بسمه تعالی

پروژه پایانی درس بازیابی اطلاعات

- **عنوان :** پیاده سازی و ارزیابی سیستم بازیابی اطلاعات

در این پروژه قصد داریم یک سیستم بازیابی اطلاعات پیاده‌سازی کرده و با استفاده از مجموعه داده برچسب گذاری شده (در این مجموعه اسناد مرتبط با هر پرس‌وجو مشخص شده‌اند)، دقت سیستم پیاده‌سازی شده را بررسی نمائیم. پارامترهای مورد استفاده شامل دقت، فراخوانی و معیار F می‌باشند. در این پروژه مراحل کار شامل موارد زیر است:

۱) دریافت متون اسناد(CISI.ALL)، نرمال‌سازی، توکن‌سازی و ساخت جدول posting list و آمار فراوانی کلمات در اسناد.
۲) اعمال پرس و جوهای موجود در مجموعه داده (CISI.QRY) و محاسبه معیارهای ارزیابی برای هر کدام و میانگین‌گیری برای محاسبه میزان دقت کلی مدل‌های پیاده‌سازی شده. اسناد مرتبط با هر پرس‌وجو در فایل (CISI.REL) نگهداری می‌شوند. با توجه به اینکه در روش‌های برداری لیستی از اسناد به همراه امتیاز برگردانده می‌شود، برای ارزیابی می‌توانید با در نظر گرفتن محدودیت بر روی تعداد اسناد بازیابی شده (مثلا ۱۰ سند اول) ارزیابی را انجام داد.
۳) بررسی تاثیر هر کدام از بخش‌های پیش‌پردازش در نتیجه نهایی.
- به عنوان مثال بخش آیا حذف کلمات توقف تاثیر مثبت دارد یا خیر. برای پیدا کردن جواب می تواند ابتدا با حذف کلمات توقف میانگین امتیاز F را محاسبه کرد و در ادامه این کار را بدون حذف کلمات انجام داد و نتیجه را با هم مقایسه کرد.
۴) بررسی تاثیر روش‌های مختلف وزن دهی بر روی نتیجه نهایی.
- به عنوان مثال اگر تنها از معیار فرکانس کلمات استفاده کنیم با زمانی که فرکانس معکوس سند را هم استفاده کنیم چه تغییری در نتایج خواهیم داشت.

درنهایت جدول پیوست را تکمیل نمائید.

- **مجموعه داده CISI** :

لینک جهت دانلود: http://ir.dcs.gla.ac.uk/resources/test_collections/cisi

لینک جهت توضیحات : https://www.pragmalingu.de/docs/guides/data-comparison/

- **تاریخ تحویل:**

بارگزاری در سامانه VU: یکشنبه ۹ بهمن ساعت ۲۳:۵۹

تحویل حضوری : دوشنبه ۱۰ بهمن ساعت ۱۰:۰۰

پیوست:

روش استفاده شده	پیش پردازش	Precision	recall	F-measure
فقط TF	حذف S.W با ریشه یابی	0.25	0.006	0.009
	عدم حذف S.W با ریشه یابی	0.1	0.003	0.005
	حذف S.W بدون ریشه یابی	0.25	0.006	0.009
	عدم حذف S.W بدون ریشه یابی	0.01	0.004	0.005
TF*IDF بدون نرمال سازی	حذف S.W با ریشه یابی	0.25	0.006	0.009
	عدم حذف S.W با ریشه یابی	0.01	0.004	0.005
	حذف S.W بدون ریشه یابی	0.25	0.006	0.009
	عدم حذف S.W بدون ریشه یابی	0.1	0.004	0.005
TF*IDF نرمال سازی شده	حذف S.W با ریشه یابی	0.25	0.006	0.009
	عدم حذف S.W با ریشه یابی	0.1	0.004	0.005
	حذف S.W بدون ریشه یابی	0.01	0.004	0.005
	عدم حذف S.W بدون ریشه یابی	0.25	0.006	0.009

با ارزوی موفقیت

سلطانی