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. (۱۰٪) [**تحلیل معنایی پنهان**] در یک پیکره متنی دو سند یک جمله‌ای ۱ و ۲ به صورت زیر وجود دارد. یک کاربر، پرسش "توجه خبری" را جستجو می‌کند. پس از حذف ایست واژگان، میزان شباهت پرسش کاربر را با جملات ۱ و ۲ 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	دیروز خبری منتشر شد.
۲	به بخش خبری منتشر شده، توجه نمایید.

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. (۲۵٪) [**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.

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

h.veisi@ut.ac.ir

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 شبکه عصبی MLP] یک برنامه رایانه‌ای بنویسید که با استفاده از دادگان مجموعه *زبرا*، متون فارسی را با روش شبکه عصبی MLP دسته‌بندی کند. برای این کار، هر کدام از ۷ پوشه دادگان زبرا را به عنوان ۷ موضوع (دسته) در نظر بگیرید. دو فایل اول هر پوشه را به عنوان داده تست و هشت فایل دیگر را به عنوان داده آموزش به کار ببرید. در

h.veisi@ut.ac.ir

به نام خدا

این تمرین تعداد ۱۵۰ واژه پرتکرار را (بعد از حذف ایست واژه‌ها) به عنوان ویژگی هر سند در نظر بگیرید. درصد دقت (Accuracy) را روی مجموعه آزمون و مجموعه آموزش گزارش کنید. مقادیر اولیه را به صورت تصادفی بین ۰.۵+ و ۰.۵- انتخاب کرده و از نرخ یادگیری ۰.۰۱ استفاده کنید.

الف) در این شبکه از ویژگی‌های TF (نرمال شده) و TF-IDF استفاده کنید. شبکه را برای هر کدام از ویژگی‌ها، برای حداقل دو تعداد مختلف از نرون‌های لایه مخفی آموزش داده و نتیجه را در هر حالت ارائه دهید. نمودار خطای شبکه در حین آموزش را برای هر حالت رسم کنید.

ب) بر روی داده تست، ماتریس درهم‌ریختگی (Confusion Matrix) را محاسبه کنید

[نمره اضافی ۵۰٪ نمره این سوال] در پیاده‌سازی این الگوریتم می‌توانید از کدهای آماده استفاده کنید. در صورت پیاده‌سازی الگوریتم MLP توسط خود شما، نمره اضافی به این منظور در نظر گرفته می‌شود.

---

۴. (۳۰٪) **[پیاده‌سازی- دسته‌بندی متون فارسی با شبکه عصبی MLP و Word2Vec]** در این تمرین می‌خواهیم دسته‌بندی متون دادگان زبرا را با ویژگی‌های استخراج شده از کلمات با روش Word2Vec و دسته‌بند MLP انجام دهیم. برای این کار بردارهای از قبل آموزش داده شده را از سایت زیر دریافت کنید (با روش skip gram) و به ازای هر سند میانگین بردارهای کلمات را به عنوان نمایش بردار آن سند محاسبه کنید.

https://nlpdataset.ir/farsi/pre-trained_embeddings.html

یک شبکه MLP با یک لایه مخفی (با تعداد نرون‌های مخفی برابر با میانگین تعداد نرون‌های ورودی و خروجی) ایجاد کرده و با داده مذکور آموزش دهید و بر روی داده تست سوال قبل ارزیابی کنید و نتایج آن را با سوال قبل مقایسه کنید.

h.veisi@ut.ac.ir