مبانی بازیابی اطلاعات و جست و جوی وب

تمرین شماره 2

در این تمرین باید با استفاده از indexing و توکن هایی که در تمرین قبلی بدست آورده اید، برای Query که خودتان در مورد داستان فیلم ها می نویسید، شباهت کسینوسی را محاسبه کنید. در ابتدا 10 کوئری برای این تمرین می نویسید. سپس امتیاز tf.idf را برای هر کلمه داخل این کوئری ها حساب می کنید و 10 خلاصه داستان فیلم که بیشترین شباهت کسینوسی را با کوئری دارند بر می گردانید.

کوئری هایی که در ابتدا تعریف می کنید باید پس از پیش پردازش طولی بیشتر از 4 کلمه و به صورت جمله سوالی کامل باشند. برای مثال "Does the movie take place in America" یک کوئری مناسب است.

در آخر یک داکیومنت شامل توضیح اعمالی که روی داده انجام دادید، قطعه کد مربوط به هر قسمت از برنامه، توضیح کد، نمونه ای از خروجی جدول tf.idf برای یک متن و 10 کوئری نوشته شده و خروجی مربوط به آن ها را بنویسید. منظور از جدول tf.idf نوشتن امتیاز tf.idf هر کلمه مربوط به plot است. فایل ارسالی شما در vu باید یک فایل zip یا rar شامل این داکیومنت و یک فایل txt از برای هر کوئری، id فیلم و متن خلاصه ده فیلم مرتبط تر بر اساس شباهت کسینوسی باشد.