نشریه علمی «پدافند الکترونیکی و سایبری»
سال هشتم، شماره ۲، تابستان ۱۳۹۹؛ ص ۱۱۴- ۱۰۵

علمی-پژوهشی

استخراج خودکار کلمات کلیدی متون کوتاه فارسی با استفاده از word2vec

امید حاجی‌پور[۱]، سعیده سادات سدیدپور[۲*]

۱- دانشجوی دکتری هوش‌مصنوعی، دانشگاه صنعتی امیرکبیر، ۲- استادیار، دانشگاه صنعتی مالک‌اشتر

(دریافت: ۹۸/۴/۵، پذیرش: ۹۸/۸/۱)

چکیده

با رشد روز افزون اسناد و متون الکترونیکی به زبان فارسی، به کارگیری روش‌هایی سریع و ارزان برای دسترسی به متون مورد نظر از میان مجموعه وسیع این مستندات، اهمیت بیشتری می‌یابد. برای رسیدن به این هدف، استخراج کلمات کلیدی که بیانگر مضمون اصلی متن باشند، روشی بسیار مؤثر است. تعداد تکرار یک کلمه در متن نمی‌تواند نشان‌دهنده اهمیت یک کلمه و کلیدی بودن آن باشد. همچنین در اکثر روش‌های استخراج کلمات کلیدی مفهوم و معنای متن نادیده گرفته می‌شوند. از طرفی دیگر بدون ساختار بودن متون جدید در اخبار و اسناد الکترونیکی، استخراج این کلمات را مشکل می‌سازد. در این مقاله روشی بدون نظارت و خودکار برای استخراج این کلمات در زبان فارسی که دارای ساختار مناسبی نمی‌باشد، پیشنهاد شده است که نه تنها احتمال رخ دادن کلمه در متن و تعداد تکرار آن را در نظر می‌گیرد، بلکه با آموزش مدل word2vec روی متن، مفهوم و معنای متن را نیز درک می‌کند. در روش پیشنهادی که روشی ترکیبی از دو مدل آماری و یادگیری ماشین می‌باشد، پس از آموزش word2vec روی متن، کلماتی که با سایر کلمات دارای فاصله کمی بوده استخراج شده و سپس با استفاده از هم‌رخدادی و فرکانس رابطه‌ای آماری برای محاسبه امتیاز پیشنهاد شده است. درنهایت با استفاده از حدآستانه کلمات با امتیاز بالاتر به‌عنوان کلمه کلیدی در نظر گرفته می‌شوند. ارزیابی‌ها بیانگر کارایی روش با معیار F برابر ۵۳,۹۲٪ و با ۱۱٪ افزایش نسبت به دیگر روش‌های استخراج کلمات کلیدی می‌باشد.

کلیدواژه‌ها: استخراج کلمات کلیدی، زبان فارسی، متن‌کاوی، شباهت کلمات، word2vec

۱- مقدمه

ازآنجایی‌که تعداد مستندات الکترونیکی و درعین‌حال فارسی به سرعت رو به افزایش است، به کارگیری روش‌های کارآمد جهت بازیابی اطلاعات بسیار اهمیت دارد. کلمات کلیدی مجموعه‌ای از لغات مهم در یک سند هستند که توصیفی از محتوای سند را ارائه می‌دهند و برای اهداف مختلف مورد استفاده قرار می‌گیرند. کلمات کلیدی اطلاعات نحوی مفیدی را برای بسیاری از کارهای پردازش متن فراهم می‌کنند. به عبارتی استخراج کلمات کلیدی، فرآیند شناسایی خودکار کلمات به‌کاررفته در یک سند است که مفهوم و معنای متن را به خواننده منتقل کرده و می‌تواند در وظایف مختلف پردازش زبان طبیعی مانند رده‌بندی متون [۱]، خوشه‌بندی متون [۲]، خلاصه‌سازی [۴-۳]، تجزیه و تحلیل متون [۶-۵] و مانند آن مورد استفاده قرار گیرد.

در مجموع، کلمات کلیدی اطلاعات مفیدی برای جستجوی حجم زیادی از مستندات در زمان کوتاه هستند. استخراج کلمات کلیدی به طور دستی فرآیندی بسیار دشوار و زمان‌بر است.

بنابراین، یک فرآیند خودکار که کلمات کلیدی را از مستندات استخراج نماید مورد نیاز می‌باشد.

استخراج خودکار عبارت‌های کلیدی، یک متن بلند را به خلاصه‌ای کوتاه تبدیل می‌کند. به‌عنوان‌مثال، می‌توان از این ویژگی در مرورگرهای وب [۷] استفاده کرد؛ بدین ترتیب که کاربر با فشار دادن یک دکمه، عبارت‌های کلیدی متن را مشاهده و در نتیجه به حوزه‌ی موضوعی متن مورد نظر پی می‌برد.

استخراج کلمات کلیدی یک مسئله بسیار مهم در پردازش زبان فارسی است. در زبان فارسی کلمات دارای صورت‌های نگارشی پیچیده هستند و پوشش کلیه‌ی حالات دستوری کلمات با به کارگیری یک سری قواعد معین، ناممکن است. به همین دلیل استخراج کلمات کلیدی به طور خودکار از متون فارسی دشوار و پیچیده است.

یکی از مشکلاتی که همواره در بسیاری از روش‌های استخراج کلمات کلیدی به چشم می‌خورد، این است که اکثر این روش‌ها برخی از کلماتی که فرکانس پایین دارند را نادیده گرفته و به عنوان کلمه کلیدی شناسایی نمی‌کنند، درحالی‌که ممکن است آن کلمه رساننده مفهوم و معنای واقعی متن بوده و با سایر کلمات دارای رابطه مفهومی و معنایی نزدیکی باشد. از طرفی

*رایانامه نویسنده پاسخگو sadidpour@mut.ac.ir

دیگر بر اساس قانون Zipf، انسان‌ها تمایل دارند کارهای خود را به گونه‌ای ساده‌تر انجام دهند و در نوشتن متون سعی دارند بیشتر از کلمات تکراری استفاده کنند [۸]. در واقع بر اساس این قانون، ارزش کلماتی که تعداد تکرار آن‌ها کمتر است، چندین برابر می‌شود. در این مقاله برای ارزش‌دهی به این نوع کلمات، از word2vec و فاصله آن‌ها با سایر کلمات استفاده شده است تا ارزش این کلمات به مراتب بالا برود. روش پیشنهادی این مقاله، یک روش امتیازدهی بر اساس فاصله و فرکانس کلمات می‌باشد.

در روش پیشنهادی با آموزش word2vec بر روی متن، برای تمامی کلمات جدید و تازه‌وارد در زبان فارسی مانند "برجام" و "FATF"، ارزشی در نظر گرفته می‌شود تا بتوان آن‌ها را از متن استخراج کرد.

در این مقاله روشی پیشنهاد شده است که علاوه بر بدون نظارت بودن و بهره‌مندی از سادگی و دقت، مفهوم متن را نیز توسط مدل word2vec درک می‌کند. از طرفی دیگر کلمات با فرکانس پایین نیز دیده شده و تنها به کلمات با فرکانس بالا اهمیت نمی‌دهد.

بخش دوم این مقاله، به بررسی روش‌های استخراج کلمات کلیدی پرداخته و در بخش سوم مفاهیم استفاده شده به اختصار توضیح داده شده است. سپس، روش پیشنهادی در بخش چهارم شرح داده می‌شود. در بخش پنجم، نتایج بیان گردیده و در نهایت در بخش آخر نتیجه‌گیری ارائه می‌شود.

۲- استخراج کلمات کلیدی

در این بخش به‌طور خلاصه به بررسی روش‌های استخراج کلمات کلیدی پرداخته می‌شود. استخراج این کلمات بر اساس مطالعات انجام شده به چهار دسته کلی تقسیم می‌شوند: روش‌های آماری، یادگیری ماشین، روش‌های مبتنی بر گراف و روش‌های ترکیبی.

۲-۱- روش‌های آماری

روش‌های استخراج کلمات کلیدی آماری بر اساس اطلاعات آماری استخراج شده از متن به‌دست می‌آیند. ازجمله این روش‌ها که تا حدودی قدیمی نیز می‌باشند می‌توان به استفاده از N-تایی‌ها [۹]، TF-IDF [۱۰]، هم‌رخدادی کلمات [۱۱]، فرکانس کلمات [۱۲]، میدان‌های تصادفی شرطی [۱۳] اشاره کرد.

سادگی، مستقل از زبان بودن، پیچیدگی محاسباتی پایین و سرعت به‌نسبت بالا، از مزیت‌های این روش‌ها است؛ اما ضعف عمده این روش‌ها در نظر نگرفتن معنا می‌باشد.

۲-۲- روش‌های یادگیری ماشین

در این روش‌ها، از الگوریتم‌های یادگیری ماشین برای استخراج کلمات کلیدی استفاده می‌شود. از جمله الگوریتم‌هایی که مورد استفاده قرار گرفته‌اند می‌توان به بیزین [۱۴]، ماشین‌های بردار پشتیبان [۱۵] و درخت تصمیم اشاره کرد.

یکی از مشکلات این روش‌ها این است که اکثراً به صورت با نظارت بوده و نیاز به مجموعه‌ی داده برچسب‌خورده دارند. همچنین از این روش‌ها برای استخراج کلمات و ارزیابی روش پیشنهادی در تحقیقات نیز استفاده می‌شود. به‌عنوان‌مثال Gross و Masa انواع الگوریتم‌های یادگیری ماشین را برای ارزیابی روش استخراج کلمات کلیدی استفاده کرده‌اند که الگوریتم C4.5 دارای نتایج بهتری بوده است [۱۶].

۲-۳- روش‌های مبتنی بر گراف

از جمله روش‌های دیگری که برای استخراج این کلمات مطرح می‌باشد، روش‌های مبتنی بر گراف می‌باشند. یکی از مدل‌های مبتنی بر گراف TextRank است که سند را به عنوان یک گراف نشان داده و در آن، کلمات، رأس‌ها و یال‌ها، رابطه‌ی هم‌رخدادی بین رأس‌های متصل‌شده در یک پنجره با اندازه‌ی مشخص می‌باشند [۱۷]. نهایتاً اهمیت هر کلمه با استفاده از روش مبتنی بر گراف PageRank [۱۸] محاسبه می‌گردد. TopicRank بهبودیافته‌ی TextRank می‌باشد که با استفاده از خوشه‌بندی موضوعی عبارات اسمی، منجر به بهبود آن شده است. در این روش، سند به عنوان یک گراف کامل در نظر گرفته شده و رأس‌ها، موضوعات و یال‌ها، رابطه‌ی معنایی بین رأس‌های متصل می‌باشند [۱۹].

همچنین، Tixier و همکاران [۲۰] روشی را پیشنهاد کرده‌اند که در آن پس از ساخت گراف بین کلمات در فضای برداری، کلماتی به‌عنوان کلمه کلیدی استخراج می‌شوند که دارای ارتباط بیشتری با سایر گره‌های گراف باشند. دقت به‌دست‌آمده توسط این روش، در بهترین حالت ۴۹ درصد بوده است؛ اما تنها عامل اهمیت کلمات کلیدی، نمی‌تواند ارتباط آن با سایر کلمات باشد. استخراج کلمات کلیدی به عوامل مختلفی مانند فرکانس، مرکزیت، موقعیت و قدرت همسایگان کلمه کلیدی نیز بستگی دارد [۶].

همچنین Li و همکاران با ترکیب روش‌های مبتنی بر گراف و word2vec روش جدیدی را برای استخراج کلمات کلیدی ارائه کرده و از مقادیر آستانه ۳، ۵، ۷ و ۱۰ برای تعداد کلمات کلیدی استفاده کرده‌اند [۲۱].

در سال ۲۰۱۸ Biswas و همکاران، مدلی مبتنی بر گراف را برای تجزیه و تحلیل داده‌های توییتر پیشنهاد کردند که در آن کلمات کلیدی با استفاده از وزن کلاسیک استخراج شده و اهمیت یک کلمه کلیدی به‌صورت جمعی با استفاده از مؤلفه‌های مختلف تأثیرگذار تعیین می‌شود [۶].

یکی از اشکالات عمده شیوه‌های مبتنی بر گراف، مسئله "ثروتمندتر شدن ثروتمندان" است. بدین معنی که کلماتی که به تعداد بیشتری از کلمات دیگر متصل هستند، اعتبار بیشتری به‌دست می‌آورند. در مقابل، کلمات مهم کم تکرارتر، امکان دست‌یابی به امتیاز بالا نخواهند داشت. امکان تشخیص و به‌کارگیری ارتباطات معنایی، ساختاری و دستوری بین کلمات، مهم‌ترین ویژگی مثبت این شیوه‌ها است.

۲-۴- روش‌های ترکیبی

در این روش‌ها از ترکیب چند روش استفاده می‌شود. این رویکردها هر کدام از روش‌های مذکور را ترکیب و یا از اکتشافات استفاده می‌کنند. ازجمله این روش‌ها می‌توان به موقعیت و طول کلمات و تگ‌های HTML [۲۲] در اطراف آنها اشاره کرد. این الگوریتم‌ها برای به‌دست آوردن بهترین ویژگی‌ها از رویکردهای ذکر شده طراحی شده‌اند. به‌عنوان‌مثال در سال ۲۰۰۸، Wan و Xiao روشی را بر اساس فاصله‌ی کسینوسی نزدیک‌ترین همسایه و TF-IDF ارائه کردند. آن‌ها با استفاده از معیار شباهت فاصله کسینوسی یک یا چند سند که به سند هدف نزدیک‌تر بودند را اضافه و با TF-IDF کلمات کلیدی آن را استخراج کردند [۲۳].

در سال ۲۰۱۸ Naidu و همکاران با استفاده از فرکانس کلمات و برچسب اجزای کلام، روشی را برای استخراج کلمات کلیدی و استفاده از آن در خلاصه‌سازی ارائه دادند [۲۴].

۳- مفاهیم اولیه

در این بخش روش‌ها و مفاهیم استفاده شده، توضیح داده شده‌اند.

۳-۱- قانون Zipf

زیف، استاد زبان‌شناسی دانشگاه هاروارد، در سال ۱۹۴۹ با آزمایش کلمات کتاب جویس در مورد کلمات و میزان تکرار آن‌ها در متن، به این نتیجه دست یافتند که اگر تمام کلمات یک کتاب شمرده و از زیاد به کم مرتب شوند، فراوانی (بسامد) همان کلمه، با رتبه‌ی هر کلمه نسبت عکس خواهد داشت؛ یعنی تعداد دفعاتی که هر کلمه در متن ظاهر می‌شود، با رتبه‌ی همان کلمه در متن رابطه معکوس دارد. این نسبت برای کلمات کل متن برقرار است؛ که به قانون zipf معروف شده است[۸].

او این قضیه را با اصل کمترین کوشش توجیه کرد. انسان‌ها بر اساس این اصل تمایل دارند کارهای خود را به‌گونه‌ای ساده‌تر انجام دهند و در نوشتن متن سعی دارند بیشتر از کلمات تکراری استفاده کنند؛ هم‌چنین در هنگام صحبت کردن و سخنرانی سعی دارند کلمات کمتری را بیشتر تکرار کنند.

این رابطه بین فراوانی f و رتبه r برقرار است که رابطه‌ی لگاریتمی آن (در فرمول (۱))، شناخته‌شده‌تر بوده و کاربرد بیشتری دارد.

$$\log r + \log f = \log c \qquad (۱)$$

۳-۲- word2vec

روش word2vec، روشی برای تبدیل کلمات به بردار است که در سال ۲۰۱۳ توسط گوگل پیشنهاد شد. این روش برای نمایش لغات و متون و پردازش آن‌ها، بسیار کارآمد و مناسب است. هدف این روش، نمایش برداری کلمات است که می‌تواند در بسیاری از کاربردهای نوین پردازش متن مانند سنجش احساسات، جستجوی متون مشابه، پیشنهاد اخبار یا کالای مشابه استفاده شود [۲۵].

در این روش به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظرگرفته شده و این بردار با اعداد مناسب در فاز آموزش مدل، برای هر لغت به‌دست می‌آید. در این بردار هر ستون، نمایشگر کلمه یا ویژگی خاصی نیست و فقط یک عدد را نمایش می‌دهد. اگر این بردار ۲۰۰تایی فرض شود، یک فضای ۲۰۰ بعدی ایجاد شده است که هر لغت در این فضا یک نمایش منحصربه‌فرد خواهد داشت. بعد از ایجاد بردارهای مرتبط با هر لغت، برای نمایش برداری هر متن یا خبر، می‌توان بردار تک‌تک کلمات به‌کاررفته در آن را یافته و میانگین اعداد هر ستون را به دست‌آورد که نتیجه آن یک بردار برای هر متن یا سند خواهد بود. این الگوریتم برای ساخت بردارهای کلمات از یکی از دو روش زیر استفاده می‌کند:

- کیف لغات پیوسته (CBOW)
- اسکیپ‌گرام (skip-gram)

در روش کیف لغات پیوسته، ابتدا به‌ازای هر لغت یک بردار با طول مشخص و با اعداد تصادفی (بین صفر و یک) تولید می‌شود. سپس به‌ازای هر کلمه از یک سند یا متن، تعدادی مشخص از کلمات بعد و قبل از آن (به غیر از خود لغت فعلی) به شبکه عصبی داده شده و با عملیات ساده ریاضی، بردار لغت فعلی را تولید می‌کند.

سپس، این اعداد با مقادیر قبلی بردار لغت جایگزین می‌شوند. زمانی که این کار بر روی تمام لغات در تمام متون انجام گیرد، بردارهای نهایی لغات همان بردارهای مطلوب هستند.

روش اسکیپ‌گرام برعکس کیف لغات پیوسته کار می‌کند. هدف این روش این است که بر اساس یک لغت مشخص شده، چند لغت قبل و بعد آن تشخیص داده شود و با تغییر مداوم اعداد بردارهای لغات، نهایتاً به یک وضعیت باثبات برسد.

TF-IDF -۳-۳

در این روش به لغات یک وزن بر اساس فراوانی آن در سند داده می‌شود. در واقع این سیستم وزن‌دهی نشان می‌دهد چقدر یک کلمه برای یک سند مهم است. این مسئله کاربردهای بسیاری در بازیابی اطلاعات دارد. وزن کلمه با افزایش تعداد تکرار آن در متن افزایش می‌یابد، اما توسط تعداد کلمات در متن کنترل می‌شود، زیرا در صورت زیاد بودن طول متن، بعضی از کلمات که ممکن است چندان اهمیتی هم در معنی نداشته باشند، به طور طبیعی بیشتر از سایر کلمات تکرار خواهند شد[۲۷-۲۶].

وزن عبارات با استفاده از دو معیار تکرار اصطلاح (TF) و معکوس تکرار در سند (IDF) تعیین می‌شود. اگر فرض شود تعداد دفعاتی که کلمه t در سند d اتفاق افتاده با TF(t,d) نشان داده شود، در ساده‌ترین حالت تعداد تکرار اولیه t با f(t,d) نمایش داده خواهد شد که مطابق رابطه (۲) می‌باشد.

$$TF(t,d) = f(t,d) \qquad (۲)$$

همچنین می‌توان برای محاسبه TF از فرمول (۳) نیز استفاده کرد:

$$TF(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}} \qquad (۳)$$

IDF معیاری است برای عباراتی که در کلیه سندها بسیار متداول هستند و معمولاً تکرار می‌شوند. این پارامتر نیز از فرمول (۴) محاسبه می‌گردد.

$$IDF(t,D) = \log\left(\frac{D}{1 + (d \in D; t \in d)}\right) \qquad (۴)$$

که در آن، D تعداد کل سندها، d یک سند خاص از مجموعه‌ی اسناد و t یک کلمه از سند d می‌باشد.

براساس فرمول‌های (۲) یا (۳) و (۴)، فرمول نهایی برای محاسبه TF-IDF هر عبارت به‌صورت فرمول (۵) خواهد بود.

$$TF\_IDF(t,d,D) = TF(t,d) \times IDF(t,D) \qquad (۵)$$

TextRank -۴-۳

الگوریتم‌های رتبه‌بندی مبتنی بر گراف، در اساس روش‌های امتیازدهی به یک رأس در گراف می‌باشند. TextRank نیز جزو این دسته روش‌ها می‌باشد که در استخراج کلمات کلیدی کاربرد دارد. ایده اساسی که توسط یک مدل رتبه‌بندی مبتنی‌بر گراف اجرا می‌شود، "رأی دادن" یا "توصیه" است. هنگامی‌که یک ریشه به رأسی دیگر پیوند دارد، اساساً رأی دادن برای آن رأس دیگر است. هرچه تعداد رأی‌هایی که برای یک رأس ارزیابی می‌شوند بیشتر باشد، اهمیت رأس بالاتر است. علاوه بر این، اهمیت ارجاع به رأی، تعیین‌کننده اهمیت خود رأی بوده و این اطلاعات نیز با توجه به مدل رتبه‌بندی به‌دست می‌آید [۱۷].

به‌طورکلی، TextRank یک گراف از کلمات و روابط بین آن‌ها را از یک سند ایجاد می‌کند، سپس مهم‌ترین رأس‌های گراف (کلمات) را بر اساس امتیازات اهمیت محاسبه‌شده به‌صورت مجزا از کل گراف مشخص می‌کند. اگر G=(V,E) گراف تولید شده باشد، V مجموعه رئوس و E مجموعه یال‌ها بوده که E زیرمجموعه‌ی V*V است. برای هر رأس Vi، In (Vi) تعداد یال‌های ورودی به این رأس و Out (Vi) یال‌های خروجی از آن خواهد بود. امتیاز رأس Vi مطابق رابطه‌ی (۶) به‌دست می‌آید.

$$S(Vi) = (1-d) + d \times \sum_{j \in In(Vi)} \frac{1}{|Out(Vj)|} S(Vj) \qquad (۶)$$

که در آن، d یک عامل محرک بین ۰ و ۱ بوده و می‌تواند با پرش از یک رأس به سایر رئوس موجود در گراف، یکپارچه‌سازی در مدل را ایجاد کند. این الگوریتم برای هر رأس چندین بار تکرار شده تا در نهایت امتیاز رئوس به ثبات برسند.

Yake -۵-۳

Yake روشی جدید برای استخراج کلمات کلیدی متن بوده که بر ویژگی‌های استخراج‌شده از متن تکیه کرده و بنابراین، به اسناد موجود در بسیاری از زبان‌های مختلف بدون نیاز به دانش خارجی اعمال می‌شود. این موضوع می‌تواند برای تعداد زیادی از وظایف و در مواقع بسیاری که در آن، دسترسی به آموزش پیکره محدود است، مفید باشد [۲۸]. این سیستم علاوه بر استخراج کلمات کلیدی، قادر به استخراج عبارات کلیدی نیز می‌باشد.

۶ مرحله کلی این روش عبارتند از: ۱- پیش‌پردازش متن؛ ۲- استخراج ویژگی؛ ۳- محاسبه امتیاز تک کلمه‌ای‌ها با استفاده از فرکانس کلمه (TF)؛ ۴- تولید فهرست کلمات کلیدی کاندید؛ ۵- حذف داده‌های تکراری؛ و ۶- محاسبه امتیاز کلمات کلیدی.

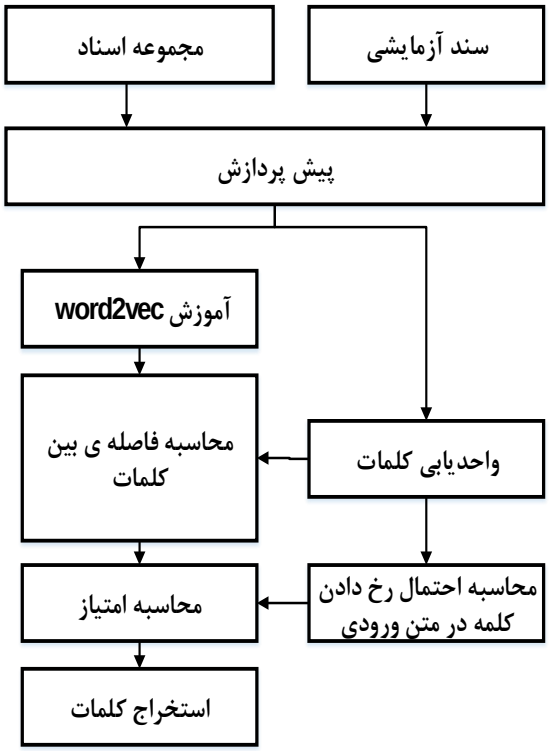
امتیاز کلمات کلیدی در این مدل بر اساس رابطه (۷) به دست می‌آید.

$$S(kw) = \frac{\prod_{w \in kw} S(w)}{TF(kw) \times (1 + \sum_{w \in kw} S(w))} \quad (۷)$$

که در آن، S(w) امتیاز عبارات با استفاده از n-تایی‌ها و با طول پنجره ۳ می‌باشد. S(kw) نیز امتیاز کلمه کلیدی کاندید بوده و هر چه امتیاز کلمه کلیدی کمتر باشد، به معنای اهمیت بیشتر آن است. S(kw) توسط اعمال S(w) بر روی آن به‌دست می‌آید که امتیاز اولین بخش از عبارت ضرب در سایر بخش‌های عبارت می‌باشد.

۴- استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار

در این مقاله روشی جدید برای استخراج کلمات کلیدی با استفاده از word2vec و احتمال رخ دادن کلمات در متن ارائه شده است. سامانه پیشنهادی شامل ۵ مرحله اصلی پیش‌پردازش، محاسبه احتمال رخ دادن کلمه در متن ورودی، آموزش word2vec، محاسبه فاصله کلمات و سپس تعیین امتیاز برای هر کلمه می‌باشد که روال کلی آن در شکل (۱) نشان داده شده است:



شکل (۱): روش پیشنهادی برای استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار کلمات.

در روش پیشنهادی پس از پیش‌پردازش مجموعه داده، مدل word2vec با اندازه بردار تعبیه ۳۰۰ آموزش دیده و سپس برای هر متن ورودی فاصله هر کلمه با سایر کلمات و احتمال رخ دادن آن کلمه در متن محاسبه می‌شود. در گام بعد، امتیاز به دست آمده و کلمات با امتیاز کمتر به‌عنوان کلمات کلیدی استخراج می‌شوند. استفاده از word2vec باعث کاهش تأثیر فرکانس شده و این مطابق قانونی است که zipf بیان کرده است؛ به عبارت دیگر در صورت زیاد بودن طول متن، بعضی از کلمات که ممکن است چندان اهمیتی هم در معنی نداشته باشند، به طور طبیعی بیشتر از سایر کلمات تکرار خواهند شد [۲۴] که تأثیر این کلمات با استفاده از word2vec کم‌رنگ می‌گردد.

۴-۱- مجموعه داده

برای ارزیابی نتایج روش پیشنهادی و سایر روش‌ها از ۲۰۰۰ سند از سایت خبری yjc که کلمات کلیدی آن مشخص بوده، استفاده شده است. از آنجایی‌که کلمات کلیدی این سایت قطعی نبوده، پس از بازبینی مجدد توسط دو فرد خبره اصلاحاتی روی این کلمات صورت گرفته است.

۴-۲- پیش‌پردازش داده‌ها

یکی از مراحل مهم داده‌کاوی، پیش‌پردازش دادگان است. پیش‌پردازش، داده را به قالب مناسب برای داده‌کاوی تبدیل کرده و روند محاسبات و استخراج اطلاعات را تسریع و ساده می‌کند [۲۹].

پردازش زبان فارسی از جهاتی با پردازش زبان انگلیسی تفاوت دارد. در زبان انگلیسی تمامی حروف و تمامی کلمات جدا از هم و با قانونی مشخص نوشته می‌شوند و این در حالی است که در زبان فارسی بعضی از حروف به هم چسبیده هستند، برخی از حروف جدا از هم نوشته می‌شوند، بعضی از کلمات یکپارچه‌اند، بعضی از کلمات با فاصله یا نیم‌فاصله به دو یا چند بخش تقسیم می‌شوند. تمامی حوزه‌های مرتبط با پردازش زبان طبیعی به نحوی با متون واقعی سروکار دارند.

اگر حروف، نشانه‌های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورد استفاده قابل تحلیل توسط سامانه‌های رایانه‌ای نخواهند بود. به‌عنوان‌مثال اگر نرمال‌سازی روی دو داده "رئیس‌جمهور" و "رییس‌جمهور" اعمال نشود، سیستم این دو را دو عبارت جدا در نظر می‌گیرد که روی نتایج تأثیر زیادی دارد [۳۰].

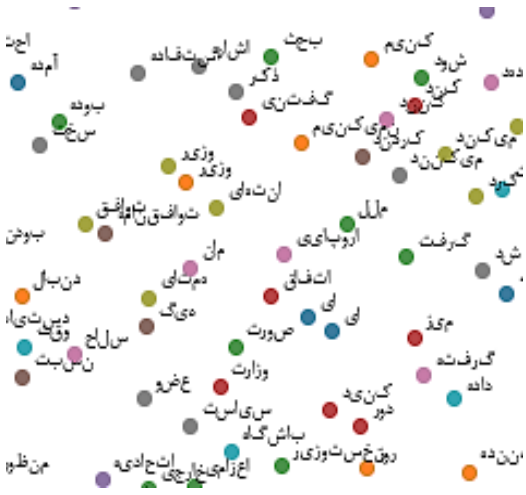
طی فرایند نرمال‌سازی، علائم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات و غیره بدون ایجاد تغییرات معنایی در متن به شکل استاندارد تبدیل می‌گردند. بنابراین، بایستی از یک استاندارد مشترک برای پیش‌پردازش و پردازش متون استفاده کرد. همچنین کارکترها و کلمات زائد و توقف مانند "و"، "به"،

"از" و امثال آن حذف شده و کاراکترها نرمال می‌شوند. به‌عنوان‌مثال، برای دو کلمه «مسئله» و «مسأله»، کل متن نرمال شده و یکی از این دو و یا یک کلمه جایگزین مانند «مساله» به عنوان کلمه مرجع انتخاب می‌شود. از این قبیل کلمات می‌توان «رییس» و «رئیس»، کلماتی که دارای حروف "ی" و "ک" عربی می‌شوند را نام برد [۳۰].

همچنین فاصله بین کلمات و یا عبارات نرمال شده و همه به یک فاصله تبدیل می‌شوند. به‌عنوان‌مثال "میرفت"، "می‌رفت" و "می-رفت" هر سه دارای یک معنا و مفهوم هستند، اما اگر پردازش روی آن‌ها صورت نگیرد، دارای نتایج متفاوتی خواهند بود [۳۰].

۴-۳- آموزش word2vec روی مجموعه داده

در این مرحله مدل word2vec روی کل مجموعه داده آموزش می‌بیند. به این صورت مدل ساخته شده تمامی کلمات موجود در مجموعه داده را مشاهده کرده و می‌تواند برای آن‌ها بردار تولید کند که در شکل (۲) می‌توان خروجی این قسمت را بر روی بخشی از این مجموعه داده مشاهده کرد.



شکل (۲): اجرای مدل word2vec بر روی مجموعه داده.

تعداد رخداد کلمات دارای اهمیت است اما این موضوع به صرف نمی‌تواند در تعیین کلمه کلیدی به تنهایی نقش داشته باشد؛ زیرا ممکن است کلمه‌ای که دارای فرکانس پایین می‌باشد، دارای مفهوم متن بوده و وابستگی با سایر کلمات پرتکرار و اصلی متن داشته باشد. به همین دلیل از مدل word2vec برای ساخت بردار تعبیه کلمات استفاده شده است؛ زیرا این مدل علاوه بر تکرار، وابستگی بین کلمات و مفهوم را نیز درک می‌کند.

علاوه بر این کلمه‌ای که می‌تواند جزو کلمات کلیدی باشد، فاصله کمتری نسبت به سایر کلمات داشته و در اطراف آن تراکم کلمات بیشتر است. به‌عنوان‌مثال با استفاده از word2vec که تنها روی داده‌های فارسی آموزش دیده است، دیگر نمی‌توان کلمات انگلیسی مانند "FATF" را استخراج کرد. به همین علت روش پیشنهادی روشی پویاست و قادر به استخراج هر نوع کلمه کلیدی می‌باشد.

۴-۴- احتمال رخ دادن کلمات در متن

دلیل استفاده از تعداد رخداد برای اندازه‌گیری رتبه اهمیت، بر این باور استوار است که نویسنده معمولاً از واژگان معینی برای پیشبرد، بحث یا تشریح دقیق جنبه‌های مختلف موضوع موردنظر استفاده و آن‌ها را تکرار می‌کند. تعداد رخداد هر واژه می‌تواند به عنوان عامل تعیین درجه اهمیت واژگان مورد استفاده قرار گیرد. هرچند که این موضوع دارای اهمیت است، اما همان‌طور که در بخش ۴-۳ بیان شد، نمی‌تواند به تنهایی در تعیین این کلمات نقش داشته باشد.

در این مرحله پس از آموزش word2vec، برای متن ورودی پس از واحدیابی کلمات، احتمال رخ دادن هر کلمه در متن به‌صورت زیر محاسبه می‌گردد.

اگر فرض شود که کلمه w از متن خارج شده و قرار است برای آن امتیاز محاسبه گردد، احتمال رخ دادن این کلمه در متن از رابطه (۸) محاسبه می‌شود.

$$p(w) = count(w) / C \qquad (۸)$$

که در آن، count(w) تعداد تکرار کلمه w در متن و C تعداد کل کلمات متن ورودی می‌باشد.

۴-۵- محاسبه فاصله کلمات و امتیاز

بعد از محاسبه احتمال رخ دادن کلمه در متن، فاصله کلمه با تمامی کلمات موجود در متن آزمایشی با استفاده از مدل word2vec ساخته شده و فاصله کسینوسی محاسبه شده و میانگین فاصله کلمه w با سایر کلمات از معادله (۹) به‌دست می‌آید.

$$dis(w) = (\sum_i (dw \cdot dw_i) / |dw| . |dw_i|) / C \qquad (۹)$$

صورت کسر برابر میانگین فاصله یک کلمه با سایر کلمات متن و مخرج آن، C تعداد کل کلمات متن ورودی است. همچنین i به ازای تمام کلمات متن ورودی تغییر می‌کند.

پس از محاسبه dis(w) و p(w)، امتیاز کلمه w مطابق رابطه (۱۰) محاسبه خواهد شد.

$$score(w) = \log_2(p(w)) \times dis(w) \qquad (۱۰)$$

گرفتن لگاریتم از p(w)، باعث متعادل‌سازی و کم کردن تأثیر تعداد تکرار کلمات روی امتیاز آن‌ها می‌شود. باید توجه داشت که در محاسبه احتمال رخ دادن کلمات، احتمال هر کلمه بین صفر و یک نرمال شده و وقتی لگاریتم گرفته شود، از نظر مفهومی نیز رابطه معکوس می‌شود. در واقع کلماتی که (log(p(w آن‌ها کمتر باشد، اهمیت بیشتری دارند. در نهایت کلمات مرتب شده و با استفاده از یک حد آستانه، ۵ کلمه‌ای که دارای امتیاز کمتر باشند، به‌عنوان کلمات کلیدی متن به‌دست می‌آیند.

۵- نتایج

برای بررسی نتایج، مدل پیشنهادی با دو روش TextRank و TF-IDF مقایسه شده است. علاوه بر این، از سیستم جدیدی به نام Yake که به‌تازگی معرفی شده استفاده گردیده است [۲۸].

برای ارزیابی نیز از سه معیار دقت، بازخوانی و معیار F- استفاده شده که دو معیار دقت و بازخوانی به‌صورت فرمول‌های (۱۱) و (۱۲) می‌باشد. فرمول (۱۱) مربوط به دقت و فرمول (۱۲) مربوط به بازخوانی است.

$$\Pr ecision = TP / (TP + FP) \qquad (۱۱)$$

$$\Re call = TP / (TP + FN) \qquad (۱۲)$$

پارامترهای این دو فرمول نیز به‌صورت زیر تعریف می‌شود:

TP: تعداد عبارت‌های کلماتی که فرد خبره از مجموعه اسناد استخراج کرده و روش مورد نظر نیز آن را به درستی استخراج می‌کند.

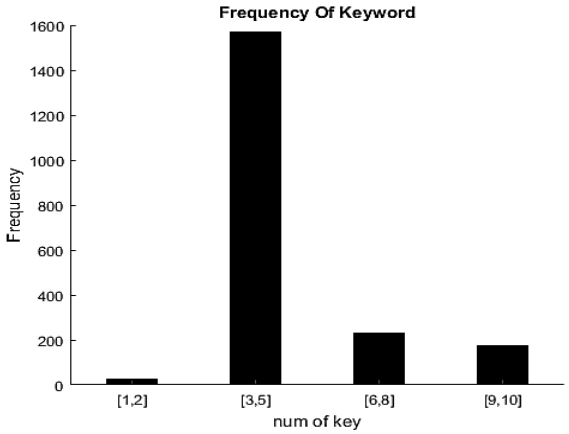
FN: تعداد کلماتی که فرد خبره از مجموعه اسناد استخراج کرده ولی روش مورد نظر قادر به استخراج آن نیست.

FP: تعداد کلماتی که در مجموعه عبارات استخراج شده توسط فرد خبره نیست، ولی روش مورد نظر آن را به‌عنوان کلمه تشخیص داده است.

همچنین معیار F- مطابق فرمول (۱۳) می‌باشد:

$$F = 2 \times \frac{\Pr ecision \times \Re call}{\Pr ecision + \Re call} \qquad (۱۳)$$

برای بررسی نتایج، از آنجا که حدآستانه مشخصی برای کلمات کلیدی در اسناد وجود ندارد، از شکل (۳) مشخص است که بهترین حد آستانه با توجه به اسناد، ۵ می‌باشد. البته در آزمایش‌ها از حد آستانه ۱۰ نیز استفاده شده است.



شکل (۳): فرکانس کلمات کلیدی اسناد.

نتایج به‌دست‌آمده از روش‌ها با حد آستانه ۵ در جدول (۱) نشان داده شده است:

جدول (۱): نتایج حاصل از روش‌ها با حدآستانه ۵٪.

مدل مورد استفاده	دقت	بازخوانی	معیار F
TF-IDF	۱۴/۵۱	۳۴/۰۷	۲۰/۳۶
TextRank	۱۹/۴۶	۲۳/۴۸	۲۱/۲۸
Yake	۳۷/۴۸	۴۸/۴۵	۴۲/۲۶
استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار	۴۹/۲۳	۶۰/۱۴	۵۳/۹۲

در استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار با توجه به توضیحات بخش ۴-۵، این مسئله برعکس بوده و ۵ کلمه‌ی با امتیاز کمتر انتخاب می‌شوند. از نتایج پیداست که روش پیشنهادی نسبت به سایر روش‌ها دارای نتایج بهتری است. در روش TF-IDF و TextRank معنا و مفهوم متن در نظر گرفته نمی‌شود و به کلماتی که تعداد تکرار بیشتری دارند، اهمیت بیشتری داده می‌شود. روش Yake نیز جزو روش‌های جدیدی است که برای هر زبانی جواب مطلوب می‌دهد، اما ازآنجایی‌که بر اساس ویژگی‌های استخراج شده از متن کار کرده و معمولاً متون فارسی دارای ساختارهای مناسبی نیستند، نمی‌توانند به خوبی سایر زبان‌ها، خروجی داشته باشد. با این حال یکی از بهترین روش‌ها برای استخراج کلمات کلیدی است.

اما در روش پیشنهادی تنها به کلمات پرتکرار اهمیت داده نشده و با فرمول (۱۰)، تأثیر این تکرار تا حدودی کم می‌شود. از طرفی دیگر این روش بر اساس این اصل کار می‌کند که کلمات مهم متن، با سایر کلمات در ارتباط بوده و می‌تواند مفهوم متن را

منتقل کند. در واقع این بدین معناست که کلمات مهم در فضای برداری، فاصله کمتری با سایر کلمات خواهند داشت. بر همین اساس مدل word2vec با توجه به مفهوم متن بردار تعبیه کلمات را می‌سازد و پس از محاسبه امتیاز، کلماتی که تکرار بیشتری داشته و فاصله آن‌ها با سایر کلمات کمتر باشد، استخراج می‌شوند.

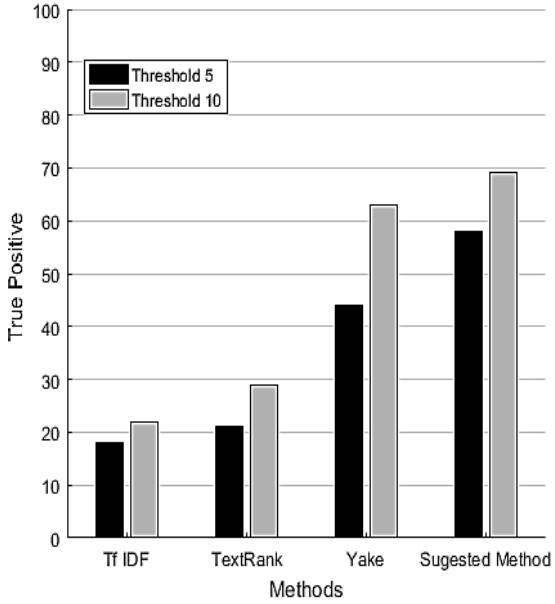
معیار دقت در روش‌های مورد آزمایش به این دلیل کم است که در مجموعه اسناد مورد استفاده، تمامی اسناد دارای تعداد کلمه کلیدی ثابت نبوده و برای هر سند متفاوت است. مثلاً برای یک سند ممکن است ۳ کلمه کلیدی و برای سندی دیگر ۷ کلمه کلیدی موجود باشد. به همین دلیل حدآستانه برابر ۵ در نظر گرفته شده تا کلمات از دست نروند و تا حد امکان کلمات کلیدی واقعی استخراج شوند.

از معیار فراخوانی نیز مشخص است که سیستم پیشنهادی با توانایی بسیار بیشتری نسبت به سایر روش‌ها قادر به استخراج این کلمات می‌باشد و در نهایت معیار F گویای بهتر بودن روش پیشنهادی نسبت به سایر روش‌ها است. برای ارزیابی بهتر، نتایج با حدآستانه ۱۰ نیز بررسی شده‌اند که در جدول (۲) نشان داده شده است.

جدول (۲): نتایج حاصل از روش‌ها با حدآستانه ۱۰ (٪).

مدل مورد استفاده	دقت	بازخوانی	معیار F
TF-IDF	۹/۷۸	۴۸/۲۱	۱۶/۲۶
TextRank	۱۱/۵۳	۴۲/۴۹	۱۸/۱۳
Yake	۲۴/۲۸	۶۲/۹۶	۳۵/۰۵
استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار	۲۶/۱۹	۶۷/۹۰	۳۷/۸۰

ازآنجا که در حد آستانه ۱۰، کلمات بیشتری انتخاب شده، تعداد کلمات کلیدی درست انتخاب شده افزایش میابد و این امر باعث افزایش معیار بازخوانی می‌گردد؛ اما با توجه به شکل (۳)، چون تعداد کلمات کلیدی اکثر استاد بین ۳ تا ۵ می‌باشد، مابقی کلمات FP بوده و باعث کاهش شدید معیار دقت می‌گردند؛ اما با توجه به معیار F مشخص است که همچنان روش پیشنهادی نسبت به سایر روش‌ها دارای نتایج بهتری است. همچنین درصد کلمات کلیدی درست تشخیص داده شده در روش‌های مورد آزمایش، در شکل (۴) نشان داده شده است.



شکل (۴): نرخ کلمات درست استخراج شده برای هر روش (٪).

با توجه به درصدها مشخص است که با افزایش حدآستانه، تعداد کلمات کلیدی درست تشخیص داده شده (TP) افزایش میابد، اما با توجه به فرکانس کلمات کلیدی هر سند در شکل (۳)، این امر باعث افزایش FP نیز می‌شود. به همین علت معیار F برای حدآستانه ۱۰ کلمه کمتر از ۵ کلمه می‌باشد.

برای درک بهتر این مسئله نمونه‌ای از کلمات کلیدی استخراج شده در جدول (۳) با دو روش Yake و پیشنهادی برای حد آستانه ۱۰ آورده شده که علت انتخاب این دو روش، معیار F بهتر آن‌ها می‌باشد.

جدول (۳): کلمات کلیدی واقعی و استخراج شده با حدآستانه ۱۰.

حدآستانه	کلمات کلیدی اصلی	Yake	استخراج کلمات کلیدی با استفاده از word2vec و تعداد تکرار
۱	نفت	نفت	نفت
۲	قیمت	سنت	دلار
۳	دلار	دلار	قیمت
۴	بازار	قیمت	ایران
۵		فروش	فروش
۶		سازندگان	سنت
۷		تجهیزات	بازار
۸		ایران	ارزش
۹		انجمن	صنعت
۱۰		صنعت	رئیس

با توجه به جدول (۳)، تعداد کلمات کلیدی اصلی این سند ۴ می‌باشد که بر اساس شکل (۳) مشخص شد تعداد کلمات کلیدی در اکثر سندها در بازه ۳ تا ۵ قرار دارد. با این‌حال اسنادی نیز وجود دارند که تعداد کلمه کلیدی بیشتر یا کمتر از این تعداد دارند.

ازآنجایی‌که ترتیب کلمات بر اساس امتیاز و اهمیت کلیدی بودن کلمه می‌باشد، اگر حد آستانه ۵ در نظر گرفته شود، روش پیشنهادی و Yake دارای یک دقت خواهند بود و هر دوی این روش‌ها ۳ کلمه از ۴ کلمه اصلی را به درستی تشخیص می‌دهند؛ اما نکته دارای اهمیت این است که روش پیشنهادی این کلمات را با اولویت بالاتر (اولویت‌ها بر اساس امتیازها مشخص می‌شوند) به‌دست آورده و تا حدودی اهمیت کلیدی بودن کلمه را در نظر گرفته است. در مثال مذکور، ازآنجایی‌که دو روش، ۳ کلمه از ۴ کلمه را درست تشخیص داده‌اند، نرخ تشخیص کلمات درست (TP) ۳ می‌باشد، اما با توجه به حدآستانه ۵ یک کلمه دیگر به‌عنوان FP و کلمه تشخیص داده نشده توسط سیستم به عنوان FN در نظر گرفته می‌شود.

حال اگر حد آستانه ۱۰ در نظر گرفته شود، کلمه ۴ام ("بازار") که جزو کلمات کلیدی اصلی می‌باشد، توسط روش پیشنهادی تشخیص داده شده، درحالی‌که روش Yake قادر به تشخیص آن کلمه نبوده است. این امر افزایش TP و کاهش FN در روش پیشنهادی را به همراه دارد (به عبارت دیگر TP به مقدار ۴ و FN به مقدار صفر تغییر می‌یابد). اما نکته مهم این است که تعداد کل کلمات کلیدی این سند ۴ بوده و در واقع ۶ کلمه دیگر FP محسوب می‌شوند. به همین علت با توجه به مقایسه جدول‌های (۱) و (۲)، معیار F با حدآستانه ۱۰ نسبت به حد آستانه ۵ کاهش یافته، در حالی که با توجه به شکل (۴) دارای TP بیشتری می‌باشد.

۶- نتیجه‌گیری

به‌طورکلی، استخراج کلمات کلیدی از دادگان متنی یکی از مهم‌ترین و پرکاربردترین مسائل در متن‌کاوی و بازیابی اطلاعات بوده و در اغلب امور مرتبط با محتوای متن، تشخیص کلمات کلیدی نقش تعیین‌کننده‌ای دارد.

روش پیشنهادی در این مقاله روشی ترکیبی از مدل‌های آماری و یادگیری ماشین بوده و مزیت اصلی آن، یک‌دستی و یک‌نواختی کلمات کلیدی است. به دلیل دخالت نداشتن توانایی‌ها و تمایلات انسان‌ها، کلمات استخراج شده از شایستگی، یکدستی و پویایی برخوردارند. هنگامی‌که کلمات کلیدی در سطح وسیع در اختیار کاربران قرار گیرد، کاربران یاد خواهند گرفت که چگونه آن‌ها را درک کنند و چگونه مفاهیم مورد نظر خود را تشخیص دهند. البته، این احتمال نیز وجود دارد که با توجه به سبک نگارش نویسنده در گسترش مطالب، کلمات درجه دو انتخاب شوند یا نتایج حاصل نامناسب باشند، اما می‌توان حالت‌های استثنائی را به سیستم آموزش داد.

همچنین سیستم پیشنهادی قادر به استخراج کلمات کلیدی جدید در متن به صورت خودکار و بدون نظارت بوده که این نوع کلمات حتی برای خود انسان نیز می‌تواند ناشناخته باشد. در واقع نیاز به دانش خارجی نداشته و سیستم به‌صورت خودکار معنا و مفهوم را می‌آموزد و کلمات کلیدی را با استفاده از روش پیشنهادی استخراج می‌کند. همچنین سیستم پیشنهادی قادر به اولویت دادن به کلمات کلیدی با اهمیت بالاتر می‌باشد.

۷- مراجع

[1] Z. Wu, et al., "An Efficient Wikipedia Semantic Matching Approach to Text Document Classification," Information Sciences, vol. 393, pp. 15-28, 2017.

[2] C. Jia, et al., "Concept Decompositions for Short Text Clustering by Identifying Word Communities," Pattern Recognition, vol. 76, pp. 691-703, 2018.

[3] S. K. Bharti and K. S. Babu, "Automatic Keyword Extraction for Text Summarization: a Survey," arXiv preprint arXiv:1704.03242, 2017.

[4] M. Yousefi-Azar and L. Hamey, "Text Summarization Using Unsupervised Deep Learning," Expert Systems with Applications, vol. 68, pp. 93-105, 2017.

[5] Z. Sepehrian and H. Shirazi, "A New Way To Summarize Persian Texts Based on User Query Expression," Electronic and Cyber Defense, 2018.

[6] S. K. Biswas, M. Bordoloi, and J. Shreya, "A Graph Based Keyword Extraction Model Using Collective Node Weight," Expert Systems with Applications, vol. 97, pp. 51-59, 2018.

[7] R. Harakawa, T. Ogawa, and M. Haseyama, "Extraction of Hierarchical Structure of Web Communities Including Salient Keyword Estimation for Web Video Retrieval," in Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, 2015.

[8] G. Zipf, "Human Behaviour and The Principle of Least-Effort," Cambridge MA edn. Reading: Addison-Wesley, 1949.

[9] B. Das, et al., "Automatic Keyword Extraction From any Text Document Using N-Gram Rigid Collocation," Int. J. Soft Comput. Eng.(IJSCE), vol. 3(2), pp. 238-242, 2013.

[10] J. Li and K. Zhang, "Keyword Extraction Based on Tf/Idf for Chinese News Document," Wuhan University Journal of Natural Sciences, vol. 12(5), pp. 917-921, 2007.

[11] Y. Matsuo and M. Ishizuka, "Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information," International Journal on Artificial Intelligence Tools, vol. 13(01), pp. 157-169, 2000.

[12] S. Rose, et al., "Automatic Keyword Extraction From Individual Documents," Text Mining: Applications and Theory, pp. 1-20, 2010.

[13] C. Zhang, "Automatic Keyword Extraction From Documents Using Conditional Random Fields," Journal of Computational Information Systems, vol. 4(3), pp. 1169-1180, 2008.

[14] E. Frank, et al., "Domain-Specific Keyphrase Extraction," In 16th International joint conference on artificial intelligence (IJCAI 99). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1999.

[15] K. Zhang, et al., "Keyword Extraction Using Support Vector Machine," In International Conference on Web-Age Information Management, Springer, 2006.

[16] Y. HaCohen-Kerner, Z. Gross, and A. Masa, "Automatic Extraction and Learning of Keyphrases From Scientific Articles," In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2005.

[17] R. Mihalcea and P. Tarau, "Bringing Order Into Text," Textrank in Proceedings of the 2004 conference on empirical methods in natural language processing, 2004.

[18] S. Brin and L. Page, "The Anatomy Of A Large-Scale Hypertextual Web Search Engine," Computer networks and ISDN systems, vol. 30(1-7), pp. 107-117, 1998.

[19] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-Based Topic Ranking For Keyphrase Extraction," In International Joint Conference on Natural Language Processing (IJCNLP), 2013.

[20] A. Tixier, F. Malliaros, and M. Vazirgiannis, "A Graph Degeneracy-Based Approach to Keyword Extraction," In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.

[21] J. Li, et al., "Key Word Extraction for Short Text Via Word2vec, Doc2vec, and Textrank," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27(3), pp. 1794-1805, 2019.

[22] J. R. Thomas, S. K. Bharti, and K. S. Babu, "Automatic Keyword Extraction for Text Summarization in E-Newspapers," In Proceedings of the International Conference on Informatics and Analytics, ACM, 2016.

[23] X. Wan and J. Xiao, "Single Document Keyphrase Extraction Using Neighborhood Knowledge," In AAAI, 2008.

[24] R. Naidu, et al., "Text Summarization with Automatic Keyword Extraction in Telugu E-Newspapers," In Smart Computing and Informatics, Springer, pp. 555-564, 2018.

[25] T. Mikolov, et al., "Distributed Representations of Words and Phrases and Their Compositionality," In Advances in neural information processing systems, 2013.

[26] W. Zhang, T. Yoshida, and X. Tang, "A Comparative Study of TF* IDF, LSI and Multi-Words for Text Classification," Expert Systems with Applications, vol. 38(3), pp. 2758-2765, 2011.

[27] J. A. Lossio-Ventura, et al., "Yet Another Ranking Function For Automatic Multiword Term Extraction," In International Conference on Natural Language Processing, Springer, 2014.

[28] R. Campos, et al., "Yake! Collection-Independent Automatic Keyword Extractor," In European Conference on Information Retrieval, Springer, 2018.

[29] M. Saraswathi and V. Balu, "Preprocessing Techniques for Effective Data Extraction and Computation," IUP Journal of Computer Sciences, vol. 7(3), p. 27, 2013.

[30] O. Hajipoor, et al., "Determine the Sentiment for Persian Words and Phrases Using Deep Learning," Computer Society of Iran Conference, vol. 24, 2019.

Journal of Electronical & Cyber Defence
Vol. 8, No. 2, 2020, Serial No. 30

Automatic Keyword Extraction from Persian short Text Using word2vec

O. Hajipoor, S. S. Sadidpour*
Malek-Ashtar University of Technology
(Received: 26/06/2019, Accepted: 23/10/2019)

ABSTRACT

With the growing number of Persian electronic documents and texts, the use of quick and inexpensive methods to access desired texts from the extensive collection of these documents becomes more important. One of the effective techniques to achieve this goal is the extraction of the keywords which represent the main concept of the text. For this purpose, the frequency of a word in the text can not be a proper indication of its significance and its crucial role. Also, most of the keyword extraction methods ignore the concept and semantic of the text. On the other hand, the unstructured nature of new texts in news and electronic documents makes it difficult to extract these words. In this paper, an automated, unsupervised method for keywords extraction in the Persian language that does not have a proper structure is proposed. This method not only takes into account the probability of occurrence of a word and its frequency in the text, but it also understands the concept and semantic of the text by learning word2vec model on the text. In the proposed method, which is a combination of statistical and machine learning methods, after learning word2vec on the text, the words that have the smallest distance with other words are extracted. Then, a statistical equation is proposed to calculate the score of each extracted word using co-occurence and frequency. Finally, words which have the highest scores are selected as the keywords. The evaluations indicate that the efficiency of the method by the F-measure is 53.92% which is 11% superior to other methods.

Keywords: Keyword Extraction, Persian Language, Text Mining, Word Similarity, Word2vec

* Corresponding Author Email: sadidpour@mut.ac.ir