ایمان خلیل الرحمانی

اجرای کد اصلی باTFIDF

- **پیش‌پردازش و تاثیر آن:**

در کد کلاس، ابتدا بخش‌هایی از پیش‌پردازش مانند پاکسازی متن، نرمال‌سازی حروف، حذف توقف‌کلمات و استمینگ/لمتایز کردن اعمال شده‌اند. حذف یا اضافه کردن این قسمت‌ها می‌تواند تاثیر چشمگیری بر نتایج داشته باشد.

  - **حذف برخی پیش‌پردازش‌ها:**

برای نمونه، اگر حذف توقف‌کلمات انجام نشود، کلمات پرتکرار و بی‌محتوا می‌توانند در محاسبه TFIDF وزن زیادی بگیرند که دقت نتایج را کاهش می‌دهد.

  - **اضافه کردن پیش‌پردازش‌های اضافی:**

افزودن تکنیک‌هایی مانند تصحیح املایی یا حذف نویز می‌تواند به بهبود کیفیت متن و در نتیجه نتایج TFIDF کمک کند.

- **فاکتور وزن‌دهی:TFIDF**

TFIDFبه عنوان یک معیار متداول برای ارزیابی اهمیت کلمات در اسناد استفاده می‌شود. در این کد، با تغییرات پیش‌پردازشی (مثلاً تغییر در حذف توقف‌کلمات یا استفاده از استمینگ متفاوت) نتایج جستجو برای یک پرس‌وجوی مشخص تغییراتی خواهند داشت. به عنوان مثال، در پرس‌وجویی مانند "آموزش ماشین لرنینگ"، اگر توقف‌کلمات حذف نشوند، ممکن است کلمات عمومی مانند "آموزش" بیش از حد برجسته شوند و نتایج دقیق کمتری ارائه شود.

مقایسه نتایج TFIDF وBM25

- **نتایج:TFIDF**

در استفاده ازTFIDF ، کلمات پرتکرار در یک سند به‌طور نسبی اهمیت کمتری می‌یابند. اما در برخی موارد ممکن است اسناد طولانی‌تر به دلیل فراوانی کلمه امتیاز بالاتری بگیرند حتی اگر از لحاظ معنا مرتبط نباشند.

- **نتایج:BM25**

BM25با در نظر گرفتن طول سند و پارامترهای تنظیمی، نتایجی دقیق‌تر ارائه می‌دهد. در پرس‌وجویی مانند "آموزش ماشین لرنینگ"، BM25ممکن است اسنادی را که به صورت متوازن درباره موضوع بحث می‌کنند، بالاتر رتبه‌بندی کند.

- **نتیجه‌گیری:**

  - **پیش‌پردازش :**اضافه یا حذف قسمت‌های پیش‌پردازش می‌تواند به بهبود کیفیت هر دو روش کمک کند، اما تاثیر آن در BM25به دلیل پارامترهای تنظیمی کمتر از TFIDF متغیر است.

  - **امتیازدهی** BM25 **:**به دلیل انعطاف‌پذیری و در نظر گرفتن طول سند، معمولاً عملکرد بهتری در بازیابی اطلاعات دارد و نتایج مرتبط‌تری ارائه می‌کند.

- **کاربرد :**در شرایطی که اسناد دارای طول‌های متفاوت هستند، BM25معمولاً ترجیح داده می‌شود.

نتیجه نهایی گزارش:

در این گزارش ابتدا کد اصلی TFIDF اجرا و بررسی شد، سپس تغییرات لازم جهت پیاده‌سازی BM25 اعمال گردید. با مقایسه نتایج به‌دست‌آمده مشاهده می‌شود که BM25 در شرایطی که تنوع طول اسناد وجود دارد، نتایج مرتبط‌تر و دقیق‌تری ارائه می‌دهد. به همین دلیل در پروژه‌های بازیابی اطلاعات مدرن، استفاده از BM25 به عنوان یک معیار رتبه‌بندی بهینه‌تر تلقی می‌شود.