به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۱-۱۴۰۰

تمرین شماره ۲

تاریخ تحویل:
۱۴۰۰/۰۹/۲۶

---

۱. (۱۵٪) [**تحلیل معنایی پنهان**] در یک پیکره متنی سه سند یک جمله‌ای به صورت زیر وجود دارد. یک کاربر، پرسش "سواد رسانه‌ای" را جستجو می‌کند. پس از حذف ایست واژگان، میزان شباهت پرسش کاربر را با جملات را با استفاده از تحلیل معنایی پنهان (LSA) و معیار تشابه کسینوسی محاسبه نمایید. دقت کنید که محاسبات به صورت گام به گام و شامل همه جزئیات باشد.

سند	محتوا
۱	سواد و دانش او در خبر مورد شک است.
۲	خبر رسانه‌ای فراگیر است
۳	دانش رسانه‌ای او بالا است

راهنمایی: برای تجزیه SVD می‌توانید از تابع SVD در محیط‌های برنامه‌نویسی استفاده کنید. همچنین، برای این تمرین می‌توانید به سایت http://www.wolframalpha.com بروید و در کادر جستجو عبارت "SVD{ {1,2,3},{3,2,1} }=" را بنویسید تا ماتریس $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$ را تجزیه کنید.

---

۲. (۵۰٪) [**پیاده‌سازی- تشابه‌یابی در متن**] به پیوست این تمرین یک پیکره شامل جفت جملات فارسی و میزان عددی تشابه آن‌ها در یک فایل اکسل آورده شده‌است. هدف از این تمرین محاسبه میزان شباهت این جفت جملات با استفاده از روش‌های مختلف استخراج ویژگی و معیارهای مختلف محاسبه شباهت می‌باشد. برنامه‌ای بنویسید که برای هر کدام از بخش‌های زیر، متوسط میزان شباهت را برای ۱۰۰ جفت محاسبه کند. میزان شباهت جفت جملات را با استفاده از معیارهای شباهت کسینوسی و جاکارد بدست آورید.

الف) دادگان را نرمال کرده و ایست واژه‌های آن را حذف کنید. خروجی این مرحله را به صورت یک فایل جدید تولید کنید. در این مرحله واژگان پیکره را استخراج و در فایل دیگری به عنوان Dic.txt قرار دهید.

h.veisi@ut.ac.ir

به نام خدا

ب) میزان شباهت با استفاده از بردار فراوانی کلمه (TF) نرمال شده و معیارهای شباهت کسینوسی و جاکارد محاسبه کنید.

ج) میزان شباهت با استفاده از بردارفراوانی عبارت-معکوس فراوانی سند (TF-IDF) و معیارهای شباهت کسینوسی و جاکارد محاسبه کنید. در این حالت هر جمله را معادل یک سند در نظر بگیرید.

د) برای محاسبه میزان ارتباط نتایج حاصل و امتیازات جفت جملات موجود در پیکره از ضریب همبستگی (correlation coefficient) استفاده می‌شود که نحوه محاسبه آن از طریق رابطه زیر امکان‌پذیر است.

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

در این رابطه x_i مقدار تشابه محاسبه شده توسط شما برای جفت جمله iام است و y_i مقدار تشابه واقعی داده شده در پیکره برای این جفت جمله است. مقدار $\bar{x}$ برابر میانگین مقادیر تشابه محاسبه شده توسط شما برای همه جفت جملات است و $\bar{y}$ میانگین مقادیر عددی تشابه واقعی نمونه‌های پیکره می‌باشد.

در نهایت جدول زیر را تکمیل کنید و نتایج را تحلیل کنید که کدام معیار و کدام روش نمایش نتایج بهتری ارائه می‌دهد. هر خانه این جدول نمایشگر میزان ضریب همبستگی با استفاده از معیار مدنظر و با استفاده از شیوه نمایش مد نظر می‌باشد. نتایج را تحلیل نمایید.

ویژگی / معیار شباهت	فراوانی کلمه	فراوانی کلمه-معکوس فراوانی سند
کسینوسی		
جاکارد		

۳. (۳۵٪) [**پیاده‌سازی- دسته‌بندی متون فارسی با بیز ساده**] قصد داریم یک مدل ساده دسته‌بندی ایمیل فارسی بسازیم که بتواند ایمیل‌های دریافتی اسپم را تشخیص دهد. برای این‌کار داده برچسب خورده زیر را در اختیار داریم.

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۱-۱۴۰۰

تمرین شماره ۲

تاریخ تحویل:
۱۴۰۰/۰۹/۲۶

ایمیل	برچسب
حراج فصل لباس در سایت تگموند شروع شد.	اسپم
خرید کفش‌های ورزشی با تخفیف آخر فصل.	اسپم
سلام، من مینا هستم، امکان صحبت با شما را دارم؟	اسپم
با سلام، لطفا مدارک مربوطه را در سایت دانشکده بارگذاری کنید.	غیراسپم
سلام آقای دکتر کلاس امروز چه ساعتی تشکیل میشه؟	غیراسپم
سلام مدارک را دریافت کردید؟ با تشکر.	غیراسپم
برای خرید بیمه تکمیلی در سایت دانشگاه اقدام کنید.	غیراسپم

با استفاده از داده بالا می‌خواهیم مدلی بسازیم که بتواند غیراسپم/اسپم بودن ایمیل ورودی زیر را تشخیص دهد:

"خرید فصل در سایت"

ابتدا stopwordها را حذف کنید و با استفاده از روش بیز ساده، برای ایمیل ورودی، هر دو احتمال اسپم/غیراسپم بودن را محاسبه کنید. با مقایسه نتایج به دست آمده، عملکرد مدل را ارزیابی کنید.

h.veisi@ut.ac.ir