به نام خدا

گزارش پروژه درس پردازش زبان طبیعی

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

استاد: دکتر سعیده ممتازی

دانشجو: محمد قربانی - ۹۷۱۳۱۰۹۹

بهار ۹۷-۹۸

فهرست مطالب:

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:

Damping Factor	0.85
Minimum Difference	0.00001
Iteration Steps	10

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 احتمالی به راس جدید خواهیم رفت. این متغیر در فرمول PageRank یک نقش متعادل[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

[1] Balance

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 نقش-های 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 ۲۰ درصد کاهش دارد البته در مناطق سردسیر و استان‌های غربی این میزان نسبت به سال گذشته ۵۰ درصد کاهش پیدا کرده است.

وی افزود: متأسفانه گندم تولیدی خوراک دام می‌شود این در حالی است که هر کیلو جو ۲۳۵۰ تومان بوده و هر کیلو گندم ۱۷۰۰ 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 ٤ هزار تا ٤۵۰۰ 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 ۵ میلیون تن از خارج وارد کند این در حالی است که به اعضای هر کیلوگندم وارداتی باید ٤ 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 ۱۶۰۰ 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 ۱۷۰۰ تومان رساندند.

وی افزود: دولت اگر قیمت خرید تضمینی گندم را تا چند روز آینده اصلاح نکرده و به آن را به ۲۳۰۰ 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 ۲۳۰۰ تومان خریداری کند باید مجبور شود در چند ماه آینده آن را به نرخ ٤ هزار تومان از خارج وارد کند و عنایتی هم به کشاورز ایرانی نداشته باشد.

وی افزود: دولت اگر فقط افزایش ۵۰۰ تومانی به نرخ خرید تضمینی گندم را اعمال کند ۵ 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» «خواستار پیگیری این موضوع شدند .

کلمات کلیدی استخراج شده:

گندم - ۶٫۳۳۲۱۳۳۱۸۴۷۵۷۰۵۳
دولت - ۳٫۲۸۲۷۵۸۶۳۸۱۰۵۱۹
تومان - ۳٫۲۱۸۸۴۲۱۹۴۵۷۵۰۶۱
نرخ - ۲٫۶۳۴۴۸۳۳۸۵۵۷۹۹۳۷
کشاورز - ۲٫۲۷۸۵۴۱۶۶۶۶۶۶۶۷
مجلس - ۱٫۸۰۷۷۱۹۰۹۶۶۷۷۹۸۶۶
فارس - ۱٫۷۳۲۹۴۷۹۱۶۶۶۶۶۶۶۸
تحویل - ۱٫۶۵۶۴۲۶۰۸۱۹۴۰۰۹۰۴
خرید - ۱٫۶۴۶۱۷۷۹۶۲۲۷۳۵۹۸۴
خریداری - ۱٫۵۶۸۴۷۶۳۲۱۴۰۳۶۹۲
صرفه جویی - ۱٫۴۸۸۷۵
سال - ۱٫۴۷۳۳۹۱۶۷۵۳۷۴۴۳۴

عبارات کلیدی استخراج شده:

بودجه سال جاری نرخ خرید تضمینی گندم - ۱۴٫۸۶۹۲۲۶۰۹۵۰۰۱۷۴۱
نرخ خرید تضمینی گندم اشتباه - ۱۲٫۳۹۴۱۱۱۹۷۴۲۶۸۵۴۷
نرخ خرید تضمینی گندم - ۱۱٫۶۵۸۴۱۹۱۱۸۹۹۱۶۴
افزایش نرخ گندم - ۱۰٫۲۸۹۴۵۹۱۹۳۷۶۹۵۹۱
هیچ وجه کاهش تولید گندم - ۹٫۹۱۷۷۷۰۷۱۳۰۵۷۲۹۸
قیمت خرید تضمینی گندم - ۹٫۶۲۵۸۰۲۴۳۹۲۶۳۳۲۳
گندم تولیدی خوراک دام - ۹٫۲۹۹۰۱۱۳۸۴۸۶۱۵۴۵
کشور خرید گندم - ۸٫۷۰۶۳۹۰۸۰۶۷۷۴۶۴۲
تحویل گندم - ۷٫۹۸۸۵۵۹۲۶۶۶۹۷۱۴۳
دلیل کمبود گندم - ۷٫۶۸۸۵۲۶۳۰۷۷۹۷۸۰۵
نرخ فعلی خرید تضمینی - ۶٫۴۸۲۰۸۸۵۰۴۶۵۸۶۵۵
خبرنگار پارلمانی خبرگزاری فارس - ۴٫۹۴۵۴۴۷۹۱۶۶۶۶۶۶۶

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 - ۸۵۷۳۸۴۸۰۳۸٫۷۷۶۸۴۳
تهدید - ۸۰۴۸۳۹۸۷۰۰٫۰۴۱۲۹۷
ثروتمند - ۷۲۸۰۰۹۷۱۵۰٫۸۹۱۷۳۸
یادداشتی - ۶۸۱۴۰۶۱۶۱۳٫۴۲۰۱۹۷۵
نخواهد_کرد - ۶۷۰۸۷۰۵۶۶۵٫۶۶۱۵۷۵
عضو - ۶۶۷۴۹۳۶۴۹۵٫۹۶۵۸۶۹
رئیس - ۶۳۵۳۲۹۸۵۸۵٫۲۷۵۹۲۴
رقابت - ۶۱۷۶۹۴۴۳۲۴٫۷۳۲۳۷
ورشکسته - ۶۱۰۷۶۷۹۰۱۵٫۰۴۴۰۵۱
چارچوب - ۶۰۳۲۰۸۰۷۴۰٫۸۲۳۲۸
نوشت - ۵۸۲۵۷۰۸۳۶۷٫۳۸۳۰۴۷
کافی - ۵۶۳۱۵۵۶۹۸۹٫۲۰۲۰۰۷

عبارات کلیدی استخراج شده:

کشورهای ثروتمند عضو سازمان تجارت جهانی - ۱۷۴۷۶۲۰۸۶۱۹٫۱۳۰۸۸۶
رئیس جمهوری آمریکا هدف - ۱۶۳۴۷۸۵۲۸۷۹٫۱۴۲۸۹
نماینده تجاری آمریکا دستور - ۱۵۳۴۶۲۷۷۸۸۰٫۳۲۴۱۳۳
رئیس جمهوری آمریکا - ۱۳۸۱۲۰۳۶۱۳۳٫۸۹۶۷۱۹
اساس قوانین سازمان تجارت جهانی - ۱۳۲۷۶۰۰۵۷۹۳٫۲۰۰۶۸
آینده سازمان تجارت جهانی اقدامی - ۱۲۵۲۱۶۵۹۳۵۷٫۱۴۶۳۹
چارچوب سازمان همکاری - ۱۱۷۵۱۵۳۴۷۹۶٫۴۹۱۹۷
مزایای سازمان تجارت جهانی سوء - ۱۱۱۸۹۴۲۱۹۶۱٫۸۵۰۵۳۶
حالی توسعه معرفی - ۱۰۹۷۴۱۶۹۶۱۰٫۲۸۱۳۹۹
تعاملات اقتصادی خاص - ۱۰۹۳۵۲۶۰۹۱۷٫۴۳۲۵۷۳
مزایا دست - ۱۰۵۴۳۴۲۱۴۱۳٫۱۸۸۸
هدف اعطای امتیازات تجاری - ۱۰۴۹۴۰۰۴۲۲۵٫۱۷۰۴۵

مقالات مطالعه شده

مقاله اول

عنوان مقاله:

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

در این مقاله که در سال ۲۰۱۱ 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:

Used POS patterns:

A) POS patterns for 3-grams:
- (N or named entity), (V or A or stop word), (N or named entity)
- 3x (named entity)

example: Tim Berners Lee

B) POS patterns for 2-grams
- A, (N or named entity)
- 2x (named entity)

example: Bill Gates

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 ۲۰۱۸ 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 ۵ 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.

$$semantic(c_j^{d_i}, c_k^{d_i}) = \frac{1}{1 - cosine(c_j^{d_i}, c_k^{d_i})} \quad (1)$$

$$cooccur(c_j^{d_i}, c_k^{d_i}) = PMI(c_j^{d_i}, c_k^{d_i}) \quad (2)$$

$$sr(c_j^{d_i}, c_k^{d_i}) = semantic(c_j^{d_i}, c_k^{d_i}) \times cooccur(c_j^{d_i}, c_k^{d_i}) \quad (3)$$

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 ۱ و ۲)

$$R(c_j^{d_i}) = (1-d)w_{c_j}^{d_i} + d \times \sum_{c_k^{d_i} \in \varepsilon(c_j^{d_i})} \left(\frac{sr(c_j^{d_i}, c_k^{d_i})}{\left|out(c_k^{d_i})\right|}\right) R(c_k^{d_i})$$

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:

SemEval 2010 (Combined)	Key2Vec	SGRank (Danesh et al., 2015)	HUMB (Lopez and Romary, 2010)	TopicRank (Bougouin et al., 2013)
Micro Avg. F1@10	**29.04 %**	26.07 %	22.50 %	12.1 %

مقاله سوم

عنوان مقاله:

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

در این مقاله که در سال ۲۰۱۳ 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[۲] یک راس برای محاسبه امتیاز آن نتیجه بهتری در مقابل TextRank دارد و همچنین معیار نزدیکی[۳] 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[۴] 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:

$$C_D(V_i) = \frac{|\mathcal{N}(V_i)|}{|V| - 1} \qquad (1)$$

[2] Simple degree centrality
[3] Closeness centrality
[4] Co-occurrence

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.

$$C_C(V_i) = \frac{|V| - 1}{\sum_{V_j \in V} distance(V_i, V_j)} \quad (2)$$

معیار میانه‌ای [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:

$$C_B(V_i) = \frac{\sum_{V_i \neq V_j \neq V_k \in V} \frac{\sigma(V_j, V_k | V_i)}{\sigma(V_j, V_k)}}{(|V| - 1)(|V| - 2)/2} \quad (3)$$

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:

$$C_E(V_i) = \frac{1}{\lambda} \sum_{V_j \in \mathcal{N}(V_i)} w_{ji} \times C_E(V_j) \quad (4)$$

در این فرمول لاندا یک مقدار ثابت می‌باشد.

معیار TextRank

این معیار در واقع بر مبنای معیار Eigenvector می‌باشد که مفهوم "رای‌دادن[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:

---

[5] Betweenness centrality

[6] Voting

$$S(V_i) = (1-d) + \left(d \times \sum_{V_j \in \mathcal{N}(V_i)} \frac{w_{ji} \times S(V_j)}{\sum_{V_k \in \mathcal{N}(V_j)} w_{jk}} \right) \quad (5)$$

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:

$$\text{score}(k) = \frac{\sum_{\text{word} \in k} \text{Score}(\text{word})}{\text{length}(k) + 1} \quad (6)$$

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:

Centrality	Inspec			Semeval			DEFT		
	P	R	F	P	R	F	P	R	F
Degree	31.4	37.6	32.2	**11.4**	**8.0**	**9.3**	**7.7**	**14.8**	**10.0**
Closeness	**32.8**‡	**38.6**†	**33.3**‡	4.1	2.8	3.3	2.6	5.2	3.4
Betweenness	31.5	37.7	32.3	10.0	7.1	8.2	7.3	13.9	9.5
Eigenvector	29.5	35.0	30.2	10.7	7.4	8.7	6.2	12.1	8.1
TextRank	31.5	37.7	32.2	10.7	7.4	8.7	7.6	14.5	9.9

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 ۲۰۱۹ منتشر شده است نویسندگان روشی مبتنی بر بردارها ارائه می‌دهند که F-Measure بالاتری نسبت به روش‌های مبتنی بر گراف ارائه می-کند. موضوع مهمی که نویسندگان در مورد روش خود برجسته می‌کنند اطلاع دهندگی[۷] و متنوع بودن[۸] 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

[7] Informativeness

[8] Diversity

می‌شود تنوعیت در عبارات برگردانده شده کم باشد. آن‌ها در ادامه برای حل این مشکل معیاری به نام MMR[9] را معرفی می‌کنند که در واقع یک میانه‌ای بین اطلاع دهندگی[10] و تنوعیت[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.

N	Method	Inspec			DUC			NUS		
		P	R	F_1	P	R	F_1	P	R	F_1
5	TextRank	24.87	10.46	14.72	19.83	12.28	15.17	5.00	2.36	3.21
	SingleRank	38.18	23.26	28.91	30.31	19.50	23.73	4.06	1.90	2.58
	TopicRank	33.25	19.94	24.93	27.80	18.28	22.05	16.94	8.99	11.75
	Multipartite	34.61	20.54	25.78	29.49	19.42	23.41	**19.23**	**10.18**	**13.31**
	WordAttractionRank	38.55	23.55	29.24	30.83	19.79	24.11	4.09	1.96	2.65
	EmbedRank d2v	**41.49**	**25.40**	**31.51**	30.87	19.66	24.02	3.88	1.68	2.35
	EmbedRank s2v	39.63	23.98	29.88	**34.84**	**22.26**	**27.16**	5.53	2.44	3.39
	EmbedRank++ s2v ($\lambda = 0.5$)	37.44	22.28	27.94	24.75	16.20	19.58	2.78	1.24	1.72
	EmbedRank$_{positional}$ s2v	38.84	23.77	29.49	39.53	25.23	30.80	15.07	7.80	10.28

[9] Maximal Marginal Relevance

[10] Informativeness

[11] Diversity

مقاله پنجم

عنوان مقاله:

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

این مقاله که در سال ۲۰۱۲ 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 توییت-ها 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[۱۲] نامیدند که فرمول امتیاز دهی آن به صورت زیر می‌باشد:

$$R(V_i) = (1-d) \cdot W(V_i) + d \cdot W(V_i) \cdot \sum_{j:V_j \to V_i} \frac{w_{ji}}{\sum_{k:V_j} w_{jk}} R(V_j)$$

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

$$W(V_i)_{TFIDF} = tf(V_i) \cdot log_2 \frac{N}{df(V_i)}$$

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):

$$R(k) = \sum_{R \in k} R(V_i)$$

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

$$R(k) = \sum_{R \in k} logR(V_i)$$

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.

[12] Node-Edge Rank

Table I
TOP 10 KEYWORDS

	Precision	Bpref
PageRank	0.54	0.750
TextRank	0.68	0.754
NE-Rank	0.76	0.857

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.