مبانی بازیابی اطلاعات و جست و جو وب

تمرین شماره 1

در این تمرین به شما دیتاست خلاصه داستان یک سری فیلم به زبان انگلیسی در قالب یک فایل txt داده شده است. فایل مورد نظر تمرین را می توانید از این لینک دانلود کنید. برای تمرین اول باید متون مربوط به فیلم های مختلف در این فایل را پردازش کنید. خلاصه داستان هر فیلم در داخل فایل plot_summaries به صورت زیر است. ابتدا یک شماره یکتا برای فیلم آمده است و با کمی فاصله، متن خلاصه داستان نوشته شده است.

23890098 Shlykov, a hard-working taxi driver and Lyosha, a saxophonist, develop a bizarre love-hate relationship, and despite their prejudices, realize they aren't so different after all.

در تمرین اول باید متن مربوط به هر فیلم را جدا کرده، آن را پیش پردازش کنید و token های مربوط به هر کدام را استخراج کنید. پس از بدست آوردن token ها و انجام پیش پردازش های لازم، باید با استفاده از این token ها تمام متون را به صورت معکوس نمایه گذاری کنید. برای نمایه گذاری کردن از id که در اول هر متن نوشته شده بود برای نمایش فیلم استفاده کنید.

پیش پردازش هایی که می توانید انجام دهید، شامل حذف stopword ها، stemming، lemmatization و دیگر اعمالی که به نظر خودتان منطقی و لازم است می باشد. برای تست سیستم، در آخر باید بتوان یک term را به برنامه داد و لیست فیلم هایی که این کلمه را در خلاصه داستان خود داشتند به همراه تعداد تکرار این کلمات به عنوان خروجی دریافت کرد. بعدا یک تست کیس برای تست کردن برنامه به شما داده میشود.

برای پیاده سازی کد این برنامه، می توانید از هر زبان برنامه نویسی و کتاب خانه ای که خواستید استفاده کنید.

در آخر یک داکیومنت شامل توضیح اعمالی که روی داده انجام دادید، قطعه کد مربوط به هر قسمت از برنامه، توضیح کد و نمونه ای از خروجی را بنویسید. فایل ارسالی شما در vu باید یک فایل zip یا rar شامل این داکیومنت و یک فایل CSV از نمایه گذاری معکوسی که انجام داده اید باشد.