تمرین سوم پردازش زبان طبیعی

توجه:

- **کد باید فقط در زبان پایتون باشد.**
- **استفاده از کتابخانه‌های از پیش آماده مجاز است. در گزارش کتابخانه ها ذکر شوند.**
- **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**
- **فایل گزارش به فرمت pdf و به زبان فارسی باشد.**

در این تمرین شما یک مدل استخراج اطلاعات از متن را بر پایه الگوریتم TF-IDF طراحی می کنید. دادگانی که در اختیار قرار گرفته شده است، دادگان Lisa است، که شامل مجموعه ای از مستندات و درخواست ها است. فایل های LISA.QUE و LISA.REL مربوط به درخواست ها و نتایج استخراج است، که شما برای ارزیابی استفاده خواهید کرد؛ سایر فایل ها مستندات هستند.

۱. مستندات و درخواست ها را از هر فایل اسخراج کنید، و پیش پردازش های لازم را بر روی آن ها انجام دهید (حذف علائم، اعداد، stopwords). دو نمونه اول از مستندات پس از پیش پردازش در فایلی به نام 1.txt ذخیره و ضمیمه کنید.

۲. تعداد کلمات منحصر به فرد، مستندات و درخواست را مشخص کنید.

۳. الگوریتم TF-IDF را پیاده کنید (توجه داشته باشید که مقادیر فقط برپایه مستندات باید استخراج شوند) . برای درخواست ها ۵، ۱۰، ۲۰ و ۴۰ نزدیک ترین مستندات را استخراج کنید. معیار precision، recall و f1score را برای هر مقدار گزارش کنید و در رابطه با الگوی موجود در آن صحبت کنید. بهترین تعداد مستندات مرتبط را گزارش دهید. (توجه داشته باشید که از فاصله کسینوسی استفاده کنید.)

۴. با استفاده از الگوریتم LSI ابعاد را به دو کاهش دهید. سپس بردار مربوط به ۱۰ درخواست های اول را به همراه نزدیک ترین مستندات شون را در فضای دو بعدی رسم کنید (تعداد نزدیک ترین مستندات بر پایه مسئله قبل است). درخواست ها و مستند هایشان باید هم رنگ باشند، همچنین درخواست ها را با نماد متفاوت از مستندات نشان دهید. (marker و color را در رابطه با matplotlib جست‌جو کنید.)