استخراج چندسندی کلمات کلیدی با استفاده از ترکیب الگوریتم‌های TFIDF و PageRank

ناصر گرامی نیا[1]
[1]دانشگاه آزاد اسلامی واحد مشهد، Geraminia@Gmail.com

چکیده – در این مقاله روشی جدید برای استخراج کلمات کلیدی اسناد متنی پیشنهاد شده است. در روش پیشنهادی، از الگوریتم *PageRank* برای رتبه دهی به هر کدام از کلمات سند متنی و سپس تعیین کلمات کلیدی آن استفاده شده است. با توجه به نواقص موجود در روش های قبلی، با در نظر گرفتن ماهیت الگوریتم *PageRank* ، نحوه نگاشت عناصر مورد استفاده در حوزه جستجوی وب به عناصر معادل آن در حوزه استخراج کلمات کلیدی، به صورتی موثر اصلاح شده است. برای این منظور، پس از تشکیل گراف مربوط به متن، به جای استفاده از شاخص های مرسوم قبلی مثل همجواری یا میزان شباهت گره ها، از شاخص *TFIDF* در وزندهی به یال های گراف استفاده شده است تا انطباق بیشتری با الگوریتم *PageRank* ایجاد شده و برآورد مناسب تری از میزان ارتباط هر کدام از جفت گره ها به دست آید. مقایسه کلمات کلیدی بدست‌آمده توسط روش پیشنهادی با نتایج روش‌های قبلی، بهبودی محسوس در کیفیت و کمیت کلمات کلیدی استخراج شده را به صورت توام نشان می‌دهد.

.

کلید واژه- استخراج کلمات کلیدی، الگوریتم PageRank، الگوریتم TFIDF.

۱- مقدمه

حجم روز افزون منابع اطلاعاتی در دسترس بشر، ضرورت ارایه روش هایی موثرتر برای سازماندهی و بازیابی این منابع را بیش از بیش نمایان کرده است. در این میان، تخصیص کلمات کلیدی به اسناد متنی، یکی از موثر ترین روش ها برای تسریع و تسهیل دسترسی به این منابع بوده است. کلمه کلیدی[1] یا عبارت کلیدی[2] به کلمه یا مجموعه‌ای از کلمات یک سند متنی اطلاق می‌شود که اهمیت مفهومی ویژه ای در آن داشته و حامل پیام اصلی موجود در متن است. با توجه به این خاصیت ذاتی کلمات کلیدی، استفاده از آنها در حوزه های مختلفی از علوم بازیابی اطلاعات[3] و پردازش زبان طبیعی[4] مثل نمایه سازی[5] کتاب‌ها و دسته بندی[6]، خوشه بندی[7] و خلاصه سازی[8] اسناد متنی رایج شده است.

کاربرد رو به گسترش کلمات کلیدی متون در حوزه های مختلف از یک طرف، و تخصصی بودن، گران بودن و زمان‌بر بودن فرآیند استخراج دستی کلمات کلیدی از طرف دیگر، باعث نیاز به روش‌هایی دقیق و کارا در استخراج خودکار کلمات کلیدی شده است.

روش‌های استخراج خودکار کلمات کلیدی به دو دسته کلی استخراجی[9] و انتسابی[10] تقسیم می‌شوند. در روش‌های استخراجی، کلمات کلیدی دقیقاً از بین کلمات موجود در سند و بدون هیچ‌گونه تغییری گزینش و سپس رتبه‌دهی[11] می‌شوند. اما در روش انتسابی، معمولاً کلمات کلیدی استخراج شده با کلمات مناسب دیگری مثل شکل ریشه‌ای یا مترادف آنها جایگزین می-شود. اگر در فرآیند استخراج کلمات کلیدی از الگوریتم‌های یادگیری ماشین برای رتبه‌دهی به کلمات استفاده شود، فرایند استخراج یک فرآیند استخراج نظارت شده[12] خواهد بود. در غیر این صورت به این فرایند، استخراج بدون نظارت[13] گفته می‌شود. در این مقاله از یک روش استخراجی بدون نظارت استفاده شده است.

۱-۱- TFIDF

روش های پایه استخراج کلمات کلیدی بر اساس اطلاعات آماری مربوط به هر کدام از کلمات موجود در عمل می کنند. الگوریتم TFIDF یکی از اصلی ترین الگوریتم های آماری مورد استفاده در این مورد است. این الگوریتم روی یک مجموعه اسناد متنی عمل کرده و با دریافت یک یا چند کلمه جستجو، تمامی اسناد مجموعه را بر اساس میزان انطباق با این کلمه رتبه بندی

کرده و سپس تعداد یک یا چند تا از منطبق ترین اسناد را به عنوان نتیجه جستجو برمی گرداند. این پارامتر، با استفاده از حاصل‌ضرب نرخ تکرار کلمه[14] در سند اصلی و نیز معکوس تعداد اسناد[15] مجموعه که شامل این کلمه هستند، میزان انطباق این کلمه با هر کدام از اسناد مجموعه را برآورد می کند.. در این مقاله، TFIDF با استفاده از رابطه (۱) محاسبه شده است [1]:

$$TFIDF(t,d) = TF(t,d) \times \log \frac{|D|}{|\{d : t \in d\}|} \qquad (۱)$$

که در آن t کلمه مورد جستجو، d سند متنی مورد نظر، D مجموعه اسنتاد متنی، TF(t,d) نسبت تعداد تکرار کلمه t در سند d به تعداد کل کلمات این سند، |D| تعداد اسناد مجموعه D و |{d:t∈d}| تعداد اسناد شامل کلمه t در مجموعه D است.

هر چند معیار TFIDF، حاوی اطلاعات مفیدی در مورد میزان انطباق یک کلمه و یک سند متنی است، اما نمی توان از آن به عنوان شاخصی دقیق برای شناسایی میزان کلیدی بودن هر کدام از کلمات یک سند استفاده کرد، زیرا این شاخص روی مجموعه‌ای از اسناد عمل کرده و در صورت اعمال آن روی یک سند جداگانه، فقط تعداد تکرار کلمه در آن سند بدست می آید. لازم به توضیح نیست که تعداد تکرار کلمه نمی تواند برآورد دقیقی از کلیدی بودن یک کلمه در سند باشد.

PageRank -۲-۱

یکی از روش هایی که برای رفع این مشکل ارایه شده، استفاده از الگوریتم های رتبه دهی مبتنی بر گراف مثل PageRank است. الگوریتم PageRank [8] بر مبنای مدل زنجیره مارکوف عمل کرده و با نگاشت صفحات وب و لینک های موجود بین آنها یه یک زنجیره مارکوف، سعی در یافتن مهم‌ترین صفحات با استفاده از حالت سکون این زنجیره دارد. در الگوریتم PageRank، نحوه مشاهده صفحات وب توسط کاربران، یک زنجیر مارکوف در نظر گرفته شده و بر این فرض استوار است که هر کاربر در هنگام مرور صفحات وب، به احتمال زیاد صفحه بعدی را فقط از روی پیوندهای صفحه جاری انتخاب می‌کند. با این فرض می‌توان وب را به صورت یک زنجیر مارکوف در نظر گرفت که صفحات وب گره‌ها، و پیوندهای بین صفحات، یال‌های گراف جهت دار مربوط به زنجیره مارکوف هستند. توزیع ایستایی که بر اساس آن یک رتبه عددی نهایی به هرکدام از گره ها نسبت داده می شود، طبق رابطه (۲) محاسبه می شود:

$$\text{Score}(v_i) = \frac{1-d}{|V|} + d \times \sum_{\forall j \in \text{In}(v_i)} \frac{1}{\text{Out}(v_j)} \times \text{Score}(v_j) \qquad (۲)$$

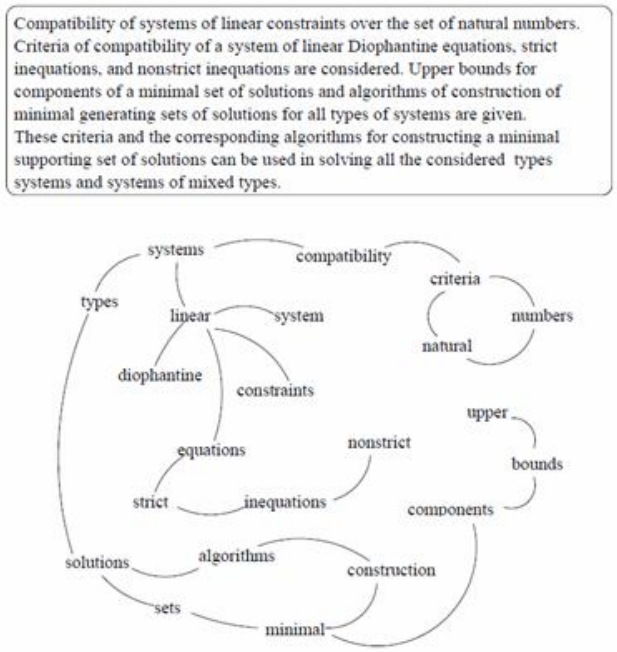
$\text{In}(v_i)$ مجموعه تمام گره‌هایی است که پیوندی به گره v_i دارند و $\text{Out}(v_j)$ مجموعه همه گره‌هایی است که از گره v_j پیوندی به آنها وجود دارد. همچنین d احتمال انتخاب صفحه بعدی از میان لینک های صفحه جاری است که معمولاً برابر ۰,۸۵ در نظر گرفته می شود.

در این مقاله، روش جدیدی برای استخراج کلمات کلیدی یک سند ارایه شده که بر اساس ترکیب الگوریتم‌های PageRank و TFIDF عمل می‌کند. برای اثبات کارایی روش پیشنهادی نسبت به روش پایه، آزمایش‌هایی روی مجموعه داده‌های مورد استفاده در روش‌های قبلی انجام و ارزیابی‌های لازم نیز انجام شده است. نتایج به دست آمده نشان دهنده برتری کمی و کیفی روش پیشنهادی در استخراج کلمات کلیدی است.

۲- کارهای مشابه انجام شده

میهالکا [2] روشی جدید به نام TextRank برای استخراج کلمات کلیدی اسناد متنی (به صورت تک‌سندی) ارایه کرده که ابتدا بر اساس همجواری کلمات، گراف مربوط به متن را ایجاد کرده و سپس الگوریتم PageRank را روی آن اعمال می کند. در این روش کلمات اصلی موجود در متن (که اسامی و صفات فرض شده‌اند)، به عنوان گره های گراف در نظر گرفته شده و در صورتی که هر جفت از این کلمات در فاصله ای به طول حداکثر w کلمه (w پیش فرض برابر ۲ است) از یکدیگر قرار گرفته باشند، بین آنها یالی رسم می شود. در نهایت با اعمال الگوریتم PageRank روی گراف به دست آمده و به دست آوردن حالت سکون گراف، گره های دارای بیشترین رتبه عددی به عنوان کلمات کلیدی متن شناخته می شوند. در شکل ۱ نحوه ایجاد گراف متناظر با یک متن نمونه نشان داده شده است [2].
تساترونیس [3] از روش پیشنهادی میهالکا استفاده کرده و شباهت میان کلمات هر یال با استفاده از Wikipedia و WordNet محاسبه و در وزندهی یال‌ها استفاده کرده است. لیو [4] با ادغام کلمات مترادف در یک گره، نحوه ساخت گراف را تغییر داده و سپس الگوریتم PageRank را روی گراف اعمال کرده است. لیو [5] [6] با استفاده از WordNet و HowNet گراف مربوط به متن را با استفاده از معانی مختلف هر کدام از کلمات آن را تشکیل داده و سپس با دو بار اعمال الگوریتم PageRank، ابتدا از بین معانی مختلف کلمات معنی اصلی را

یافته و سپس با حذف کلمات هم معنی اضافی، کلمات کلیدی را پیدا کرده است.



شکل ۱: نحوه ساخت گراف مربوط به یک متن نمونه

۳- روش پیشنهادی

در این مقاله روشی جدید و موثر برای استخراج کلمات کلیدی هر سند جداگانه متعلق به یک مجموعه اسناد متنی ارایه شده است. در این روش، دو الگوریتم PageRank و TFIDF به صورتی موثر با یکدیگر ترکیب شده اند تا هر کدام، مشکلات دیگری را به شکلی مناسب پوشش دهند.

هر چند الگوریتم PageRank، موفقیت فراوانی در حوزه جستجو و به خصوص موتور جستجوی گوگل به همراه داشته و به یک الگوریتم رتبه‌دهی ایده‌آل تبدیل شده است، اما استفاده از این الگوریتم در فرآیند استخراج کلمات کلیدی نتوانسته باعث تکرار نتایج موفقیت آمیز قبلی شده و بهبودهای به دست آمده عمدتاً جزیی و وابسته به مجموعه داده مورد استفاده بوده است [7]. به عقیده ما، این مشکل از ناحیه الگوریتم PageRank نبوده و ناشی از عدم نگاشت صحیح عناصر و ارتباطات موجود در دو حوزه جستجوی وب و استخراج کلمات کلیدی استاد متنی، به یکدیگر است. این موضوع در جدول ۱ نشان داده شده است.

همانطور که در جدول ۱ مشاهده می شود، الگوریتم PageRank روی مجموعه ای از صفحات وب عمل کرده و بر اساس لینک های موجود بین آنها (که به عنوان اهمیت هر کدام از صفحات مقصد از دید صفحات مبدا لینک دهنده محسوب می شود)، با اهمیت ترین صفحات محاسبه و صفحاتی که بیشترین و بهترین لینک ها را به دست آورده اند، شناسایی می کند، و این معادل صفحاتی است که بیشترین اهمیت را از دید دیگر صفحات دارند. اما از آنجا که در الگوریتم TextRank این نگاشت به درستی انجام نشده و کلمات به عنوان گره های گراف و یال ها معادل همجواری کلمات در نظر گرفته شده اند، از اعمال الگوریتم PageRank روی این گراف، تنها می توان انتظار شناسایی کلماتی را داشت که بیشترین همجواری را با سایر کلمات آن سند داشته اند، که به هیچ وجه نشان دهنده میزان کلیدی بودن این کلمه نخواهد بود.

جدول ۱: مقاسیه نحوه نگاشت عناصر و ارتباطات موجود از الگوریتم PageRank به الگوریتم TextRank

---	PageRank	TextRank
گره	**صفحه وب**	**کلمه**
نحوه ایجاد یال	**لینک بین صفحات**	**هم جواری بین کلمات**
مجموعه کلی	**کل صفحات وب**	**یک سند متنی**
نتیجه نهایی به دست آمده	**صفحاتی که بیشترین و بهترین لینک ها را به خود اختصاص داده اند**	**کلماتی که بیشترین همجواری را با سایر کلمات دارند**

در روش پیشنهادی، برای رفع این معضل ملاک ایجاد یال بین کلمات تغییر کرده و به دلایلی که گفته شد، به جای همجواری کلمات از TFIDF آنها استفاده شده است. در صورتی که حاصل ضرب TFIDF هر جفت از کلمات اصلی متن (اسم ها و صفت ها) از حد آستانه مشخصی عبور کند، یک یال بین آن دو کلمه ایجاد می شود که وزن آن، همان حاصل ضرب TFIDF جفت کلمه است. در این حالت، از آنجا که لینک بین جفت کلمات بر اساس ارزش توام TFIDF آنها ایجاد شده و نشان دهنده میزان کلیدی بودن این دو کلمه از دیدگاه یکدیگر است، لذا با اعمال الگوریتم PageRank می توان انتظار داشت که کلیدی ترین کلمات متن مشخص شود. البته در صورتی که تنها یک سند متنی موجود باشد، برای تمامی جفت کلمات، قسمت دوم موجود در پارامتر TFIDF (معکوس تعداد اسناد) فاقد تاثیر و برابر ۱ بوده و وزن لینک ایجاد شده صرفاً برابر حاصل ضرب تعداد تکرار هر کدام ازکلمات در متن سند خواهد بود.

از آنجا که گراف ایجاد شده در روش پیشنهادی دارای یال های وزندار است و الگوریتم TextRank بر اساس گراف های بدون وزن ارایه شده، اصلاح رابطه (۲) ضروری است. رابطه ۳ شکل اصلاح شده رابطه (۲) است:

$$\text{Score}(v_i) = \frac{1-d}{|V|} + d \times \sum_{\forall j \in \text{In}(v_i)} \frac{w_{ji}}{\sum_{k \in \text{Out}(v_j)} w_{jk}} \times \text{Score}(v_j) \quad (۳)$$

۴- ارزیابی روش پیشنهادی

از آنجا که عمده کارهای انجام شده در این حوزه، با الگوریتم TextRank مقایسه شده اند، برای امکان ارزیابی و مقایسه مستقیم کارایی روش پیشنهادی، مجموعه داده INSPEC که در [2] استفاده شده، عیناً به‌کار برده شده است. این مجموعه شامل ۲۰۰۰ خلاصه مقالات علمی است که برای هر کدام دو سری عبارات کلیدی تهیه شده است. کلمات کلیدی سری اول به صورت کنترل شده و از میان مجموعه عبارات از پیش تعریف شده موجود در یک فرهنگ لغت استخراج شده، اما در سری دوم، کلمات کلیدی به صورت دستی و کاملاً آزاد تهیه شده است. در این مقاله، از روش دوم برای انتخاب کلمات کلیدی استفاده شده و کلمات، بدون هر گونه کنترلی عیناً از متن اسناد انتخاب شده‌اند. این ۲۰۰۰ خلاصه مقاله در سه گروه آموزشی شامل ۱۰۰۰ مقاله، آزمایشی شامل ۵۰۰ مقاله و ارزیابی شامل ۵۰۰ مقاله دسته بندی شده‌اند که مجدداً مطابق [] تنها از ۵۰۰ خلاصه موجود در مجموعه ارزیابی استفاده شده است و نتایج حاصل از روش پیشنهادی، صرفاً با استفاده از مقالات موجود در مجموعه ارزیابی، تحلیل و بررسی شده است.

شاخص‌های ارزیابی مورد استفاده در این مقاله، شاخص‌های ارزیابی رایج مورد استفاده در حوزه بازیابی اطلاعات، یعنی دقت[۱۶]، بازخوانی[۱۷] و F-measure بوده است. رابطه های (۴)، (۵) و (۶) نحوه محاسبه این شاخص‌ها را نشان می دهد.

$$\text{Precision} = \frac{\text{Correct Extracted}}{\text{Total Extracted}} \quad (۴)$$

$$\text{Rcall} = \frac{\text{Correct Extracted}}{\text{Total Manual}} \quad (۵)$$

$$\text{F-measure} = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۶)$$

که Correct Extracted تعداد کلمات (و عبارات) کلیدی استخراج شده صحیح هر سند، Total Extracted تعداد کلمات کلیدی استخراج شده برای هر سند و Total Manual تعداد کلمات کلیدی موجود در مجموعه داده برای هر سند است.

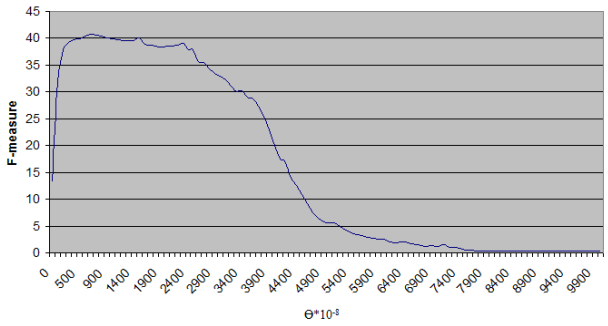
برای ارزیابی نتایج حاصل از روش پیشنهادی، کلمات و عبارات کلیدی استخراج شده سند، تنها در صورتی که عیناً در مجموعه کلمات کلیدی آزاد آن سند موجود باشند، در شکل یک انتخاب درست فرض شده‌اند. در جدول ۲ نتایج حاصل از روش پیشنهادی با الگوریتم پایه TextRank مقایسه شده است:

جدول ۲: مقایسه نتایج به دست آمده توسط الگوریتم TextRank و روش پیشنهادی

روش مورد استفاده	Precision	Recall	F-measure
الگوریتم TextRank	۳۱,۲	۴۳,۱	۳۶,۲
روش پیشنهادی	۳۷,۴	۴۴,۶	۴۰,۶۸

از آنجا که که دو روش فوق با تنظیمات مختلفی انجام شده و هر کدام از این تنظیمات نتایج مخصوص خود را به همراه داشته، در مقایسه انجام شده، مقادیر مربوط به بهترین F-measure هر کدام از آنها ذکر شده است. مهم ترین پارامتر موثر در الگوریتم TextRank، طول پنجره همجواری است که با طول ۲ بهترین نتایج به دست آمده است. همچنین پارامتر عمده موجود در روش پیشنهادی، پارامتر آستانه حاصلضرب TFIDF جفت کلمات برای ایجاد یال بین آنهاست که با Θ نشان داده شده است. چنانچه در جدول ۲ دیده می‌شود، روش پیشنهادی هم از لحاظ دقت و هم از لحاظ بازخوانی، برتری دارد.

همچنین برای بررسی دقیق میزان تأثیر پارامتر آستانه حاصلضرب TFIDF (Θ)، آزمایش‌هایی با مقادیر مختلف Θ انجام شده و در هر مورد هر سه پارامتر دقت، بازخوانی و F-measure ارزیابی شده‌اند. شکل ۲ منحنی کارایی روش پیشنهادی بر اساس مقادیر مختلف Θ را نشان می دهد:



شکل ۲: مقایسه کارایی روش پیشنهادی بر حسب مقادیر مختلف Θ

طبق شکل ۲، با افزایش مقدار Θ، ابتدا کارایی افزایش پیدا کرده و سپس ابتدا با شیبی ملایم و در نهایت با شیب بسیار زیادی کاهش می یابد. این موضوع از آنجا می‌شود که هر چند با افزایش اولیه آستانه حاصلضرب TFIDF مورد نیاز برای تشکیل یال بین کلمات (Θ) کارایی الگوریتم به دلیل حذف تعداد زیادی از کلمات نامرتبط افزایش پیدا می کند، اما در ادامه و با افزایش بیش از حد این پارامتر، تعداد گره های حذف شده گراف به

شدت زیاد شده و امکان حذف کلمات کلیدی اصلی از گراف بیشتر می شود. مطابق شکل، بهترین کارایی الگوریتم با مقدار $\Theta=630*10^{-8}$ به دست آمده است.

۵- نتیجه گیری

در این مقاله برای استخراج کلمات کلیدی اسناد متنی، روش جدیدی بر مبنای ترکیب الگوریتم های TFIDF و PageRank ارایه شده است. در این روش، نحوه ایجاد گراف مربوط به متن بر اساس حاصلضرب TFIDF هر کدام از چفت کلمات تغییر کرده است. آزمایش‌هایی برای بررسی کمیت و کیفیت روش پیشنهادی، انجام شده و شاخص‌های دقت، بازخوانی و F-measure ارزیابی شده‌اند. این آزمایش‌ها با استفاده از مجموعه داده مورد استفاده در روش قبلی انجام شده است. نتایج به دست آمده نشان‌دهنده برتری کمی و کیفی محسوس روش پیشنهادی نسبت به روش قبلی است.

مراجع

[1] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, First Edition, Cambridge University Press, pp.257, 2008.

[2] R. Mihalcea, P. Traau, "TextRank: Bringing order into texts", *EMNLP'04*, pp. 404-411, 2004.

[3] G. Tsatsaronis, I. Varlamis, K. Nørvåg, "SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs", *COLING '10*, Beijing, pp.1074-1082 , 2010.

[4] L. Zhengyang, L. Jianyi, , Y. Wenbin, W. Cong "Keyword Extraction using PageRank on Synonym Networks", *E-Product E-Service and E-Entertainment (ICEEE*), pp. 1-4,2010

[5] W. Jinghua, L. Jianyi, W. Cong "Keyword Extraction Based on PageRank", *Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pp. 857-864, Berlin, 2007.

[6] W. Jinghua, L. Jianyi, W. Cong "Keyword Indexing System with HowNet and PageRank", *Networking, Sensing and Control*, pp. 389-392, 2008.

[7] K. Saidul, Ng. Vincent, "Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art", *COLING '10*, Beijing, pp. 365-373, 2010.

[8] S. Brin, L. Page, "The Anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, pp.107-117, 1998.

[1] Keyword
[2] Keyphrase
[3] Information Retrieval
[4] Natural Language Processing
[5] Indexing
[6] Classification
[7] Clustering
[8] Summarization
[9] Extractive
[10] Abstractive
[11] Ranking
[12] Supervised
[13] Unsupervised
[14] Term Frequency
[15] Inverse Document Frequency
[16] Precision
[17] Recall