تشخیص سرقت علمی متون فارسی با رویکرد مبتنی‌بر بردار کلمات

محبوبه گلچین پور[۱]، هادی ویسی[۲] مصطفی صالحی[۳]

[۱] دانشجوی کارشناسی ارشد دانشگاه تهران، m.golchinpor@ut.ac.ir

[۲] استادیار، عضو هیئت‌علمی دانشگاه تهران ، h.veisi@ut.ac.ir

[۳] استادیار، عضو هیئت‌علمی دانشگاه تهران ، mostafa_salehi@ut.ac.ir

چکیده

گسترش اینترنت و دسترسی سریع و آسان به انبوه داده‌های متنی، سرقت علمی را به معضلی جدی و روبه رشد تبدیل کرده است. ازاین‌رو در این مقاله تابع فاصله جدیدی به نام فاصله‌ی برداری کلمات که مبتنی بر یادگیری عمیق[۱] است، برای تشابه‌یابی و تشخیص سرقت علمی متون فارسی پیشنهاد می‌گردد. این روش کلمات را به‌صورت بردارهایی در فضای N بعدی تعبیه[۲] و تشابه دو سند متنی را به‌صورت میانگین فاصله کسینوسی موردنیاز برای حرکت از کلمات تعبیه‌شده‌ی سند اول، برای رسیدن به کلمات مشابه‌شان در سند دوم تعریف می‌کند. روش فاصله‌ی برداری کلمات به‌آسانی می‌تواند تشابه اسناد متنی با کلمات مختلف ولی با مفهوم مشابه را تشخیص دهد. با استفاده از این روش دو سند متنی که حداکثر تشابه کسینوسی را نسبت به هم داشته باشند، مشابه نامیده و سرقت علمی تشخیص داده می‌شود. یکی از ضعف‌های روش ارائه‌شده عدم در نظر گرفتن طول رشته‌های متنی مورد مقایسه می‌باشد، از این رو با توجه به مزیت روش لونشتاین در بررسی تطابق کاراکتری رشته‌های متنی با طول‌های مختلف، در این مقاله از روش لونشتاین به‌منظور کاهش خطای روش فاصله‌ی برداری کلمات استفاده شده‌است. نتایج استفاده از ترکیب این دو روش تشابه‌یابی، برای تشخیص سرقت علمی متون فارسی روی پیکره متنی PAN2015 دارای معیار F- ۹۷/۹٪ می‌باشد.

واژگان کلیدی

یادگیری عمیق، بازنمایی برداری کلمات، تشابه‌یابی، سرقت علمی، بردار کلمه

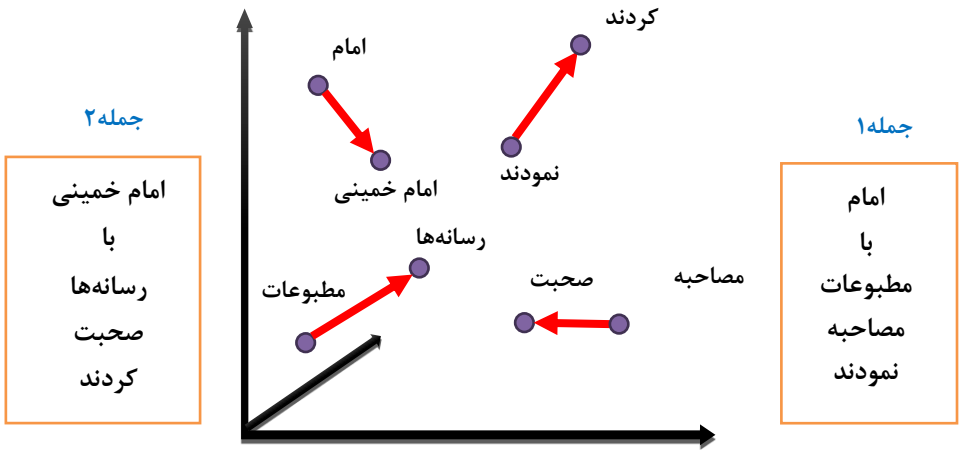
۱- مقدمه

امروزه با حجم انبوه داده‌های متنی روبه‌رو هستیم. در داده‌های متنی به دلیل سهولت تغییر متن، داده‌های مشابه به‌وفور تولید می‌شوند. درنتیجه سرقت علمی به‌عنوان مشکلی جدی بین اسناد متنی مطرح است[۱]. سرقت علمی استفاده از ایده‌ها و نتایج دیگران به‌عنوان اندیشه یا اثر خود بدون انتساب نام مالک اثر است. امروزه شناسایی سرقت علمی به کمک نرم‌افزارها آسان‌تر شده است. اما انواع مختلفی از سرقت همچنان موضوع پردردسری است. یکی از جدی‌ترین کاربردهای تشابه‌یابی، تشخیص سرقت علمی اسناد متنی است. از تشابه‌یابی متون در موارد دیگری نظیر بازیابی اطلاعات، خوشه‌بندی اسناد، امتیازدهی خودکار مقالات، طبقه‌بندی پاسخ‌های کوتاه، ترجمه ماشینی و خلاصه‌سازی متون استفاده می‌شود. با توجه به گسترش منابع متنی و کاربردهای مختلف بیان‌شده از تشابه‌یابی متون در زمینه‌های مختلف، به‌وضوح می‌توان به اهمیت تشابه‌یابی پی برد. درنتیجه استفاده از روش‌های مناسبی که به‌راحتی بتوانند تشابه متون را تشخیص دهند، از اهمیت زیادی برخوردار است. استفاده از روش‌های سنتی پردازش زبان طبیعی، برای نمایش کلمات و یافتن تشابه معنایی و نحوی واژگان، علیرغم کاربردهای فراوان و جالب‌توجه، موانع و مشکلات فراوانی در حوزه استخراج ویژگی دارد. این روش‌ها اغلب دارای دو مشکل ابعاد بالا و تنک بودن بردار ویژگی کلمات هستند. در مواقعی که یک مفهوم را می‌توان با روش‌های متفاوتی بیان کرد، روش‌های سنتی به دلیل اینکه تشابه معنایی و نحوی واژگان را در نظر نمی‌گیرند، قادر به تشخیص تشابه نخواهند بود.

---

۱ Deep Learning

۲ Embedding



شکل۱. مثالی از فاصله بین دو مستند با استفاده از معیار فاصله‌ی برداری کلمات [۲]

این روش‌ها نمی‌توانند تشابه جملاتی با اجزای متفاوت، ولی با معانی یکسان را تشخیص دهند. به‌طور مثال دو جمله‌ی زیر با وجود اینکه کلمات متفاوتی دارند، اما مفهوم تقریباً مشابه ای را بیان می‌کنند و روش‌های سنتی قادر به تشخیص تشابه این نوع از جملات نیستند.

جمله۱: امام با مطبوعات مصاحبه نمودند.

جمله۲: امام خمینی با رسانه‌ها صحبت کردند.

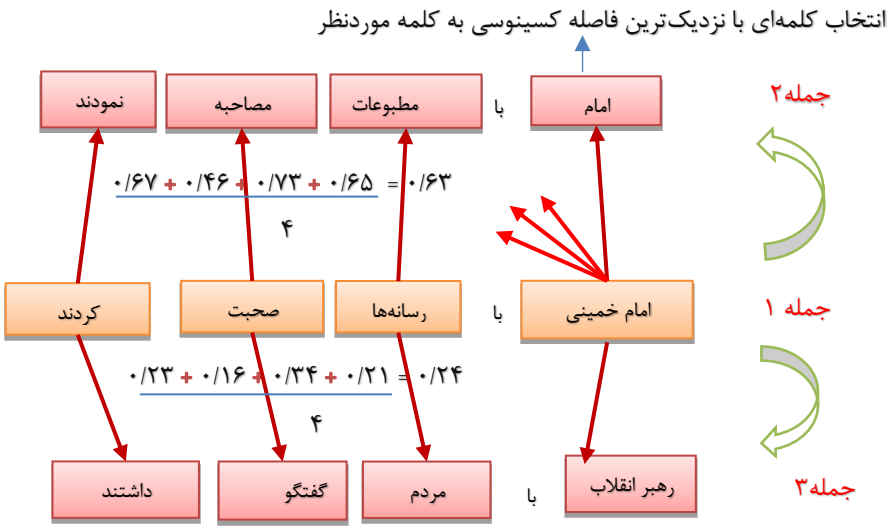
مشکلات موجود در روش‌های سنتی پردازش زبان طبیعی، حجیم بودن داده-های متنی و بحث افزایش دقت و کارایی سامانه‌های تشابه‌یابی، لزوم استفاده از روش‌های نوین پردازش و استخراج دانش را درزمینه‌ی تشابه‌یابی متون آشکار می‌سازد. یکی از روش‌های جدید تشابه‌یابی، استفاده از یادگیری عمیق در پردازش زبان طبیعی، برای بازنمایی‌های برداری کلمات است. با استفاده از بازنمایی کلمات در فضای برداری، می‌توان نشان داد که کلمات مشابه، بردارهایی نزدیک به هم و کلمات غیرمشابه بردارهایی دور از هم خواهند داشت. درنتیجه از طریق محاسبه فاصله‌ی بردارهای کلمات، می‌توان تشابه کلمات را تشخیص داد. این روش طول رشته‌های متنی را در تشابه‌یابی در نظر نمی‌گیرد. از این رو در برخی موارد برای تشخیص تشابه جملاتی با طول مختلف به‌درستی عمل نمی‌کند. لونشتاین، روشی است که تشابه دو رشته متنی را از طریق بررسی تطابق کاراکتری رشته‌ها، تشخیص می‌دهد. از این رو در این مقاله، به‌منظور بهبود روش ارائه شده از ترکیب روش فاصله‌ی برداری کلمات و روش لونشتاین استفاده می‌گردد. در بخش دوم پیشینه تحقیق مورد مطالعه و بررسی قرار می‌گیرد و به طور خلاصه روش‌های تشابه-یابی موجود مطرح می‌گردد. سپس روش ارائه شده‌ی تشابه‌یابی که مبتنی بر بازنمایی برداری کلمات و روش لونشتاین می‌باشد، شرح داده می‌شود. در بخش آخر نیز نتایج حاصل از ترکیب دو روش بیان می‌گردد.

۲- مطالعه پیشین

در این بخش روش‌های تشابه‌یابی متون مورد معرفی قرار می‌گیرند. برخی از روش‌های تشابه‌یابی به شرح ذیل می‌باشند.

- تشابه مبتنی بر واژگان: این روش، تشابه متون را بر اساس تشابه تعداد کاراکترها و اصطلاحات مشترک متون ارزیابی می‌کند. روش‌های مبتنی بر n-gram و روش‌های مبتنی‌بر طول بزرگ‌ترین زیر دنباله‌ی مشترک جز این روش هستند[۶].
- تشابه مبتنی بر روابط نحوی: این روش، تشابه متون را بر اساس تشابه واحدهای نحوی موجود در متون تشخیص می‌دهد. این روش از ویژگی‌هایی همچون برچسب‌های کلمات موجود در متون برای مقایسه تشابه متون استفاده می‌کند[۶].
- تشابه مبتنی بر روابط معنایی متون: این روش از تشابه معنایی واحدهای متنی برای مقایسه تشابه متون استفاده می‌کند. روش‌هایی که از مترادف، متضاد و شمول معنایی استفاده می-کنند در این دسته قرار می‌گیرند[۶].

در مرجع [۴] سه روش تشابه مبتنی بر رشته، تشابه مبتنی بر پیکره متنی و تشابه مبتنی بر دانش نیز برای اندازگیری تشابه معرفی شده است. مرجع[۵] روش مبتنی بر گراف را برای تشابه‌یابی و تشخیص سرقت علمی فارسی پیشنهاد داده است. نتایج حاصل از این روش روی مجموعه داده PAN2015 دارای معیار دقت ۹۱٪ و فراخوانی ۸۹٪ می‌باشد. مرجع [۶] از روش مبتنی بر یادگیری عمیق برای تشخیص سرقت علمی استفاده نموده است.



شکل۲. نمایش استفاده از روش فاصله‌ی برداری کلمات برای تعیین تشابه جمله‌ی ۱ با جمله‌ی ۲و۳[۱۲]

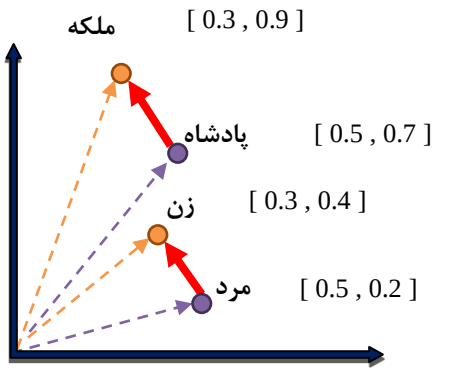
نتایج حاصل از این روش روی داده PAN2016 دارای معیار دقت ۹۵/۹ ٪ و فراخوانی ۸۵/۸ ٪ می‌باشد. در مراجع [۷, ۸] از الگوریتم‌های اثرانگشت برای تشخیص سرقت علمی متون فارسی استفاده شده است. الگوریتم‌های اثر انگشت، متن را به‌عنوان مجموعه‌ای از کاراکترها در نظر گرفته، سپس کاراکترها را در دسته‌های n کاراکتری تقسیم می‌نمایند، معروف‌ترین آن‌ها ۱۶-گرام، ۸-گرام و ۵-گرام می‌باشند.

روش‌های اخیر ترتیب کلمات در بافت‌های متنی را در نظر نمی‌گرفتند و همچنین قادر به تشخیص تشابه بافت‌های متنی مشابه ولی با کلمات مختلف نبودند. همچنین این روش‌ها مشکل دقت پایین در تشخیص تشابه را دارا بودند. از این رو در این مقاله به‌منظور افزایش دقت و کارایی روش‌های تشابه یابی، روش فاصله‌ی برداری کلمات که مبتنی بر word2vec است، برای تشابه‌یابی و تشخیص سرقت علمی متون فارسی ارائه می‌گردد[۲].

۳- روش پیشنهادی برای تشابه یابی متون

۳-۱ معناشناسی توزیعی و بازنمایی‌های برداری کلمات

فرضیه توزیعی که توسط هریس ارائه‌شده است بیان می‌کند که کلمات بامعنای مشابه در متون مشابه ظاهر می‌شوند[۹]. هدف معناشناسی توزیعی یافتن نمایش برداری است که معنای یک کلمه را تخمین بزند. کلمات با معنای مشابه در فضای برداری، بردارهایی نزدیک به هم و کلمات غیرمشابه، بردارهایی دور از هم خواهند داشت. درنتیجه از طریق بررسی تشابه بردارهای کلمات می‌توان تشابه کلمات را به‌راحتی تشخیص داد. Word2Vec روش بسیار کارآمد و مناسبی برای نمایش برداری کلمات و متون و پردازش آن‌ها است که توسط میکولو و همکارانش توسعه‌یافته است[۱۰].



شکل ۳. محاسبه بردار کلمه ملکه از طریق جبر برداری

ملکه = زن + مرد - پادشاه [۱۳]

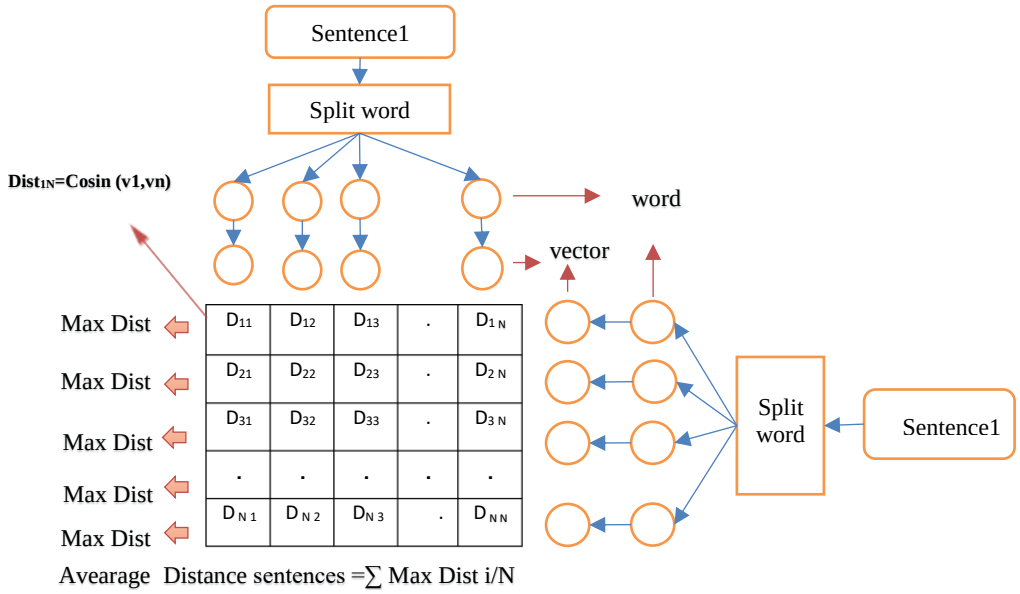
Word2Vec، ابزاری است که با استفاده از دو روش مختلف سبد پیوسته کلمات[۳] و پرش پیوسته چندتایی[۴][۶] که مبتنی بر شبکه عصبی[۵] و یادگیری عمیق هستند، به نمایش برداری کلمات می‌پردازد[۱۰, ۱۱]. با استفاده از

---

[۳] Continuous bag-of-words(CBOW)

[۴] Continuous Skip-Gram

[۵] Deep learning



شکل۴. محاسبه تشابه دو جمله با استفاده از تشابه کسینوسی بردار کلمات دو جمله

روش Word2Vec می‌توان دریافت که نمایش برداری کلمات قادر خواهد بود، روابط زبانی و معنایی واژگان را در فضای برداری به‌خوبی حفظ نماید[۱۱]. نکته شگفت‌آور روش نمایش برداری کلمات این است که فاصله‌ها در فضای برداری به‌نوعی معانی را جابه‌جا می‌کنند. در فضای برداری فاصله کسینوسی بین دو بردار می‌تواند تفسیری از نحوه تشابه کلمات نسبت به هم باشد. طبق شکل ۳ اگر در فضای برداری از بردار کلمه شاه به‌اندازه فاصله بین بردار کلمات مرد و زن و درهمان جهت حرکت کنیم به بردار کلمه ملکه خواهیم رسید، درنتیجه این سیستم به‌راحتی می‌تواند نشان دهد که پادشاه معادل ملکه است.

ملکه = زن + مرد - پادشاه

این روش برای صفت‌های عالی هم درست عمل می‌کند.

بهترین = بهتر + سریع – سریع‌ترین

برای افعال نیز

شد = می‌شود + رفت – می‌رود

با استفاده از این روش می‌توان ورزش ملی آلمان را از ورزش ملی ایران استنباط کرد.

فوتبال = آلمان + ایران – کشتی

با استفاده از نام خانوادگی نخست‌وزیر فعلی انگلستان می‌توان به نام خانوادگی رئیس‌جمهور ایران پی برد.

روحانی = ایران + انگلستان – کمرون [۱۴, ۱۵]

میکولو و همکارانش نشان دادند که چطور با انجام جستجوی نزدیک‌ترین همسایه در فضای پیوسته نمایش کلمات می‌توان به پاسخ چنین سؤالاتی دست‌یافت [۴, ۱۰].

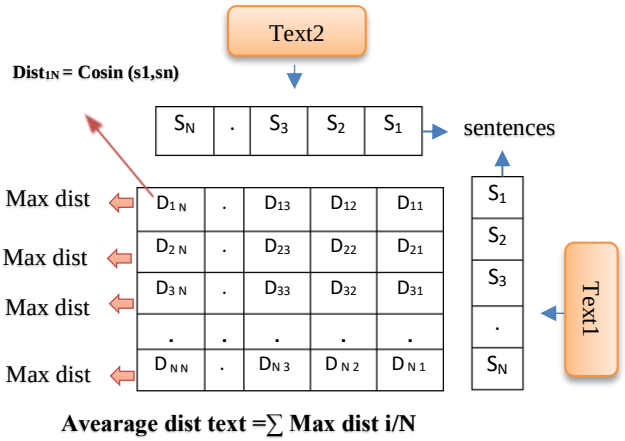
۳-۲ روش فاصله‌ی برداری کلمات

در این مقاله از روش فاصله‌ی متحرک کلمات که مبتنی بر بازنمایی‌های کلمات در فضای برداری است به‌منظور تشخیص سرقت علمی اسناد متنی استفاده می‌شود. فاصله‌ی متحرک کلمات به‌آسانی می‌تواند تشابه جملاتی با کلمات متفاوت، ولی با معانی مشابه را تشخیص دهد. در این روش کلمات اصلی سند متنی به‌صورت بردار کلمه در فضای برداری تعبیه می‌شوند و فاصله بین دو سند متنی، میانگین فاصله کسینوسی خواهد بود که بردار کلمات سند A نیاز به حرکت برای رسیدن به تطابق با بردار کلمات سند B دارند[۱۲, ۱۶]. شکل ۱ نمونه‌ای از فاصله دو مستند متنی با استفاده از این روش را نشان می‌دهد. دو جمله زیر موجود است.

- امام با مطبوعات مصاحبه نمودند.
- امام‌خمینی با رسانه‌ها صحبت کردند.

این دو جمله درحالی‌که کلمات مختلفی دارند ولی مفهوم تقریباً مشابه‌ای را بیان می‌کنند. فاصله‌ی برداری کلمات تشابه چنین جملاتی را با استفاده از در نظر گرفتن تشابه بردارهای کلمات در فضای تعبیه حل می‌کند. همه کلمات دو جمله در فضای برداری تعبیه می‌شوند و فاصله کسینوسی بردار کلمات هر کلمه موجود در جمله اول با تمام کلمات جمله دوم محاسبه می‌شود، سپس کلمه‌ای که بیشترین تشابه کسینوسی با کلمه مورد نظر را داشته باشد به‌عنوان کلمه مشابه تشخیص داده می‌شود. بدین ترتیب مشابه‌ترین کلمات دو جمله مشخص می‌شوند. در ادامه با استفاده از میانگین فاصله مشابه‌ترین

کلمات دو جمله، فاصله تشابه دو جمله محاسبه می‌شود. شکل ۲ اعمال روش فاصله‌ی برداری کلمات روی دو جمله‌ی ۲و ۳ که با جمله‌ی ۱ مورد مقایسه قرارگرفته‌اند را نشان می‌دهد. در ابتدا ایست واژگان حذف می‌شوند، روش فاصله‌ی برداری کلمات بر اساس فاصله موجود بین کلمات و اسناد، جریان هر کلمه را به سمت کلمه‌ای که از لحاظ معنایی تشابه بیشتری با آن کلمه دارد، انتقال می‌دهد. هزینه‌ی انتقال از رسانه ها به مطبوعات بسیار کم‌هزینه‌تر از هزینه‌ی انتقال از مردم به مطبوعات است. این مسئله به این خاطر است که Word2Vec بردار کلمه‌ی مطبوعات را نسبت به بردار کلمه‌ی مردم، در فاصله‌ی نزدیک‌تری از بردار کلمه‌ی رسانه‌ها تعبیه می‌کند. فاصله‌ی کسینوسی جمله‌ی ۱ و جمله ۲، ۰/۶۳ است که به‌صورت قابل‌ملاحظه‌ای از فاصله‌ی جمله ۱ و جمله ۳ بهتر می‌باشد. درنتیجه جمله‌ی ۲ نسبت به جمله ۳ دارای تشابه بیشتری با جمله۱ است[۱۲].



شکل۵. محاسبه تشابه دو متن از طریق تشابه کسینوسی جملات

مراحل تشخیص سرقت علمی بین دو سند متنی با استفاده از روش مطرح‌شده به شرح زیر است.

۱. **نرمال‌سازی متن:** در ابتدا دو متن منبع و متن مشکوک به سرقت به‌عنوان ورودی سیستم تشابه‌یابی دریافت می‌شوند، سپس متون نرمال‌سازی می‌شوند و علائم نگارشی و ایست واژگان از متون حذف می‌گردد.

۲. **تقطیع جمله‌ای متون:** متن منبع و متن مشکوک به سرقت به جملات تشکیل‌دهنده‌شان تقطیع می‌شوند.

۳. **استخراج بردار کلمات:** توسط الگوریتم Word2Vec بردار کلمات موجود در هر یک از جملات متن منبع و متن مشکوک به سرقت استخراج می‌شوند.

۴. **مقایسه تشابه جملات:** هر جمله از متن مشکوک به سرقت با همه‌ی جملات متن منبع مقایسه می‌شود و مشابه‌ترین جملات دو سند متنی مشخص می‌گردد. برای مقایسه تشابه دو جمله، فاصله کسینوسی بردار هر کلمه‌ی جمله‌ی مشکوک به سرقت با همه‌ی بردارهای کلمات جمله‌ی منبع محاسبه می‌شود.

$$Cosine\ Similarity = \frac{v_1 . v_2}{\| v_1 \| \| v_2 \|} \quad (۱)$$

V_1 بردار کلمه موجود در جمله مشکوک به سرقت و V_2 بردار کلمه‌ی موجود در جمله منبع است. سپس کلمه‌ای که دارای نزدیک‌ترین فاصله کسینوسی به کلمه موردنظر است به‌عنوان کلمه مشابه انتخاب می‌شود. متغیر Dist، فاصله کسینوسی کلمه‌ی موجود در جمله مشکوک به سرقت با کلمه‌ی جمله منبع است. N، تعداد کلمات جمله‌ی منبع می‌باشد.

$$Max\ Dist = Max\left(Dist_{11}, Dist_{12}, ..., Dist_{1N}\right) \quad (۲)$$

۵. **محاسبه تشابه جملات:** پس از مشخص کردن کلمات مشابه دو جمله، با توجه به شکل۴ با استفاده از میانگین فاصله کلمات مشابه دو جمله از هم، فاصله تشابه دو جمله به دست می‌آید.

۶. **محاسبه تشابه دو متن:** هر جمله‌ی متن مشکوک به سرقت با همه‌ی جملات متن منبع مقایسه شده و جمله مشابه با آن تعیین می‌شود. طبق شکل۵ پس از تعیین مشابه‌ترین جملات دو سند، با میانگین‌گیری فاصله مشابه‌ترین جملات دو سند، میزان تشابه دو سند متنی مشخص می‌شود.

$$Texts\ Distance = \frac{\sum_{i=1}^{i=M} Max\ Dist_i}{M} \quad (۳)$$

۷. **تشخیص سرقت علمی:** پس از مشخص کردن مشابه‌ترین جملات دو سند مشکوک به سرقت و سند منبع، فاصله‌ی مشابه‌ترین جملات دو سند متنی با سطح آستانه مقایسه می‌گردد و درصورتی که از سطح آستانه فاصله کسینوسی ۰/۶۷ بیشتر باشد سرقت علمی تشخیص داده می‌شود.

۳-۳ استفاده از روش لونشتاین به‌منظور بهبود روش فاصله‌ی برداری کلمات

روش فاصله‌ی برداری کلمات به‌راحتی قادر است تشابه جملاتی با کلمات مختلف ولی با مفهوم مشابه را تشخیص دهد ولی این روش طول رشته‌های متنی را در تشابه‌یابی در نظر نمی‌گیرد. از این رو در مواردی دچار خطای عدم تشخیص صحیح تشابه اسناد متنی می‌گردد.

در این مقاله ابتدا با استفاده از روش فاصله‌ی برداری کلمات، تشابه اسناد مشکوک به سرقت با اسناد مرجع موجود در پیکره محاسبه شده و با بررسی دستی اسناد مقایسه شده، نتایج نشان داده شد، جفت جملاتی از دو سند مشکوک به سرقت و سند مرجع که دارای تشابه کسینوسی

بیشتر از ۰/۶۷ هستند، مشابه بوده و هر چه این درصد تشابه به یک نزدیک‌تر می‌شود میزان تشابه افزایش می‌یابد. دو جمله‌ای که دارای تشابه کسینوسی یک و یا نزدیک به یک هستند تقریباً کپی هستند، جفت جملاتی که تشابه کسینوسی آن‌ها بین ۰/۶۷ و ۰/۸۰ بود اگرچه دارای کلمات مختلفی بودند ولی از نظر مفهومی مشابه‌ای داشتند. در برخی از موارد نیز مشاهده نشان داد که جملاتی که تشابه کسینوسی آن‌ها از ۰/۵۷ بیشتر است نیز در مواردی تا حدی مشابه‌اند.

با توجه به مشاهدات در ابتدا تشابه کسینوسی۰/۶۷ به‌عنوان سطح آستانه‌ی سرقت علمی برای دو جمله در نظر گرفته شد. اما نتایج نشان داد که با استفاده از روش فاصله‌ی برداری کلمات تشابه جملاتی که طول یکی از آن‌ها دو برابر طول دیگری است و مفهوم جمله‌ی کوچک‌تر یا حتی خود جمله‌ی کوچک‌تر به‌صورت کامل در جمله‌ی بزرگ‌تر وجود دارد در برخی از موارد قابل تشخیص نمی‌باشد. این روش در مواقعی که جمله‌ی کوچک‌تر به‌صورت کامل در جمله‌ی بزرگ‌تر وجود دارد، فاصله‌ی تشابه دو جمله را نزدیک به سطح آستانه سرقت علمی در نظر می‌گیرد ولی چون پایین‌تر از حد آستانه می‌باشد باعث اشتباه در تشخیص تشابه می‌شود.

بدین منظور حد آستانه‌ی ۰/۵۷ به‌عنوان حد آستانه‌ی تشخیص سرقت علمی در نظر گرفته شد و با استفاده از روش فاصله‌ی برداری، فاصله‌ی تشابه جملات دو سند مشکوک به سرقت و سند مرجع مشخص گردید، در ادامه بین جفت جملاتی از دو سند که فاصله تشابه کسینوسی بالاتر از سطح آستانه داشتند به‌منظور بررسی تشابه دقیق‌تر و رفع مشکل بیان شده، تشابه لونشتاین محاسبه گردید. فاصله‌ی لونشتاین تشابه دو رشته‌ی متنی را با استفاده از بررسی تعداد عملیات درج، حذف و یا جایگزینی که می‌توان یک رشته‌ی متنی را به رشته‌ی دیگر تبدیل کرد تعریف می‌کند. همان‌گونه که بیان شد روش فاصله‌ی برداری کلمات طول رشته‌های متنی را در تشابه‌یابی در نظر نمی‌گیرد و فقط از طریق بررسی تشابه بردارهای کلمات تشابه را تشخیص می‌دهد ولی روش لونشتاین با بررسی تطابق کاراکتری رشته‌های متنی به‌نوعی طول رشته‌ها را در تشابه یابی در نظر می‌گیرد از این رو با توجه به اینکه روش لونشتاین از تطابق کاراکتری برای بررسی تشابه استفاده می‌کند، می‌توان از این روش برای بررسی تطابق کاراکتری و در نظر گرفتن طول جملات یا رشته‌های مورد مقایسه استفاده کرد.[۱۷]. مراحل انجام کار به شرح ذیل می‌باشد.

۱. **تعیین فاصله مشابه‌ترین جملات دو سند متنی:** در ابتدا با استفاده از روش فاصله‌ی برداری کلمات، تشابه کسینوسی جملات دو سند متنی محاسبه می‌شود، سپس درصورتی‌که این فاصله از سطح آستانه‌ی فاصله‌ی برداری جملات که نشان دهنده‌ی سرقت علمی است، بیشتر باشد، دو جمله برای بررسی تشابه لونشتاین انتخاب می‌شوند.
۲. **محاسبه فاصله لونشتاین دو کلمه:** تشابه لونشتاین هر کلمه‌ی جمله‌ی مشکوک به سرقت با همه‌ی کلمات موجود در جمله‌ی منبع محاسبه می‌شود.
۳. **تعیین کلمات مشابه:** برای هر کلمه‌ی جمله‌ی مشکوک به سرقت، کلمه‌ای از جمله منبع که کمترین فاصله‌ی لونشتاین با کلمه‌ی مورد نظر را دارد، به عنوان کلمه مشابه با کلمه مورد نظر انتخاب می‌شود. در این مقاله ما فاصله‌ی لونشتاین صفر و یک بین دو کلمه را که نشان دهنده‌ی تطابق دقیق دو کلمه است به عنوان تشابه لونشتاین در نظر گرفته‌ایم.
۴. **محاسبه تشابه جملات:** پس از محاسبه تشابه لونشتاین کلمات دو جمله، تعداد کلماتی از دو جمله که تطابق لونشتاین صفر و یک با هم دارند در هر یک از جملات شمرده می‌شوند. سپس بررسی می‌شود که اگر تعداد کلمات مشابه دو جمله از تعداد کلمات کوچک‌ترین جمله بیشتر باشد، این دو جمله مشابه در نظر گرفته می‌شوند.
۵. **تشخیص سرقت علمی:** جفت جملاتی که با استفاده از روش لونشتاین نیز مشابه تشخیص داده شوند، برچسب سرقت علمی به آن‌ها تخصیص داده می‌شود.

در این مقاله با استفاده از روش لونشتاین خطای روش فاصله‌ی برداری کلمات به‌طور چشم‌گیری کاهش یافته‌است. مراحل روش تشخیص سرقت علمی ارائه شده در شکل۷ قابل مشاهده می‌باشد.

۴- نتایج

در این بخش دادگان، پارامترها، معیارهای ارزیابی و نتایج پیاده‌سازی و اجرای روش پیشنهادی روی مجموعه داده سرقت علمی اسناد فارسی مطرح می‌گردد.

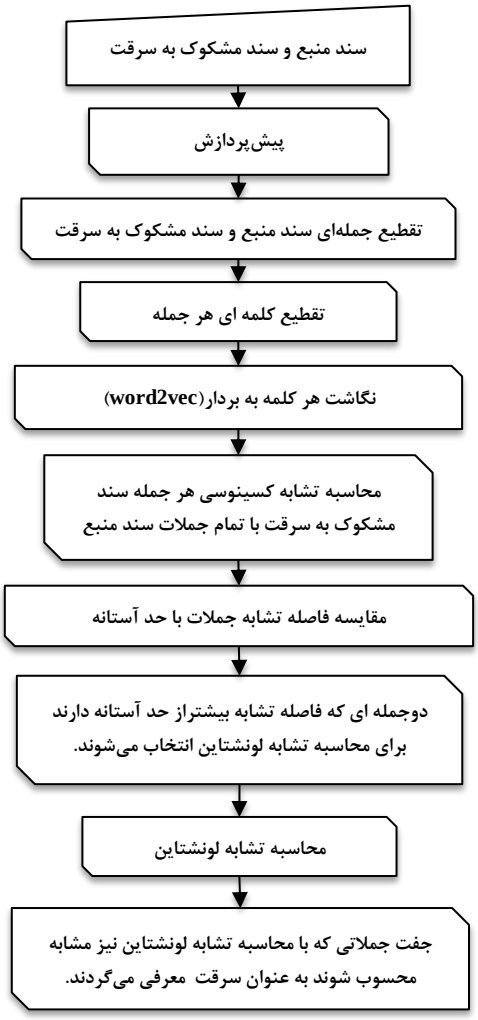
۴-۱ دادگان

روش تشابه‌یابی ارائه‌شده در این مقاله روی پیکره متنی PAN2015-khoshnavataher مورد ارزیابی قرارگرفته است. این پیکره شامل ۲۸۲ جفت سند با سرقت علمی تصادفی، ۱۲۹ جفت سند با سرقت علمی بدون ابهام و ۶۵۲ جفت سند بدون سرقت علمی است. این مجموعه داده به‌طور کل شامل ۷۲۰ سند مشکوک به سرقت و ۸۰۲ سند مرجع می‌باشد[۱۹] .

۴-۲ پارامترها

در این مقاله پارامترهای ذیل می‌توانند بهینه شوند.

۱. برای تعیین تشابه کلمات با استفاده از محاسبه فاصله اقلیدسی و یا فاصله‌ی تکان‌دهنده‌ی زمین ممکن است بتوان به دقت بالاتری در تشخیص تشابه رسید.
۲. با تغییر سطح آستانه فاصله کسینوسی دو جمله مورد مقایسه، درصد خطای تشابه‌یابی قابل بهبود می‌باشد.
۳. مراحل تشابه‌یابی به علت محاسبه تشابه همه‌ی کلمات دو سند متنی کمی زمان‌بر است. بدین منظور با اجرای پردازش موازی می‌توان سرعت تشابه‌یابی را بهبود بخشید.
۴. با تغییر سطح آستانه تشابه لونشتاین دو کلمه نیز ممکن است دقت بهتری حاصل شود.



شکل۶. مراحل انجام روش تشخیص سرقت علمی

۴-۳ معیارهای ارزیابی

معیارهای ارزیابی شامل معیار دقت[۶]، فراخوانی[۷]، معیارF[۸]، Granularity و plagdet هستند[۱۸].

$$Perc(S,R)=\frac{1}{|R|}\sum_{r\in R}\frac{|U_{s\in S}(S\cap r)|}{|r|} \quad (۴)$$

S تعداد موارد سرقت علمی موجود در پیکره می‌باشد و R تعداد موارد سرقت علمی شناسایی شده‌ی موجود در پیکره می‌باشد.

$$\mathrm{Rec}(S,R)=\frac{1}{|S|}\sum_{s\in S}\frac{|U_{r\in R}(S\cap r)|}{|s|} \quad (۵)$$

معیار F نوعی میانگین بین معیار دقت و فراخوانی است.

$$F-mesure=2*\frac{Perc*\mathrm{Re}c}{Perc+\mathrm{Re}c} \quad (۶)$$

از معیار Granularity برای تشخیص همپوشانی و سرقت علمی‌های چندگانه استفاده می‌شود.

$$gran(\mathrm{S},\mathrm{R})=\frac{1}{|\mathrm{S}_R|}\sum_{s\in S_R}\|\mathrm{R}_S\| \quad (۷)$$

معیار plagdet ترکیبی از همه‌ی معیارهای بیان شده می‌باشد که به‌صورت زیر محاسبه می‌گردد.

$$\mathrm{plagdet}(\mathrm{S},\mathrm{R})=\frac{F_1}{\log_2(1+\mathrm{gran}(\mathrm{S},\mathrm{R}))} \quad (۸)$$

۴-۴ نتایج روش ارائه شده

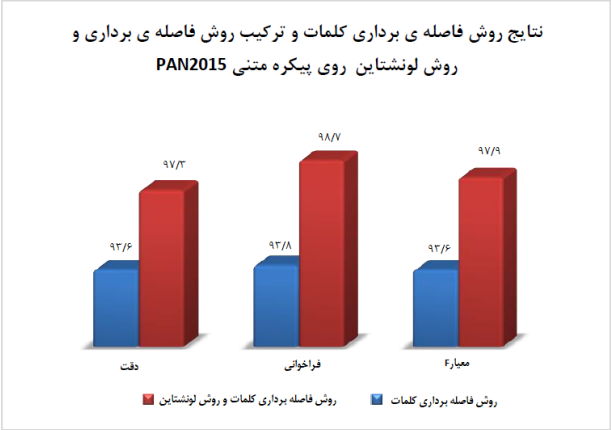
نتایج حاصل از ارزیابی روش پیشنهادی روی پیکره متنی معرفی شده در شکل ۷ قابل مشاهده می‌باشد.

---

۶ Percision

۷ Recall

۸ F-mesure



شکل۷. نتایج روش پیشنهادی روی پیکره متنیPAN2015

نتایج حاصل از مقایسه روش پیشنهادی با روش مبتنی‌بر شبکه عصبی[۶] روی پیکره متنی PAN2015 در جدول ۱ قابل مشاهده می‌باشد. همان‌طور که مشاهده می‌شود روش پیشنهادی دارای کارایی بهتری نسبت به روش مبتنی بر یادگیری عمیق برای تشخیص تشابه متون می‌باشد. روش مذکور برای هر جمله با استفاده از میانگین بردارهای کلمات موجود، یک بردار جمله محاسبه می‌کند، سپس برای مقایسه تشابه دو جمله از مقایسه تشابه بردارهای دو جمله استفاده می‌کند. این روش برای دو جمله غیر مشابه که میانگین بردار جملات آن‌ها تقریباً نزدیک است دچار خطا می‌شود. در حالی که روش ارائه شده در مقاله فعلی با مقایسه تشابه تمام بردارهای کلمات موجود در جملات به‌راحتی عدم تشابه چنین جملاتی را تشخیص می‌دهد.

جدول۱. نتایج روش پیشنهادی و روش مبتنی بر یادگیری عمیق

روش	پیکره متنی	دقت	فراخوانی	Granularity	Plagdet
فاصله‌ی برداری کلمات و لونشتاین	PAN2015	۹۷/۳	۹۸/۷	۱/۰۰۸	۰/۹۳۱
روش مبتنی بر یادگیری عمیق[۶]	PAN2016	۹۵/۹	۸۵/۸	۱/۰۰۰	۰/۹۰۶

۵- نتیجه‌گیری

در این مقاله روش فاصله‌ی برداری کلمات که مبتنی بر بازنمایی برداری کلمات با استفاده از روش word2vec است، برای تشابه‌یابی و تشخیص سرقت علمی متون فارسی معرفی گردید. این روش، تشابه مستندات را بر اساس تشابه بردارهای کلمات دو سند متنی محاسبه می‌نماید. روش فاصله‌ی برداری کلمات به‌آسانی قادر است تشابه جملاتی با کلمات مختلف ولی با مفهوم مشابه را تشخیص دهد. با توجه به اینکه روش پیشنهادی طول رشته-های متنی را درتشابه‌یابی در نظر نمی‌گیرد، در برخی موارد به درستی عمل نمی‌کند. در نتیجه از روش لونشتاین به منظور بررسی تطابق رشته‌های متنی با طول مختلف استفاده می‌شود. نتایج حاصل از ترکیب روش فاصله‌ی برداری کلمات و روش لونشتاین برای تشخیص سرقت علمی متون فارسی روی پیکره متنی PAN2015 دارای معیارF ۹۷/۹٪ می‌باشد.

۶- مراجع

1. Broder, A.Z., et al., *Indexing shared content in information retrieval systems*. Lecture Notes in Computer Science, *3896: p.* 313,2006.
2. Kusner, M.J., et al. *From word embeddings to document distances*. in *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*.
3. Ha, T.-L., *Lexical-Syntactic Approaches for English-Dutch Cross-lingual Textual Entailment*, Master's thesis, University of Groningen, 2011.
4. Gomaa, W.H. and A.A. Fahmy, *A survey of text similarity approaches*. International Journal of Computer Applications, 68(13), 2013.
5. Momtaz, M., et al. *Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents*. in *FIRE (Working Notes)*. 2016.
6. Gharavi, E., et al. *A Deep Learning Approach to Persian Plagiarism Detection*. in *FIRE (Working Notes)*. 2016.
7. Zini, M., et al. *Plagiarism detection through multilevel text comparison*. in *Automated Production of Cross Media Content for Multi-Channel Distribution, AXMEDIS'06. Second International Conference on* ,IEEE. 2006.
8. Nahnsen, T., O. Uzuner, and B. Katz, *Lexical chains and sliding locality windows in content-based text similarity detection*. 2005.
9. Mihalcea, R., C. Corley, and C. Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity*. in *AAAI*. 2006.
10. Heuer, H., *Semantic and stylistic text analysis and text summary evaluation.* Master's Thesis. Stockholm, July 20, 2015.
11. Goldberg, Y. and O. Levy, *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. arXiv preprint arXiv:1402.3722, 2014.
12. Kusner, M., et al. *From word embeddings to document distances*. in *International Conference on Machine Learning*. 2015.
13. Socher, R., Y. Bengio, and C.D. Manning. *Deep learning for NLP (without magic)*. in *Tutorial Abstracts of ACL*. Association for Computational Linguistics, *2012*.
14. Rink, B. and S. Harabagiu. *UTD: Determining relational similarity using lexical patterns*. in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2012.

15. Mikolov, T., W.-t. Yih, and G. Zweig. *Linguistic regularities in continuous space word representations*. in *hlt-Naacl*. 2013.
16. Huang, G., et al. *Supervised Word Mover's Distance*. in *Advances in Neural Information Processing Systems*. 2016.
17. Yujian, L. and L. Bo, *A normalized Levenshtein distance metric*. IEEE transactions on pattern analysis and machine intelligence. 29(6): p. 1091-1095 , 2007
18. Potthast, M., et al. *An evaluation framework for plagiarism detection*. in *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics, 2010.
19. ttp://www.uni-weimar.de/medien/webis/events/pan- 15/pan15-web/about.html

انجمن فناوری اطلاعات و ارتباطات ایران

IKT 2017
کنفرانس بین المللی فناوری اطلاعات و دانش

دانشگاه صنعتی امیرکبیر
۲۶ - ۲۷ مهر ۱۳۹۶

فراخوان مقاله

نهمین کنفرانس بین المللی فناوری اطلاعات و دانش

انجمن فناوری اطلاعات و ارتباطات ایران و دانشگاه صنعتی امیرکبیر، **نهمین کنفرانس بین المللی فناوری اطلاعات و دانش** را با هدف رشد و توسعه دانش و فناوری اطلاعات و انتقال آخرین تجارب و دست آوردهای پژوهشی و صنعتی در این زمینه و فراهم کردن بستری مناسب برای توسعه زمینه‌های مطرح در کنفرانس، برگزار می‌کنند.

از تمامی محققان، دانشگاهیان و صاحبان نظر دعوت می شود تا با ارسال مقاله و شرکت در کنفرانس با تبادل یافته های علمی، فناوری و کاربردی خود، زمینه ساز اشاعه و توسعه عمومی و ارائه قابلیتهای فناوری اطلاعات در زمینه‌های مختلف کنفرانس شوند (مقالات برگزیده در اولویت انتشار در فصلنامه علمی-پژوهشی ارتباطات و فناوری اطلاعات ایران قرار خواهد گرفت).

محورهای کنفرانس

- کلان داده ها
- اینترنت اشیاء
- سامانه های شبکه ای
- تعامل انسان و کامپیوتر
- مدیریت اطلاعات و دانش
- سیستمهای اطلاعات سلامت
- رایانش توزیعی،موازی و توری
- سامانه های آموزش الکترونیکی
- متن کاوی، فرایند کاوی، داده کاوی
- سامانه های تصمیم یار وسیستم های خبره
- سامانه های دولت الکترونیکی وتجارت الکترونیکی
- هوش تجاری
- شبکه های اجتماعی
- سامانه های هوشمند
- رایانش ابری و فراگیر
- حریم خصوصی و امنیت اطلاعات
- سامانه های اطلاعات سازمانی و مدیریتی
- مدیریت و آینده پژوهی فناوری اطلاعات
- سامانه ها و برنامه های کاربردی چند رسانه ای
- استخراج و بازیابی اطلاعات و موتورهای جستجو
- سامانه های باز مهندسی و مدیریت فرایندهای سازمانی

تاریخ های مهم

مهلت ارسال مقالات	۱۷ تیر
ارسال مقالات تصحیح شده و ثبت نام	۱ مهر
اعلام نهایی برگزاری کارگاه ها	۱۵ شهریور
اعلام نتایج داوری	۱۶ شهریور
ارسال پیشنهاد برگزاری کارگاه	۱ شهریور

کمیته برگزار کننده

نشانی دبیرخانه کنفرانس: تهران- خیابان حافظ- دانشگاه صنعتی امیرکبیر- دانشکده مهندسی کامپیوتر و فناوری اطلاعات- دفتر دانشکده تلفن: ۰۲۱-۶۴۵۴۲۷۰۰ فکس: ۰۲۱-۶۶۴۹۵۵۳۱

نشانی دبیرخانه انجمن: تهران- خیابان حافظ- دانشگاه صنعتی امیرکبیر- ساختمان ابوریحان- طبقه ششم- اتاق ۶۱۲ تلفن: ۰۲۱-۶۶۴۸۵۸۵۶ تلفکس: ۰۲۱-۶۶۴۹۵۲۳۳

وبسایت کنفرانس: http://ikt2017.aut.ac.ir ikt2017@aut.ac.ir ikt2017

نهمین کنفرانس بین المللی فناوری اطلاعات و دانش (IKT 2017)

۲۶ تا ۲۷ مهر ۱۳۹۶ - دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)

گواهی ارائه مقاله

بدین وسیله گواهی می گردد، مقاله علمی با عنوان:

تشخیص سرقت علمی متون فارسی با رویکرد مبتنی بر بردار کلمات

توسط نویسندگان محترم:

محبوبه گلچین پور، هادی ویسی و مصطفی صالحی

در نهمین کنفرانس بین المللی فناوری اطلاعات و دانش (IKT 2017) مورد پذیرش قرار گرفته است و به صورت شفاهی ارائه شده است.

دبیر علمی کنفرانس
دکتر بابک صادقیان

دبیر کنفرانس
دکتر مهدی دهقان