به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

# درس پردازش زبان‌های طبیعی

تمرین ۲

اسفند ماه ۱۴۰۱

***فهرست**

## سوال ۱

در این سوال هدف حل مسئله تشخیص احساسات (Sentiment Analysis) است. برای این منظور ابتدا دادگان **Snapfood** را دریافت کنید. این دیتاست دارای دو کلاس احساسات Positive, Negative است. مدلی که برای تشخیص احساسات در این تمرین آموزش خواهید داد، یک مدل Naive Bayes است. برای این سوال تنها از ۲۰٪ از دادگان ذکر شده را استفاده کنید (دقت کنید که نسبت تعداد نمونه های دو کلاس احساسات را حفظ کنید). همچنین از دادگان نمونه گرفته شده ۱۰٪ را برای ارزیابی و باقی را برای آموزش مدل در نظر بگیرید.

الف) ابتدا دادگان را با استفاده از پیش پردازش های مورد نیاز (Normalization،Tokenization و موارد دیگر) مهیا کنید. می‌توانید از کتابخانه‌های آماده زبان فارسی هم استفاده کنید.

برای آموزش Naive Bayes Classifier نیاز به استخراج ویژگی از متن دارید. برای اینکار یک بار از روش Tf-Idf و یک بار از روش PPMI استفاده کنید. برای پیاده سازی این دو روش مجاز به استفاده از کتابخانه‌های آماده نیستید.

در مرحله نهایی با استفاده از کتابخانه‌ی Scikit-Learn یک طبقه بند Naive Bayes برای تشخیص احساسات پیاده سازی کنید. پس از آموزش طبقه بند، مدل خود را بر اساس معیار های Precision, Recall و F1-score برای هر دو حالت استفاده از Tf-Idf و PPMI ارزیابی کنید سپس نتایج را بررسی کنید.

## سوال ۲

در این سوال هدف تولید بردار های معنا (Vector Semantic) برای جانمایی کلمات یک دادگان با روش مشابه word2vec است. مدلی که از آن برای تولید این بردار های معنا استفاده می‌کنید Skipgram است. برای آشنایی با این مدل میتوانید این ویدیو را ببینید یا مقاله اصلی و این لینک را مطالعه کنید.

الف) دادگان T8.Shakespeare را دریافت کنید. ابتدا پردازش های لازم را روی این دادگان انجام دهید. دو ماتریس جانمایی و زمینه (Context) را در نظر بگیرید که به تعداد کلمات یا اندازه‌ی دیکشنری سطر بردار ویژگی داشته باشند. طول بردار ها برای هر دو ماتریس را برابر با ۱۰۰ در نظر بگیرید. با استفاده از تکنیک Negative Sampling به ازای هر نمونه مثبت ۴ نمونه منفی تولید کنید. مدل Skipgram را پیاده سازی و آموزش دهید. پس از آموزش بردار ویژگی کلمات را از جمع ماتریس های جانمایی و زمینه بسازید.

ب) مطابق آنچه در مقاله توضیح داده شده است، با استفاده از تبدیل PCA بردار ویژگی کلمات را در دو بعد تصویر کنید. سپس چهار بردار تفاضل زیر را رسم کنید:

۱ -king - man (the difference between the representation vectors of the words king and man)

۲ -queen - woman

۳ -brother - sister

٤ -uncle - aunt

تحلیل خود را از نتایج به دست آمده بیان کنید.

## سوال ۳

در این سوال هدف حل مسئله تشخیص احساسات برای دادگان FinancialPhraseBank-v1.0 با روش logistic Regression است. ابتدا این دادگان را دریافت کنید. این دادگان دارای سه کلاس احساسات Positive, Negative و Neutral است. پیش پردازش های مورد نیاز را روی دادگان انجام دهید. همچنین ۱۰٪ از داده ها را برای test مدل خود و باقی را برای آموزش استفاده کنید.

الف) یک مدل ساده Logistic Regression شامل یک ماتریس وزن W و بردار بایاس b را برای این مسئله طراحی کنید. برای بازنمایی کلمات خود از بردار های معنای GloVe نسخه ۶b بهره بگیرید (اگر در محیط Google Colab کد می‌نویسید می‌توانید با دستور wget از همین لینک GloVe را دانلود کنید) و وزن های آن را آموزش ندهید. با استفاده از تابع هزینه Cross Entropy مدل خود را برای تشخیص احساس نظرات بیان شده این دادگان آموزش دهید. می‌توانید برای پیاده سازی از کتابخانه‌های آماده هم استفاده کنید. عملکرد مدل خود را بر اساس معیارهای Precision, Recall و F1-score ارزیابی کنید. نتایج را بررسی کنید.

ب) توضیح دهید که توزیع نامتوازن (Imbalanced) کلاس های این دادگان چه تاثیری بر عملکرد مدل می‌گذارد.

ج) به نظر شما با توجه به توزیع نامتوارن دادگان، اگر به جای طبقه بند Logistic Regression برای این سوال از روش Naive Bayes استفاده کنیم، عملکرد بهتر یا بدتری خواهیم داشت؟ علت را توضیح دهید.

## ملاحظات (حتما مطالعه شود)

**تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA2_StudentID تحویل داده شود.**

- **خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.**
- **کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.**
- **توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.**
- **در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:**

p.baghershahi@ut.ac.ir

**مهلت تحویل: ۱۴۰۲/۲/۴**