به نام خدا

گزارش پروژه درس پردازش زبان طبیعی

عنوان پروژه: **استخراج کلمات کلیدی (Keywords Extraction)**

استاد: دکتر سعیده ممتازی

دانشجو: محمد قربانی - 97131099

بهار 97-98

فهرست مطالب:

[فهرست مطالب: 2](#_Toc15377117)

[روند کلی انجام پروژه 3](#_Toc15377118)

[فرضهای در نظر گرفته شده 4](#_Toc15377119)

[خروجیهای بدست آمده 6](#_Toc15377120)

[مقالات مطالعه شده 11](#_Toc15377121)

[مقاله اول 11](#_Toc15377122)

[مقاله دوم 13](#_Toc15377123)

[مقاله سوم 15](#_Toc15377124)

[مقاله چهارم 18](#_Toc15377125)

[مقاله پنجم 20](#_Toc15377126)

[منابع 23](#_Toc15377127)

روند کلی انجام پروژه

در ابتدا به مطالعه مقاله ارجاع شده در شرح پروژه پرداختیم تا یک دید اولیه از الگوریتم PageRank و TextRank به دست بیاوریم. سپس به پیاده سازی اولیه از یک طرح ساده از الگوریتم TextRank پرداختیم. در این پروژه نیاز به یک POS Tagger داشتیم که بتوانیم نقش زبانی هر کلمه را به دست آوریم. برای همین به نصب و یادگیری کتابخانه Hazm پرداختیم. در پیاده­سازی اولیه برنامه فقط قادر به استخراج کلمات کلیدی بود. در ادامه با مطالعه مقالات دیگر در این زمینه به دنبال یافتن ایده­هایی در زمینه چگونگی یافتن عبارات کاندید و همچنین چگونگی امتیاز دادن به آن­ها بودیم. در نهایت به کامل سازی برنامه خود پرداختیم و تلاش در بهبود نتایج حاصل از الگوریتم خود را داشتیم.

فرض­های در نظر گرفته­ شده

**فرض اول:** در این پروژه نیاز به شناسایی نقش زبانی کلمات می­باشد که ما برای رسیدن به این موضوع از کتابخانه­ی آماده Hazm استفاده کردیم هر چند می­توانستیم از مدلی که در تمرین سوم تهیه کرده­ بودیم نیز استفاده کنیم.

**فرض دوم:** مقادیر مهم در این پروژه را به صورت زیر در نظر گرفتیم:

|  |  |
| --- | --- |
| 0.85 | Damping Factor |
| 0.00001 | **Minimum Difference** |
| 10 | **Iteration Steps** |

متغیر **Damping** نشان دهنده­ی این است که با چه احتمالی ما در گراف از یک راس به راسی که قبلا مشاهده شده می­رویم و مقدار (1-d) نشان می­دهد که با چه احتماللی به راس جدید خواهیم رفت. این متغیر در فرمول PageRank یک نقش متعادل[[1]](#footnote-1) کننده دارد. در واقع این مقدار کمک می­کند که اگر وارد راسی شدیم که یال خروجی نداشت بتوان با یک احتمالی از آن خارج شویم. در مقالات پیشنهاد شده­است که مقدار این متغیر برابر 0.85 در نظر گرفته شود.

متغییر **Minimum Difference** بیانگیر کمینه مقداری است که باید به آن برسیم تا در واقع به یک همگرایی در امتیاز رئوس رسیده باشیم.

**فرض سوم:** در این پروژه ما یک گراف بدون جهت در نظر گرفته­ایم. همچنین یال­ها در این گراف فاقد وزن می­باشند. در مقاله ارجاع شرح پروژه گفته شده­بود که استفاده از گراف وزن­دار یا بدون­وزن در تعداد گام­ها برای رسیدن به همگرایی و شکل همگرایی بی­تاثیر است اما در امتیاز نهایی راس اثر گذار می­باشد. همچنین در این مقاله گفته شده بود که آن­ها بهترین نتیجه را برای گراف بدون جهت به­دست آورده­اند. به همین دلیل نیز ما گراف بدون جهت را برای پروژه خود انتخاب کردیم.

**فرض چهارم:** ما برای استخراج کلمات و عبارات کلیدی از کلماتی که دارای نقش زبانی خاصی هستند استفاده کردیم. برای استخراج کلمات کلیدی، کلماتی که دارای نقش­های اسم، صفت و فعل بودن را به عنوان کلمات کلیدی در نظر گرفتیم. همچنین برای استخراج عبارات کلیدی، عباراتی که از اسم و صفت تشکیل شده­باشند را به عنوان عبارات کاندید در نظر می­گیریم.

**فرض پنجم:** ما برای امتیاز دهی به کلمات کلیدی از الگوریتم TextRank استفاده کردیم اما برای امتیاز دهی به عبارات از جمع امتیاز کلمات تشکیل دهنده آن عبارت استفاده می­کنیم. چرا که استفاده از co-occurrence برای عبارات کمی بی­معنا به نظر می­رسد و در واقع گراف حاصل از این رابطه گرافی با یال­های کم و راس­های منفصل زیاد می­باشد.

خروجی­های بدست آمده

در این قسمت خروجی حاصل از اجرای برنامه بر روی برخی از متون که برگرفته شده از دو خبرگزاری فارس و همشهری می­باشند قرار داده می­شود:

متن اول

شادمهر کاظم‌زاده نماینده مردم دهلران در مجلس شورای اسلامی در گفت‌وگو با خبرنگار پارلمانی خبرگزاری فارس گفت: در کل کشور خرید گندم توسط دولت در وضعیت فعلی 20 درصد کاهش دارد البته در مناطق سردسیر و استان‌های غربی این میزان نسبت به سال گذشته 50 درصد کاهش پیدا کرده است.

وی افزود: متأسفانه گندم تولیدی خوراک دام می‌شود این در حالی است که هر کیلو جو 2350 تومان بوده و هر کیلو گندم 1700 تومان است.

نماینده مردم دهلران در مجلس  تصریح کرد: دامدار، گندم را که ارزان‌تر از جو است خریداری کرده و به دامش می‌دهد.

کاظم‌زاده خاطرنشان کرد: حجم زیادی از گندم به کشورهای همسایه قاچاق می‌شود زیرا گندم در این کشورها بین 4 هزار تا 4500 تومان است.

وی گفت: برخی کشاورزان که توانایی انبار کردن گندم را دارند محصول خود را دپو کرده و حاضر نیستند آن را به خرید فعلی تضمینی که دولت اعلام کمرده تحویل دهند و بنا دارند تا هنگام بی‌ثباتی و افزایش نرخ گندم آن را به بازار عرضه کنند.

نماینده مردم دهلران در مجلس افزود: دولت در سال جاری باید به دلیل کمبود گندم 5 میلیون تن از خارج وارد کند این در حالی است که به اعضای هر کیلوگندم وارداتی باید 4 هزار تومان پرداخت کند.

وی خاطرنشان کرد: به هیچ‌وجه کاهش تولید گندم در سال جاری نداریم بلکه شاهد افزایش تولید هم بوده‌ایم اما چون سیاست‌های دولت در اعلام نرخ خرید تضمینی گندم اشتباه و غلط است شاهد آن هستیم که کشاورزان کمتر به تحویل گندم خود به نرخ فعلی خرید تضمینی به دولت تمایل دارند.

کاظم‌زاده تصریح کرد: دولت  در بودجه سال جاری نرخ خرید تضمینی گندم را کیلویی 1600 تومان به مجلس پیشنهاد داده بود اما نمایندگان، یارانه نان صنعتی را حذف کرده و با صرفه‌جویی هزار میلیارد تومانی ایجاد شده این نرخ را به 1700 تومان رساندند.

وی افزود: دولت اگر قیمت خرید تضمینی گندم را تا چند روز آینده اصلاح نکرده و به آن را به 2300 تومان نرساند گندمی که از چرخه تحویل به دولت خارج شده باز نخواهد گذشت و به سمت قاچاق یا خوراک دام سوق  پیدا خواهد کرد.

نماینده مردم دهلران در مجلس تصریح کرد: چرا دولت به جای اینکه گندم را به نرخ پیشنهادی 2300 تومان خریداری کند باید مجبور شود در چند ماه آینده آن را به نرخ 4 هزار تومان از خارج وارد کند و عنایتی هم به کشاورز ایرانی نداشته باشد.

وی افزود: دولت اگر فقط افزایش 500 تومانی به نرخ خرید تضمینی گندم را اعمال کند 5 میلیون تن گندم به سیلوها بازخواهد گشت و اصلاً نیازی به واردات نخواهیم داشت و با این کار صرفه‌جویی خواهیم کرد و به جای آنکه مابه‌التفاوت به جیب کشاورز خارجی برود بهتر است به سفره کشاورز داخلی هدایت شود.

به گزارش خبرگزاری فارس، مخاطبان گرامی با ثبت سوژه‌ای در بخش «فارس من» با عنوان «[گندمکاران را دریابید](https://my.farsnews.com/c/2719)» خواستار پیگیری این موضوع شدند.

**کلمات کلیدی استخراج شده:**

گندم - 6.332133184757053

دولت - 3.28275863810519

تومان - 3.218842194575061

نرخ - 2.634483385579937

کشاورز - 2.278541666666667

مجلس - 1.8077190966779866

فارس - 1.7329479166666668

تحویل - 1.6564260819400904

خرید - 1.6461779622735984

خریداری - 1.568476321403692

صرفه جویی - 1.48875

سال - 1.473391675374434

**عبارات کلیدی استخراج شده:**

بودجه سال جاری نرخ خرید تضمینی گندم - 14.869226095001741

نرخ خرید تضمینی گندم اشتباه - 12.394111974268547

نرخ خرید تضمینی گندم - 11.65841911899164

افزایش نرخ گندم - 10.289459193769591

هیچ وجه کاهش تولید گندم - 9.917770713057298

قیمت خرید تضمینی گندم - 9.625802439263323

گندم تولیدی خوراک دام - 9.299011384861545

کشور خرید گندم - 8.706390806774642

تحویل گندم - 7.988559266697143

دلیل کمبود گندم - 7.688526307797805

نرخ فعلی خرید تضمینی - 6.482088504658655

خبرنگار پارلمانی خبرگزاری فارس - 4.945447916666666

متن دوم

به گزارش خبرگزاری رویترز، دونالد ترامپ در یادداشتی ویژه به «رابرت لایت هایزر» نماینده تجاری آمریکا دستور داد تا در سازمان تجارت جهانی برای اصلاح عنوان کشورهای ثروتمندی که خود را کشورهای «در حال توسعه» می خوانند و از مزایای سازمان تجارت جهانی سوء استفاده می کنند، اقدام کند.

ترامپ در این یادداشت هشدار داده است چنانچه عنوان این کشورها اصلاح نشود و در فهرست کشورهای ثروتمند قرار نگیرند، اقدامی یک جانبه اتخاذ خواهد کرد.

بر اساس این گزارش، رئیس جمهوری آمریکا از شماری از اعضای سازمان تجارت جهانی و به صورت مشخص ۱۰ کشور نام برده که خود را به صورت یک طرفه کشورهای در حالی توسعه می دانند و این در حالی است که به گفته ترامپ آن ها کشورهای ثروتمند هستند و نه در حال توسعه.

ترامپ از کشورهای برونئی، هنگ کنگ، کویت، ماکائو، قطر، سنگاپور، امارات، مکزیک، کره جنوبی و ترکیه به عنوان کشورهای ثروتمند عضو سازمان تجارت جهانی نام برده است.

رئیس جمهوری آمریکا هدف از اصلاح عنوان این کشورها را جلوگیری از سوء استفاده آن ها از مزایای تجارت یاد کرده که به دلیل قوانین و اصول سازمان تجارت جهانی به این مزایا دست پیدا می کنند.

وی تهدید کرد: چنانچه ظرف ۹۰ روز آینده سازمان تجارت جهانی اقدامی در جهت اصلاح این مورد خاص نکند، آمریکا همکاری اش با کشورهای مذکور را به عنوان کشورهای در حال توسعه متوقف خواهد کرد و از هیچ یک از این کشورها در چارچوب سازمان همکاری و توسعه اقتصادی حمایت نخواهد کرد.

ترامپ همچنین در توییتی نوشت: سازمان تجارت جهانی ورشکسته شده در حالی که ثروتمندترین کشورهای جهان به خاطر دور زدن این سازمان و دستیابی به تعاملات اقتصادی خاص، خود را در حالی توسعه معرفی می کنند. کافی است.

بر اساس قوانین سازمان تجارت جهانی، برای کمک به رقابت محصولات تولیدی در کشورهای در حال توسعه، برقراری نظام تعرفه‌های ترجیحی با هدف اعطای امتیازات تجاری به بعضی از فراورده‌های این کشورها مجاز است.

**کلمات کلیدی استخراج شده:**

اقتصادی - 8573848038.776843

تهدید - 8048398700.041297

ثروتمند - 7280097150.891738

یادداشتی - 6814061613.4201975

نخواهد\_کرد - 6708705665.661575

عضو - 6674936495.965869

رئیس - 6353298585.275924

رقابت - 6176944324.73237

ورشکسته - 6107679015.044051

چارچوب - 6032080740.82328

نوشت - 5825708367.383047

کافی - 5631556989.202007

**عبارات کلیدی استخراج شده:**

کشورهای ثروتمند عضو سازمان تجارت جهانی - 17476208619.130886

رئیس جمهوری آمریکا هدف - 16347852879.14289

نماینده تجاری آمریکا دستور - 15346277880.324133

رئیس جمهوری آمریکا - 13812036133.896719

اساس قوانین سازمان تجارت جهانی - 13276005793.20068

آینده سازمان تجارت جهانی اقدامی - 12521659357.14639

چارچوب سازمان همکاری - 11751534796.49197

مزایای سازمان تجارت جهانی سوء - 11189421961.850536

حالی توسعه معرفی - 10974169610.281399

تعاملات اقتصادی خاص - 10935260917.432573

مزایا دست - 10543421413.1888

هدف اعطای امتیازات تجاری - 10494004225.17045

مقالات مطالعه شده

مقاله اول

عنوان مقاله:

**Automatic Keyphrase Extraction based on NLP and Statistical Methods** [1]

در این مقاله که در سال 2011 منتشر شده­است نویسندگان رویکردی مبتنی بر روش­های آماری و الگوهای مبتنی Wordnet برای استخراج عبارات کلیدی ارائه داده­اند. روش آن­ها در واقع یک روش جدید برگرفته شده از دو روش TextRank و TF\*IDF می­باشد. آن­ها روش پیشنهادی خود را بر روی مقالات خبری ارزیابی کردند. آن­ها رویکرد خود را در سه مرحله تدوین کرده­اند: **پیش­پردازش**، **استخراج کلید­واژه** و **استخراج عبارت**.

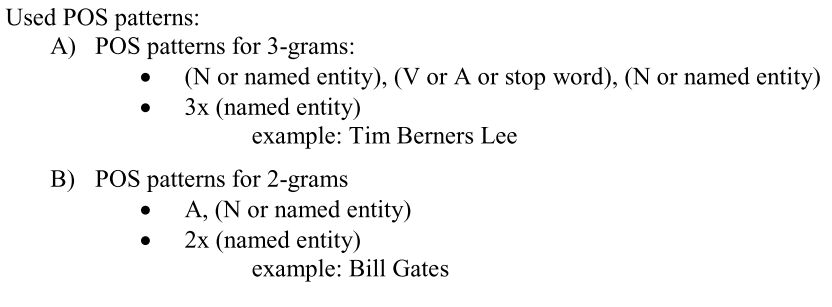
**پیش­پردازش**در مرحله پیش­پردازش بعد از حذف کاراکترهای غیر مهم، کلمه­هایی که اسم یا صفت باشند نگه داشته می­شوند. نویسندگان با این استدلال که کلمات کلیدی برای خوشه­بندی متون استفاده می­شوند، کلماتی که دارای تعداد تکرار کم هستند را حذف می­کنند. در نهایت برای کلمات باقی مانده، مقدار TF\*IDF را محاسبه کرده­اند.

**استخراج کلیدواژه**در این مرحله تنها یک تابع فراخوانی می­شود و با این فراخوانی کلید­هایی که مقدار TF\*IDF آن­ها از یک­پنجم مقدار بیشینه کمتر باشد حذف می­شود. این محدوده می­تواند بر اساس تعداد کلید واژه­های مورد نیاز هم مقدار­دهی شود.

**استخراج عبارت**

در این مرحله ابتدا عبارات مورد علاقه استخراج می­شود. این عبارات می­توانند عباراتی باشند که دارای الگوی POS خاصی می­باشند. سپس امتیاز این عبارت محاسبه می­شود که می­تواند شامل تعداد تکرار عبارت و مقدار TF\*IDF هر کلمه باشد. در نهایت اگر عبارت کاندید در یک همسایگی با کلمه کلیدی و یا یک موجودیت باشد با هم ادغام می­شوند.

برخی از الگوهای به کار رفته شده در به دست آوردن عبارات کاندید به صورت زیر می­باشد:



در آخر نیز کلمات و عبارات با بیشترین امتیاز TF\*IDF به عنوان خروجی الگوریتم برگردانده می­شوند. در پایان نیز آن­ها این گونه ادعا کرده­اند که روش پیشنهادی آن­ها حداقل بر روی داده­هایی با حجم کوچک، از دو روش معروف TextRank پایه و Rake چه در precision و چه در recall بهتر عمل می­کند.

مقاله دوم

عنوان مقاله:

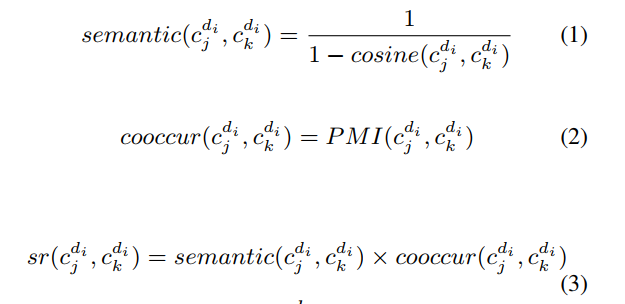
Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific  
Articles using Phrase Embeddings [2]

این مقاله که در سال 2018 چاپ شده­است با استفاده از بردار­های تعبیه­شده اقدام به استخراج عبارات کلیدی متن می­کند. آن­ها محدوده­ی کار خود را بر روی دامنه­ی ­خاصی قرار داده و یک مدل بر روی داده­های آن محدوده ایجاد کردند. آن­ها مقالات علمی که در سایت arxiv.com قرار دارد را به عنوان منابع داده­ای خود انتخاب کردند و با استفاده از آن­ها یک مدل ایجاد کردند. آن­ها در ابتدا عبارت­های کاندید خود را با استفاده از الگوهایی که در نظر گرفته بودند استخراج کردند. از جمله این الگوها می­توان به این موارد اشاره کرد:

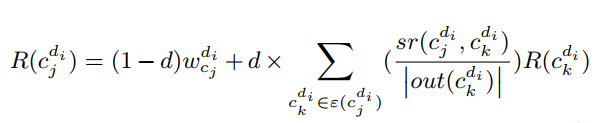
* گزاره­های اسمی و موجودیتی که شامل اعداد باشند حذف می­شوند
* موجودیت­هایی که در رابطه با زمان، تاریخ، درصد، پول، مقدار و ... باشند حذف می­شوند
* کلمات ایست استاندارد حذف می­شوند
* صفت­های معمول و فعل­های گزارش دهنده اگر در ابتدا و انتهای گزاره­های اسمی و موجودیت­ها باشند حذف می­شوند
* ...

نویسندگان برای ایجاد مدل از کتابخانه Fasttext استفاده کرده­اند. آن­ها دلیل عدم استفاده از Word2Vec و Glove را این طور مطرح کرده­اند که این مدل­ها تنها بعد معنایی کلمات را در نظر می­گیرند و به بعد لفظی (ساختاری) کلمات توجه ندارد. نویسندگان بعد از پیش­پردازش داده­ها، اقدام به آموزش مدل خود کرد­ه­اند. آنها­ در آموزش مدل خود اندازه پنجره­ را 5 در نظر گرفتند.

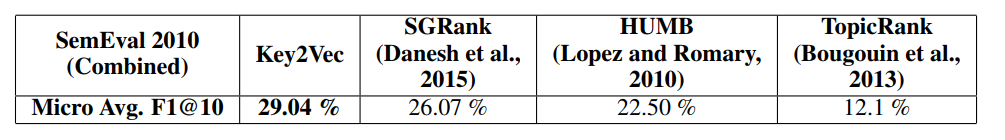
نویسندگان در روش خود به هر سند یک بردار (theme vector) انتصاب می­کنند. این بردار از جمع بردارهای حاصل از بخش­های مهم متن همچون عنوان، و عبارات کاندید به دست می­آید. سپس نویسندگان برای عبارت­های کاندید نیز بردار­هایی استخراج می­کنند.



در ادامه آن­ها گراف خود را ساخته که بتوانند الگوریتم TextRank را بر روی آن اجرا کنند و راس­های با ارزش را استخراج کنند. آن­ها در روش خود هم از ارتباط معنایی و هم ارتباط محلی بین دو راس استفاده کردند. (فرمول 1 و 2)



برای بدست آوردن امتیاز یا­ل­های بین هر دو راس نیز از رابطه­ی بالا استفاده کردند. سپس الگوریتم TextRank را اجرا کرده و ارزش هر راس را بدست آورده­اند. در انتها نیز روش خود را با برخی از روش­های معروف دیگر مقایسه کرده­اند که نتیجه­ی آن به صورت زیر است:

**

مقاله سوم

عنوان مقاله:

**A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction** [3]

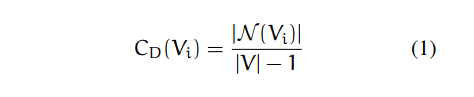
در این مقاله که در سال 2013 منتشر شده­است به مقایسه برخی از روش­های جایگزین برای محاسبه امتیاز یک راس در یک گراف پرداخته است. آن­ها پس از مطالعات خود به این نتیجه گیری رسیدن که استفاده از درجه[[2]](#footnote-2) یک راس برای محاسبه امتیاز آن نتیجه بهتری در مقابل TextRank دارد و همچنین معیار نزدیکی[[3]](#footnote-3) بهترین نتیجه را برای متن­های کوتاه ارائه می­دهد.

آن­ها در استخراج عبارات کلیدی سه مرحله را دنبال می­کنند: اول یک گراف از کلمات ایجاد می­شود، دوم اهمیت هر راس که یک کلمه باشد تعیین می­شود، سوم عبارات کاندید استخراج می­شود و براساس کلمات تشکیل­دهنده امتیاز دهی می­شوند.

نویسندگان بر در روش پیشنهادی خود برای ساخت گراف تنها از کلماتی که اسم یا صفت باشند استفاده کرده­اند. همچنین یال­ها نیز براساس هم­رویدادی[[4]](#footnote-4) بین کلمات تعیین شده­است. در ادامه به راه­های جایگزین که برای امتیاز دهی به راس­ها استفاده شده­است­می­پردازیم:

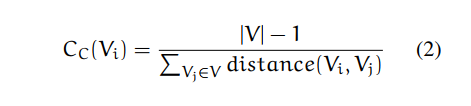
**معیار درجه**

این معیار وابسته به تعداد یال­های یک راس می­باشد و براساس فرمول زیر محاسبه می­شود:



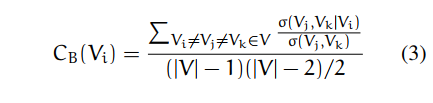
**معیار نزدیکی**

این معیار بدین صورت تعریف می­شود: جمع کوتاه­ترین مسیرها بین یک راس و تمامی راس­های دیگر. فرمول محاسبه این معیار در زیر آورده شده­است.



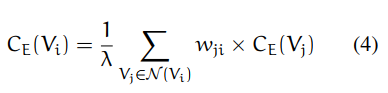
**معیار میانه­ای [[5]](#footnote-5)**

این معیار در واقع بیانگر تعداد باری است که یک راس به عنوان یک پل در کوتاه­ترین مسیری که در بین دو راس دیگر است واقع شده است. فرمول محاسبه به صورت زیر می­باشد:



**معیار Eigenvector**

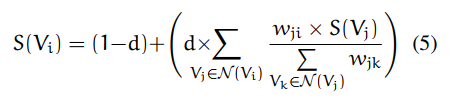
این معیار که در واقع مبنای الگوریتم TextRank می­باشد اهمیت یک راس را بر مبنای اهمیت رئوس همسایه محاسبه می­کند. فرمول محاسبه آن به صورت زیر می­باشد:



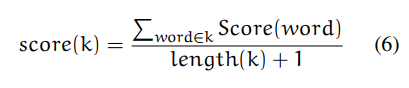
*در این فرمول لاندا یک مقدار ثابت می­باشد.*

**معیار TextRank**

این معیار در واقع بر مبنای معیار Eigenvector می­باشد که مفهوم "رای­دادن[[6]](#footnote-6)" نیز به آن اظافه شده است. در این روش ارزیابی برای هر راس در ابتدا یک مقدار اولیه در نظر گرفته می­شود. فرمول محاسبه این روش به صورت زیر می­باشد:

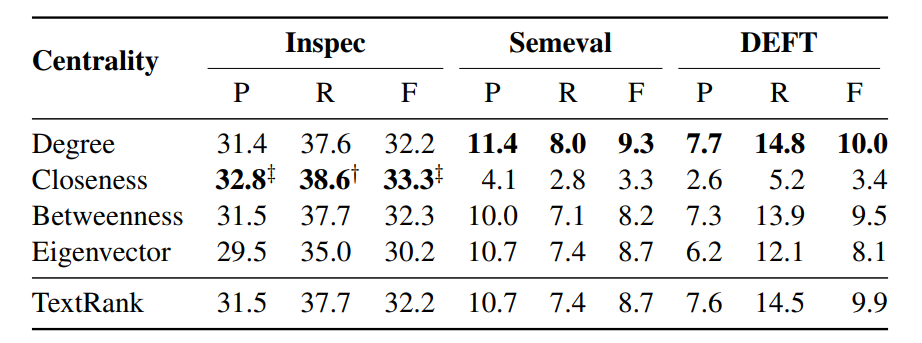


نویسندگان در روش پیشنهادی خود امتیاز عبارت کاندید را بر مبنای کلمات تشکیل دهنده­ی آن به دست می­آورند. فرمول محاسبه امتیاز عبارت به صورت زیر می­باشد:



در نهایت عبارات کاندید امتیاز دهی می­شوند و عبارات کاندید اضافی حذف می­شوند. دو عبارت کاندید در صورتی که stemming یکسانی داشته باشند یکسان تشخیص داده می­شوند و حذف می­شوند.

نویسندگان برای ارزیابی از سه مجموعه دادگان استفاده کرده­اند که نتایج آن به صورت زیر می­باشد:



بر اساس این نتایج این گونه استدلال کردند که ارزیابی رئوس صرفا بر مبنای اندازه­ی درجه­ی آنها در حالی که ساده­ترین ارزیابی می­باشد اما به بهترین نتایج می­رسد. علاوه بر این آن­ها این گونه استدلال کرده­اند که معیار نزدیکی نیز برای متون کوتاه مناسب می­باشد.

مقاله چهارم

عنوان مقاله:

**Simple Unsupervised Keyphrase Extraction using Sentence Embeddings** [4]

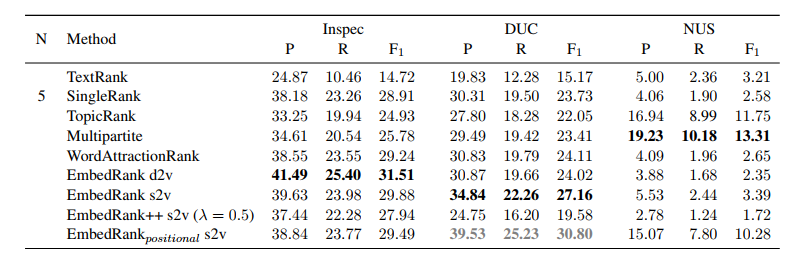
در این مقاله که در سال 2019 منتشر شده است نویسندگان روشی مبتنی بر بردارها ارائه می­دهند که F-Measure بالاتری نسبت به روش­های مبتنی بر گراف ارائه می­کند. موضوع مهمی که نویسندگان در مورد روش خود برجسته می­کنند اطلاع دهندگی[[7]](#footnote-7) و متنوع بودن[[8]](#footnote-8) عبارت استخراج شده ­می­باشد. آن­ها دلیل این امر را امکان محاسبه فاصله بردار عبارت استخراجی و عبارت سند می­دانند که این موضوع باعث می­شود بتوان ارزیابی­هایی از اطلاع دهندگی عبارت داشت. همچنین فاصله برداری بین عبارات کاندید نیز می­تواند نشان دهنده­ی متنوع بودن عبارت­های در نظر گرفته شده باشد.

نویسندگان روش پیشنهادی خود را در سه مرحله بیان می­کنند: در گام اول عبارات کاندید استخراج می­شود که این عبارات دنباله­ایی از صفر یا چند صفت که به دنبال آن یک یا چند اسم می­آید هستند. در گام دوم از sentence embedding برای استخراج بردار برای عبارات و سند استفاده کرده­اند. در گام آخر نیز به امتیاز دهی به عبارات استخراج شده می­پردازند.

نویسندگان در برداری که برای سند استخراج می­شود فقط کلماتی از داکیومنت را نگهداری می­کنند که یا اسم باشند و یا صفت. برای امتیاز دهی به عبارات کاندید از اختلاف کسینوسی بردار کاندید و بردار سند استفاده می­کنند. در نهایت N عبارت با بیشترین امتیاز به عنوان خروجی برگردانده می­شود.

یک مشکلی که نویسندگان به آن اشاره می­کنند این است که عباراتی که برگدانده می­شوند ممکن است که دارای معانی یکسانی با یکدیگر باشند. یعنی اینکه چندین عبارت دارای شکل مختلف باشند اما یک معنای یکسان رو بیان کنند که این باعث می­شود تنوعیت در عبارات برگردانده شده کم باشد. آن­ها در ادامه برای حل این مشکل معیاری به نام MMR[[9]](#footnote-9) را معرفی می­کنند که در واقع یک میانه­ای بین اطلاع دهندگی[[10]](#footnote-10) و تنوعیت[[11]](#footnote-11) عبارات برگردانده شده برقرار می­کند.

آن­ها در ارزیابی خود نشان دادند که روششان در بیشتر دیتاست­های معروف، بهتر از برخی از روش­ها عمل می­کند.



مقاله پنجم

عنوان مقاله:

**NE-Rank: A Novel Graph-based Keyphrase Extraction in Twitter** [5]

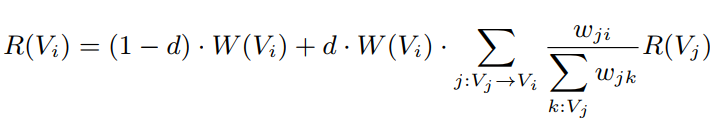
این مقاله که در سال 2012 منتشر شده­است روشی برای استخراج عبارات کلیدی موضوعی از توییت­ها است که در واقع نشان دهنده موضوع توییت­ها می­باشد. آن­ها در روش خود به دلیل غیررسمی بودن توییت­ها، دارای نویز بودن و اندازه کوتاه توییت­ها با چالش­های بیشتری نسبت به سایر داده­ها همراه بوده­اند.

نویسندگان روشی جدید مبتنی بر گراف به نام NE-Rank پیشنهاد کرده­اند که علاوه بر درنظر گرفتن وزن برای هر یال، برای هر راس نیز وزنی در نظر می­گیرد که از آن در امتیازدهی به راس­ها استفاده می­کند. آن­ها نوآوری خود را در دو قسمت قرار داده­اند: یکی ارائه یک رویکرد جدید در الگوریتم مبتنی بر گراف که باعث بهبود امتیازدهی به رئوس می­شود و دیگری یک رویکرد جدید برای استفاده بهینه از هشتگ­ها در توییت­ها برای شناسایی عبارات کلیدی که در توییت پنهان شده­است ارائه داده­اند.

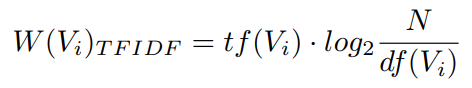
آن­ها در رویکردشان ابتدا از دیتاست خود که شامل تعداد زیادی توییت است، موضوعات مختلف را استخراج کرده­اند. استخراج موضوعات مختلف نیاز است چرا که بتوانند عبارت­کلیدی­موضوعیتی را استخراج کنند. سپس با استفاده از رویکرد مشهور استخراج عبارات­کلیدی که شامل استخراج کلمات­کلیدی، ایجاد عبارات کاندید و امتیازبندی عبارات کاندید می­شود، لیستی از عبارات مهم به­دست ­می­آورند. در انتها نیز آن­ها به استخراج هشتگ­ها می­پردازند.

یکی از نوآوری آن­ها ارائه فرمولی جدید برای امتیاز دهی به نودها برا اساس فرمول TextRank بود که ما در این بخش به آن می­پردازیم. آن­ها در امتیاز دهی خود به هر راس امتیازی نسبت داده­اند که در فرمول قبلی این وجود نداشت. آن­ها گفته­اند که صرف اکتفا به هم­رویدادی بین کلمات برای امتیاز دهی به کلمات کافی نمی­باشد چرا که به طور مثال اگر دو کلمه دارای اهمیت کمی هستند به طور مکرر در کنار هم ظاهر شوند در الگوریتم TextRank امتیاز بالایی می­گیرند. همچنین اگر کلمه­ی مهمی با بعضی از کلمات به طور مکرر ظاهر نشود نمی­تواند با این الگوریتم امتیاز خوبی بگیرد.

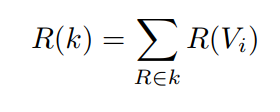
آن­ها این روش امتیازدهی NE-Rank[[12]](#footnote-12) نامیدند که فرمول امتیاز دهی آن به صورت زیر می­باشد:



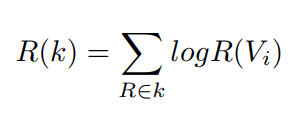
در این فرمول وزن هر راس برابر با TFIDF آن کلمه به صورت زیر می­باشد.



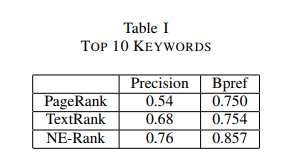
آن­ها بعد از استخراج N کلمه­کلیدی مهم، آن کلماتی که در مجاورت هم قرار گرفته بودند را با هم ادغام کرده و از این طریق عبارات­کلیدی مهم را نیز ساخته­اند. آن­ها برای امتیاز دهی به عبارات کلیدی استخراج شده، به جای اینکه صرفا امتیاز کلمات تشکیل­دهنده­ی آن­ها را با هم جمع کنیم (که به صورت فرمول زیر می­باشد):



بیاییم و لگاریتم امتیاز کلمات تشکیل دهنده آن­ها را با هم جمع کنیم (که به صورت فرمول زیر می­باشد):

**

آن­ها مدعی شده­اند که استفاده از فرمول دوم نتایج بهتری ایجاد کرده­است. آن­ها نتایج ارزیابی خود را به صورت زیر نشان داده­اند. بر طبق این نتایج، روش پیشنهادی آن­ها از دو الگوریتم دیگر بهتر عمل نموده است.



منابع

[1] M. Dostal and K. Ježek, “Automatic Keyphrase based on NLP and statistical methods and Statistical Methods,” *Proc. Dateso 2011 Annu. Int. Work. DAtabases, TExts, Specif. Object*, pp. 140–145, 2011.

[2] D. Mahata, J. Kuriakose, R. R. Shah, and R. Zimmermann, “Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings,” pp. 634–639, 2018.

[3] F. Boudin and L. U. M. R. Cnrs, “A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction,” *Ijcnlp*, no. October, pp. 834–838, 2013.

[4] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi, “Simple Unsupervised Keyphrase Extraction using Sentence Embeddings,” pp. 221–229, 2019.

[5] A. Bellaachia and M. Al-Dhelaan, “NE-Rank: A novel graph-based keyphrase extraction in Twitter,” *Proc. - 2012 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2012*, pp. 372–379, 2012.

1. Balance [↑](#footnote-ref-1)
2. Simple degree centrality [↑](#footnote-ref-2)
3. Closeness centrality [↑](#footnote-ref-3)
4. Co-occurrence [↑](#footnote-ref-4)
5. Betweenness centrality [↑](#footnote-ref-5)
6. Voting [↑](#footnote-ref-6)
7. Informativeness [↑](#footnote-ref-7)
8. Diversity [↑](#footnote-ref-8)
9. Maximal Marginal Relevance [↑](#footnote-ref-9)
10. Informativeness [↑](#footnote-ref-10)
11. Diversity [↑](#footnote-ref-11)
12. Node-Edge Rank [↑](#footnote-ref-12)