به نام خدا

گزارش پروژه درس پردازش زبان طبیعی

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

استاد: دکتر سعیده ممتازی

دانشجو: محمد قربانی

بهار 97-98

فهرست مطالب:

[فهرست مطالب: 2](#_Toc15320873)

[روند کلی انجام پروژه 3](#_Toc15320874)

[خروجیهای بدست آمده 4](#_Toc15320875)

[فرضهای در نظر گرفتهشده 5](#_Toc15320876)

[مقالات مطالعه شده: 6](#_Toc15320877)

[مقاله اول 6](#_Toc15320878)

[مقاله دوم 8](#_Toc15320879)

[مقاله سوم 10](#_Toc15320880)

[مقاله چهارم 13](#_Toc15320881)

[مقاله پنجم 15](#_Toc15320882)

[منابع 18](#_Toc15320883)

روند کلی انجام پروژه

در ابتدا به مطالعه مقاله ارجاع شده در شرح پروژه پرداختیم که یک دید اولیه از الگوریتم PageRank و TextRank به دست آوردیم. 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. در این پیاده­سازی اولیه برنامه فقط قادر به استخراج کلمات کلیدی بود. در ادامه با مطالعه مقالات دیگر در 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.

خروجی­های بدست آمده

فرض­های در نظر گرفته­شده

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**.

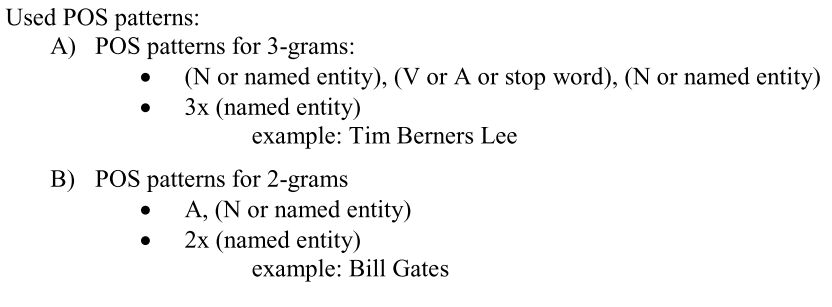
**پیش­پردازش**در مرحله پیش­پردازش بعد از حذف کاراکترهای غیر مهم، کلمه­هایی که اسم یا صفت باشند نگه داشته می­شوند. نویسندگان با این استدلال که کلمات کلیدی برای خوشه­بندی متون استفاده می­شوند، کلماتی که دارای تعداد تکرار کم هستند را حذف می­کنند. در نهایت برای کلمات باقی مانده، مقدار TF\*IDF را محاسبه کرده­اند.

**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:

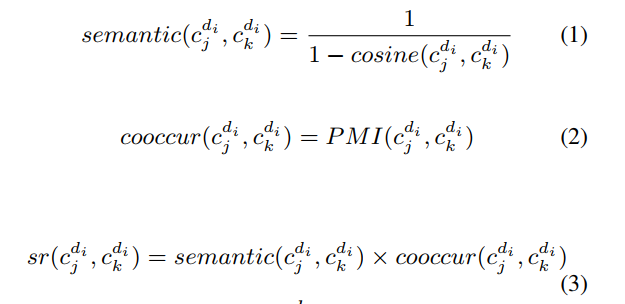
Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific  
Articles using Phrase Embeddings [2]

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 الگوها اشاره کرد:

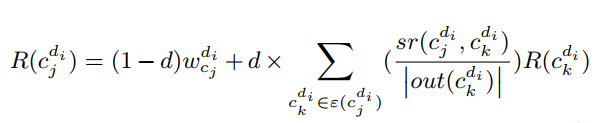
* گزاره­های اسمی و موجودیتی که شامل اعداد باشند حذف می­شوند
* موجودیت­هایی که در رابطه با زمان، تاریخ، درصد، پول، مقدار و ... باشند حذف می­شوند
* کلمات ایست استاندارد حذف می­شوند
* صفت­های معمول و فعل­های گزارش دهنده اگر در ابتدا و انتهای گزاره­های اسمی و موجودیت­ها باشند حذف می­شوند
* ...

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.

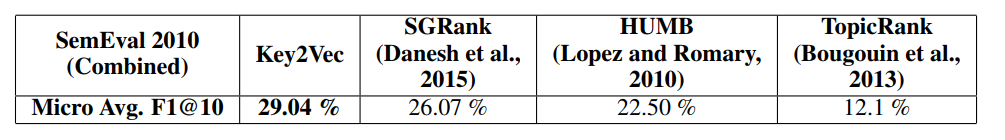
نویسندگان در روش خود به هر سند یک بردار (theme vector) انتصاب می­کنند. این بردار از جمع بردارهای حاصل از بخش­های مهم متن همچون عنوان، و عبارات کاندید به دست می­آید. سپس نویسندگان برای عبارت­های کاندید نیز بردار­هایی استخراج می­کنند.



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]

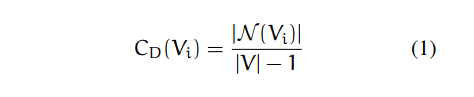
در این مقاله که در سال 2013 منتشر شده­است به مقایسه برخی از روش­های جایگزین برای محاسبه امتیاز یک راس در یک گراف پرداخته است. آن­ها پس از مطالعات خود به این نتیجه گیری رسیدن که استفاده از درجه[[1]](#footnote-1) یک راس برای محاسبه امتیاز آن نتیجه بهتری در مقابل TextRank دارد و همچنین معیار نزدیکی[[2]](#footnote-2) 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[[3]](#footnote-3) بین کلمات تعیین شده­است. در ادامه به راه­های جایگزین که برای امتیاز دهی به راس­ها استفاده شده­است­می­پردازیم:

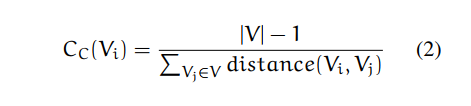
**معیار درجه**

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



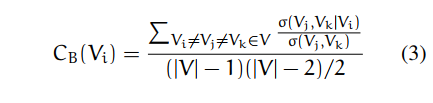
**معیار نزدیکی**

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



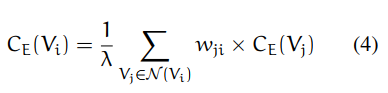
**معیار میانه­ای [[4]](#footnote-4)**

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**

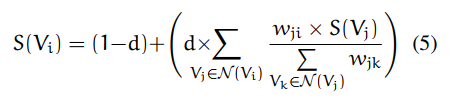
این معیار که در واقع مبنای الگوریتم TextRank می­باشد اهمیت یک راس را بر مبنای اهمیت رئوس همسایه محاسبه می­کند. فرمول محاسبه آن به صورت زیر می­باشد:



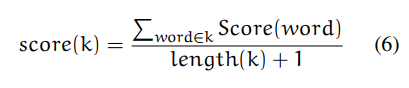
*در این فرمول لاندا یک مقدار ثابت می­باشد.*

**معیار TextRank**

این معیار در واقع بر مبنای معیار Eigenvector می­باشد که مفهوم "رای­دادن[[5]](#footnote-5)" 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:

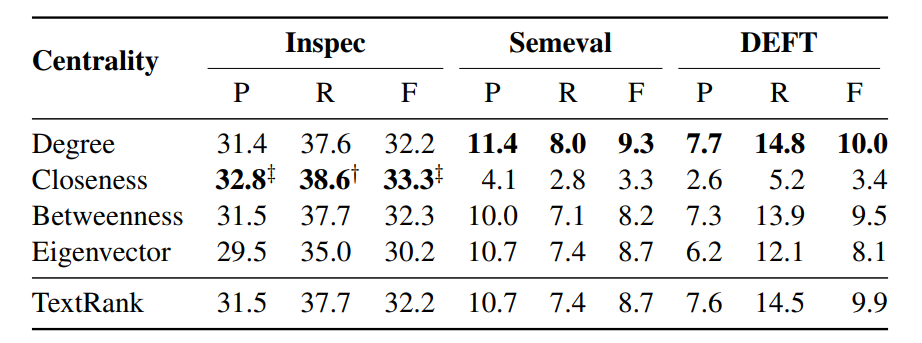


نویسندگان در روش پیشنهادی خود امتیاز عبارت کاندید را بر مبنای کلمات تشکیل دهنده­ی آن به دست می­آورند. فرمول محاسبه امتیاز عبارت به صورت زیر می­باشد:



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]

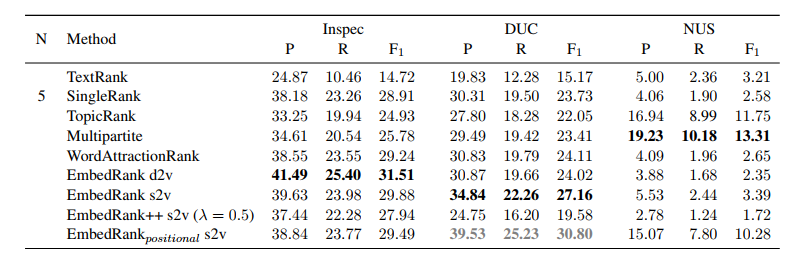
در این مقاله که در سال 2019 منتشر شده است نویسندگان روشی مبتنی بر بردارها ارائه می­دهند که F-Measure بالاتری نسبت به روش­های مبتنی بر گراف ارائه می­کند. موضوع مهمی که نویسندگان در مورد روش خود برجسته می­کنند اطلاع دهندگی[[6]](#footnote-6) و متنوع بودن[[7]](#footnote-7) عبارت استخراج شده ­می­باشد. آن­ها 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[[8]](#footnote-8) را معرفی می­کنند که در واقع یک میانه­ای بین اطلاع دهندگی[[9]](#footnote-9) و تنوعیت[[10]](#footnote-10) 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]

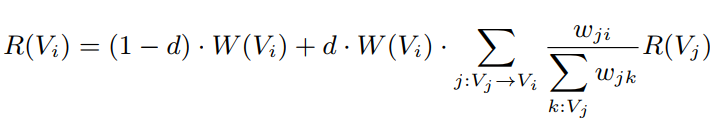
این مقاله که در سال 2012 منتشر شده­است روشی برای استخراج عبارات کلیدی موضوعی از توییت­ها است که در واقع نشان دهنده موضوع توییت­ها می­باشد. آن­ها در روش خود به دلیل غیررسمی بودن توییت­ها، دارای نویز بودن و اندازه کوتاه توییت­ها با چالش­های بیشتری نسبت به سایر داده­ها همراه بوده­اند.

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.

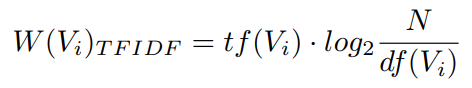
آن­ها در رویکردشان ابتدا از دیتاست خود که شامل تعداد زیادی توییت است، موضوعات مختلف را استخراج کرده­اند. استخراج موضوعات مختلف نیاز است چرا که بتوانند عبارت­کلیدی­موضوعیتی را استخراج کنند. سپس با استفاده از رویکرد مشهور استخراج عبارات­کلیدی که شامل استخراج کلمات­کلیدی، ایجاد عبارات کاندید و امتیازبندی 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.

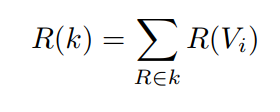
آن­ها این روش امتیازدهی NE-Rank[[11]](#footnote-11) نامیدند که فرمول امتیاز دهی آن به صورت زیر می­باشد:



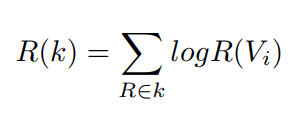
در این فرمول وزن هر راس برابر با TFIDF آن کلمه به صورت زیر می­باشد.



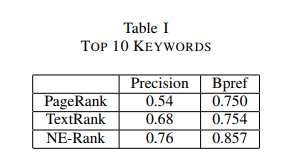
آن­ها بعد از استخراج N کلمه­کلیدی مهم، آن کلماتی که در مجاورت هم قرار گرفته بودند را با هم ادغام کرده و از این طریق عبارات­کلیدی مهم را نیز ساخته­اند. آن­ها برای امتیاز دهی به عبارات کلیدی استخراج شده، به جای اینکه صرفا امتیاز کلمات تشکیل­دهنده­ی آن­ها را با هم جمع کنیم (که به صورت فرمول زیر می­باشد):



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