به نام خدا

گزارش فاز دوم پروژه‌ی بازیابی اطلاعات

امیرحسین بینش 9531012

بخش اول | محاسبه‌ی امتیاز tfidf

تیتر خبر: ۱۵۰ راس گوسفند در یک حادثه آتش سوزی در پارس آباد تلف شد.

۵تای اول از اول: طویله، بهمنی، علوفه، پاسگاه، گوسفند

۵تای آخر از آخر (بعد از حذف امتیازات کم در Index Elimination): افزود، کرده، انتهای، اینکه، بیان

زیاد بودن امتیاز ممکن است صرفا بخاطر کم تکرار شدن آن کلمه باشد، مثلا غلط املایی یا تفاوت در نگارش که از دست normalizer در رفته باشد می‌تواند idf زیادی داشته باشد و باعث افزایش ورن شود. در اینجا همانطور که مشخص است "بهمنی" چون در کل اسناد کم تکرار شده (چون خیلی کم پیش می‌آید "بهمنی" در کوه‌های ایران اتفاق بیافتد یا فامیلی کسی "بهمنی" باشد) امتیاز بالایی آورده.

حدس من این بود که بهمنی گله‌دار بوده ولی با جستجو فهمیدم این بهمنی رئیس همون پاسگاه است.

بخش دوم | پاسخگویی به پرسمان در فضای برداری

این بخش با کنسول انجام می‌شود و فقط تیتر نتایج نمایش داده می‌شوند. مشکل برعکس بودن فارسی حل نشده (خروجی برعکس چام می‌شود ولی بازهم حروف جدا از هم هستند)

بخش سوم | افزایش سرعت با champions list

از آنجا که champions list با کمک tf-idf ها تعیین می‌شود، بین نتایج هر دو روش تغییر زیادی حس نمی-شود.

سرعت استفاده از tf-idf برای کوئری دو کلمه‌ای به طور میانگین ۲.۲ ثانیه بود ولی با روش champions list به ۳۰۰ میلی ثانیه کاهش پیدا کرد.

فرهنگیان:

- اجرای سند تحول و تامین مهیشتی فرهنگیان دو محور اصلی بودجه سال آینده(سه بار)
- پنجمین همایش ملی تربیت معلم اول آبان برگزار می‌شود
- احداث پارکینگ اداری در آموزش و پرورش ناحیه یک اهواز

ابریشم:

- سی امین قطار گردشگری بین المللی عقال طلایی به یزد رسید
- قطار گردشگری عقاب طلایی به یزد رسید
- نخستین برداشت پاییزه پیله‌تر در گلستان
- فرصت ثبت نهایی قنات قاسم آباد از بین رفت

آرامش:

- عظیم زاده – آخرین تحولات در بیروت
- اخبار اربعین ۹۸ | موج بازگشت زائران کربلای معلی از مرز مهران
- آخرین جزئیات ناآرامی های شهر بیروت
- جهده کرج – چالوس یک طرفه شد

پیاده‌سازی

۱. ساخت شاخص معکوس

a. خواندن فایل ورودی: اسناد بصورت لیستی از رشته‌ها ذخیره می‌شوند

b. Linguistic Modules

- در روش ساده، ابتدا علایم اضافی مانند علامت سوال و تگ‌های html حذف می‌شوند و سپس کلمات با فاصله از هم تبدیل به توکن‌ها می‌شوند و در نهایت کلمات پرتکرار از توکن‌های نهایی حذف می‌شوند.
- در روش پیشرفته که بخاطر تعداد زیاد افعال و چک کردن تا دو کلمه‌ی بعد برای یافتن کلمه‌ی خاص خیلی کندتر انجام می‌شود. در این روش ابتدا علایم اضافی مانند قبل حذف می‌شود و بعد با lookahead دوتایی توکن‌ها جدا می‌شوند و در نهایت کلمات پرتکرار حذف می‌شوند. بعد از آن عملیات نرمال سازی(حذف اعراب و تبدیا ک عربی به فارسی) و ریشه‌یابی انجام می‌شود که از Persian stemmer برای این کار کمک گرفتم.

  بعد از ساخت توکن‌ها توکن‌های تکراری حذف می‌شوند و وارد مرحله‌ی ساخت شاخص معکوس می‌شوند.

c. ساخت شاخص

در این مرحله ابتدا از کل توکن‌ها یک دیکشنری(مجموع کل کلمات نه ساختار داده دیکشنری) می‌سازیم و بعد به ازای هر کلمه‌ی دیکشنری در شاخص معکوس که با دیکشنری پیاده‌سازی شده، یک کلید اضافه می‌کنیم و تمام توکن‌های ساخته شده در هر سند را iterate می‌کنیم تا شاخص معکوس ساخته شود. در اینجا هر کلمه کلید یک لیست به عنوان مقدار خودش دارد که اولین عنصر آن تعداد اسنادی است که آن کلمه در آن تکرار شده‌اند.

۲. امتیازدهی

در این مرحله با روش tf-idf و بعد نرمال کردن آن، برای هر سند یک tf-idf می‌سازیم که کلمات و امتیازات در آن نگهداری می‌شود.

برای Index Elimination کلماتی که امتیاز آن‌ها در این سند کمتر از threshold است را حذف می‌کنیم.

۳. جستجو

در این مرحله ابتدا اطلاعات ذخیره‌شده در مراحل قبل را می‌خوانیم و سپس از کاربر کوئری می-گیریم و با مقایسه امتیازات Champions list برای واژه‌های کوئری‌ها نزدیکی کسینوسی را می‌سنجیم و در نهایت ده سند برتر را به ترتیب باز می‌گردانیم.
این قسمت با کلاس Heap انجام شده که نودهای آن شامل یک اندیس سند و یک امتیاز tf-idf است.