بسمه تعالی

تاریخ تحویل نهایی ۱۰ و ۱۷ و ۲۰ آبان ماه ۱۴۰۱

تمرین شماره ۱

عنوان: آشنایی با فرایند اندیس‌گذاری و مدل‌های بازیابی

توضیحات: هدف از انجام این تمرین آشنایی با ابزارهای پیش پردازش متن و همچنین فرایند اندیس‌گذاری ساده در بحث مقدمات بازیابی اطلاعات است. همچنین در این تمرین با دو مدل بازیابی اولیه (بولی و برداری) بصورت عملی آشنا خواهید شد. در این تمرین شما باید تعدادی از اخبار متنی خبرگزاری را پیمایش کرده و عملیات اندیس‌گذاری را برای آنها انجام دهید. سپس با استفاده از خروجی قسمت اندیس‌گذاری، مدل ساده‌ای برای بازیابی اطلاعات پیاده سازی نمائید. این عملیات شامل مراحل زیر خواهد بود:

فاز اول: (انجام در کلاس و تحویل در سامانه vu تا ساعت ۱۳:۵۹ روز ۱۰ آبان ۱۴۰۱)

۱) دریافت متن خبر
۲) انجام پیش پردازش های لازم (جداسازی توکن ها و ریشه یابی و ...) که می‌تواند با استفاده از ابزارهای آماده موجود برای پردازش زبان فارسی، انجام شود.
۳) استخراج تعداد رخداد کلمات در اسناد
۴) ایجاد لغت‌نامه و posting لیست ها
۵) دریافت posting لیست به عنوان ورودی از مراحل قبل

فاز دوم: (انجام در کلاس و تحویل در سامانه vu تا ساعت ۱۳:۵۹ روز ۱۷ آبان ۱۴۰۱)

۶) پیاده سازی مدل بولی
۷) پیاده سازی مدل برداری (tf-idf) و امکان انجام پرس و جوی ساده .

فاز سوم: (تحویل در سامانه vu ساعت ۱۳:۵۹ روز ۲۰ آبان ۱۴۰۱)

۸) پیاده سازی انواع مدل های وزن‌دهی و نرمال سازی برای اسناد و پرس و جو و امکان انجام جستجو بر اساس انتخاب نوع پارامتر ها

منابع: دیتاست ورودی مجموعه اخبار از خبرگزاری باشگاه خبرنگاران جوان و یا خبرگزاری فارس است که از طریق لینک زیر بصورت رایگان قابل دانلود است. این فایلها بصورت json ذخیره شده و هر خط شامل یک خبر به همراه فراداده متناسب با آن است.

https://github.com/Text-Mining/Useful-Corpora-for-Text-Mining-in-Persian-Language

با توجه به حجم بالای دیتاست مورد نظر، می‌توان در colab و با استفاده از دستورات زیر ابتدا دیتاست رو دانلود نموده و سپس در مسیر جاری از حالت فشرده خارج کنیم:

```
!git clone https://github.com/Text-Mining/Useful-Corpora-for-Text-Mining-in-Persian-Language.git
!unrar x '/content/Useful-Corpora-for-Text-Mining-in-Persian-Language/News/FarsNews 97/farsnews.part01.rar'
```

برای فرایندهای پیش پردازش متنی می‌توانید از ابزارهای آماده فارسی استفاده نمائید. دو نمونه از آنها از آدرس های زیر قابل دسترسی است:

https://www.sobhe.ir/hazm/
https://text-mining.ir/persian-nlp-textmining

محدودیت ها: زبان برنامه نویسی پایتون است.

نحوه تحویل: بخشی از تمرین در کلاس بصورت حضوری تحویل گرفته می‌شود و گروه ها باید گزارش از نحوه انجام کار و همچنین کد پایتون مربوط به تمرین خود به همراه لغت نامه و posting لیست های خود را در سامانه vu بارگذاری نمایند.