درس مبانی بازیابی اطلاعات و جستجوی وب

استاد: دکتر محمود نشاطی

ترم اول سال تحصیلی 1400/1399

دانشکده مهندسی و علوم کامپیوتر

تمرین دوم: آشنایی و تحقیق درباره مفاهیم

آخرین تاریخ تحویل: 15 آذر ماه

مقدمه:

در این تمرین درس شما با برخی مفاهیم آشنا خواهید شد. هدف این تمرین علاقه مند کردن و آشنایی شما با برخی مفاهیم پرکاربرد است.

خواسته ها:

در این تمرین از شما انتظار می‌رود در مورد مفاهیمی مانند QL(Query likelihood) و LM و NER را به صورت مختصر و ساده توضیح دهید(با تصاویر و گام به گام) و روش های مختلف NER و ابزار های مختلف پایتون آن را بررسی و مقایسه کنید و بهترین روش را بیابید.

در دو جلسه گذشته در مورد ElasticSearch و وزن دهی به قسمت های مختلف یک query آموختیم. روی صد خبر از اخبار عصر ایران کوئری ای که شامل عنوان چهار دسته از دسته بندی های اخبار است را بررسی کنید و با وزن دادن به هر دسته بندی(هر قسمت از کوئری) در معیار map نتایج را بهبود دهید این وزن دهی میتواند با روشی برای هر سند شخصی سازی شود(روش عام منظوره برای این هدف نمره اضافی دارد) یا یک وزن دهی بهینه برای کل اسناد استخراج شود. کد پایتون الگوریتم را ارسال کنید. توجه کنید که query شما باید از حداقل چهار قسمت تهیه شده باشد و با یک روش بهینه وزن مناسب برای هر کدام از چهار قسمت را بیابید.

معیار ارزیابی:

الگوریتمی مناسب تر است که برای آن سند، به قسمتی از کوئری که نام دسته بندی آن سند است وزن بیشتری بدهد.

جستجو بر روی فیلد دسته بندی ها مجاز نمیباشد و صرفا بر روی متن و عنوان خبر جستجو انجام شود.

نکات مهم:

- این تمرین تماما به صورت گروهی انجام شود
- استفاده از elasticsearch 7.6 در این تکلیف توصیه میشود.
- برای ارزیابی معیار و رسم نمودار نتایج از لینک زیر استفاده شود: https://github.com/joaopalotti/trectools
- گرفتن نمره ی کامل منوط به **حضور** در جلسه‌ی احتمالی تحویل اسکایپی و تسلط بر روی کد می باشد که زمان آن متعاقبا اعلام خواهد شد. در این جلسه شما باید کدی که در کورس ویر آپلود کردید را بر روی رایانه ی خود، مجددا کامپایل کرده و از برنامه خروجی بگیرید.
- همچنین هر دو عضو گروه باید بر روی کد کاملا مسلط باشند. **دقت داشته باشید که عدم شرکت در جلسه اسکایپی در صورت تشکیل، موجب از دست دادن درصد زیادی از نمره خواهد شد.**
- تمیزی و خوانایی کد، الزامی است و ترجیحا فایل با پسوند ipynb با توضیحات باشد.
- سورس کد برنامه ، خروجی برنامه(فایل اجرایی) و گزارش را به صورت ir_proj1_{groupNO}.zip بر روی کورس ویر آپلود کنید.
- در صوورتی که از زبان پایتون استفاده می کنید، حتماً باید پکیج هایی که استفاده کردید را در فایل requirements.txt ذخیره کرده باشید. همچنین در مستندات خوود به ورژن پایتونی که استفاده کرده‌اید، اشاره کنید.
- ترجیحا ازآخرین ورژن پکیج های کرالر و زبان های برنامه نویسی استفاده نمایید.
- از خروجی خود یک گزارش با فرمت pdf. تهیه کنید. تعداد صفحات و فرمت گزارش مهم نیست صرفا توضیحاتی کوتاه به همراه چند اسکرین شات از اجرای برنامه و خروجی آن در گزارش گنجانده شود.
- در صورت داشتن هرگونه سوال مرتبط با این پروژه، می توانید با ایمیل omidomkk@gmail.com در ارتباط باشید. لطفا در نظر داشته باشید که سوال هایتان را نهایتاً تا سه روز مانده به ددلاین مطرح کنید، چرا که پس از آن احتمال پاسخ به شما بسیار کم می باشد.

با آرزوی موفقیت و سربلندی
محمدی