به نام ایزد منان

تمرین اول درس بازیابی اطلاعات، «روش‌های سنتی بازیابی اطلاعات»

استاد درس: دکتر ممتازی

پاییز ۹۹ – دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

نکاتی در مورد این تمرین که نیاز به توجه و دقت دوستان دارد:

۱- در جدول زیر نحوه اعمال جریمه تاخیر در ارسال تمرین‌ها ذکر شده است.

میزان تاخیر (روز)	میزان جریمه
۱ الی ۲ روز	هر روز ۵٪
۲ الی ۶ روز	هر روز ۱۰٪

در صورتی که بین ۷ تا ۱۴ روز تاخیر داشته باشید، نمره شما از ۵۰٪ محاسبه می‌شود و پس از این بازه با توجه به سایر تمرین‌ها و زمان تحویل، به تمرین ارسالی نمره‌ای تعلق نمی‌گیرد.

۲- هرگونه کپی کردن باعث عدم تعلق نمره به تمامی افراد مشارکت کننده در آن می‌شود.

۳- آخرین مهلت ارسال تمرین، ساعت **۲۳:۵۵** دقیقه روز **چهارشنبه ۱۴ آبان ماه** می‌باشد. این زمان با توجه به جمع‌بندی‌های صورت گرفته، شرایط و با توجه به سایر تمرین‌ها در نظر گرفته شده است و **قابل تمدید نمی‌باشد**.

۴- دوستان فایل ارسالی خود را به صورت فشرده و به صورت «شماره دانشجویی_HW01» مانند HW01_99131123 نام گذاری کنید. در این فایل باید مواردی نظیر کدها، فایل گزارش و سایر موارد مورد نیاز در هنگام بررسی وجود داشته باشد و صرفا این فایل در روز ارائه در نظر گرفته می‌شود.

۵- این تمرین دارای تحویل در محیط اسکایپ می‌باشد. زمان آن پس از یک هفته از پایان مهلت تمرین از طریق مودل درس اعلام می‌شود.

۶- زبان برنامه‌نویسی این تمرین می‌تواند پایتون، سی‌پلاس‌پلاس و یا جاوا باشد. (پیشنهاد ما پایتون است).

۷- کدهای خود را به صورت مناسب کامنت گزاری کنید. به صورتی که بتوان حداقل روال اجرا و موارد مورد نیاز را درک کرد.

۸- سعی کنید ابتدا تمامی سوالات و بخش‌ها را مطالعه کنید.

۹- استفاده از هیچ کتابخانه آماده‌ای به جز موارد مطرح شده در تمرین مجاز **نمی‌باشد** و شما باید تمامی موارد را پیاده‌سازی کنید.

۱۰- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یار، درس از طریق ایمیل در ارتباط باشید.

amirmohammadrostami@yahoo.com

بخش اول – معرفی دادگان

دادگان[1] ارائه شده در این تمرین شامل ۴ فایل Corpus، judgement، Query و Retrieval می‌باشد که در ادامه به شرح جزئیات هر کدام می‌پردازیم.

دو فایل Corpus و Retrieval از نمونه‌هایی با فرمت زیر تشکیل شده‌اند:

DID	شماره یا آیدی نمونه (سند[2])
Date	تاریخ ارائه و انتشار سند
Cat	دسته بندی آن سند
Main Text	متن سند

فایل *Query* شامل کوئری[3]های مورد استفاده در این تمرین می‌باشد که هر کوئری یا نمونه این فایل از فرمت زیر تشکیل شده است.

QID	شماره یا آیدی کوئری
Title	عنوان و متن اصلی
Description	توضیحی در مورد کوئری
Narrative	اطلاعات و توضیح مختصر

فایل judgement از سطرهای مانند نمونه زیر و برای ارزیابی و آزمایش مدل‌ها تشکیل شده است که بیانگر سند مناسب هر کوئری می‌باشد.

```
QID    DID    1
```

[1] Dataset
[2] Document
[3] Query

بخش دوم – بازیابی با استفاده از مدل فضای برداری[4] (۴۵ امتیاز)

در این قسمت شما با استفاده از اسناد موجود در فایل `Retrieval` اسناد را به صورت `TF-IDF` نمایش دهید. (برای محاسبه `IDF` از `Corpus` که شامل تعداد بیشتری سند می‌باشد، استفاده کنید.) برای این امر دو حالت زیر را در نظر بگیرید.

۱ـ بردار `TF` اسناد و کوئری‌های به صورت باینری باشد.
۲ـ بردار `TF` اسناد و کوئری‌های به صورت شمارشی باشد.

به طور مثال بردار حاصل برای کوئری «`f a a a b c d`» به صورت زیر می‌باشد:

نوع بردار	بردار	[a,b,c,d,e,f]					
باینری	1	1	1	1	0	1	
شمارشی	1	3	1	1	0	1	

سپس با استفاده از دو معیار فاصله[5] `Cosine` و `Jaccard` ۱۵سند مشابه به کوئری را گزارش کنید. اگر در کوئری‌ها کلمه‌ای جدید وجود داشت که در دادگان فایل `Retrieval` وجود نداشت، آن کلمه را در نظر نگیرید. در گزارش خود تحلیل خود را از معیارهای فاصله و ۲ حالت استفاده از بردار `TF` را شرح دهید.

بخش سوم – بازیابی با استفاده از مدل احتمالاتی BM25[6] (۳۰ امتیاز)

با استفاده از مدل *BM25* ۱۵ سند مرتبط برای هر کوئری را بازیابی و گزارش کنید. تاثیر هایپر پارامترها b و k را تحلیل کنید. (حداقل سه مقدار برای هر یک از آن‌ها در نظر بگیرید.)

بخش چهارم – ارزیابی روش‌ها (۲۵ امتیاز)

تمامی روش‌های مطرح شده در بخش دوم و سوم را با استفاده از معیارهای *P@10*، *P@5*، *MAP*[7] و *MRR*[8] را ارزیابی و گزارش کنید. تحلیل خود را از نتایج بدست آمده به صورت مختصر بیان کنید.

---

[4] Vector Space
[5] Distance Metric
[6] Best Match
[7] Mean Average Precision
[8] Mean Reciprocal Rank

بخش آخر- برخی نکات در مورد گزارش و تمرین

- دادگان مطرح شده در این تمرین و تمامی بخش‌ها را می‌توانید از [لینک][9] پانوشت دانلود کنید.

- در این تمرین شما مجاز به استفاده کتابخانه‌های زیر و موارد مشابه و هم‌کاربرد با آن‌ها می‌باشد (تمامی کتابخانه‌های استاندارد پایتون، مجاز می‌باشند.):

```
numpy, scipy, pandas, genism, pickle, hazm
```

- در این تمرین سعی شده است علاوه بر آشنایی شما با کاربرد مباحث ارائه شده در کلاس و لمس بهتر آن، خلاقیت و حل چالش شما نیز ارزیابی شود. لذا در صورتی که در این تمرین چالشی وجود دارد که شما راه حلی برای آن ارائه دادید و استفاده کردید، آن را در گزارش بیان کنید. اما اگر مشکلی بزرگ وجود دارد که نیاز به بررسی مجدد دارد، آن را از طریق ایمیل با تدریس‌یاران درس مطرح کنید.

- در صورتی که هر گونه پیش پردازش بر روی دادگان انجام دادید آن را در گزارش خود بیان کنید.

- این تمرین **۱** نمره از بارم کلی شما از تمرینات را با توجه به پوشش مباحث و حجم تمرین دارد. امتیاز این تمرین از ۱۰۰ محاسبه می‌شود که بارم هر بخش مشخص شده است.

- در تمامی بخش‌ها، میزان نتایج شما در ارزیابی شما تاثیر چندانی ندارند (مگر اینکه بسیار دور باشد). بلکه میزان تسلط، دیدگاه و پیاده سازی، تحلیل‌ها و خلاقیت شماست که در نمره شما تاثیر مستقیم دارد و بر اساس این موارد مورد ارزیابی قرار می‌گیرید.

موفق باشید – رستمی

[9] https://drive.google.com/drive/folders/1cX-6tLr076TtbdnhcPCOYlhCRsiyEVCq?usp=sharing