای کلیدواژه‌ها بیفزاید.

این ویژگی‌ها، حداقل الزامات را برای ایجاد مجموعه داده استاندارد فارسی مشخص می‌کند. بنابراین در این پژوهش، مجموعه داده فارسی علمی با ویژگی‌های زیر ایجاد می‌شود:

- هر سند دارای عنوان، چکیده، و کلیدواژه است.
- سندها باید از اسناد علمی منتشر شده معتبر انتخاب شده باشند.
- کلیدواژه‌های هر سند براساس کلیدواژه‌های نویسنده و نیز نظر متخصصین موضوعی و نمایه‌سازان مشخص می‌شود.
- برای افزایش تعداد کلیدواژه‌های هر سند، از چندین روش استخراج کلیدواژه استفاده می‌شود و مجموعه متنوعی از انواع کلیدواژه‌ها برای هر سند تشکیل می شود. در نهایت با استفاده از نظر متخصص نمایه‌ساز این کلیدواژه‌ها می‌توانند مجدد بررسی شوند.

فصل چهارم:

روش پیشنهادی ایجاد مجموعه داده استاندارد

۴. روش پیشنهادی ایجاد مجموعه داده استاندارد

هدف اصلی از این پژوهش ایجاد مجموعه داده استاندارد فارسی برای مساله استخراج کلیدواژه از مستندات علمی فارسی است. برای این منظور پس از مطالعه ادبیات موضوع و مجموعه داده‌های موجود، در فصل‌های قبل، در این بخش روش پیشنهادی برای این منظور تشریح می‌گردد.

٤- ١. داده‌های اولیه

برای ایجاد مجموعه داده استاندارد، نیاز است که مجموعه‌ای اولیه از اسناد فارسی علمی در نظر گرفته شوند تا از آنها، مجموعه داده استاندارد ایجاد شود. داده‌های اولیه مورد استفاده در این پژوهش ۲۸۷۷ پایان‌نامه یا رساله فارسی است که از پایگاه اطلاعات علمی ایران، گنج، جمع‌آوری شده‌اند. این داده‌ها مربوط به پایان‌نامه‌هایی هستند که طی دو سال اخیر در این پایگاه اضافه شده‌اند. از این داده‌ها، اطلاعات فرادادهای آنها در دسترس است که این اطلاعات عبارتند از:

- شناسه سند
- عنوان
- چکیده
- موضوع
- کلیدواژه‌های نویسنده و کلیدواژه‌های نمایه‌ساز

این داده‌ها در چهار حوزه مختلف فنی و مهندسی، علوم انسانی، علوم و علوم کاربردی، و هنر براساس موضـوع‌بندی موجود در پایگاه گنج هسـتند. توزیع موارد موجود در هر موضـوع به شـرح جدول زیر است.

جدول ٤-١-تعداد اسناد در هر حوزه موضوعی

حوزه موضوعی	تعداد سند
علوم انسانی	۱۳۹۹
فنی و مهندسی	۹۰۱
علوم و علوم کاربردی	۳۵۴
هنر	۱۸۴

٤- ۲. روش پژوهش

با فرض در دسترس بودن مجموعه داده اولیه، ایجاد مجموعه داده استاندارد، سه مرحله اصلی دارد:

۱. آماده‌سازی داده‌های اولیه و پیاده‌سازی الگوریتم‌های منتخب استخراج کلیدواژه روی آنها
۲. یکپارچه‌سازی نتایج الگوریتم‌های استخراج کلیدواژه
۳. ارزیابی کلیدواژه‌های استخراج شده، توسط متخصصین و نمایه‌سازان

براسـاس اسـتانداردهای موجود در مجموعه داده‌ها، در نهایت نیاز اسـت که برای هر سـند، علاوه بر عنوان و چکیده آن، مجموعه‌ای از کلیدواژه‌های مختص آن سند وجود دا شته با شد. در این پژوهش، علاوه بر این موارد، مو ضوع سند، و نیز امتیاز هر کلیدواژه برا ساس نظر متخصص نیز مشخص خواهد شد و تعداد کلیدواژه‌های تخصیص یافته بیش از ۱۰ کلیدواژه خواهد بود. در ادامه این مراحل تشریح می‌شود.

٤- ۳. آماده‌سازی و پیاده‌سازی الگوریتم‌ها

یکی از مراحل اصلی در الگوریتم‌های استخراج کلیدواژه، آماده سازی و نرمال سازی داده است. برای پردازش متون فارسـی، مرحله نرمال‌سـازی متون، بسـیار اهمیت دارد که در آن عموما فعالیت‌های زیر انجام می‌شود:

- اصلاح نویسه‌ها

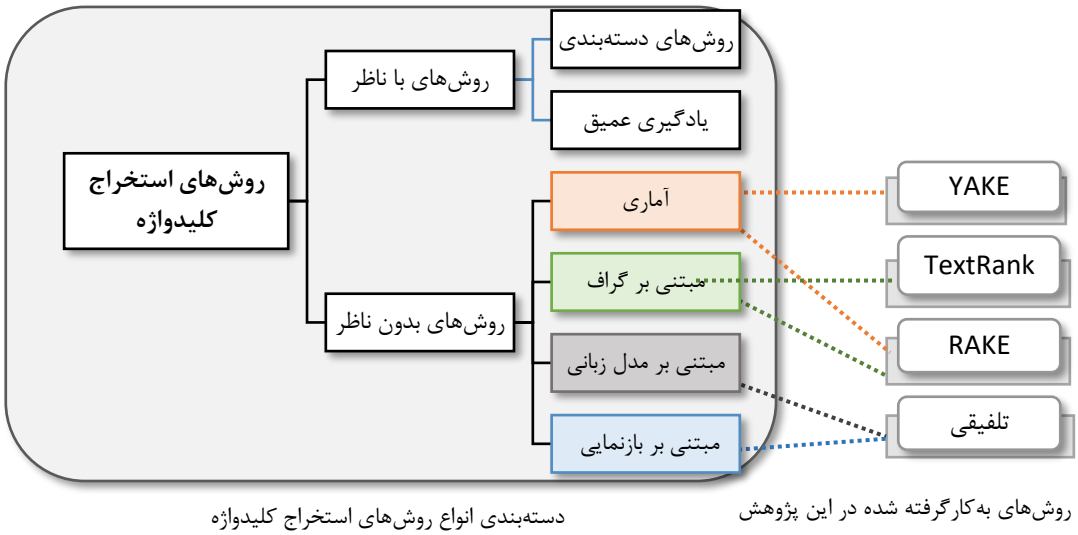
- اصلاح نشانه‌گذاری
- رعایت فاصله و نیم‌فاصله
- حذف اعراب و تشدید
- حذف علایم نگارشی (غیر از نقطه)
- استخراج جملات
- ریشه‌یابی واژه‌ها

برای انجام بخشی از این فعالیت‌ها، ازکتابخانه ابزار hazm که برای پردازش زبان طبیعی فارسی کاربرد دارد، استفاده می‌شود. پس از آن داده‌ها در یک فایل جیسون ذخیره می‌شود.

پس از آماده‌سازی داده‌ها، ۴ الگوریتم استخراج کلیدواژه روی داده‌ها پیاده‌سازی می‌شوند. این الگوریتم‌ها عبارتند از:

- الگوریتم RAKE
- الگوریتم YAKE
- الگوریتم TextRank
- الگوریتم تلفیقی (محبی و جلالی‌منش، ۱۳۹۷)

روش‌های استخراج کلیدواژه بسیار متنوع هستند. از بین این روش‌ها، سعی شده از هر نوع از روش‌های بدون ناظر، یک نمونه از روش‌های استخراج کلیدواژه در نظر گرفته شود. در شکل ۴-۱ جایگاه هر یک از روش‌های استفاده شده در بین انواع روش‌ها نمایش داده شده است.



شکل ٤-١- روش‌های به‌کارگرفته شده در این پژوهش و جایگاه آنها در انواع روش‌های استخراج کلیدواژه

چند نمونه از نتایج پیاده‌سازی هر یک از روش‌ها در جدول ۴-۲ آمده است.

جدول ۴-۲- نمونه کلیدواژه‌های استخراج شده براساس روش‌های مختلف در حوزه علوم انسانی

عنوان: ماهیت، شرایط و آثار نصب داور در حقوق ایران	
حوزه موضوعی: علوم انسانی	
چکیده: داوری فرایندی است که به موجب آن اختلاف طرفین نسبت به حقوق و تکالیف قانونی خود از طریق انتخاب یک یا چند نفر داور به جای دادگاه حل وفصل گردیده و رای لازم الاجرا صادر می گردد. در فصل اول کلیاتی راجع به مفهوم و مقایسه‌ی داوری با مفاهیم مشابه و اقسام داوری،داوری در مسائل خاص، مبانی حقوقی داوری در حقوق ایران بیان شده است و در فصل دوم به بررسی ماهیت و شرایط نصب داور و ممنوعیت‌های مرتبط با نصب داور پرداخته‌ایم و درفصل سوم به بر رسی آثار نصب داور پرداخته شده ا ست.النهایه به این نتیجه ر سیده‌ایم که ماهیت نصب داور بانظریه مختلط تطبیق بیشتری دارد و در خصوص شرایط نصب داور می‌توان به موارد زیر اشاره کرد: داورپذیری، استقلال و بی‌طرفی داور اشاره کرد و در خصوص ممنوعیت نصب داور می‌توان به ممنوعیت‌های مرتبط با داور و ممنوعیت‌های مرتبط با موضوع اختلاف و همچنین ممنوعیت‌های مرتبط با طرفین اختلاف اشاره کرد و در خصوص آثار نصب داور می‌توان گفت که هم برای طرفین اختلاف و هم برای دادگاه‌ها و هم برای داور تعهداتی را ایجاد می‌کند و از جمله‌ی آن‌ها می‌توان به لازم التباع بودن رأی داور توسط طرفین و اعمال قاعده‌ای امر قضاوت شده اشاره کرد و برای دادگاه‌ها مهم‌ترین اثرآن این است از دادگاه‌ها سلب صلاحیت می‌کند و برای داور الزام به رسیدگی و صدور رأی را ایجاد می‌کند.	
کلیدواژه‌ها:	
RAKE	ماهیت نصب داور بانظریه مختلط تطبیق,حقوق ایران، خصوص ممنوعیت نصب داور، فصل اختلاف، نصب داور، اصول حقوقی، استقلال، خصوص آثار نصب داور، خصوص شرایط نصب داور، شرایط نصب داور، آثار نصب داور، دادگاه حل وفصل، مبانی حقوق داور
YAKE	نصب داور، داور الزام، رای لازم الاجرا، اختلاف طرفین نسبت، لازم الاجرا صادر، آثار نصب داور، طرفین اختلاف اشاره، داور، داور تعهداتی، داور اشاره، اختلاف اشاره، طرفین اختلاف، نصب داور بانظریه، شرایط نصب داور، ماهیت نصب داور، ممنوعیت نصب داور
TextRank	آثار نصب داور، نصب داور، مفهوم، النهایه، رسیدگی ، موجب ، لازم التباع ، درفصل ، کلیات راجع ، انتخاب ، جمله ، مقایسه داور ، دادگاه سلب صلاحیت ، طرفین اختلاف ا شاره ، مسائل خاص ، تکالیف قانون ، رأی داور ، حقوق ایران بیان ، خصوص شرایط نصب داور ، ممنوعیت مرتبط ، فصل ، صدور رأی ، دادگاه مهم اثرآن ، لازم الاجرا صادر ، داور فرایند ، داورپذیری ، ماهیت نصب داور بانظریه مختلط تطبیق ، دادگاه حل وفصل گردیده ، نفر داور ، خصوص ممنوعیت نصب داور ، مبانی حقوق داور ، داور الزام ، بی‌طرف داور ا شاره ، ایجاد ، شرایط نصب داور ، اقسام داور ، اعمال قاعده امر قضاوت ، مفاهیم مشابه ، طرفین اختلاف ، قلال ، نتیجه ، خصوص آثار نصب داور ، اختلاف طرفین نسبت ، موضوع اختلاف ، داور تعهدات ، ممنوعیت‌های مرتبط
تلفیقی	داوری، ایران، انصاف، کمیسیون حقوق تجارت بین‌الملل سازمان ملل متحد، شخص ثالث، حقوق، قانون و قانونگزاری، بی‌طرفی، تجارت بین‌المللی، نظام حقوقی، استقلال، حقوق بین‌الملل، چالش، قرارداد، قانون، داوری بین‌المللی، مدیریت زمان, مدیریت هزینه، اتاق بازرگانی بین المللی، اجرا، اعتراض، داوری بین المللی، حقوق بین‌المللی, زمان‌بندی, اتاق بازرگانی بین‌المللی، داوری نامه، هویت، قدرت سیاسی، شناسایی، دعاوی تجاری بین المللی، صنایع پایین دستی، صنعت نفت، حقوق ایران، قرارداد (موافقتنامه) داوری، تجاری، حقوق (علم)، صلاحیت، روابط دیپلماتیک، ملی‌گرایی، سرمایه‌گذاری خارجی، نفت، شهود، دارایی عمومی، حل اختلاف، قانون داوری تجاری بین‌المللی، اختلاف، قانون اساسی، قرارداد داوری، بررسی تطبیقی، انگلستان، اجرای حکم، بین الملل، رای قضایی، قانون آیین دادرسی مدنی، بین المللی، اموال دولتی، نظم عمومی، نظارت قضایی، قانونگذار (مقننه)، دادگاه، مقامات دولتی، هیات حل اختلاف، افغانستان، قوه قضاییه، حقوق موضوعه، قضاوت، داوری تجاری بین المللی، مقررات، دادرسی، کامن‌لا، حقوق اسلامی، شوراهای حل اختلاف، رژیم حقوقی، قضات، مرکز بین‌المللی حل و فصل اختلافات سرمایه‌گذاری، قانون‌گذاری، حقوق تجارت، ابطال (فسخ کردن)، انحلال، رسیدگی کیفری، بطلان، ابطال، رأی داوری، صلاحیت قضایی، انحلال قرارداد، شرط داوری، ابطال رأی، ابطال رای داوری، ابطال (فقه)، مقررات داوری آنسیترال

جدول ٤-۳-نمونه کلیدواژه‌های استخراج شده براساس روش‌های مختلف در حوزه فنی و مهندسی

عنوان: تحلیل لرزه‌ای موج‌شکن‌های قائم و بررسی پایداری آنها از نظر واژگونی و لغزش	
حوزه موضوعی: فنی و مهندسی	
چکیده: دیوارهای ساحلی از مهم ترین سازه های حفاظت از ساحل به شمار می آیند که از لحاظ ایستایی کاملا متکی به خود بوده و جهت حفاظت ساحل از اثرات امواج مورد استفاده قرار می گیرند. سازه‌های دریایی همواره در معرض نیروهای محیطی از جمله نیروی باد، امواج، زلزله و ... قرار دارند. نیروی امواج مهم‌ترین نیروی وارد بر این سازه‌ها می‌باشد که روش‌های مختلفی برای محاسبه این نیرو و بررسی اثرات ضربه ای آن ، بر روی دیوارهای ساحلی تو سط محققان ارائه شده ا ست. در این تحقیق ابتدا مبانی کلی طراحی دیوارهای ساحلی مورد بررسی و ارزیابی قرار می گیرد. در گام بعد روش های طراحی و ساخت بررسی شده طرح های اقتصادی، جانمایی پروژه مورد مطالعه برر سی شده. نیروهای وارد بر دیواره ساحلی نا شی از نیروی زلزله تعیین شده و تنش های بحرانی محا سبه می شود. میزان تغیرات مکانی تمام نقاط دیواره ساحلی با کانتور های رنگی و تحلیلی ارائه می شود. همچنین پایداری دیوار ساحلی بتنی تحت اثر واژگونی و لغزش با استفاده نرم افزار Abacus بررسی می شود. و مقطع بهینه معرفی می گردد. کلیه نیروهای هیدرواستاتیکی و هیدرودینامیکی با توجه به روابط و مدلهای ریاضی به شبیه سازی اضافه می شود. در گام نهایی نتایج تحلیل های نرم افزاری و شبیه سازی با مطالعات تئوری بررسی شده و نتایج ارائه خواهد شد.	
کلیدواژه‌ها:	
RAKE	گام نهایی نتایج تحلیل نرم افزار، پایدار دیوار ساحل بتنی، نرم افزار، نقاط دیواره ساحل، طراح دیوار ساحل، مقطع بهینه معرف، تنش بحران ، سازه حفاظت، معرض نیرو محیط، نیرو زلزله
YAKE	ایستایی ، مهم , شمار , متکی , لحاظ ایستایی , مهم سازه , ساحل , ساحل شمار , حفاظت , ایستایی متکی , دیوارهای ساحلی , سازه حفاظت , ساحلی , حفاظت ساحل , سازه
TextRank	شبیه , تغیر مکان , نقاط دیواره ساحل , ساخت , تحلیل ارائه , مقطع بهینه معرف , تحقیق ابتدا مبانی , رنگ , طرح , نیرو امواج مهم نیرو وارد , بحران محاسبه , اثر واژگون , روابط , هیدرودینامیک , جانمایی پروژه , اثرات ضربه , جمله نیرو , دیواره ساحل , اضافه , طراح دیوار ساحل , متکی , نرم افزار , گام نهایی نتایج تحلیل , مطالعه , اثرات امواج , نیرو وارد , لحاظ ایستا , ارزیاب قرار , شمار , دیوار ساحل , مطالعات تئوری , سازه دریا , معرض نیرو محیط , نیرو زلزله تعیین , نیرو هیدرواتیکی , مدلهای ریاضی , تنش , پایدار دیوار ساحل بتنی , اقتصاد , نتایج ارائه , توجه , محقق ارائه , نرم افزار
تلفیقی	زلزله, اندرکنش, دیوار ساحلی, جدا سازی پایه, نشین سازه, شتاب سنج, ذخیره سازی آب, گرانی, استاندارد ۲۸۰۰, سفتی, تحلیل تاریخچه زمانی, جزیره قشم, ظرفیت تحمل, مقاوم سازی, روش اجزای محدود, قابلیت اطمینان, بویه دریایی, سازه دریایی, عمق, ضریب اطمینان, پو شش تونل, اجزای محدود, برهم کنش ساختار خاک, مقاومت بر شی, تغییر شکل, سازه فولادی, متغیر تصادفی, بار ایستا, کلاف, روش اجزای گسسته, ربات ماشین ابزار, تغییر مکان, طراحی سازه, شیب, بتن مسلح, شرط مرزی, مقاوم‌سازی, ظرفیت باربری, قاب مقاوم خمشی, سربار, آنالیز عددی, خاک مسلح, پمپاژ, فشار خاک, هیدرودینامیک, بلوک بتنی, برهم کنش سیال-سازه, برهم کنش خاک-سازه, مقاومت مصالح, دیواره بالادست, تجزیه و تحلیل دینامیکی, مصالح بنایی, شاخص اعتماد, حفاظت ساحل, نظریه اعتمادپذیری, تحلیل سازه, قاب خمشی فولادی, اندرکنش خاک و سازه, لغزش, قاب بتنی, مدلسازی عددی, سازه غیرمسلح, المان محدود, بدنه, مدلسازی عددی, تنش نهایی (سازه), بار جانبی, سد بتنی, عضو فشاری, حفاری, مقاومت در برابر زلزله, نیروی محوری, ساخت بنایی, دینامیک غیرخطی, عملکرد لرزه‌ای, ساختمان بتنی, مخازن مدفون بتنی, نیروی موج, ساختمان مجاور, نیروی زلزله, تحلیل دینامیکی, جریان آب, کشتی, مخزن بتنی, قاب خمشی, تنش مجاز (سازه), تحلیل پایداری, رفتار دینامیکی, طراحی براساس ضرایب بار و مقاومت, دیوار آجری, بارگذاری دینامیکی, دیوار بارانداز, واژگونی, بهسازی لرزه‌ای, پایداری سازه‌ای, روش عددی, سدهای بتنی, تحلیل استاتیکی, گودبرداری, پهنه لرزه‌ای, جابجایی, موج‌شکن, میخکوبی, پایدارسازی, تحلیل دینامیکی افزایشی, بارگذاری جانبی, بار برکنش, تحلیل قابلیت اطمینان, دیوار دیافراگمی, شمع(در ساختمان), موج شوکی, مدل سه بعدی, رفتار لرزه‌ای, سازه بلند, ماشین‌ابزار, بتن‌پاشی

جدول ٤-٤-نمونه کلیدواژه‌های استخراج شده براساس روش‌های مختلف در حوزه هنر

عنوان: بررسی شاخص‌های کیفیت در معماری بناهای مسکونی در دوره معاصر تهران	
حوزه موضوعی: هنر	
چکیده: مطالعه آثار و اندیشه‌های معماران معاصر از جمله حوزه‌های نسبتا نوپا و البته ضروری در مباحث معماری در دوران معاصر می‌باشد. بررسی و تحلیل دیدگاه‌های معمارانی که هم در عرصه عمل و هم در عرصه نظر در دوران معاصر به فعالیت مشغول بوده‌اند، می‌تواند چراغ راه آینده معماران جوان کشور قرار گیرد. بی‌هویتی و عدم کیفیت مساله مورد توجه در معماری معاصر ایران است، لذا ضروری است شاخص‌هایی برای سنجش کیفیت معماری در دوره معاصر مشخص شود، تا بتوان با بهره‌گیری از آن‌ها به معماری با کیفیت قابل قبولی د ست پیدا کرد. از این رو، نو شتار حا ضر با مرور ادبیات مو ضوعی که درباره نظرات اندیشمندان حوزه معماری و شهر سازی ا ست، به مطالعه آراء و نظرات معماران معا صر ایران پرداخته است و از دریچه نگاه آنان که سالیان طولانی عمر خود را برای ارتقاء حرفه و هنر معماری صرف کرده‌اند، به مسائل پیرامون معماری معا صر، نیازها و مشکلات معماری امروز پرداخته ا ست، سپس با برر سی و تحلیل گفته‌های آنان شاخص‌های کیفی به‌منظور سنجش کیفیت در معماری معا صر با تمرکز روی معماری مسکونی در تهران شنا سایی و معرفی گردیده است. اساس پژوهش حاضر بر پایه روش تحقیق توصیفی-تحلیلی محتوا شکل گرفته و روش و ابزارهای مورد استفاده شامل روش اسنادپژوهی می‌باشد که با مطالعات منابع کتابخانه‌ای حاصل شده است. تجزیه و تحلیل داده‌ها نیز با استفاده از روش آمار توصیفی و استنباطی و همچنین مدل استراتژیک سوآت انجام گرفته است. بعد از تحلیل داده‌های استخراج شده از پرسشنامه‌ای که با توجه به نقطه‌نظرات معماران پیشکسوت کشور تهیه گردید، شاخص‌های شنا سایی شده از نظر پا سخ‌دهندگان به این صورت الویت‌بندی شدند: شاخص عملکردی، شاخص اجرایی، شاخص هویتی، شاخص زیبایی‌شناسی و شاخص کالبدی.	
کلیدواژه‌ها:	
RAKE	نظرات معمار پیشکسوت کشور تهیه,آینده معمار جوان کشور قرار,نظرات اندیشمند حوزه معمار,تحقیق توصیفی-تحلیلی محتوا شکل,شهری,هویت,کیفیت,معماری معاصر,ساختمان مسکونی,شاخص کیفیت,دوره,نظرات معمار معاصر ایران,معمار معاصر ایران,هنر معمار صرف,سنجش کیفیت معمار,مدل استراتژیک سوآت
YAKE	جمله حوزه‌های , دوران معاصر , حوزه‌های , حوزه‌های نوپا , معماران معاصر ایران , دوران معاصر می‌باشد , معاصر ایران , آثار اندیشه‌های , اندیشه‌های معماران معاصر , معماری , معماری معاصر , مباحث معماری , معاصر ایران ، , اندیشه‌های معماران , نوپا , نظرات معماران معاصر , معماران معاصر , معاصر , معماری معاصر ایران
TextRank	شاخص شنا سایی , نو شتار , ارتقاء حرفه , شاخص هویت , تمرکز , شاخص عملکرد , نظرات اندیشمند حوزه معمار , نظر پاسخ‌دهندگان , سنجش کیفیت معمار , اساس پژوهش , الویت‌بندی , پرسشنامه , دوره معاصر مشخص , آینده معمار جوان کشور قرار , تهران شنا سایی , چراغ , بهره‌گیری , اندیشه معمار معا صر , مطالعات منابع کتابخانه حا صل , قبولی دست , نوپا , مرور ادبیات موضوع , دریچه , جمله حوزه , شاخص اجرایی , معرف گردیده , معمار معاصر ایران , عرصه نظر , اسنادپژوهی , سال طولانی عمر , معمار مسکونی , فعالیت مشغول , نیاز , نظرات معمار معاصر ایران , شاخص کالبد , کیفیت مساله , تجزیه , مدل راتژیک سوآت , نباط , نقطه‌نظر معمار پیشکسوت کشور تهیه , دوران معا صر , شاخص زیبایی شنا سی , مطالعه آراء , شاخص کیفی به‌منظور سنجش کیفیت , مسائل , شامل , پایه , ضروری شاخص , تحقیق توصیفی-تحلیلی محتوا شکل , آمار توصیف , خراج , شهرسازی , مطالعه آثار , توجه , معمار معاصر , تحلیل دیدگاه معمارانی , مشکلات معمار , مباحث معمار , ابزار , بی‌هویت , هنر معمار صرف
تلفیقی	معماری معاصر, ایران, بومی گرایی, اسلامی, اصل, آموزش معماری, زبان فارسی, دوره پهلوی, هویّت, طراحی معماری, کاشانه, معماری ایرانی, آثار هنری, سنّت گرایی, نوشتار, ادراک, اسلام, صدرالدین شیرازی, محمدبن ابراهیم, معماری اسلامی, فعالیت فرهنگی, تاریخ خوش‌نویسی, شعر, مطالعه پارامتری, ادبیات فارسی, تخیّل, معماری پارامتریک, پدیدارشناسی, نمای ساختمان, حکمت متعالیه, کرج, ساختمان مسکونی, علم, موزه, مدرن گری, دانشگاه اصفهان, آرایه, بنای تاریخی, قم, معماری, ساخت پیمانه‌ای, درکه (تهران), آب, طبیعت, خلّاقیت, خانه سازی, فرهنگ, طراحی, خلاقیت, نما (معماری), مدرنیسم, : عرفان, هنر, کیفیت زندگی, اصالت وجود, عرفان, معماری سنتی, نوآوری, هنر معاصر, هخامنشی, مشهد, سبک هنری, تهران (شهر), مدرنیته, انسان گرایی, مدرنیّت, مراکز فرهنگی, مجموعه فرهنگی,

	آمریکا, قم (شـهر), هنر اسـلامی, تمدن, آسـیا, دنباله فیوناتچی, اروپا, فضـای آموزشـی, برنامه‌ریزی شـهری, میراث فرهنگی, سرای محله, مجموعه مسکونی, توسعه شهری, مجتمع تجاری, تاریخ, باشگاه معماران, کاخ, مشارکت مدنی, مجتمع مسکونی, مسکن, کاشان, تکنیک, نقوش هندسی, ستون(سازه‌ای), هندسه, دانشکده معماری, زیبایی شناسی, پوشش گیاهی, معماری نوین, برنامه‌ریزی پایدار, انتزاع, سیاست, اموزش معماری, مجتمع فرهنگی, فناوری, اقلیم

جدول ٤-٥-نمونه کلیدواژه‌های استخراج شده براساس روش‌های مختلف در حوزه علوم و علوم کاربردی

عنوان: بررسی ترکیبات شیمیایی اسانس بذر دو گیاه دارویی بومی استان یزد (زیره سیاه، زیره سبز)	
حوزه موضوعی: علوم و علوم کاربردی	
چکیده: ایران یکی از غنی ترین منابع گیاهان دارویی جهان است که دارای تنوع بالای شرایط زیستگاهی برای انواع این گیاهان می باشد. از مهمترین گیاهان دارویی خانواده چتریان درکشورمان، زیره سیاه، زیره سبز می باشد که در مناطق مختلفی از استان یزد رویشگاه های طبیعی آنها به چشم می خورد. از اینرو این دو گیاه را انتخاب کرده و به بررسی ترکیبات شیمیایی اسانس بذر پرداخته شد. بذور جمع آوری شده از مناطق بومی استان یزد، به روش تقطیر با آب اسانس گیری شدند. جدا سازی و شناسایی ترکیبات تشکیل دهنده با استفاده از روش کروماتوگرافی گازی متصل به طیف سنج جرمی (GC/MS) انجام شد. نتایج آنالیز اسانس بذر دو گیاه نشان داد که مهترین ترکیبات زیره سیاه گاماترپینن وکومین آلدئید و پاراسیمن ، زیره سبز پروپانل و بنزن متانول و۱- فنیل ۱- بوتانل و گاماترپینن، بودند.	
کلیدواژه‌ها:	
RAKE	زیره سبز,ترکیبات شیمیایی-اسانس,زیره سیاه,ترکیبات شیمیایی,گیاهان دارویی,بذر,گیاهان,مه گیاه دارویی خانواده چتر درکشورمان,مه ترکیبات زیره سیاه گاماترپینن وکومین,غنی منابع گیاه دارویی جهان,ترکیبات شیمیایی اسانس بذر,شناسایی ترکیبات تشکیل دهنده,استان یزد رویشگاه طبیعی,نتایج آنالیز اسانس بذر,مناطق بومی استان یزد,آب اسانس
YAKE	منابع گیاهان , شرایط زیستگاهی , انواع گیاهان , منابع , انواع , گیاهان دارویی جهان , غنی , تنوع , شرایط , دارای تنوع , غنی منابع , تنوع شرایط , زیستگاهی انواع , زیستگاهی , ایران , منابع گیاهان دارویی
TextRank	ان یزد رویشگاه , زیره سیاه , ایران , طبیعی , مه گیاه دارویی خانواده چتر درکشورمان , پاراسیمن , آب اسانس , تقطیر , شرایط زیستگاه , مناطق بومی ان یزد , انتخاب , زیره سبز پروپانل , بذور جمع , ترکیبات شیمیایی اسانس بذر , بنزن متانول و۱- فنیل ۱- بوتانل , نتایج آنالیز اسانس بذر , چشم , زیره سبز , مه ترکیبات زیره سیاه گاماترپینن وکومین آلدئید , طیف سنج جرم , منابع گیاه دارویی جهان
تلفیقی	گیاه دارویی, فرازا, اجزای تشکیل‌دهنده عملکرد, ترکیبات اسانس, قابلیت جذب آب, آنالیز چند متغیره, خواص پاد زیوه‌ای, در آزمایشگاه, مراتع چهارباغ, روزمارینوس افیسینالیس, فاکتورهای زیست‌محیطی, تجزیه و تحلیل بافت

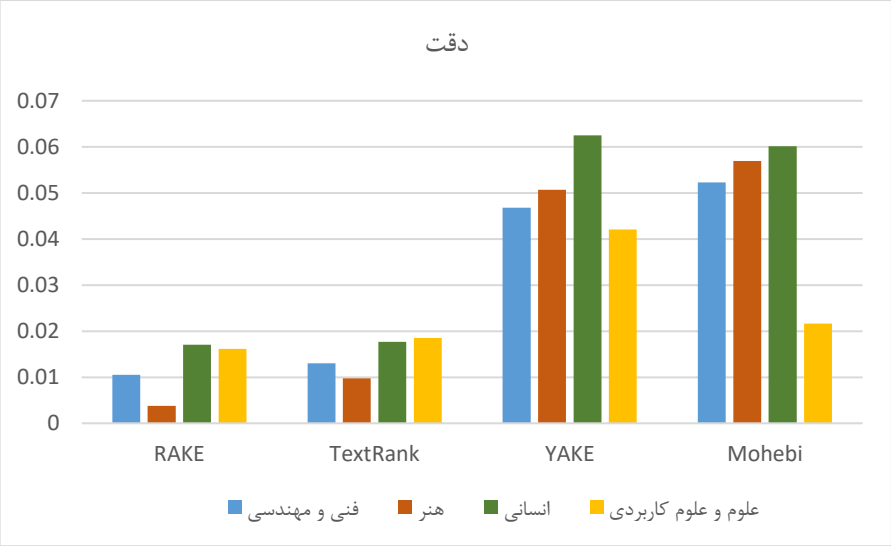
مقدار دقت و بازخوانی هر سند و نیز میانگین آنها نیز محاسبه شده است. بازخوانی براساس رابطه زیر حساب می‌شود:

$$Recall = \frac{\#\{recommended\ keywords\} \cap \{original\ keywords\}}{\#\{original\ keywords\}} \quad (٤-١)$$

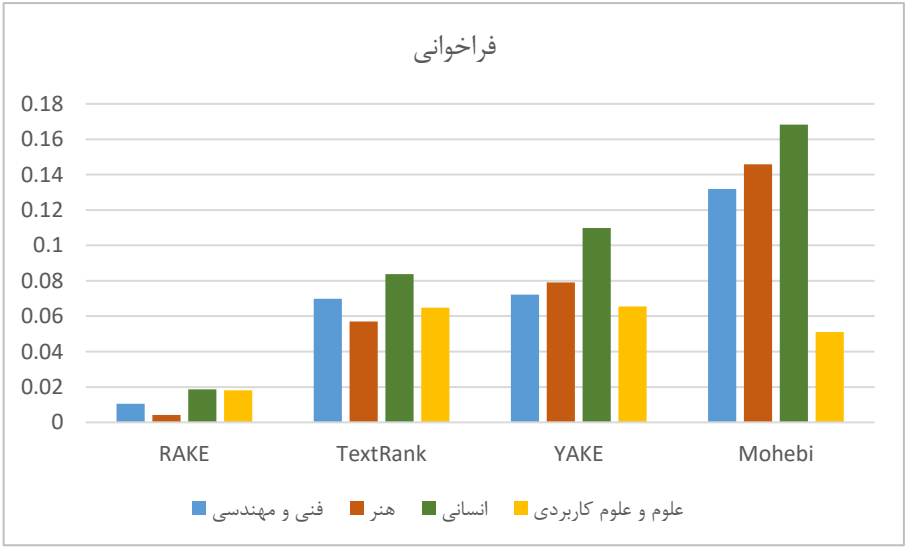
و دقت براساس روش زیر محاسبه می‌شود:

$$Precision = \frac{\#\{recommended\ keywords\} \cap \{original\ keywords\}}{\#\{recommended\ keywords\}} \quad (۲-٤)$$

که منظور از *original keywords*، همان کلیدواژه‌های اولیه سند است که توسط کاربر ارائه شده است. مقدار دقت و بازخوانی هر روش در چهار حوزه موضوعی علوم انسانی، فنی و مهندسی، هنر، و علوم و علوم کاربردی، در شکل ۴-۲ و شکل ۴-۳ نمایش داده شده است. این مقادیر براساس کلیدواژه‌های اولیه موجود برای هر سند که متشکل از کلیدواژه‌های نویسنده و نمایه‌ساز است، محاسبه شده است.



شکل ٤-۲- دقت روش‌های استخراج کلیدواژه



شکل ٤-۳- فراخوانی روش‌های استخراج کلیدواژه

٤- ٤. یکپارچه‌سازی نتایج

قدم بعدی یکپارچه کردن نتایج چهار الگوریتم است. در نهایت نیاز است که پس از یکپارچه کردن نتایج، هرسند ۲۰ تا ۳۰ کلیدواژه داشته باشد. نحوه یکپارچه کردن نتایج چند الگوریتم استخراج کلیدواژه در تعدادی از پژوهش‌های پیشین مورد توجه قرار گرفته است (Pay and Lucci 2017). در پژوهش حاضر، ابتدا کلیه کلیدواژه‌های اصلی سند که در ابتدا در داده‌های اولیه وجود داشته، در فهرست کلیدواژه قرار می‌گیرند. سپس به آن فهرست کلیدواژه‌هایی اضافه می‌شوند که حداقل دو روش آنها را استخراج کرده باشند. پس از آن در صورتیکه همچنان تعداد کلیدواژه‌ها به ۲۵ نرسیده باشد، براساس یک روش رتبه‌بندی، کلیدواژه‌های باقی مانده امتیازدهی می‌شوند و کلیدواژه‌های دارای بیشترین امتیاز به فهرست اضافه می‌شوند. روش رتبه‌بندی که در نهایت استفاده می‌شود براساس روش رتبه‌بندی تاپسیس[۱۶] است. روش تاپسیس یک روش تصمیم‌گیری و رتبه‌بندی گزینه‌های تصمیم است که توسط هوانگ و یون (۱۹۸۱) پیشنهاد شده و در بسیاری از مسایل تصمیم‌گیری و رتبه‌بندی کاربرد دارد (Hwang and Yoon 1981).

برای تعیین رتبه کلیدواژه‌های باقی مانده با استفاده از روش تاپسیس ابتدا باید معیارهای تصمیم‌گیری را مشخص کنیم. برای هر کلیدواژه k_i از سند d_j، از معیارهای زیر استفاده می‌کنیم:

- امتیاز نرمال شده کلیدواژه $\boldsymbol{k_i}$ در روش m، $score_m(k_i)$
- دقت روش m روی سندی d_j، $precision_m(d_j)$
- بازخوانی روش m روی سندی d_j، $recall_m(d_j)$
- میانگین دقت روش m، P_m
- میانگین فراخوانی روش m، R_m

که در آن، $m \in \{RAKE, YAKE, TextRank, COMB\}$ و منظور از $COMB$ همان روش تلفیقی است. بنابراین ماتریس معیارها و گزینه‌های تاپسیس به صورت زیر است:

$$\begin{matrix} & score_m(k_i) & precision_m(d_j) & recall_m(d_j) & P_m & R_m \\ k_1 \\ \vdots \\ k_n \end{matrix} \begin{bmatrix} & & & & & \end{bmatrix} \quad (۳-۴)$$

[۱۶] TOPSIS

که وزن معیارها نیز یکسان در نظر گرفته شده است.

پس از به‌کارگیری این روش، برای هر سند، فهرست ۲۵ تا ۳۰ تایی از کلیدواژه‌ها برای هر سند به‌دست می‌آید. روش انجام کار به صورت خلاصه در شبه کد زیر آمده است.

شبه کد یکپارچه کردن نتایج

برای هر سند d_j، کلیه کلیدواژه‌های به‌د ست آمده از همه روش‌ها به‌علاوه کلیدواژه‌های ا صلی را در مجموعه W_j قرار بده.

برای هر سند d_j، فهرست کلیدواژه‌های نهایی (F_j) را در ابتدا مساوی مجموعه تهی قرار بده و قدم‌های زیر را تکرار کن:

- کلیدواژه‌های اصلی اولیه هر سند را به F_j اضافه کن و از فهرست W_j حذف کن.
- در بین کلیدواژه‌های باقی مانده در W_j، کلیدواژه‌هایی را که حداقل دو روش آنها را پیشنهاد کرده، به مجموعه F_j اضافه کن و از مجموعه W_j حذف کن.
- اگر تعداد کلیدواژه‌های مجموعه F_j، ۲۵ یا بیشتر شد، توقف کن. در غیر اینصورت قدم‌های زیر را انجام بده:
  - براساس ۵ معیار روش تاپسیس، کلیدواژه‌های باقی مانده در مجموعه W_j را رتبه‌بندی کن.
  - کلیدواژه‌های دارای بیشترین رتبه را تا جایی به مجموعه F_j اضافه کن که تعداد کلیدواژه‌های آن به ۲۵ برسد.

در جدول ۴-۶ ، برای چند نمونه سند کلیدواژه‌ها پس از یکپارچه‌سازی نتایج با استفاده از روش پیشنهادی نمایش داده شده است.

جدول ٤-٦-کلیدواژه‌های پیشنهادی بعد از یکپارچه‌سازی نتایج الگوریتم‌ها برای چند نمونه سند

عنوان	حوزه موضوعی	کلیدواژه‌ها
بررسی و تحلیل نقوش نمادین و مذهبی مهرهای ساسانی	هنر	مهر,ساسانیان,نگرش,تمدن,دین,رمز,نقشمایه,سبک شناسی مُهرهای ساسانی,دین زردشت,مُهرهای ساسانی نمادگرایی,اسطوره,جایگاه اجتماع مالکین,جایگاه اجتماع افراد مرتبط,منسوجات,نقش‌های تزئینی,نقوش مُهرهای ساسانی,محراب,مانده,روزگار پی‌برد,مذهبی مُهرهای دوره,آثار هنر دوران ساسان,بیانگر ارتباط نقوش,تحقیق,مربوط
بررسی پارامتریک اثر مشخصات شمع در گودبرداری‌های عمیق مهار شده با شمع‌های بتنی درجا	فنی و مهندسی	اتصال گیردار,گودبرداری مهارشده,شمع بتنی درجا,گودبرداری,روش پارامتری,شمع بتنی,شمع(در ساختمان),جابجایی افق گود کاهش,تامین,اجزا محدود گود,تحلیل مدل اجزا محدود پایدار گودبرداری,موجبات جابجایی جانبی,خصوص نرم افزار اجزا محدود,دلیل هزینه محاسبات,تغییرات پارامترهای موثر,ارتفاع گود,قابلیت تشخیص خطا,پارامترهای موثر,کارگیری شمع,رکورد زلزله تفت,قطر کوچک,شمع افزایش,گود,نتایج حاصل
بررسی رابطه بین مسئولیت اجتماعی شرکت و بازده سرمایه‌گذاری در شرکت‌های پذیرفته شده بورس اوراق بهادار تهران	علوم انسانی	کارایی سرمایه گذاری,شرکت ثبت شده در بورس,بورس اوراق بهادار تهران,بازده سرمایه‌گذاری,مسئولیت اجتماعی,مسئولیت اجتماعی شرکتی,رگرسیون چندگانه خط,افشا مسئولیت اجتماع,بورس اوراق بهادار تهران رابطه مثبت,نرم افزارExcel محاسبه,شرکت سرمایه‌گذاری, پذیرفته,بورس اوراق,هدف,نمونه آمار تحقیق ۹۹ شرکت,نرم افزارEviews9,سال,نتایج,مسئولیت اجتماع,معنی,تحلیل قرار,هاسمن,لیمر,۱۳۸۹ لغایت ۱۳۹۳
بررسی احتمال افزایش بردباری گیاه گندم به تنش شوری به وسیله‌ی اسید آسکوربیک	علوم کاربردی	تنش اسمزی,پارامترهای رشدی,آسکوربیک اسید,شوری,تحمل نمک,اسیدآسکوربیک,گونه‌های اکسیژن فعال,مقاومت به فاکتورهای آسیب‌رسان,آنتی اکسیدان,آنتی‌اکسیدان,گندم,مقابله,مقاومت به خشکی,کاهش اثرات نامطلوب تنش شوری,پارامترهای رشد,تحمل (خواص زیستی),ارقام,فراوان محلول,فسفر,کاربرد اسیدآسکوربیک,کاهش محتوا پرولین,کاتالاز,اطلاعات نسبتاً,آنتی اکسیدان‌های مهم,نتیجه شوری افزایش,کشت گندم

٤- ٥. ارزیابی و رتبه‌بندی کلیدواژه‌ها توسط متخصص

برای ارزیابی و رتبه‌بندی نهایی کلیدواژه‌های به‌دست آمده برای هر سند پس از یکپارچه‌سازی، توسط متخصص نمایه‌ساز، سامانه ارزیابی طراحی شده است. نمایی از صفحات مختلف این سامانه در شکل‌های شکل ۴-۴ تا ۴-۸ نمایش داده شده است.

فرایند ارزیابی شامل دو مرحله است:

- مرحله ارزیابی اولیه
- مرحله بازبینی نهایی

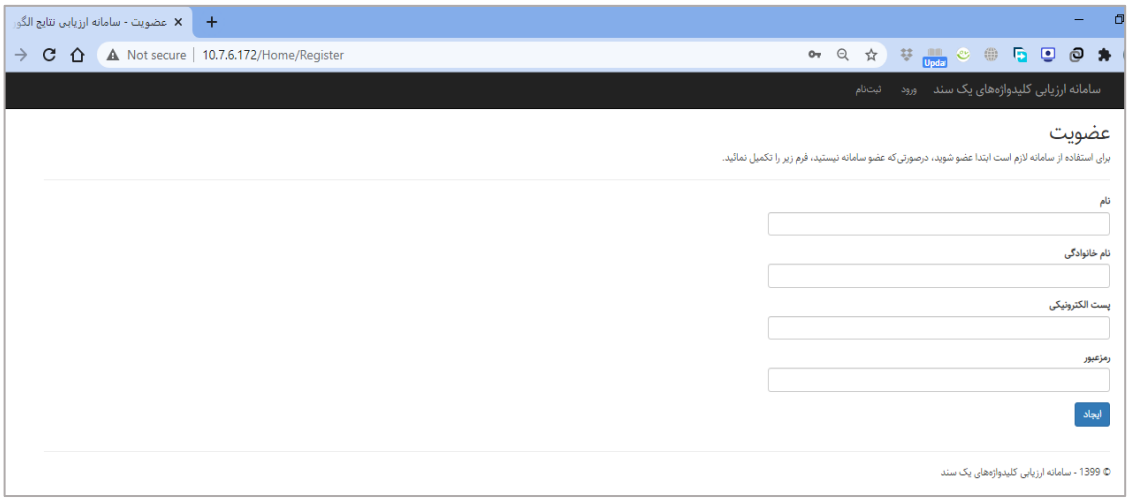
در هر مرحله، ۴ متخصص نمایه ساز مشارکت دارند که از هر حوزه موضوعی، به تعداد مساوی به هر یک، اسناد اختصاص داده شده است. بنابراین هر سند را تنها یک متخصص ارزیابی و بازبینی می‌کند.

در ادامه توضیحات بیشتری درباره نحوه ارزیابی در هر مرحله و نیز چند نمونه از داده‌های ارزیابی شده آورده شده است.

٤- ٥- ١. نحوه ارزیابی کلیدواژه‌ها

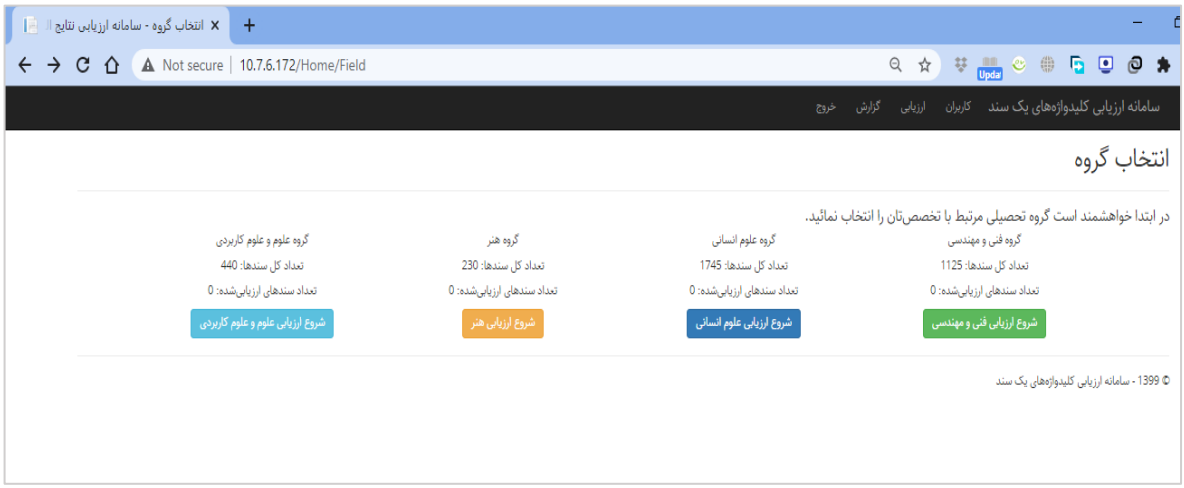
در مرحله ارزیابی اولیه، برای ارزیابی هر سند، کلیدواژه‌های پیشنهادی، به همراه چکیده و عنوان سند نمایش داده می‌شود. فرد متخصص برای هر کلیدواژه مشخص می‌کند که آیا مرتبط است یا خیر و در صورت مرتبط بودن امتیازی بین ۱ تا ۳ را برای آن در نظر می‌گیرد که ۳ بیانگر بیشترین ارتباط است. همچنین می‌تواند در صورت نیاز کلیدواژه‌های جدیدی را نیز پیشنهاد دهد. در نهایت کلیدواژه‌های مناسب برای هر سند به همراه امتیاز هر یک از لحاظ میزان ارتباط، مشخص شود.

الف) صفحه اول سامانه ارزیابی

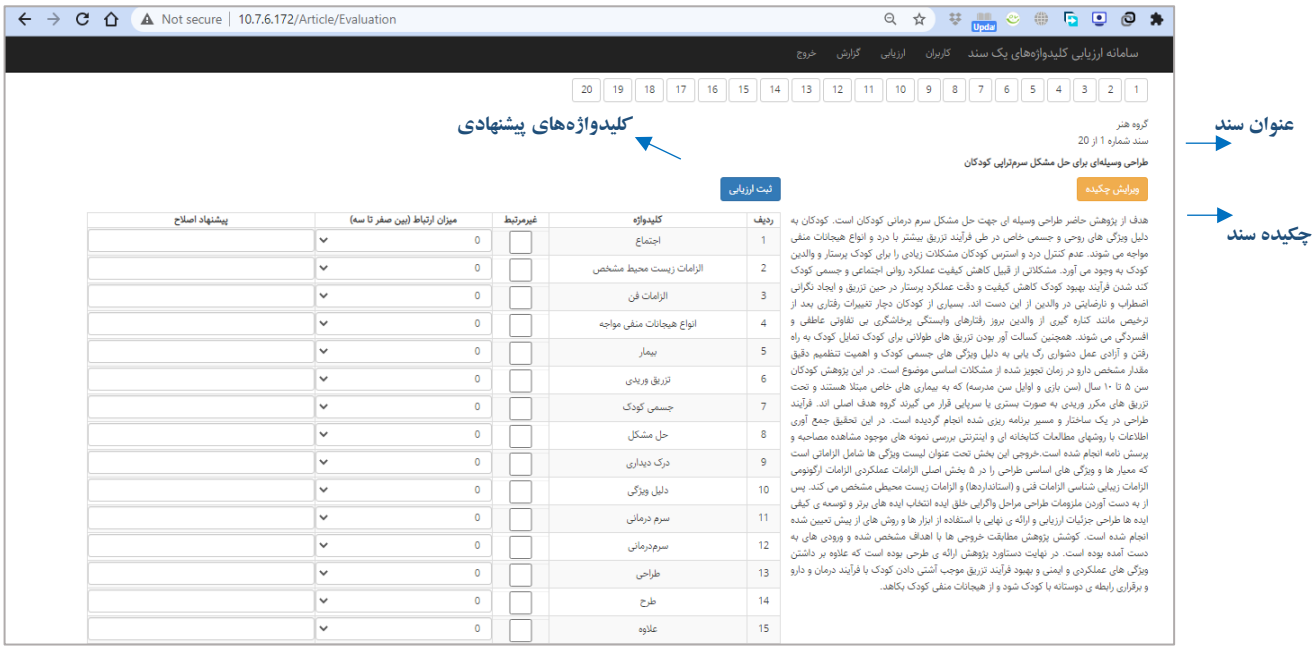


ب) صفحه ثبت‌نام/ورود کاربر

شکل ٤-٤- صفحات اول و ورود به سامانه ارزیابی کلیدواژه‌ها براساس نظرات متخصص

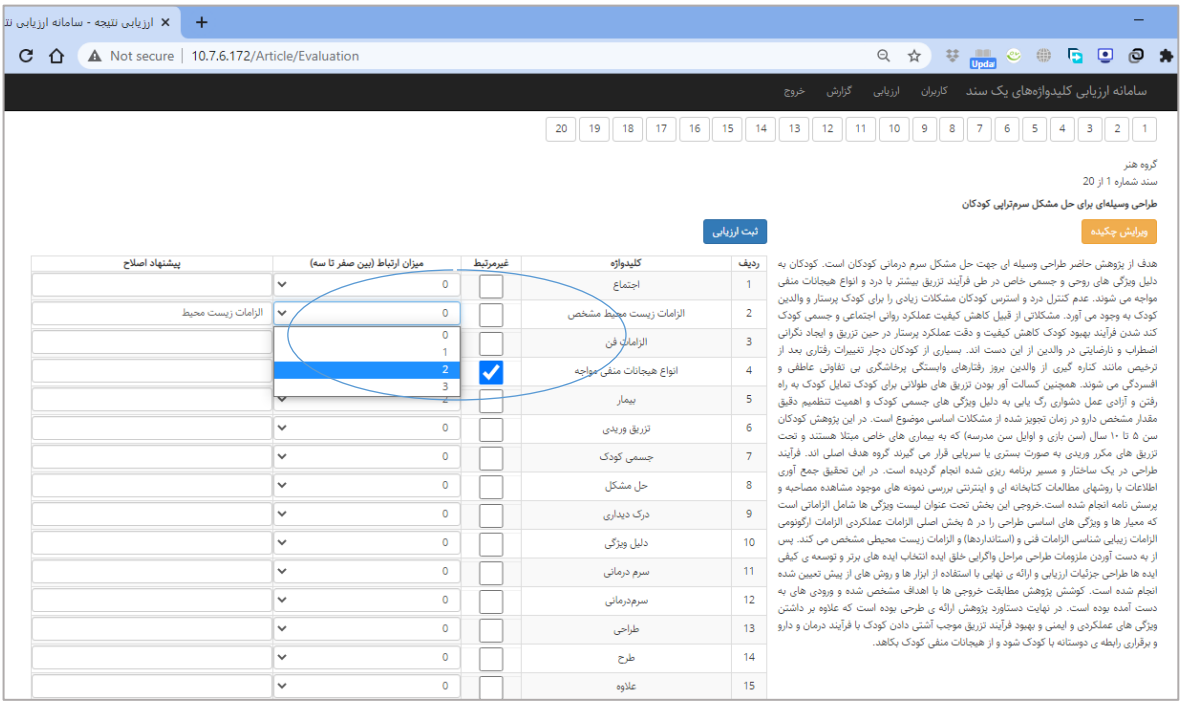


الف) صفحه انتخاب حوزه موضوعی

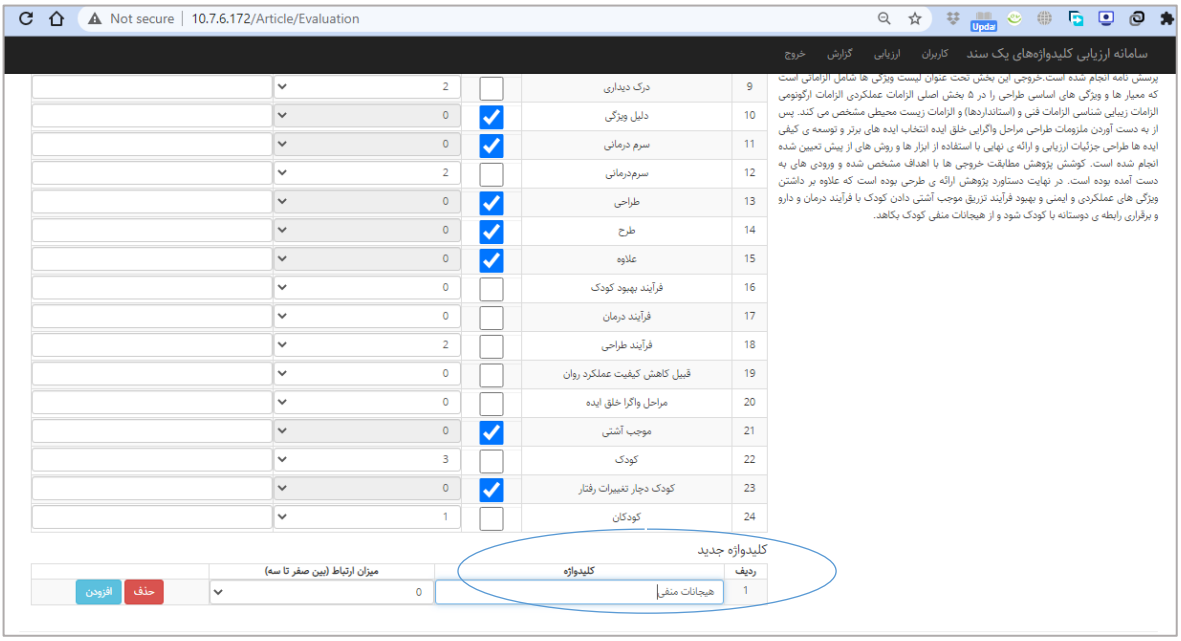


ب) صفحه ارزیابی کلیدواژه‌ها برای هر سند

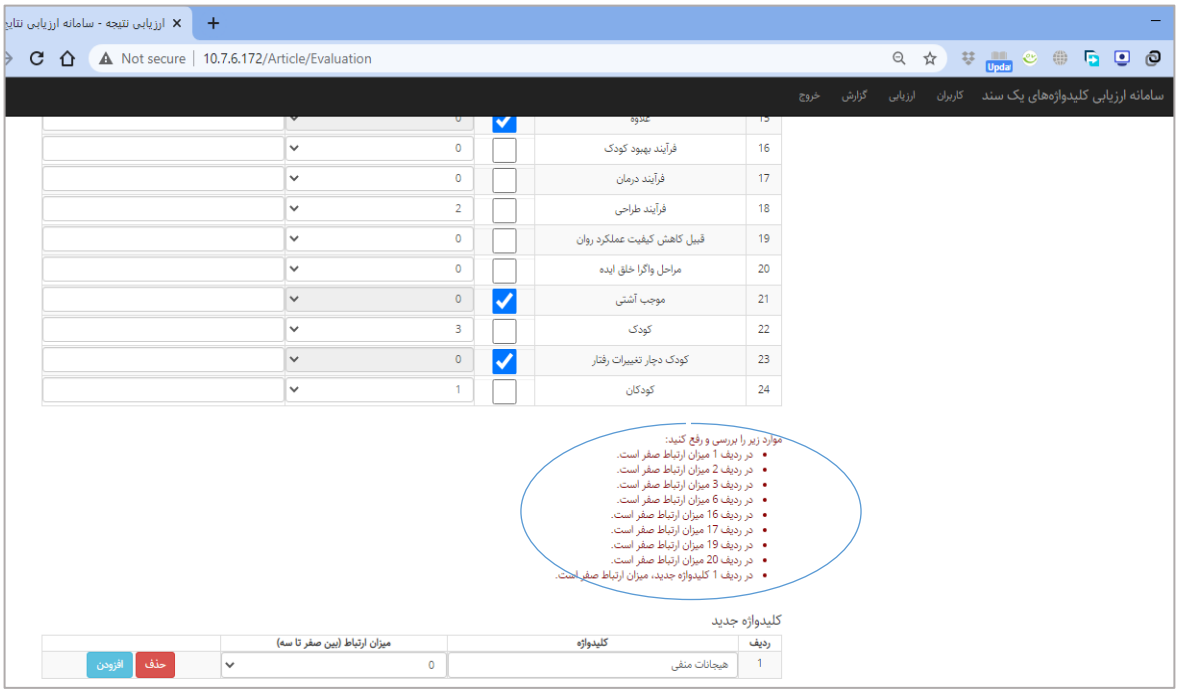
شکل ٤-٥- صفحات انتخاب حوزه موضوعی و ارزیابی کلیدواژه‌ها



شکل ٤-٦- امتیازدهی برای کلیدواژه مرتبط که برای هر یک می توان امتیاز ۱ تا ۳ را براساس میزان ارتباط تعیین کرد



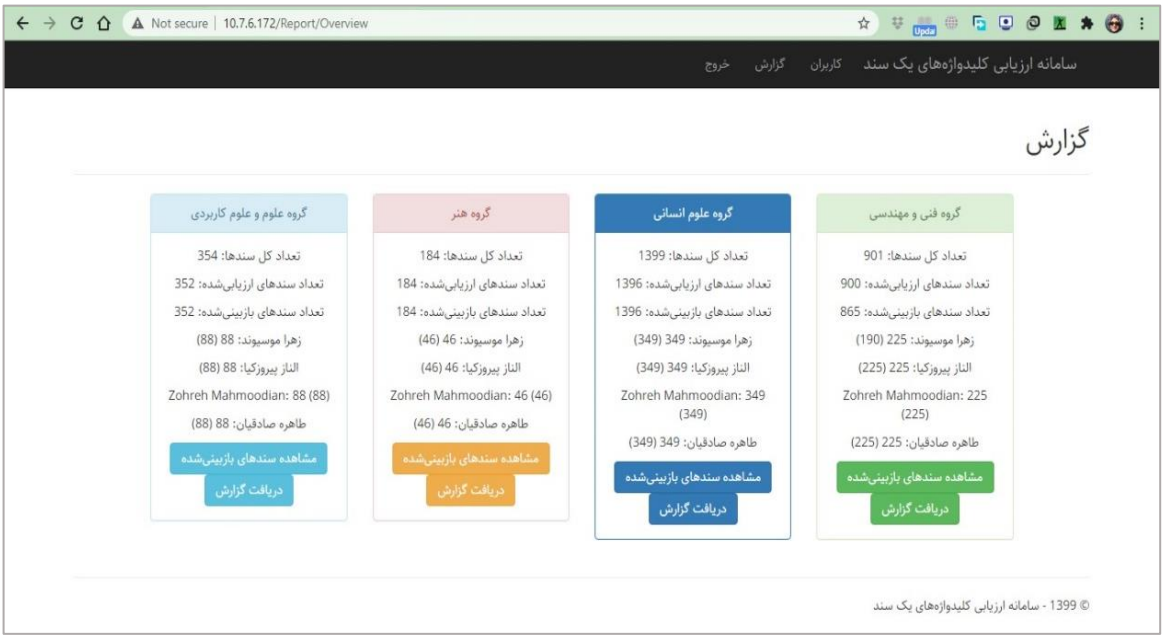
شکل ٤-٧-افزودن کلیدواژه جدید، که در انتهای جدول کلیدواژه‌های پیشنهادی، امکان افزودن کلیدژه جدید هم وجود دارد



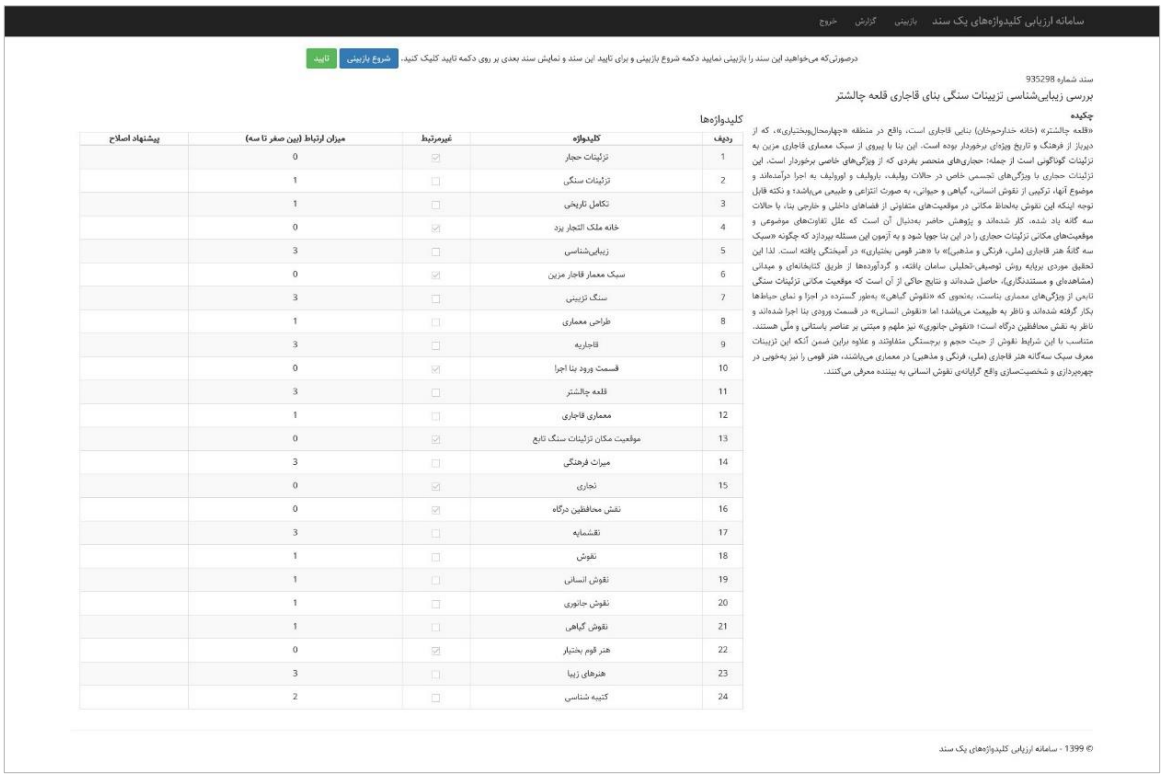
شکل ٤-٨- اخطار سامانه به کاربر در صورتیکه به درستی جدول را تکمیل نکرده باشدیا هنگام افزودن کلیدواژه جدید، امتیاز آن را مشخص نکرده باشد

در مرحله بازبینی نهایی، مجددا ا سنادی که ارزیابی شده‌اند برای متخ صص بارگذاری می شوند تا در صورت نیاز اصلاحات لازم روی آنها انجام شود. نحوه انجام این کار در شکل ۴-۹ آمده است.

در نهایت خروجی مرحله بازبینی نهایی، اسناد به همراه کلیدواژه‌های ارزیابی شده برای آنهاست. علت ارزیابی دو مرحله‌ای اینست که نمایه‌ساز ممکن است پس از بازبینی مدارک بیشتر مدارک، بخواهد با توجه به شناختی که پیدا کرده، ا صلاحاتی در کلیدواژه‌های پیشنهادی یا جدید انجام دهد. بنابراین در مرحله بازبینی نهایی، این امکان وجود دارد که نمایه‌ساز مجددا اسناد ارزیابی شده را ببیند تا در صورت نیاز آنها را اصلاح کند.



الف) صفحه اول برای مرحله بازبینی مجدد



ب) صفحه بازبینی مجدد

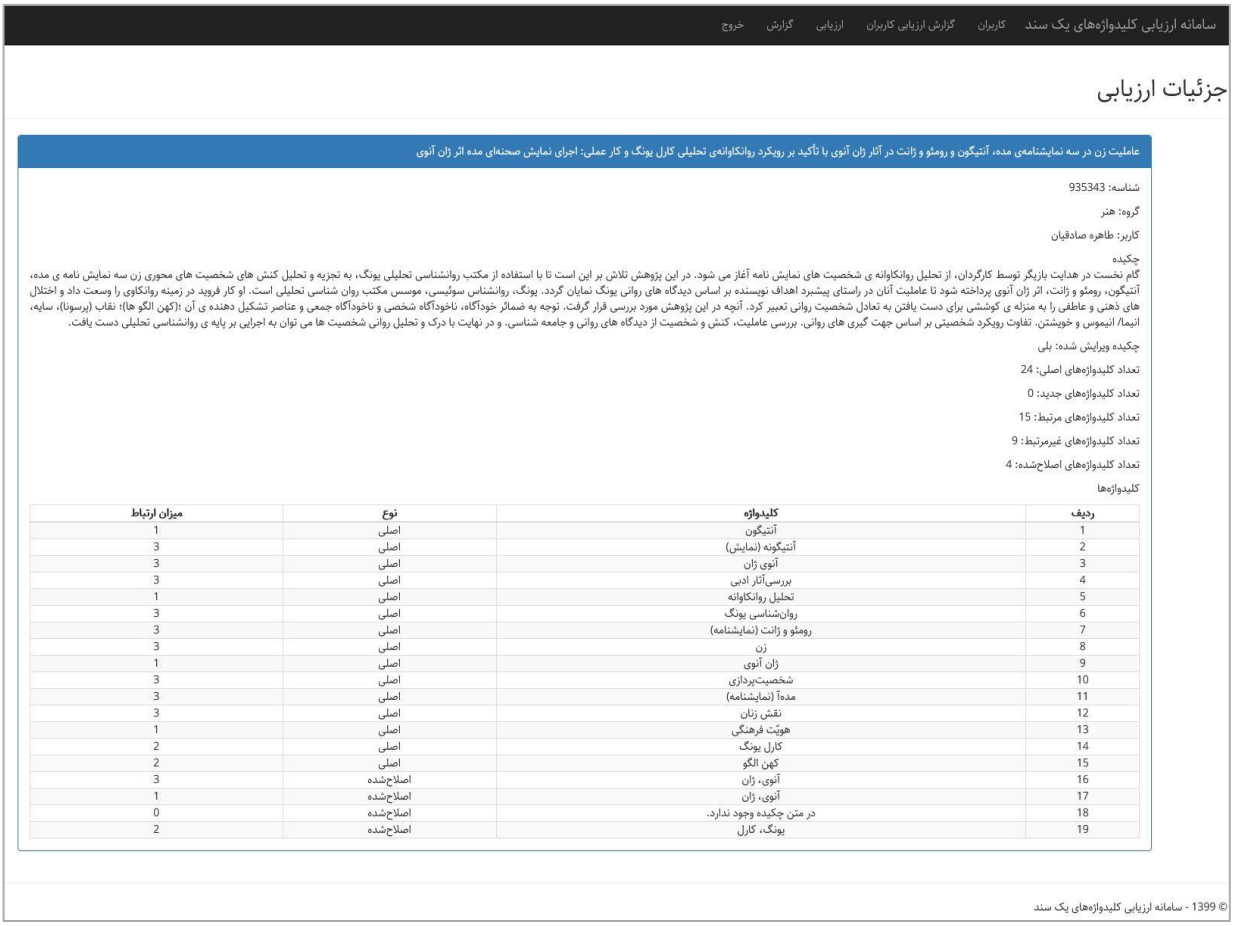
شکل ٤-۹- نحوه انجام بازبینی نهایی

٤- ٥- ۲. نمونه‌ای از داده‌های نهایی

در این بخش نمونه‌ای از داده‌های نهایی پس از دو مرحله ارزیابی نمایش داده شده است. در این مرحله موارد زیر برای هر سند مشخص می‌شود:

- کدام کلیدواژه‌ها مرتبط بودند
- کدام غیرمرتبط بودند
- امتیاز هر یک از کلیدواژه‌ای مرتبط چقدر است
- چه واژه‌های جدیدی با چه امتیازی اضافه شده‌اند
- کدام کلیدواژه‌ها اصلاح شده‌اند.

در شکل ۴-۱۰، نمونه‌ای از گزارش سامانه برای هر سند نمایش داده شده است و در جدول‌های ۴-۷ تا ۴-۱۰ چند نمونه از داده‌های نهایی به همراه امتیاز هر کلیدواژه آمده است.



شکل ٤-۱۰- نمونه گزارش سامانه برای هر سند بازبینی شده

جدول ٤-۷- نمونه‌ای از داده‌های نهایی علوم انسانی

عنوان: پیش‌بینی قیمت سهام با روش‌های براونی و خود رگرسیون میانگین متحرک انباشته (ARDL)	موضوع: علوم انسانی

چکیده:

بازارهای مالی توسعه‌یافته یکی از خصیصه‌های اقتصادهای پیشرفته امروزی است. فعالان بازارهای مالی به هدف کسب سود اقدام به خرید دارایی‌های مالی می‌کنند. پیش‌بینی قیمت‌های آتی نقش مهمی‌در تصمیمات خرید آن‌ها دارد. عموماًروش‌های اقتصادسنجی سری زمانی تک متغیره برای پیش‌بینی قیمت‌ها بکار می‌روند که در آن تلاش می‌شود متغیرهای مالی را تنها با استفاده از اطلاعات موجود در مقادیر گذشته‌شان و مقادیر فعلی و گذشته جزء اختلال، پیش‌بینی و الگوسازی کرد. موضوع مانایی و تفاضل گیری در این روش‌ها جزئی از فرآیند پیش‌بینی است. با توجه به‌اینکه با عمل تفاضل گیری بخش مهمی‌از اطلاعات متغیرها از دست می‌رود.از طرفی مدل‌های گشت تصادفی در پیش‌بینی قیمت سهام به‌صورت قابل‌توجهی بکار گرفته می‌شوند. ازنظر مدل گشت تصادفی، حرکت قیمت‌ها در یک بازارکارا مستقل از حرکات قبلی است و یا به‌عبارت‌دیگر، بین حرکات پیاپی قیمت‌ها هیچ‌گونه وابستگی مشاهده نمی‌شود. البته باید توجه داشت که تغییر قیمت سهام ناشی از مجموعه‌ای از اطلاعات است که اطلاعات تاریخی جزئی از آن محسوب می‌گردد. لذا در تحقیق حاضر سعی شده است تا قیمت سهام ۵۰ شرکت فعال در بورس اوراق بهادار تهران طی بازهء ۱۳۹۱-۱۳۹۲ با دو روش ARIMA و براونی پیش‌بینی گردد. سپس با مقایسه بیرون از نمونه با استفاده از معیارهای MSE وMAE قدرت پیش‌بینی این دو روش مورد آزمون قرار گرفت. نتایج هر دو معیار مشابه بوده و هر دو برتری پیش‌بینی‌های روش براونی را تأیید می‌کنند. همچنین نتایج تحلیل یکجای قیمت ۵۰ نماد معاملاتی تأییدکننده این نتیجه هستند.

آمار کلیدواژه‌ها

کلیدواژه‌های ارزیابی شده	کلیدواژه‌های مرتبط	کلیدواژه‌های غیرمرتبط	کلیدواژه‌های اضافه شده
۲۳	۱۳	۱۰	۰

کلیدواژه‌ها

کلیدواژه	امتیاز	کلیدواژه	امتیاز	کلیدواژه	امتیاز
ARIMA	۱	پیش‌بینی قیمت سهام	۱	حرکت براونی	۳
اطلاعات تاریخ جزئی	۱	بازار مالی	۲	روش ARIMA	۱
بازار سهام	۳	بازار مالی توسعه‌یافته	۱	قیمت آتی	۳
پیش‌بینی قیمت	۳	بورس اوراق بهادار تهران	۳		

جدول ٤-٨- نمونه‌ای از داده‌های نهایی علوم و علوم کاربردی

عنوان: اثر نانوذره نقره استخراج شده از گیاه پولیکاریا بر روی تشکیل آمیلوئید در آلفا-لاکتالبومین و عمل چاپرونی بتا-کازئین	موضوع: علوم و علوم کاربردی

چکیده:

تجعات غیرطبیعی پروتئین به شکل فیبریل های آمیلوئیدی علت اصلی بسیاری از بیماری های نورودژنراتیو مانند آلزایمر، پارکینسون،بیماری کروتزفیلدجاکوب،دیابت نوع دو و دیگر بیماریهاست. چاپرون های مولکولی، پروتئین هایی هستند که از تا خوردن ناقص پروتئین و بنابراین تجمع آن جلوگیری می کنند. در این تحقیق اثر نانوذرات نقره استخراج شده از گیاه Pulicaria undulate L بر روی تشکیل آمیلوئید در آلفا-لاکتالبومین و همچنین خاصیت چاپرونی بتا-کازئین مورد ارزیابی قرار گرفت. اثر نانوذرات نقره و چاپرون بتا-کازئین بر روی تجمعات آمیلوئیدی آلفا-لاکتالبومین با استفاده از طیف سنجی نور مرئی، فلورسانس تیوفلاوین تی، فلورسانس تریپتوفان ذاتی و ANS، طیف سنجی دو رنگ نمایی دورانی و الکتروفورز ژل پلی آکریل آمید و میکروسکوپ الکترونی گذاره TEM ارزیابی گردید. داده های حاصل از فلورسانس تیوفلاوین، نشرذاتی تریپتوفان، ANS و طیف سنجی نور مرئی نشان دادند که افزودن نانوذرات نقره باعث کاهش معنی داری در شدت نشر و جذب نمونه آلفا-لاکتالبومین احیاء شده می شود که بیانگر اثر مثبت نانوذرات نقره در مهار تجمعات پروتئینی می باشند، همچنین در حضور چاپرون بتا-کازئین در مقایسه با عدم حضور آن کاهش نشر بیشتری مشاهده شد و به میزان بیشتری تجمعات پروتئینی را مهار کردند، اندازه گیری فلورسانس نمونه های حاوی بتا-کازئین و نانوذرات نقره نشان دادند نانوذرات نقره تغییرات چشمگیری در ماکزیمم فلورسانس بتا-کازئین بوجود نمی‌آورند بنابراین اثری روی عمل چاپرونی بتا-کازئین ندارند. نتایج حاصل از TEM نیز نشان داد در حضور نانوذرات نقره از فیبریل های آمیلوئیدی آلفا-لاکتالبومین احیاء شده کاسته شده و تجمعات آمورف تشکیل شده اند، همچنین با افزودن بتا-کازئین و نانوذرات نقره نیز تجمعات آمورف کمتری مشاهده شد. طیف سنجی دو رنگ نمایی حلقوی نشان داد که نانوذرات نقره باعث حفظ پایداری ساختار دوم و سوم پروتئین شده و درحضور بتا-کازئین این پایداری بیشتر می شود همچنین یافته های حاصل از SDS-PAGE نشان داد که در حضور نانوذرات نقره، تحرک الکتروفورزی آلفا-لاکتالبومین احیاء شده افزایش و جرم مولکولی آن کاهش می یابد و در حضور بتا-کازئین تحرک الکتروفورزی به میزان بیشتری افزایش یافت. نتایج ما نشان داد که نانوذرات نقره تأثیری روی عمل چاپرونی بتا-کازئین ندارند و هرکدام از این دو عامل اثر مهاری خود را به‌صورت مستقل بر روی پروتئین هدف اعمال می کنند.

آمار کلیدواژه‌ها

کلیدواژه‌های ارزیابی شده	کلیدواژه‌های مرتبط	کلیدواژه‌های غیرمرتبط	کلیدواژه‌های اضافه شده
۲۶	۱۷	۹	۰

کلیدواژه‌ها

کلیدواژه	امتیاز	کلیدواژه	امتیاز	کلیدواژه	امتیاز
آلفا	۱	زیست‌سنتز	۲	نانوذرات نقره	۱
آلفالاکتالبومین	۳	شپرون (پروتئین)	۳	نقره	۳
آمیلوئید	۳	طیف سنج نور مرئی	۱	کازئین	۱
بتا	۱	عصب‌تباهی	۳	کک‌کش (سرده)	۳
بتاکازئین	۳	لاکتالبومین	۳		
تجعات غیرطبیعی پروتئین	۱	نانوذرات	۱		

جدول ٤-۹- نمونه‌ای از داده‌های نهایی فنی و مهندسی

عنوان: شبیه‌سازی عددی جریان سیال و انتقال حرارت در لوله مارپیچی شکل یک مبدل حرارتی با مقاطع مختلف	موضوع: فنی و مهندسی

چکیده:

مبدل های حرارتی تقریباً پرکاربردترین عضو در فرآیندهای شیمیایی اند و می توان آن ها را در بیشتر واحدهای صنعتی ملاحظه کرد. آنها وسایلی هستند که امکان انتقال انرژی گرمایی بین دو یا چند سیال در دماهای مختلف را فراهم می کنند این عملیات می تواند بین مایع-مایع، گاز-گاز و یا گاز-مایع انجام شود مبدل های حرارتی به منظور خنک کردن سیال گرم و یا گرم کردن سیال با دمای پایین تر و یا هردو مورد استفاده قرار می گیرند. مبدل های حرارتی در محدوده وسیعی از کاربردها استفاده می شوند این کاربردها شامل نیروگاه-ها پالایشگاه ها، صنایع پتروشیمی، صنایع ساخت و تولید، صنایع فرآیندی، صنایع غذایی و دارویی، صنایع ذوب فلز، گرمایش، تهویه مطبوع، سیستم های تبرید و کاربردهای فضایی می-باشند. مبدل های حرارتی در دستگاه های مختلف نظیر دیگ بخار، مولد بخار، کندانسور، اوپراتور، تبخیرکننده ها، برج خنک کن، پیش گرم کن فن کوپل، خنک کن و گرم کن روغن، رادیاتورها، کوره ها و..... کاربرد فراوان دارند.

آمار کلیدواژه‌ها

کلیدواژه‌های ارزیابی شده	کلیدواژه‌های مرتبط	کلیدواژه‌های غیرمرتبط	کلیدواژه‌های اضافه شده
۲۴	۱۲	۱۲	۰

کلیدواژه‌ها

کلیدواژه	امتیاز	کلیدواژه	امتیاز	کلیدواژه	امتیاز
انتقال گرما	۳	شبیه‌سازی رایانه‌ای	۳	گرمایش	۳
تهویه مطبوع	۱	صنایع پتروشیمی	۱	لوله مارپیچی	۳
جریان سیال	۳	صنایع ذوب فلز	۱	مبادله‌کن گرما	۳
سرمایش	۳	صنایع غذا	۱	مبدل های حرارتی	۱

جدول ٤-١٠- نمونه‌ای از داده‌های نهایی هنر

عنوان: عاملیت زن در سه نمایشنامه‌ی مده، آنتیگون و رومئو و ژانت در آثار ژان آنوی با تأکید بر رویکرد روانکاوانه‌ی تحلیلی کارل یونگ و کار عملی: اجرای نمایش صحنه‌ای مده اثر ژان آنوی	موضوع: هنر

چکیده:

گام نخست در هدایت بازیگر توسط کارگردان، از تحلیل روانکاوانه ی شخصیت های نمایش نامه آغاز می شود. در این پژوهش تلاش بر این است تا با استفاده از مکتب روانشناسی تحلیلی یونگ، به تجزیه و تحلیل کنش های شخصیت های محوری زن سه نمایش نامه ی مده، آنتیگون، رومئو و ژانت، اثر ژان آنوی پرداخته شود تا عاملیت آنان در راستای پیشبرد اهداف نویسنده بر اساس دیدگاه های روانی یونگ نمایان گردد.یونگ، روانشناس سوئیسی، موسس مکتب روان شناسی تحلیلی است. او کار فروید در زمینه روانکاوی را وسعت داد و اختلال های ذهنی و عاطفی را به منزله ی کوششی برای دست یافتن به تعادل شخصیت روانی تعبیر کرد. آنچه در این پژوهش مورد بررسی قرار گرفت: توجه به ضمائر خودآگاه، ناخودآگاه شخصی و ناخودآگاه جمعی و عناصر تشکیل دهنده ی آن ؛(کهن الگو ها)؛ نقاب (پرسونا) ، سایه، انیما/ انیموس و خویشتن. تفاوت رویکرد شخصیتی بر اساس جهت گیری های روانی. بررسی عاملیت، کنش و شخصیت از دیدگاه های روانی و جامعه شناسی. و در نهایت با درک و تحلیل روانی شخصیت ها می توان به اجرایی بر پایه ی روانشناسی تحلیلی دست یافت.

آمار کلیدواژه‌ها

کلیدواژه‌های ارزیابی شده	کلیدواژه‌های مرتبط	کلیدواژه‌های غیرمرتبط	کلیدواژه‌های اضافه شده
۲۴	۱۵	۹	۰

کلیدواژه‌ها

کلیدواژه	امتیاز	کلیدواژه	امتیاز	کلیدواژه	امتیاز
آنتیگون	۱	روان‌شناسی یونگ	۳	مدهآ (نمایشنامه)	۳
آنتیگونه (نمایش)	۳	رومئو و ژانت (نمایشنامه)	۳	نقش زنان	۳
آنوی، ژان	۳	زن	۳	هویّت فرهنگی	۱
بررسی‌آثار ادبی	۳	ژان آنوی	۱	یونگ، کارل	۲
تحلیل روانکاوانه	۱	شخصیت‌پردازی	۳	کهن الگو	۲

فصل پنجم:

تحلیل مجموعه داده و تست الگوریتم‌ها

۵. تحلیل مجموعه داده و تست الگوریتم‌ها

در این فصل، ابتدا مجموعه داده ایجاد شده از نظر تعداد کلیدواژه‌های بدست آمده، کلیدواژه‌های اضافه شده، کلیدواژه‌های اصلی و سایر معیارها، تحلیل و ارزیابی می‌شود. سپس نتایج پیاده‌سازی الگوریتم‌های مختلف استخراج کلیدواژه روی این مجموعه داده ارائه می‌گردد.

۵- ۱. تحلیل و ارزیابی مجموعه داده

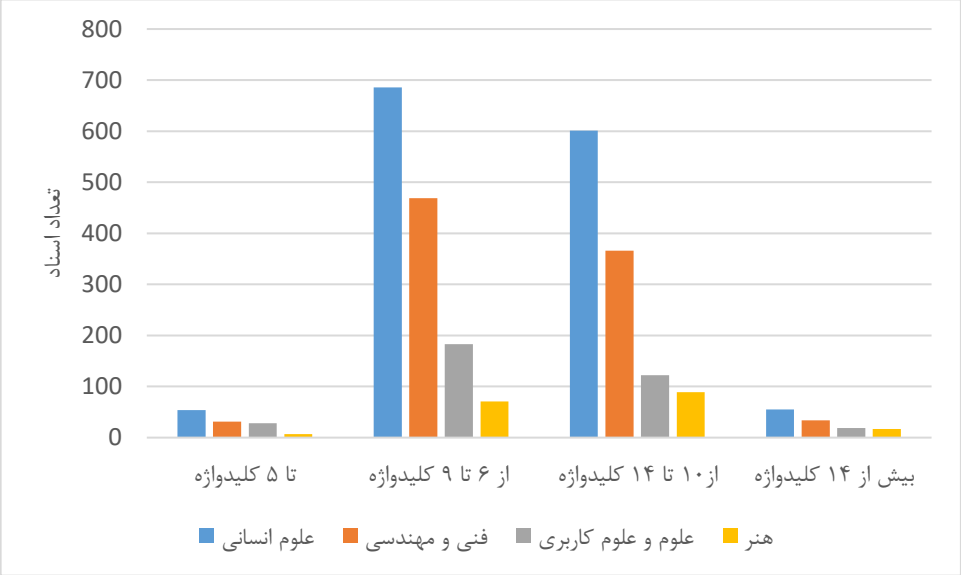
مجموعه داده بد ست آمده شامل ۲۸۳۸ سند در چهار حوزه مو ضوعی فنی و مهند سی، علوم و علوم کاربردی، هنر، و علوم انسانی است. در فرایند ارزیابی کلیدواژه‌های پیشنهادی که در فصل قبل تشریح شد، تعدادی از کلیدواژه‌های پیشنهادی از نظر متخصص نامرتبط شنا سایی شد و تعداد محدودی نیز کلیدواژه جدید برای هر سند پیشنهاد شد.

در جدول زیر اطلاعات آماری مجموعه داده بدست آمده از نظر تعداد کلیدواژه‌های پیشنهادی و کلیدواژه‌های نهایی آمده است.

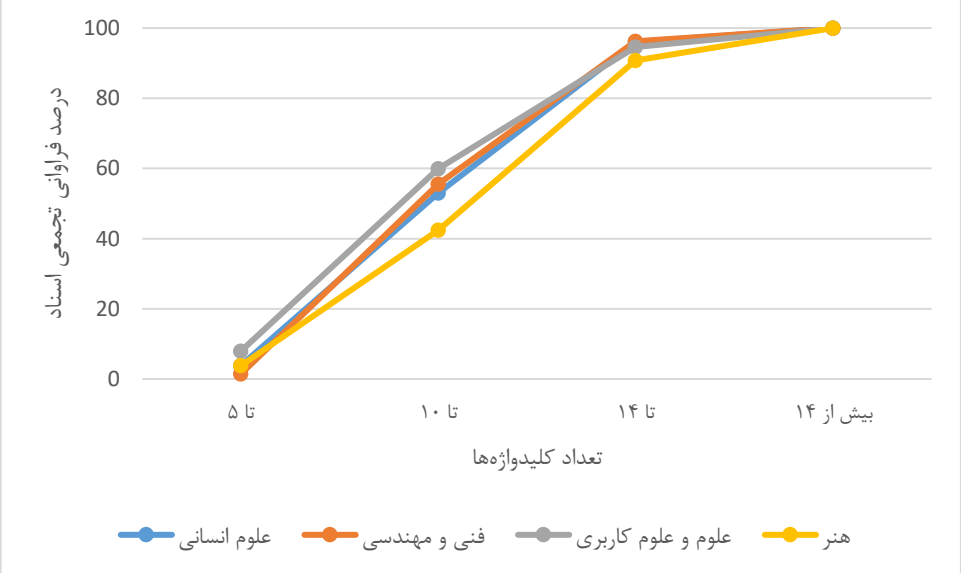
جدول ۵-۱- اطلاعات آماری تعداد کلیدواژه‌های مجموعه داده پس از ارزیابی

حوزه موضوعی / اطلاعات آماری	فنی و مهندسی	علوم انسانی	علوم و علوم کاربردی	هنر
تعداد اسناد	۹۰۱	۱۳۹۹	۳۵۴	۱۸۴
میانگین تعداد کلیدواژه‌های نهایی	۹/۴	۹/۴	۹/۲	۱۰/۳
تعداد اسناد با کلیدواژه جدید	۲۳۲	۴۱۶	۸۳	۳۵
تعداد اسناد با بیش از ۱۰ کلیدواژه	۴۰۰	۶۵۵	۱۴۱	۱۰۶
تعداد اسناد با بیش از ۱۵ کلیدواژه	۳۴	۵۵	۱۹	۱۷
بیشترین تعداد کلیدواژه	۲۱	۲۰	۲۰	۱۸
کمترین تعداد کلیدواژه	۴	۳	۳	۵

در ادامه توزیع کلیدواژه‌ها اسناد در حوزه‌های موضوعی مختلف در شکل ۵-۱ نمایش داده شده است. همچنین فراوانی تجمعی اســناد با تعداد کلیدواژ های مختلف در هر حوزه موضــوعی در شــکل ۵-۲ آمده است.



شکل ۵-۱- توزیع فراوانی کلیدواژه‌های اسناد در حوزه‌های موضوعی مختلف



شکل ۵-۲- فراوانی تجمعی اسناد با تعداد کلیدواژه‌های مختلف

علاوه بر اطلاعات آماری درباره تعداد کلیدواژه‌ها، می‌توان امتیاز کلیدواژه‌های پیشنهادی را نیز تحلیل کرد که نتایج آن در جدول ۵-۲ آمده اســت. باتوجه به اطلاعات این جدول می‌توان گفت که بیش از ۶۵ درصد کلیدواژه‌های هر سند دارای امتیاز ۳ هستند و بالاترین ارتباط را با سند دارند.

جدول ۵-۲- اطلاعات آماری امتیاز کلیدواژه‌های مجموعه داده پس از ارزیابی

حوزه موضوعی / اطلاعات آماری	فنی و مهندسی	علوم انسانی	علوم و علوم کاربردی	هنر
درصد کلیدواژه‌های با امتیاز ۱	۲۶	۳۰	۲۷	۳۰
درصد کلیدواژه‌های با امتیاز ۳	۷۱	۶۹	۷۱	۶۸
متوسط امتیاز کلیدواژه‌ها	۲/۱۹	۲/۱۳	۲/۱۶	۲/۱۳

مقدار دقت و بازخوانی روش یکپارچه ســازی نتایج براســاس نظرات متخصــصــین در ســامانه ارزیابی کلیدواژه‌ها نیز محاسبه شده است. دقت (P) و بازخوانی (R) برای سند m صورت زیر محاسبه شده است:

$$R_m = \frac{\#\{recommended\ keywords \cap accepted\ keywords\}}{\#\{accepted\ keywords \cap new\ keywords\}} \quad (۵-۱)$$

$$P_m = \frac{\#\{recommended\ keywords \cap accepted\ keywords\}}{\#\{recommended\ keywords\}} \quad (۵-۲)$$

که در آن داریم:

- Recommended keywords : کلیدواژه‌های پیشنهادی روش یکپارچه‌سازی
- Accepted keywords: کلیدواژه هایی که از بین کلیدواژه های پیشــنهادی، حین ارزیابی توسط متخصصین پذیرفته شده‌اند.
- New keywords: کلیدواژه‌های جدیدی که حین ارزیابی توسط متخصصین پیشنهاد شده‌اند.

در نهایت میانگین مقدار دقت و بازخوانی روی همه اســناد در هر حوزه موضــوعی بیانگر دقت و بازخوانی روش یکپارچه‌سازی است که درجدول ۵-۳ آمده است.

جدول ۵-۳-دقت و بازخوانی روش یکپارچه‌سازی براساس نظر متخصصین در سامانه ارزیابی

	فنی و مهندسی	علوم انسانی	علوم و علوم کاربردی	هنر
دقت	۰/۳۹۳	۰/۳۹۲	۰/۳۷۸	۰/۴۱۹
بازخوانی	۰/۳۸۷	۰/۳۸۵	۰/۳۷۱	۰/۴۱۴

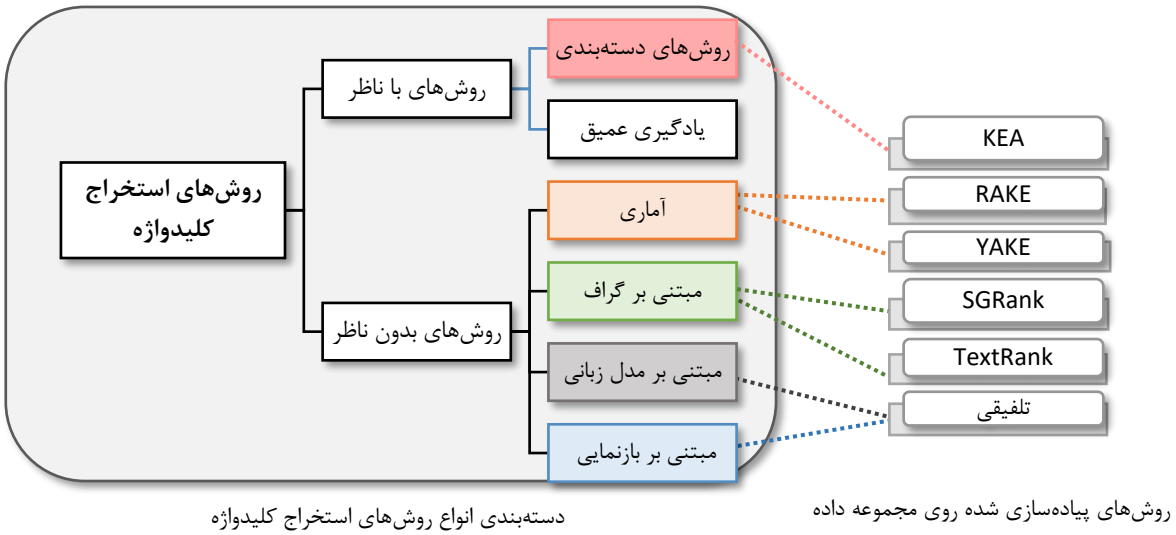
با توجه به رابطه‌های (۵-۱) و (۵-۲)، تفاوت بین دقت و بازخوانی در جدول ۵-۲، در کلیدواژه‌های جدید است و از آنجاییکه تعداد اسنادی که کلیدواژه جدید برای آنها پیشنهاد شده (۱۹٪ در هنر، ۳۰٪ در علوم انسانی، ۲۴٪ در علوم و علوم کاربردی، و ۲۶٪ در فنی و مهندسی) محدود است، مقدار دقت و بازخوانی تفاوت معناداری با هم ندارند.

۵- ۲. نتایج پیاده‌سازی الگوریتم‌های استخراج کلیدواژه

در این بخش نتایج پیاده‌سازی روش‌های مختلف استخراج کلیدواژه روی مجموعه داده ایجاد شده، نمایش داده می‌شود.

برای این منظور سعی شده که از هر دسته از انواع روش‌های استخراج کلیدواژه، حداقل یک روش پرکاربرد در نظر گرفته شود. روش‌های پیاده‌سازی شده به شرح زیر هستند:

- روش‌های با ناظر:
  - روش KEA (Witten et al. 1999)
- روش‌های بدون ناظر:
  - روش‌های مبتنی بر گراف: الگوریتم TextRank (Mihalcea and Tarau 2004) و الگوریتم SGRank (Danesh, Sumner, and Martin 2015)
  - روش‌های آماری: الگوریتم YAKE (Campos et al. 2020) و الگوریتم RAKE (Rose et al. 2010)
  - روش‌های مبتنی بر مدل زبانی و روش‌های مبتنی بر بازنمایی: الگوریتم تلفیقی محبی و همکاران (۱۳۹۵)



شکل ۵-۳- روش‌های پیاده‌سازی شده روی مجموعه داده، براساس دسته‌بندی انواع روش‌های استخراج کلیدواژه

برای پیاده‌سازی این روش‌ها از زبان برنامه‌نویسی پایتون و کتابخانه‌های NLTK، Spacy ، Gensim و PKE و HAZM استفاده شده و بخش اصلی کدهای هر الگوریتم نیز به پیوست آمده است.

نتایج پیاده‌سازی این الگوریتم‌ها براساس شاخص‌های دقت، و فراخوانی ارزیابی شده‌اند.

برای محاسبه دقت و بازخوانی در روش‌های استخراج کلیدواژه، دو رویکرد اصلی وجود دارد (Zesch and Gurevych 2009):

- رویکرد مقایسه دقیق: در این رویکرد مبنای مقایسه، یکسان بودن دقیق کلیدواژه‌های پیشنهادی با کلیدواژه‌های اصلی است. به عنوان مثال در این رویکرد دو کلیدواژه «فناوری اطلاعات» و «فناوری اطلاعات و ارتباطات» با هم یکسان نیستند.
- رویکرد مقایسه تقریبی: در این رویکرد مبنای مقایسه، اشتراک بین کلیدواژه اصلی و پیشنهادی است. به عنوان مثال اگر کلیدواژه اصلی «فناوری اطلاعات» باشد و کلیدواژه پیشنهادی «فناوری اطلاعات و ارتباطات» باشد، آنگاه کلیدواژه پیشنهادی درست است زیرا عبارت کلیدواژه پیشنهادی، دربردارنده کلیدواژه اصلی است.

بر همین اساس، نتایج مربوط به دقت و فراخوانی براساس رویکرد مقایسه دقیق، در جدول ۵-۴ آمده است. همچنین در جدول ۵-۵ نتایج دقت و بازخوانی براساس رویکرد تقریبی نمایش داده شده است.

جدول ۵-٤-دقت و بازخوانی روش‌های استخراج کلیدواژه در حوزه‌های موضوعی مختلف، براساس مقایسه دقیق

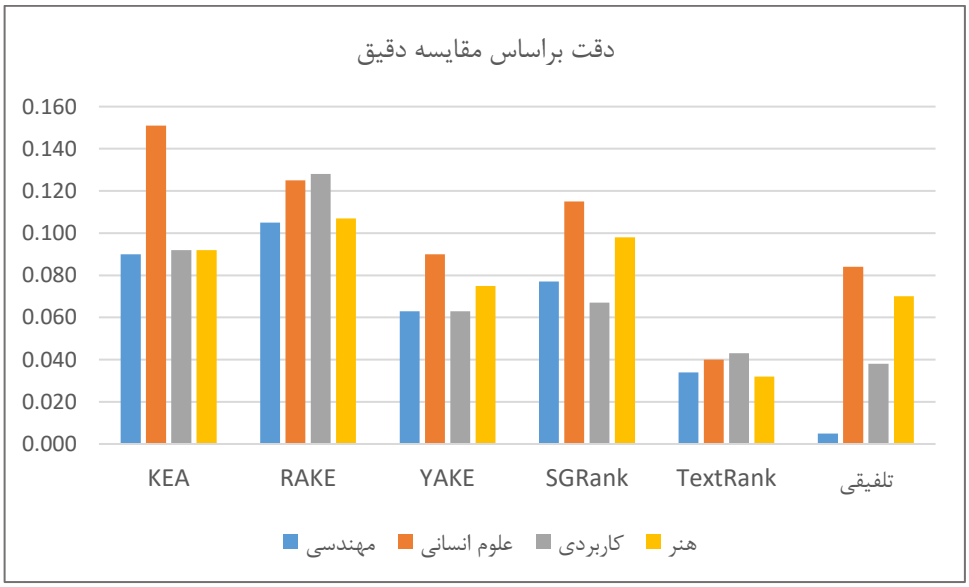
	KEA		RAKE		YAKE		SGRank		TextRank		تلفیقی	
	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی
مهندسی*	۰/۰۹	۰/۰۸۹	۰/۱۰۵	۰/۱۵۷	۰/۰۶۳	۰/۱۰۲	۰/۰۷۷	۰/۱۲	۰/۰۳۴	۰/۱۷۵	۰/۰۰۵	۰/۰۰۹
علوم انسانی	۰/۱۵۱	۰/۱۴۸	۰/۱۲۵	۰/۱۸	۰/۰۹	۰/۱۵	۰/۱۱۵	۰/۱۶۷	۰/۰۴	۰/۱۹۴	۰/۰۸۴	۰/۱۲۶
کاربردی*	۰/۰۹۲	۰/۰۹۱	۰/۱۲۸	۰/۲	۰/۰۶۳	۰/۱۰۲	۰/۰۶۷	۰/۱۱۴	۰/۰۴۳	۰/۱۶۳	۰/۰۳۸	۰/۰۲۸
هنر	۰/۰۹۲	۰/۰۹۱	۰/۱۰۷	۰/۱۵۳	۰/۰۷۵	۰/۱۱	۰/۰۹۸	۰/۱۳۱	۰/۰۳۲	۰/۱۷	۰/۰۷	۰/۱

* منظور از مهندسی، حوزه فنی و مهندسی و منظور از کاربردی، حوزه علوم و علوم کاربردی که به اختصار نوشته شده است.

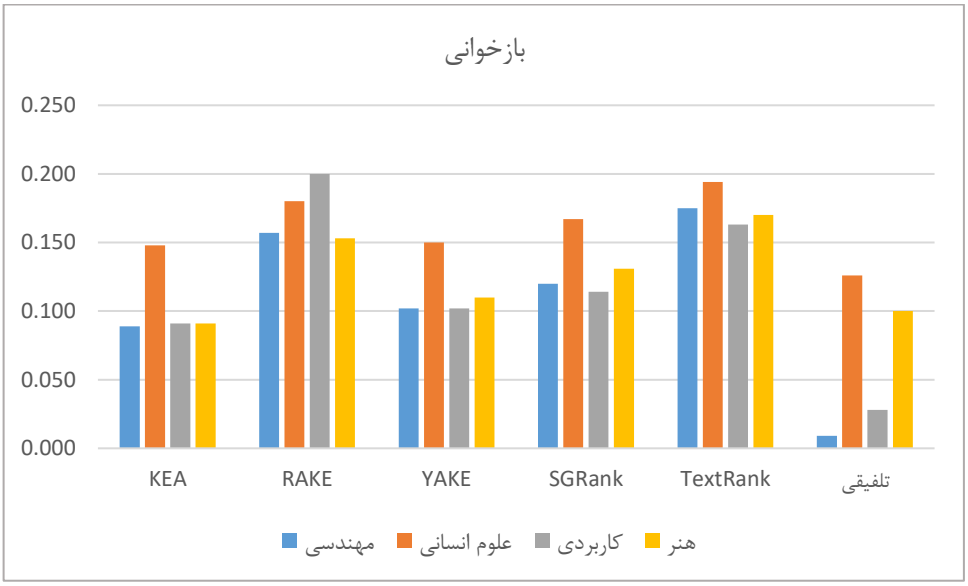
جدول ۵-۵-دقت و بازخوانی روش‌های استخراج کلیدواژه در حوزه‌های موضوعی مختلف، براساس مقایسه تقریبی

	KEA		RAKE		YAKE		SGRank		TextRank		تلفیقی	
	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی	دقت	بازخوانی
مهندسی	۰/۰۹۷	۰/۰۹۷	۰/۲۵۲	۰/۳۴۴	۰/۱۵۲	۰/۲۵۵	۰/۱	۰/۱۶۳	۰/۱۱۴	۰/۶۸	۰/۰۰۸	۰/۰۱۲
علوم انسانی	۰/۱۶	۰/۱۵۷	۰/۳۲۷	۰/۴۴۳	۰/۲۴۴	۰/۴۱۹	۰/۱۴۶	۰/۲۲	۰/۱۳۹	۰/۶۹۷	۰/۱	۰/۱۵
کاربردی	۰/۰۹۸	۰/۰۹۹	۰/۲۶۵	۰/۳۸۸	۰/۱۶	۰/۲۷۷	۰/۰۹۳	۰/۱۶۴	۰/۱۴	۰/۶	۰/۰۵	۰/۰۳۷
هنر	۰/۰۹۸	۰/۹۹۰	۰/۳۵۷	۰/۲۷۹	۰/۲۱۷	۰/۳۴۴	۰/۱۳۳	۰/۱۹۶	۰/۱۰۵	۰/۵۹۷	۰/۱۲۵	۰/۰۰۸

در شکل ۵-۴ و ۵-۵ نمودار دقت و بازخوانی براساس مقایسه دقیق، برای بررسی بیشتر عملکرد روش‌های مختلف نمایش داده شده است.



شکل ۵-٤- دقت روش‌های استخراج کلیدواژه در حوزه‌های موضوعی مختلف، براساس مقایسه دقیق



شکل ۵-۵- بازخوانی روش‌های استخراج کلیدواژه در حوزه‌های موضوعی مختلف، براساس مقایسه دقیق

از بین روش‌های پیاده‌سازی شده، دو روش SGRank و روش KEA در فرایند ایجاد مجموعه داده استفاده نشده‌اند. همانگونه که در شکل‌های فوق مشاهده می شود، دقت و بازخوانی روش‌ها به حوزه موضوعی نیز مرتبط است. به عنوان مثال روش KEA به طرز معناداری در حوزه علوم انسانی بهتر عمل می‌کند. اما روش RAKE در حوزه علوم کاربردی عملکرد بهتری دارد. تقریبا همه روش‌ها در حوزه مهندسی نسبت به سایر حوزه‌ها، عملکرد ضعیف‌تری دارند که وجود عبارات تخصصی ترکیبی ریاضی و لاتین می‌تواند یکی از دلایلی احتمالی ضعیف بودن نتایج روی این حوزه موضوعی باشد.

در بین همه روش‌ها، روش تلفیقی و روش KEA هر دو از منابع داده‌ای دیگری استفاده می‌کنند. غنی بودن منابع داده‌ای تاثیر به سزایی روی عملکرد این روش‌ها می‌گذارد. به عنوان مثال در روش KEA از یک مجموعه داده برای آموزش مدل استفاده شده است که هر چه این مجموعه داده وسیع‌تر و بزرگ‌تر باشد، نتایج بهتری حاصل می‌شود. روش تلفیقی نیز بر مبنای استدلال نمونه-محور[۱۷] عمل می‌کند، به گونه‌ای که کلیدواژه‌های پیشنهادی این روش وابسته به نمونه‌های قبلی مشابه است که در منابع داده‌ای موجود هستند. بنابراین بسته به اینکه منابع داده‌ای این روش تا چه میزان غنی هستند و نمونه‌های متعدیی را در بردارند، نتیجه این الگوریتم نیز متفاوت خواهد بود.

معمولا در روش‌های استخراج کلیدواژه، کلیدواژه‌های پیشنهادی دارای امتیاز یا رتبه‌ای هستند که بیانگر امتیاز آنها در الگوریتم استخراج کلیدواژه است و در نهایت براساس همین امتیاز، کلیدواژه‌های

[۱۷] Case-based reasoning

نهایی انتخاب می‌شوند. بنابراین در مواردی که ترتیب کلیدواژه‌های پیشنهادی براساس رتبه مرتب می‌شوند، معیارها «میانگین حد متوسط دقت» [۱۸]برای سنجش عملکرد الگوریتم می‌تواند بکارگفته شود. به طور خاص در سیستم‌های پیشنهاددهنده از این معیار هم برای سنجش میزان دقت نیز استفاده می‌شوند، زیرا معمولا در این دست از سیستم‌ها فهرستی از موارد، بیش از حد نیاز کاربر، به وی پیشنهاد می‌شود تا از بین آنها انتخاب نماید. بنابراین پیشنهاد موارد زیاد، که ممکن است تعدادی از آنها هم کاملا مرتبط نباشند، لزوما یک امتیاز منفی محسوب نمی‌شود، بلکه در این گونه از موارد، سیستمی مناسب خواهد بود که بتواند موارد مرتبط را در N پیشنهاد اول قرار دهد(Manning and Raghavan 2009; Shani and Gunawardana 2011).

در روش‌های کلاسیک اندازه‌گیری دقت، به عنوان مثال اگر سیستمی ۲۰ نمونه پیشنهاد دهد و از بین این ۲۰ نمونه، ۱۰ نمونه مرتبط باشند، آنگاه، فارغ از اینکه این ۱۰ نمونه در کجای فهرست پیشنهادی قرار دارند، دقت سیستم ۵۰ درصد است. در صورتیکه در روش «میانگین حد متوسط دقت»، رتبه نمونه‌های مرتبط نیز اهمیت دارد. به عنوان مثال اگر سیستمی بتواند ۱۰ نمونه مرتبط را در اول فهرست پیشنهادها قرار دهد، دقت بالاتری نسبت به سیستمی که این ۱۰ نمونه را انتها قرار می‌دهد، خواهد داشت.

معیار «میانگین حد متوسط دقت» (MAP) براساس پیشنهاد K کلیدواژه، میانگین $AP@K$ برای N سند مختلف است. مقدار $AP@K$ (حد متوسط دقت درK، *یعنی برای* Kکلیدواژه مرتبط) برای یک سند d_i به صورت زیر محاسبه می‌شود:

$$AP@K(d_i) = \frac{1}{m}\sum_{k=1}^{K} P(k).rel(k) \qquad (۳-۵)$$

که $rel(k)$ مساوی مقدار یک است در صورتیکه کلیدواژه kام مرتبط با d_i باشد و صفر است در صورتیکه کلیدواژه مرتبط نباشد و m تعداد کل کلیدواژه‌های مرتبط درکل فضای مسئله است. $P(k)$ دقت را برای k کلیدواژه اول نشان می‌دهد که به صورت زیر محاسبه می‌شود:

$$P(k) = \frac{\#\{recommended\ keywords\} \cap \{original\ keywords\}}{k} \qquad (۴-۵)$$

در نهایت $MAP@K$ یعنی «میانگین حد متوسط دقت»، به صورت زیر محاسبه می‌شود:

$$MAP@K = \frac{1}{N}\sum_{i=1}^{N} AP@K(d_i) \qquad (۵-۵)$$

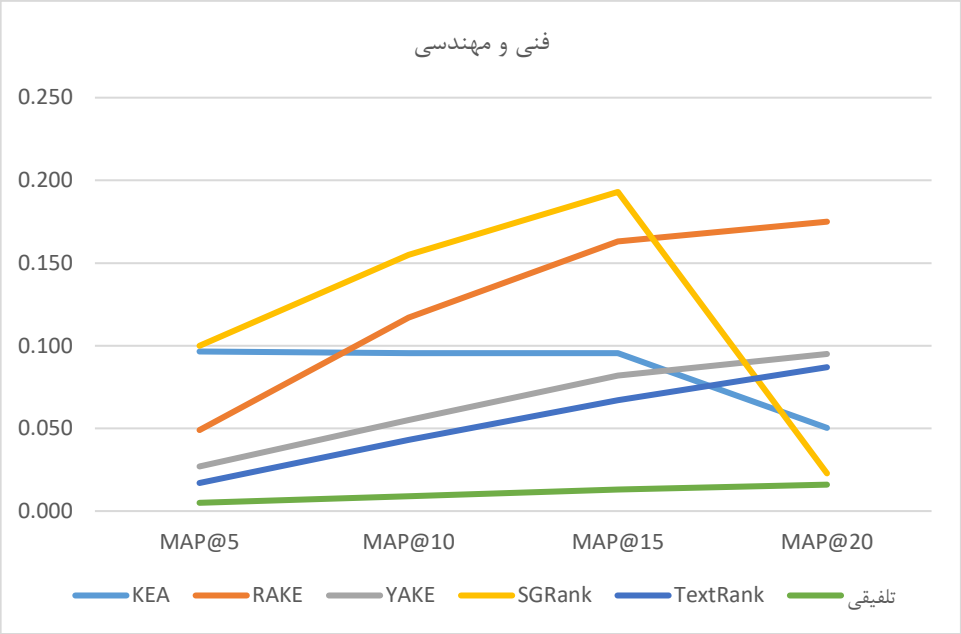
[۱۸] Mean Avergared Precision (MAP)

که N در آن تعداد اسناد مجموعه آزمایش است. در واقع $MAP@K$ میانگین مقدار $AP@K$ برای N سند مختلف است. بر همین اساس شاخص $MAP@K$ نیز برای هر روش در هر حوزه موضوعی، به ازای مقادیر مختلف K محاسبه شده است که نتایج آنها در جدول ۵-۶ آمده است.

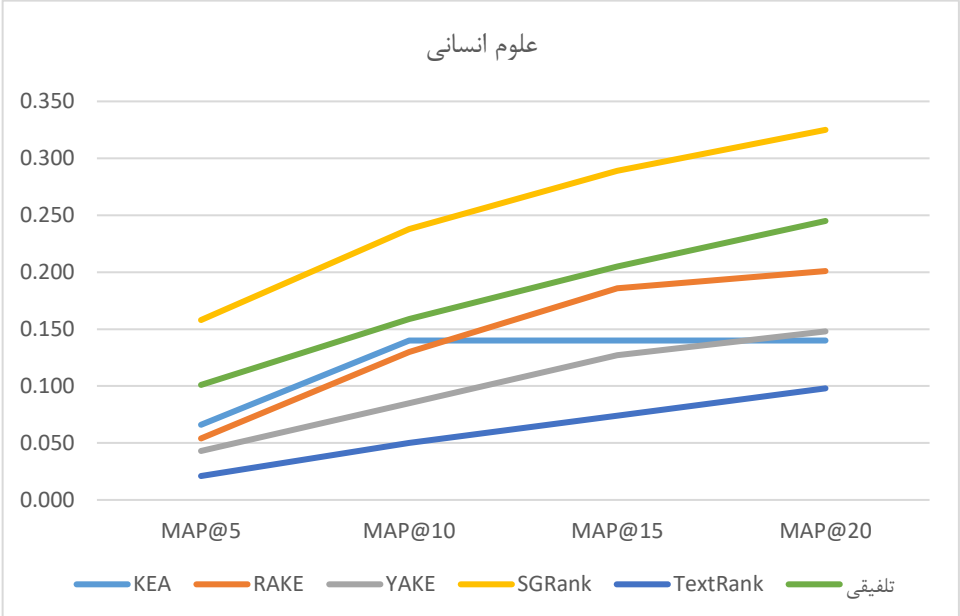
جدول ۵-٦-شاخص MAP@K روش‌های استخراج کلیدواژه در حوزه‌های موضوعی مختلف

		MAP@5	MAP@10	MAP@15	MAP@20
مهندسی	KEA	۰/۰۹۵۶	۰/۰۹۵۶	۰/۰۹۵۶	۰/۰۵۰۳
	RAKE	۰/۰۴۹	۰/۱۱۷	۰/۱۶۳	۰/۱۷۵
	YAKE	۰/۰۲۷	۰/۰۵۵	۰/۰۸۲	۰/۰۹۵
	SGRank	۰/۱	۰/۱۵۵	۰/۱۹۳	۰/۲۲۳
	TextRank	۰/۰۱۷	۰/۰۴۳	۰/۰۶۷	۰/۰۸۷
	تلفیقی	۰/۰۰۵	۰/۰۰۹	۰/۰۱۳	۰/۰۱۶
علوم انسانی	KEA	۰/۰۶۶	۰/۱۴	۰/۱۴	۰/۱۴
	RAKE	۰/۰۵۴	۰/۱۳	۰/۱۸۶	۰/۲۰۱
	YAKE	۰/۰۴۳	۰/۰۸۵	۰/۱۲۷	۰/۱۴۸
	SGRank	۰/۱۵۸	۰/۲۳۸	۰/۲۸۹	۰/۳۲۵
	TextRank	۰/۰۲۱	۰/۰۵	۰/۰۷۴	۰/۰۹۸
	تلفیقی	۰/۱۰۱	۰/۱۵۹	۰/۲۰۵	۰/۲۴۵
کاربردی	KEA	۰/۰۴۵	۰/۰۹	۰/۰۹	۰/۰۹
	RAKE	۰/۰۶۷	۰/۱۵۴	۰/۲۱۶	۰/۲۳۳
	YAKE	۰/۰۳۴	۰/۰۶۵	۰/۰۹۲	۰/۱۰۴
	SGRank	۰/۰۸۵	۰/۱۳۲	۰/۱۶۷	۰/۱۹۳
	TextRank	۰/۰۲۵	۰/۰۵۴	۰/۰۸	۰/۱۰۱
	تلفیقی	۰/۰۲۷	۰/۰۴۶	۰/۰۶۱	۰/۰۷۳
هنر	KEA	۰/۰۶۵	۰/۱۲۴	۰/۱۲۴	۰/۱۲۴
	RAKE	۰/۰۴۱	۰/۱۱۱	۰/۱۶۱	۰/۱۷۸
	YAKE	۰/۰۳۳	۰/۰۶	۰/۰۹	۰/۱
	SGRank	۰/۱۱۳	۰/۱۷۲	۰/۲۱۴	۰/۲۴۶
	TextRank	۰/۰۱۴	۰/۰۳۵	۰/۰۵۲	۰/۰۶۸
	تلفیقی	۰/۰۷۴	۰/۱۲	۰/۱۵۷	۰/۱۹

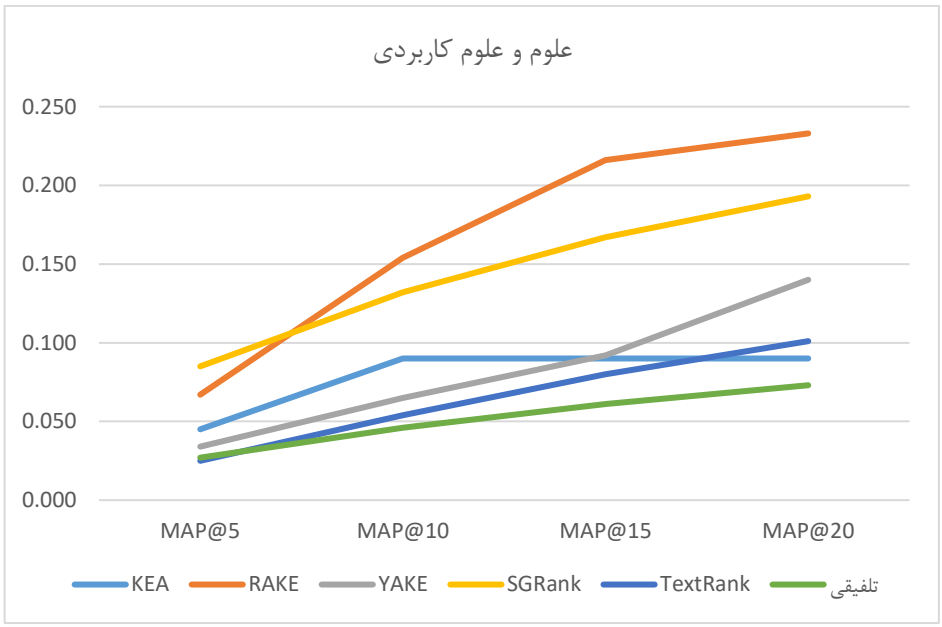
در ادامه شـــاخص $MAP@K$ به ازای مقادیر مختلف K برای مقایســـه بهتر بین روش‌های مختلف و حوزه‌های موضوعی مختلف، در شکل‌های زیر نمایش داده شده است.



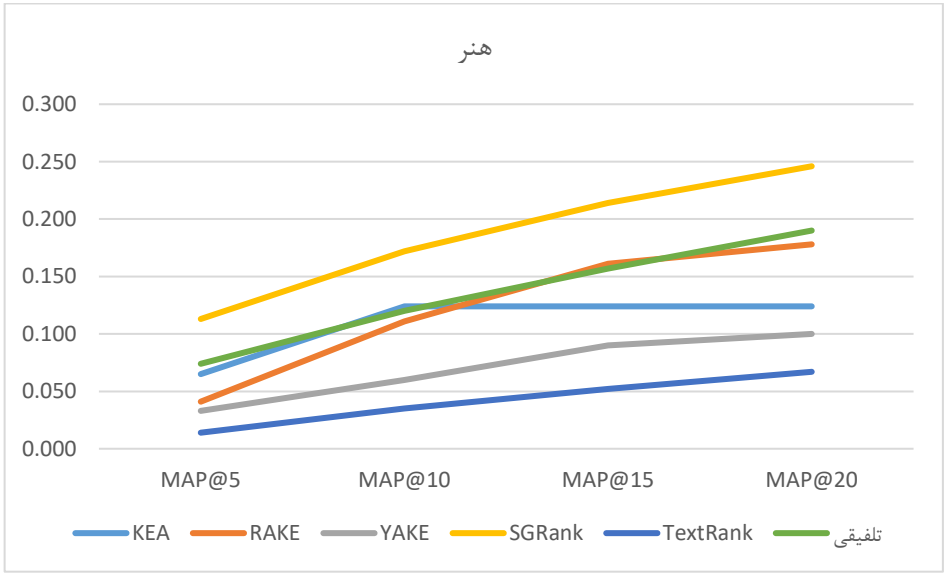
شکل ۵-٦- مقایسه شاخص MAP@K برای روش‌های استخراج کلیدواژه در حوزه موضوعی فنی و مهندسی



شکل ۵-۷- مقایسه شاخص MAP@K برای روش‌های استخراج کلیدواژه در حوزه موضوعی علوم انسانی



شکل ۵-۸- مقایسه شاخص MAP@K برای روش‌های استخراج کلیدواژه در حوزه موضوعی علوم و علوم کاربردی



شکل ۵-۹- مقایسه شاخص MAP@K برای روش‌های استخراج کلیدواژه در حوزه موضوعی هنر

فصل ششم:

راهکارهای اشاعه مجموعه داده و جمع‌بندی

۶. راهبردهای اشاعه مجموعه داده و جمع‌بندی

در این بخش ابتدا درباره مزایای تولید و اشـاعه داده‌های پژوهشـی با اشـاره به نکاتی که می‌تواند در طراحی مدل کسـب و کار برای بهره‌برداری از این مجموعه داده اسـتفاده شـود، راهبردهایی ارائه می‌شود. سپس در انتها نتایج این پژوهش جمع‌بندی می‌گردد.

٦- ١. راهبردهای اشاعه مجموعه داده

مجموعه داده‌های پژوهشـی را به روش‌ها و با اهداف گوناگونی می‌توان در دسـترس همگان قرار داد. در اینجا راهبردهای مختلفی که ایرانداک برای بهره‌برداری از این داده‌ها می‌تواند اتخاذ کند، مورد بررسی قرار می‌گیرد.

٦- ١- ١. راهبرد دسترس‌پذیر کردن به شکل داده باز و رایگان

بسیاری از موسسات پژوهشی و دانشگاهی معمولا داده‌های پژوهشی را به رایگان در وبسایت‌های خود در اختیار همگان قرار می‌دهند. هدف این موسسات از این کار می‌تواند یکی از موارد زیر باشد:

- افزایش اعتبار علمی و شناخته شدن موسسه در حوزهٔ داده‌های پژوهشی
- افزایش میزان ارجاعات علمی به موسسه با استفاده از داده‌های پژوهشی آن موسسه. در این مورد موسسات روش صحیح ارجاع به داده پژوهشی مورد نظر را در وبسایت خود درج می‌کنند.
- مرجعیت در حوزهٔ مورد نظر در طولانی مدت و با افزایش میزان اسـتفاده از داده‌های رایگان موسسه و مطرح شدن داده به عنوان استاندارد در پژوهش‌ها. در این حالت پژوهشگران برای مقایسـه نتایج الگوریتم ها و روش های داده‌محور خود می بایسـت نتایج را با داده های استانداردی که پیشتر مورد استفاده قرار گرفته‌اند، مقایسه کنند.

بنابراین یکی از راهبردهای ایرانداک برای اشاعه و دسترس‌پذیر کردن داده‌های پژوهشی به شکل رایگان، می‌تواند افزایش اعتبار علمی ایرانداک با استفاده از این داده‌های پژوهشی باشد. از آنجا که سامانه گنج ایرانداک مرجع رسمی منابع علمی متنی فارسی است، در صورتی که این داده‌ها به رایگان در وبسایت پژوهشگاه علوم و فناوری اطلاعات ایران گذاشته شود، احتمال رجوع دانشجویان و پژوهشگران به آن بالا خواهد بود. طبیعی است که زمانی طول خواهد کشید که این داده‌ها به عنوان داده‌های مرجع شناخته شوند. زمانی که میزان استفاده از یک داده پژوهشی از حد آستانه بالاتر برود و پژوهشگران شروع به مقایسه نتایج پژوهش‌های خود بر اساس آن داده کنند، رشد ارجاع به آن تضمین خواهد شد. ایرانداک می‌تواند تا زمانی که استفاده از داده‌های پژوهشی به حد آستانه نرسیده است، دربارهٔ آن داده و مزایای استفاده از آن در وبسایت خود یا سایر درگاه‌های مرجع پژوهش تبلیغ کند.

٦- ١- ٢. راهبرد برگزاری مسابقه

یکی دیگر از روش‌های دسترس‌پذیرکردن داده‌های پژوهشی برگزاری مسابقات بر پایهٔ این داده‌ها است. برگزاری مسابقه بر پایهٔ داده‌های پژوهشی می‌تواند به یکی از دلایل زیر باشد:

- برخی از موسسات، سازمان‌های پژوهشی و تجاری برای حل مشکلات خود و توسعهٔ زیر ساخت‌های محاسباتی و سرویس‌های خود، بخشی از داده‌های عملیاتی خود را در قالب پایگاه داده پژوهشی منتشر کرده و مسابقاتی را بر اساس آن داده‌ها برگزار می‌کنند. در این حالت موضوع مسابقات چالش‌هایی است که مالک داده با آن مواجه بوده و به دنبال ایده گرفتن برای حل آنها یا حل واقعی آنها توسط تیم‌های پژوهشگر و خلاق است.
- یکی دیگر از اهداف موسسات از برگزاری مسابقات برپایهٔ داده‌های پژوهشی مطرح شدن و مرجع شدن در حوزه‌های پژوهشی یا تجاری است. در این حالت موسسات در حوزه‌ای که به دنبال مرجعیت هستند، اقدام به برگزاری مسابقه می‌کنند.
- یکی دیگر از اهداف سازمان‌ها و موسسات پژوهشی و تجاری برای برگزاری مسابقات، شناسایی و جذب نیروی انسانی یا تیم‌های توانمند حل مسئله است. برای این هدف داده‌های پژوهشی ساخته شده و مسائل چالش‌برانگیز طراحی می شود و تبلیغات رویداد (مسابقه) در دانشگاه‌ها و اجتماعاتی که پتانسیل دسترسی به نیروی انسانی متخصص در آنها بالاست، گذاشته می‌شود. در این حالت طراحی مسابقه شامل فرایند برگزاری و جوایز آن، باید به گونه‌ای بازی‌وار شده باشد، که برای افراد وتیم‌های مستعد جذاب باشد.

البته تعداد داده در مجموعه داده پژوهشی باید به گونه‌ای باشد که برای مسائل حوزه مورد نظر کفایت کند. به عنوان نمونه بسیاری از الگوریتم‌ها و مدل‌های هوش‌مصنوعی همچون شبکه‌های عصبی عمیق[۱۹] نیازمند حجم بالایی از داده هستند و با داده‌های کم نمی‌توان این الگوریتم‌ها را آموزش داد.

در این راستا پژوهشگاه علوم و فناوری اطلاعات می‌تواند مسابقات/رویدادهایی را برای رسیدن به اهداف پیش‌گفته برگزار نماید. موضوعاتی همچون نمایه‌سازی ماشینی که چالشی همیشگی برای ایرانداک بوده در حالی که هر ساله هزینه‌های زیادی برای نمایه‌سازی اسناد صرف می‌شود، می‌توانند محور این مسابقات قرار گیرند. همچنین تعریف مسائلی در حوزهٔ هوش‌مصنوعی و پردازش زبان طبیعی بر پایهٔ داده‌های گنج، می‌تواند فرصت‌های بسیاری را برای توسعه خدمات ارزش افزوده هوشمند یا حل مشکلات ایرانداک ایجاد کند. پژوهشگاه می‌تواند با برگزاری این مسابقات علاوه بر تصویرسازی از خود در حوزهٔ خدمات اطلاعاتی هوشمند و خودکار، افراد، تیم‌ها و شرکت‌های مستعدی را که در حوزه مسائل پژوهشگاه کار می‌کنند، شناسایی کرده و از توانمندی‌ها و خدمات آنها بهره بگیرد.

٦- ١- ٣. راهبرد فروش داده‌ها

آخرین راهبرد برای پژوهشگاه درآمدزایی مستقیم از داده‌های پژوهشی و فروش آنها است. مسلما برای فروش داده‌های پژوهشی، قیمت‌گذاری و توجیه ارزش این داده‌ها بسیار بااهمیت است و در صورتی که قیمت‌گذاری به درستی انجام نشود، ممکن است داده‌ها جذابیتی برای پژوهشگران نداشته باشند و از آنها استفاده نشود. همچنین اگر ایرانداک بخواهد داده‌های پژوهشی خود را بفروشد، کیفیت و کفایت این داده‌ها اهمیت زیادی خواهند داشت و داده‌ها می‌بایست به خوبی پاکسازی، کدگذاری و مستندسازی شود.

در اینجا به نکاتی در مورد راهبردهای فروش و قیمت‌گذاری اشاره می‌کنیم:

- داده‌های جدید همیشه از داده‌های قدیمی ارزشمندتر هستند. بر این اساس ایرانداک می‌تواند داده‌های قدیمی را به قیمتی پایین‌تر یا به شکل رایگان در اختیار مشتریان قرار دهد و داده‌های جدید را با قیمت بالاتری بفروشد.
- داده‌هایی که فقط در دسترس ایرانداک است و نمی‌توان آن را از درگاه دیگری تهیه کرد، ارزشمندتر هستند. بر این اساس داده‌ها و فراداده‌هایی که قابلیت دسترسی از روش‌ها و درگاه‌های دیگر را دارند چندان ارزشمند نیستند و می‌توان آنها را به رایگان در دسترس قرار داد. در مقابل داده‌هایی که فقط در دسترس ایرانداک هستند، در صورتی که استاندارد و پرکاربرد باشند، به احتمال زیاد با قیمت بالاتری قابل فروش خواهند بود.

---

[۱۹] Deep neural network

- اگر برای رسیدن به دادهٔ پژوهشی مورد نظر مسیر طولانی و پرهزینه‌ای طی شده و تولید آن از ابتدا برای پژوهشگران کاری زمانبر و پرهزینه است، باز هم داده‌ها ارزشمندتر و قابل فروش خواهند بود. بر این پایه ایرانداک می‌تواند داده های خام را به رایگان یا با قیمتی اندک دسترس‌پذیر کند و در مقابل داده‌هایی که با هزینه بالاتر و در زمانی طولانی‌تر پردازش و آماده‌سازی شده‌اند را با قیمت بالاتری به فروش برساند.

همچنین پژوهشگاه می‌تواند ترکیبی از این سه راهبرد را داشته باشد. یعنی بخشی از داده‌ها را به صورت رایگان در اختیار جامعه پژوهشی قرار دهد و علاوه بر این، رویداد و مسابقاتی را بر پایهٔ داده‌های باز برگزار نماید. پس از اینکه داده‌های پژوهشی باز ایرانداک مورد توجه قرار گرفت، می‌توان داده‌هایی با حجم بیشتر، پردازش عمیق‌تر و اقلام متفاوت اطلاعاتی و نشانه‌گذاری شده را نیز به فروش رساند.

٦- ۲. جمع‌بندی

در این پژوهش، مجموعه داده‌ای برای مساله استخراج خودکار کلیدواژه از مستندات علمی ایجاد شد. این مجموعه داده مشتمل بر ۲۸۳۸ سند از حوزه‌های موضوعی علوم انسانی، فنی و مهندسی، علوم وعلوم کاربردی، و هنر است. هر سند شامل عنوان، چکیده و کلیدواژه‌های پایان‌نامه‌ها و رساله‌هایی است که در پایگاه گنج ثبت و نمایه شده‌اند.

برای تخصیص کلیدواژه به اسناد، چهار روش استخراج کلیدواژه برای هر سند بکار گرفته شده است و براساس روش پیشنهادی در این پژوهش، کلیدواژه‌های پیشنهادی همه روش‌ها با هم یکپارچه شده‌اند و در نهایت برای هر سند یک فهرست ۲۴ تا ۳۰ تایی کلیدواژه بدست آمده است.

سپس در قالب یک سامانه ارزیابی کلیدواژه‌ها، نتایج بدست آمده توسط متخصیص نمایه‌ساز ارزیابی شده‌اند و کلیدواژه‌های مناسب از بین کلیدواژه‌های پیشنهادی انتخاب شده‌اند. همچنین برای هر کلیدواژه انتخابی، میزان ارتباط آن با سند براساس سه عدد ۱، ۲، و ۳ مشخص شده‌اند. علاوه بر آن، در صورت نیاز، کلیدواژه‌های جدیدی نیز برای هر سند پیشنهاد شده‌اند.

در نهایت عملکرد الگوریتم‌های مختلف روی مجموعه داده، براساس شاخص‌های دقت، بازخوانی و MAP@K بررسی شده‌اند.

در پژوهش‌های آتی می‌توان مجموعه داده‌های متنوع‌تری برای مساله استخراج کلیدواژه ایجاد کرد و روش‌های مرسوم استخراج کلیدواژه را براساس آنها ارزیابی نمود.

منابع

۷. منابع

محبی، آزاده، و عمار جلالی‌منش.۱۳۹۸ .ارائه رو شی هو شمند برای ا ستخراج کلیدواژه از مستندات علمی زبان فار سی براساس سیستم‌های پیشنهاددهنده .تهران: پژوهشگاه علوم و فناوری اطلاعات ایران.

احمدی، عباس و طیبه حسینی‌خواه. ۱۳۹۲."اسـتخراج کلمات کلیدی یک متن با اسـتفاده از شـبکه‌های عصـبی". در دهمین کنفرانس بین‌المللی مهندسی صنایع. تهران: دانشگاه تهران.

رضـایی، وحیده، مجید محمدپور، حمید پروین و صـمد نجاتیان. ۱۳۹۶."ارائه روشـی برای اسـتخراج کلمات کلیدی و وزن‌دهی کلمات برای بهبود طبقه‌بندی متون فارسی". فصلنامه پردازش علائم و داده‌ها ۱۴(۴ :)۵۵–۷۸.

احمدی، علی اصـغر و مریم حبیبی. ۱۳۹۵."اسـتخراج کلمات کلیدی فارسـی با اسـتفاده از تکنیک‌های داده‌کاوی". در چهارمین کنفرانس بین‌المللی در مهندسـی برق و کامپیوتر. تهران: موسـسـه آموزش عالی صـالحان، دانشـکده مدیریت دانشگاه تهران.

باسره، مریم، ولی درهمی و سجاد ظریف‌زاده. ۱۳۹۶."ارائه‌روشی برای استخراج خودکار عبارات کلیدی از اخبار وب پارسی". مجله مهندسی برق دانشگاه تبریز. ۴۷(۸۱ :)۸۵۷-۸۶۶.

شیری، حسن، فاطمه کربلائی و شیرین موسوی. ۱۳۸۴."طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی" دریازدهمین کنفرانس بین‌المللی کامپیوتر. تهران: انجمن کامپیوتر ایران، پژوهشگاه دانش‌های بنیادی، پژوهشکده علوم کامپیوتر.

راد، فرهاد، حمید پروین، آتوسـا دهباشـی و بهروز مینایی. ۱۳۹۵."ارائه روشـی جدید برای شـاخص‌گذاری خودکار و اسـتخراج کلمات کلیدی برای بازیابی اطلاعات و خوشـه‌بندی متون". فصـلنامه پردازش علائم و داده‌ها ۱۳(۱ :)۸۷–۱۰۰.

گزنی، علی. ۱۳۸۵."اسـتخراج خودکار عبارتهای کلیدی از متون مقاله‌های فارسـی".کتابداری و اطلاع رسـانی ۹(۳ :) ۹۷–۱۰۸.

شکری، مسعود و محمدرضا میبدی. ۱۳۸۲" .ساخت یک نمایه‌ساز خودکار برای متون فارسی" در یازدهمین کنفرانس مهندسی برق. شیراز: دانشگاه شیراز.

ویسی، هادی و نیلوفر افلاکی. ۱۳۹۴."استخراج کلمات کلیدی متن فارسی با استفاده از آنالیز معنایی". در کنفرانس بین‌المللی مهندسی و علوم کاربردی. دوبی.

عربی‌نرئی، سمیه، مجتبی وحیدی‌ا صل و بهروز مینایی بیدگلی. ۱۳۸۶."ا ستخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی". در اولین کنفرانس داده‌کاوی ایران، ۱–۹ .تهران: دانشگاه صنعتی امیر کبیر، مؤسسه پژوهشی داده‌پردازان گیتا.

Alami Merrouni, Zakariae, Bouchra Frikh, and Brahim Ouhbi. 2020. "Automatic Keyphrase Extraction: A Survey and Trends." *Journal of Intelligent Information Systems* 54(2): 391–424.

Apté, Chidanand, Fred Damerau, and Sholom M Weiss. 1994. "Towards Language Independent Automated Learning of Text Categorization Models BT - SIGIR '94." In eds. Bruce W Croft and C J van Rijsbergen. London: Springer London, 23–30.

Aronson, Alan R et al. 2000. "The NLM Indexing Initiative." In *Proceedings of the AMIA Symposium*, , 17.

Augenstein, Isabelle et al. 2017. "SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications." In *Proceedings of the 11th International Workshop on Semantic Evaluation ({S}em{E}val-2017)*, Vancouver, Canada: Association for Computational Linguistics, 546–55. https://www.aclweb.org/anthology/S17-2091.

Beliga, Slobodan, Ana Mestrovic, and Sanda Martineie-Ipsie. 2015. "Beliga et Al. 2015 - An Overview of Graph-Based Keyword Extraction Methods and Appproaches.Pdf." *Journal of Information and Organizational Sciences* 39(1): 1–20.

Bennani-Smires, Kamil et al. 2018. "Simple Unsupervised Keyphrase Extraction Using Sentence Embeddings." In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, Brussels, Belgium: Association for Computational Linguistics, 221–29. https://www.aclweb.org/anthology/K18-1022.

Campos, Ricardo et al. 2020. "YAKE! Keyword Extraction from Single Documents Using Multiple Local Features." *Information Sciences* 509: 257–89.

Chen, Jun et al. 2018. "Keyphrase Generation with Correlation Constraints." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 4057–66. https://www.aclweb.org/anthology/D18-1439.

Danesh, Soheil, Tamara Sumner, and James H Martin. 2015. "SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction." In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, Denver, Colorado: Association for Computational Linguistics, 117–26. https://www.aclweb.org/anthology/S15-1013.

Doostmohammadi, Ehsan, Mohammad Hadi Bokaei, and Hossein Sameti. 2019. "PerKey: A Persian News Corpus for Keyphrase Extraction and Generation." *9th International Symposium on Telecommunication: IST 2018*: 460–65.

El-Beltagy, Samhaa R, and Ahmed Rafea. 2009. "KP-Miner: A Keyphrase Extraction

System for English and Arabic Documents." *Information Systems* 34(1): 132–44. http://www.sciencedirect.com/science/article/pii/S0306437908000537.

Gollapalli, Sujatha Das, and Cornelia Caragea. 2014. "Extracting Keyphrases from Research Papers Using Citation Networks." In *AAAI*, , 1629–35.

Grineva, Maria, Maxim Grinev, and Dmitry Lizorkin. 2009. "Extracting Key Terms from Noisy and Multi-Theme Documents." *WWW'09 - Proceedings of the 18th International World Wide Web Conference*: 661–70.

Hammouda, Khaled M, Diego N Matute, and Mohamed S Kamel. 2005. "CorePhrase: Keyphrase Extraction for Document Clustering BT - Machine Learning and Data Mining in Pattern Recognition." In eds. Petra Perner and Atsushi Imiya. Berlin, Heidelberg: Springer Berlin Heidelberg, 265–74.

Hasan, Kazi Saidul, and Vincent Ng. 2014. "Automatic Keyphrase Extraction: A Survey of the State of the Art." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 1262–73.

Hulth, Anette. 2003. "Improved Automatic Keyword Extraction given More Linguistic Knowledge." In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, , 216–23.

Hwang, Ching-Lai, and Kwangsun Yoon. 1981. "Methods for Multiple Attribute Decision Making." In *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey*, Berlin, Heidelberg: Springer Berlin Heidelberg, 58–191. https://doi.org/10.1007/978-3-642-48318-9_3.

Kazemi, Ashkan, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. "Biased TextRank: Unsupervised Graph-Based Content Extraction." : 1642–52.

Kim, Su Nam, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. "Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles." In *Proceedings of the 5th International Workshop on Semantic Evaluation*, , 21–26.

Krapivin, Mikalai et al. 2010. "Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing BT - The Role of Digital Libraries in a Time of Global Change." In eds. Gobinda Chowdhury, Chris Koo, and Jane Hunter. Berlin, Heidelberg: Springer Berlin Heidelberg, 102–11.

Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl. 2009. "Latent Dirichlet Allocation for Tag Recommendation." In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, New York, NY, USA: ACM, 61–68. http://doi.acm.org/10.1145/1639714.1639726.

Lazemi, Soghra, Hossein Ebrahimpour-Komleh, and Nasser Noroozi. 2019. "PAKE: A Supervised Approach for Persian Automatic Keyword Extraction Using Statistical Features." *SN Applied Sciences* 1(12): 1574. https://doi.org/10.1007/s42452-019-1627-5.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, JMLR.org, II–1188–II–1196.

Liu, Zhiyuan, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. "Automatic Keyphrase Extraction via Topic Decomposition." In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, , 366–76.

Lynn, Htet Myet, Eunji Lee, Chang Choi, and Pankoo Kim. 2017. "SwiftRank: An

Unsupervised Statistical Approach of Keyword and Salient Sentence Extraction for Individual Documents." *Procedia Computer Science* 113: 472–77. http://www.sciencedirect.com/science/article/pii/S1877050917317143.

Mallick, Chirantana et al. 2019. "Graph-Based Text Summarization Using Modified TextRank." In *Soft Computing in Data Analytics*, eds. Janmenjoy Nayak et al. Singapore: Springer Singapore, 137–46.

Manning, Christopher D., and Prabhakar Raghavan. 2009. "An Introduction to Information Retrieval." *Online* 1: 1.

Marujo, Luis, Márcio Viveiros, and João Paulo da Silva Neto. 2013. "Keyphrase Cloud Generation of Broadcast News."

Medelyan, Olena, and Ian H Witten. 2008. "Domain-Independent Automatic Keyphrase Indexing with Small Training Sets." *Journal of the American Society for Information Science and Technology* 59(7): 1026–40. https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20790.

Meng, Rui et al. 2017. "Deep Keyphrase Generation." *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1: 582–92.

Mihalcea, Rada, and Paul Tarau. 2004. "TextRank: Bringing Order into Text." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain: Association for Computational Linguistics, 404–11. https://www.aclweb.org/anthology/W04-3252.

Mikolov, Tomas et al. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, Brooklyn, NY: Curran Associates Inc., 3111--3119.

Nguyen, Thuy Dung, and Min-Yen Kan. 2007. "Keyphrase Extraction in Scientific Publications." In *International Conference on Asian Digital Libraries*, , 317–26.

Pan, Suhan, Zhiqiang Li, and Juan Dai. 2019. "An Improved TextRank Keywords Extraction Algorithm." In *Proceedings of the ACM Turing Celebration Conference - China*, ACM TURC '19, New York, NY, USA: Association for Computing Machinery, 1–7. https://doi.org/10.1145/3321408.3326659.

Papagiannopoulou, Eirini, and Grigorios Tsoumakas. 2018. "Local Word Vectors Guiding Keyphrase Extraction." *Information Processing & Management* 54(6): 888–902. http://www.sciencedirect.com/science/article/pii/S0306457317308427.

———. 2019. "A Review of Keyphrase Extraction." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(September 2018): 1–45.

Pay, T, and S Lucci. 2017. "Automatic Keyword Extraction: An Ensemble Method." In *2017 IEEE International Conference on Big Data (Big Data)*, , 4816–18.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, Doha, Qatar: Association for Computational Linguistics, 1532–43. https://www.aclweb.org/anthology/D14-1162.

Peters, Matthew et al. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for

Computational Linguistics, 2227–37. https://www.aclweb.org/anthology/N18-1202.

Radev, Dragomir R., Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. "The ACL Anthology Network Corpus." *Language Resources and Evaluation* 47(4): 919–44.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. "Automatic Keyword Extraction from Individual Documents." *Text Mining: Applications and Theory* (October 2017): 1–20.

Salton, Gerard, and Christopher Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24(5): 513–23. http://www.sciencedirect.com/science/article/pii/0306457388900210.

Schutz, Alexander Thorsten. 2008. "Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods." National University of Ireland,.

Shani, Guy, and Asela Gunawardana. 2011. "Evaluating Recommendation Systems." In *Recommender Systems Handbook*, Springer, 257–97.

Siddiqi, Sifatullah, and Aditi Sharan. 2015. "Keyword and Keyphrase Extraction Techniques: A Literature Review." *International Journal of Computer Applications* 109(2): 18–23.

Theses100. 2019. "Datasets of Automatic Keyphrase."

Tomokiyo, Takashi, and Matthew Hurst. 2003. "A Language Model Approach to Keyphrase Extraction." In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, USA: Association for Computational Linguistics, 33–40. https://doi.org/10.3115/1119282.1119287.

Wan, Xiaojun, and Jianguo Xiao. 2008. "Single Document Keyphrase Extraction Using Neighborhood Knowledge." In *AAAI*, , 855–60.

Wang, Hao, Binyi Chen, and Wu-Jun Li. 2013. "Collaborative Topic Regression with Social Regularization for Tag Recommendation." In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, AAAI Press, 2719–2725.

Witten, Ian H. et al. 1999. "KEA: Practical Automatic Keyphrase Extraction." *In Proceedings of the fourth ACM conference on Digital libraries, ACM*: 254–61.

Ye, Hai, and Lu Wang. 2018. "Semi-Supervised Learning for Neural Keyphrase Generation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 4142–53. https://www.aclweb.org/anthology/D18-1447.

You, Wei, Dominique Fontaine, and Jean Paul Barthès. 2009. "Automatic Keyphrase Extraction with a Refined Candidate Set." *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009* 1: 576–79.

Zesch, Torsten, and Iryna Gurevych. 2009. "Approximate Matching for Evaluating Keyphrase Extraction." In *International Conference Recent Advances in Natural Language Processing, RANLP*, , 484–89.

Zhang, Qi, Yang Wang, Yeyun Gong, and Xuanjing Huang. 2016. "Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter." *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*: 836–45.

Abstract

Due to the volume of data and documents created in specialized scientific databases such as Ganj (database of Persian thesis and dissertations), intelligent and automated methods are required for organizing and indexing these documents. Document indexing involves assigning appropriate keywords, to each documents, known as keyword extraction, while these keywords represent the inherent topic and concepts within the document. Various automatic keyword extraction algorithm and approaches have been developed, each with some limitations and advantages. However, development and application of such approaches requires standard criteria for evaluation and improvement. One of these criteria is having a standard dataset containing a set of documents and keywords assigned to them. Without such a dataset, it is not easy to compare and apply algorithms and methods for automatically extracting keywords from scientific documents.

Although studies on automatic keyword extraction from Persian documents, have also used some datasets to evaluate methods, either these datasets do not contain scientific documents, or they contain documents with limited keywords, i.e. author keywords. However, there are numerous English datasets containing scientific documents that are widely used in English scientific keyword extraction approaches.

In this research, a dataset of scientific documents (abstract of thesis and dissertations), from Ganj database, along with their affiliated keywords is developed. Although documents indexed in Ganj already contain keywords, their keywords are limited and incomplete. In addition, it is not clear how much each keyword is relevant to the document. For dataset generation, first a set of documents from four major subjects including human science, engineering, art, and applied science, are selected from Ganj. Then, four different algorithm of keyword extraction are applied to each document, to extract keywords. The keywords of each document extracted by different methods are aggregated through a proposed approach. Finally, the aggregated keywords, are evaluated by human experts through a web-based system, while to each keyword accepted by experts, a score representing its relationship to the document, is also assigned.
Finally, performance of various keyword extraction algorithms are evaluated by using the resulting dataset, and at the end different strategies for dataset dissemination are introduced.

Keywords: keyword extraction; dataset; Rake method; TextRank method; Information retrieval; TFIDF method; keywords scoring methods.

Iranian Research Institute
for Information Science and Technology
(I r a n D o c)

Research Project

Developing a new standard Persian dataset for automatic keyword extraction from scientific documents

By:

Azadeh Mohebi

Co Investigators:

Marzieh Zarinbal, Ammar Jalalimanesh, Zahra Dehsaraei,

Ali Salimi Sadr, Hamid Hassani

Research Group:

Amir Badamchi, Elaheh Mehrabi

Spring 2021