به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

پروژه بازیابی اطلاعات

فاز ۱ زیربخش اول

اروند درویش

۹۸۳۱۱۳۷

در این بخش ما باید اسناد موجود را که به صورت فایل JSON در اختیار داریم ابتدا با استفاده از کتابخانه pandas در پایتون خوانده و روی آن پیش پردازش های لازم را انجام دهیم.

پیش پردازش های ما عبارت است از : توکنایز کردن هر یک از اسناد (tokenizer)، نرمالایز کردن هر سند (normalization)، ریشه یابی (stemming) و حذف کلمات پرتکرار (stop words) هر سند که با استفاده از کتابخانه های از پیش تعریف شده‌ی hazm و parsivar این کار را انجام می دهیم.

دلیل هر یک پیش پردازش ها:

دلیل توکنایز کردن که بدیهی و مشخص است، ما برای داشتن دیکشنری باید متن (content) را به صورت توکن های مجزا لیست کنیم تا بتوانیم روی آنها پردازش های لازم را انجام دهیم.

اما دلیل نرمالیزیشن این است که ما می‌خواهیم همه متون استاندارد و یکپارچه شوند. برای مثال:

- تبدیل اعداد انگلیسی به فارسی
- حذف فاصله و نیم فاصله‌ها و فضاهای خالی اضافی
- حذف تشدید از کلمات(چون تشدید مخصوص زبان عربی است)
- ....

با این کار از زیاد شدن تعداد مدخل های دیکشنری جلوگیری می‌کنیم و حجم دیکشنری را کاهش می دهیم.

دلیل حذف کلمات پرتکرار و علائم نگارشی نیز کاهش حجم دیکشنری و بهبود بخشیدن سرعت و دقت پردازش روی سند های ما با تمرکز بیشتر روی کلماتی که خود معنای اصلی را دارند که معمولا مدنظر سرچ ما هستند تا stop word ها، است.

در ریشه یابی یا stemming نیز ما کلمات با ریشه های یکسان را یکی فرض می‌کنیم مثل "کتاب‌ها" که دارای ریشه "کتاب" است یا "می‌خوردن" که دارای ریشه "خورد" است. با این کار برنامه ما نتایج بهتری را برمیگرداند و باعث بهبود در اکثر موارد می شود.

پایان