# به نام خدا


تمرین سوم درس پردازش زبان طبیعی آماری

**«ابهام‌زدایی معنایی کلمات»**

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

خرداد ۱۴۰۰

**برای ارسال تمرین به نکات زیر توجه کنید.**

۱- در جدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین‌ها ذکر شده‌است.

| میزان تاخیر (روز) | میزان جریمه |
|---|---|
| ۱ الی ۲ روز | هر روز ۵٪ |
| ۲ الی ۶ روز | هر روز ۱۰٪ |

در صورتی که برای ارسال تمرین‌ها بین ۷ تا ۱۴ روز تاخیر داشته باشید، نمره شما از ۵۰٪ محاسبه می‌شود و پس از این بازه به تمرین ارسالی نمره‌ای تعلق نمی‌گیرد.

۲- هرگونه کپی‌برداری در انجام تمرین‌ها موجب کسر نمره خواهد شد.

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز دوشنبه ۳۱ خرداد** می‌باشد.

۴- فایل های ارسالی خود شامل فایل‌های پیاده‌سازی و گزارش را فشرده کنید و با عنوان «**شماره دانشجویی_HW3**» مانند `HW3_97131022` ارسال کنید.

۵- زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون یا جاوا در نظر گرفته شده‌است.

۶- کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.

۷- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یار درس از طریق ایمیل زیر در ارتباط باشید.

مریم موسویان mousavian12@gmail.com

## بخش اول: تعریف مسئله و معرفی دادگان

در این تمرین قصد داریم وظیفه ابهام‌زدایی معنایی کلمات[1] را به عنوان یک مولفه اساسی و مهم در کاربردهای مختلف پردازش زبان طبیعی با استفاده از الگوریتم‌های یادگیری ماشین انجام دهیم. مجموعه دادگان و کلیه فایل‌های مورد نیاز برای این تمرین از لینک[2] موجود در پاورقی قابل دانلود است. در فایل `train.csv` ۶۰۹۸ داده آموزش و در فایل `test.csv` ۱۱۰۶ داده آزمون قرار دارد. مجموعه کلمات و تعداد نمونه‌های آن‌ها در هر فایل در جدول زیر مشاهده می‌شوند. ویژگی‌های موجود در مجموعه داده به ترتیب کلمه (`word`)، اجزای سخن(`pos`[3])، شناسه نمونه (`instance_id`)، منبع سند (`doc_src`)، شناسه مفهوم (`sense_id`) و بافت کلمه (`context`) است. (کلمه مبهم با تگ `<head>` درون `context` مشخص شده‌است.) معنی هر مفهوم را با توجه به شناسه آن مفهوم و کلمه مورد نظر در فایل `dictionary.xml` می‌توانید بیابید.

| کلمه | داده آموزش | داده آزمون | کلمه | داده آموزش | داده آزمون | کلمه | داده آموزش | داده آزمون |
|---|---|---|---|---|---|---|---|---|
| activate | 156 | 28 | disc | 147 | 27 | organization | 92 | 17 |
| add | 192 | 35 | eat | 146 | 26 | paper | 152 | 28 |
| appear | 214 | 38 | encounter | 110 | 20 | party | 181 | 32 |
| argument | 146 | 26 | expect | 114 | 21 | performance | 132 | 24 |
| arm | 223 | 40 | express | 92 | 17 | plan | 124 | 23 |
| ask | 175 | 31 | hear | 53 | 10 | play | 88 | 16 |
| atmosphere | 104 | 19 | hot | 71 | 13 | produce | 156 | 28 |
| audience | 143 | 26 | image | 122 | 22 | provide | 115 | 21 |
| bank | 213 | 38 | important | 22 | 4 | receive | 44 | 8 |
| begin | 133 | 24 | interest | 154 | 28 | remain | 118 | 21 |
| climb | 110 | 20 | judgment | 52 | 10 | rule | 50 | 9 |
| decide | 103 | 19 | lose | 60 | 11 | shelter | 141 | 25 |
| degree | 191 | 34 | mean | 67 | 13 | simple | 27 | 5 |
| difference | 169 | 30 | miss | 49 | 9 | smell | 88 | 16 |
| different | 78 | 14 | note | 110 | 20 | solid | 45 | 8 |
| difficulty | 35 | 7 | operate | 29 | 6 | sort | 127 | 23 |
| source | 46 | 9 | treat | 94 | 17 | watch | 84 | 16 |
| suspend | 107 | 19 | use | 22 | 4 | win | 65 | 12 |
| talk | 124 | 22 | wash | 56 | 10 | write | 37 | 10 |

[1] Word Sense Disambiguation (WSD)

[2] https://drive.google.com/drive/folders/1FidQTKvT4YAyKD4ncqIizjVE4POs7VAB?usp=sharing

[3] Part Of Speech

**توجه:**

- نحوه انجام پیش‌پردازش بر روی داده‌ها شامل کتابخانه مورد استفاده و مراحل انجام‌شده را در گزارش خود مکتوب کنید.
- برای انجام این تمرین می‌توانید از کتابخانه‌های آماده استفاده کنید.

## بخش دوم: انتخاب ویژگی

در این قسمت  با استفاده از مجموعه دادگان ذکرشده در بخش قبل، بازنمایی برای کلمات دارای ابهام معنایی در مجموعه داده به دست آورید. در این تمرین می‌خواهیم از مدل بازنمایی مبتنی بر بافت BERT[4] و مدل Word2Vec برای استخراج ویژگی کمک بگیریم. معمولا برای استخراج بازنمایی کلمات از مدل BERT، وزن‌ها و پارامترهای مدل را روی پیکره مورد نظر به صورت دقیق تنظیم می‌کنند. اما در این درس به دلیل عدم آشنایی برخی دانشجویان با درس شبکه‌های عصبی، از تنظیم وزن‌ها و پارامترهای مدل صرف نظر می‌شود. بنابراین برای استخراج بازنمایی‌های مبتنی بر بافت BERT می‌توانید از کتابخانه bert-embedding[5] استفاده کنید (استفاده از سایر کتابخانه‌ها نیز مانعی ندارد.). خروجی تمام بردارها ۳۰۰ بعد در نظر گرفته شود. (با توجه به اینکه بردار خروجی مدل BERT برای هر کلمه ۷۶۸ بعد است و برای این که در هنگام آموزش دسته‌بندها  در قسمت بعد با کمبود منابع رو به رو نشوید، بردارهای خروجی برای کلمات مبهم را با استفاده از الگوریتم PCA به ۳۰۰ بعد کاهش دهید.) برای استخراج بازنمایی کلمات از مدل Word2Vec از مدل از پیش آموزش دیده گوگل بر روی مجموعه داده Google News استفاده کنید. ( این مدل در لینک حاوی داده‌ها و فایل‌های تمرین با عنوان GoogleNews-vectors-negative300.bin.gz آپلود شده‌است. برای دانلود و لود کردن این مدل نیز می‌توانید از کتابخانه gensim استفاده کنید.)

با استفاده از روش‌های زیر بازنمایی هریک از ورودی‌های مبهم را به‌دست آورید و سپس با استفاده از بازنمایی و توضیحات بخش بعد مفهوم مرتبط با آن ورودی را به‌دست آورید.

الف) استفاده از مدل از پیش آموزش داده شده BERT و استخراج بازنمایی تنها برای کلمه مبهم موجود در متن از مدل BERT بدون استخراج بازنمایی کلمات دیگر متن. (با توجه به اینکه مدل BERT مبتنی بر بافت است کلمات اطراف کلمه هدف تاثیر خود را در بازنمایی کلمه هدف به جا می‌گذارند اما نیاز به بازنمایی آن‌ها به‌صورت مجزا نیست.)

---

[4] Bidirectional Encoder Representations from Transformers (BERT)

[5] https://pypi.org/project/bert-embedding

ب) استفاده از مدل از پیش آموزش داده شده BERT و استخراج بازنمایی کلمات از مدل BERT برای کلمه هدف و سایر کلمات موجود در بافت کلمه هدف به شعاع ۳ $\pm$ کلمه و سپس استفاده از میانگین وزن‌دار بازنمایی کلمات موجود در بافت کلمه مبهم با استفاده از TF-IDF هر یک از کلمات. (منظور از میانگین وزن‌دار این است که TF-IDF هر کلمه به عنوان وزن بردار بازنمایی آن کلمه در نظر گرفته شود.).

ج) تکرار بند (ب) ولی با در نظر گرفتن تمام کلمات موجود در بافت کلمه مبهم.

د) استفاده از بازنمایی از پیش آموزش داده شده Word2Vec برای کلمه هدف و سایر کلمات موجود در بافت کلمه هدف به شعاع ۳ $\pm$ کلمه و سپس استفاده از میانگین وزن‌دار بازنمایی کلمات موجود در بافت کلمه مبهم با استفاده از TF-IDF هر یک از کلمات.

ه) تکرار بند (د) ولی با در نظر گرفتن تمام کلمات موجود در بافت کلمه مبهم.

## بخش سوم: دسته‌بندی

در این قسمت می‌خواهیم بهترین مفهوم از مجموعه مفاهیم داده شده برای کلمه مبهم را با توجه به بافت به کار رفته برای آن کلمه بیابیم. برای این کار از الگوریتم‌های یادگیری ماشین[۶] و روش‌های یادگیری گروهی[۷] استفاده می‌کنیم. (انتخاب بهترین مقدار برای پارامترهای دسته‌بندهای استفاده شده به عهده شما می‌باشد.)

الف) برای دسته‌بندی از روش‌های بازنمایی مطرح‌شده در بخش قبل و دسته‌بندهای Logistic Regression و Random Forest استفاده کنید. نتایج آزمایش‌های ذکرشده را با معیارهای Accuracy و F-Measure ارزیابی کنید. معیارهای ارزیابی را برای اسم‌ها، فعل‌ها و صفت‌ها به صورت مجزا، میانگین نتایج را گزارش کنید.

ب) با توجه به ۵ بازنمایی مختلف و دو دسته‌بند مختلف، ۱۰ حالت را در این تمرین آزمایش نموده‌اید. از میان این ۱۰ حالت ۳ نتیجه برتر را انتخاب کنید و با کمک این سه مدل یک مدل گروهی[۸] بسازید. برای پیدا کردن مفهوم مناسب برای هر کلمه مبهم بین نتایج ۳ دسته‌بند رای اکثریت بگیرید. مانند قسمت قبل معیارهای ارزیابی Accuracy و F-Measure را برای هر POS به صورت جداگانه گزارش کنید.

ج) با استفاده از ۳ تا بهترین دسته‌بند انتخاب شده در بند (ب) یک مدل گروهی دیگر بسازید. برای این کار یک دسته‌بند دلخواه را با استفاده از خروجی ۳ دسته‌بند (مفاهیم پیش‌بینی شده برای کلمات مبهم) آموزش دهید. در واقع ویژگی‌های استفاده شده

---

[6] Machine Learning
[7] Ensemble Learning
[8] Ensemble Model

برای آموزش دسته‌بند دلخواه یک بردار با ابعاد ۳ شامل برچسب‌های (مفاهیم) پیش‌بینی شده توسط ۳ دسته‌بند خواهد بود. مانند قسمت‌های قبل معیارهای ارزیابی Accuracy و F-Measure را برای هر POS به صورت جداگانه گزارش کنید.

د) در این قسمت می‌خواهیم برای ابهام‌زدایی معنایی کلمات از شبکه‌های عصبی استفاده کنیم. ابتدا بازنمایی توکن [cls] را از مدل BERT استخراج کنید و به یک شبکه عصبی پیش‌خور[9] دهید (برای استفاده از مدل BERT به ابتدای هر متن ورودی، توکن [cls] اضافه می‌شود. بازنمایی که از مدل BERT برای این توکن در هر متن استخراج می‌شود، به عنوان بازنمایی برای کل متن در نظر و از آن در وظیفه دسته‌بندی متون کمک گرفته می‌شود.). مانند قسمت‌های قبل معیارهای ارزیابی Accuracy و F-Measure را برای هر POS به صورت جداگانه گزارش کنید. **(این قسمت اختیاری و دارای نمره اضافه خواهد بود.)**

## بخش چهارم: تحلیل نتایج

در این قسمت نتایج به دست آمده از دسته‌بندهای مختلف بخش سوم را مقایسه و به صورت کامل تحلیل کنید. در صورتی که با استفاده از ایده‌ای خلاقانه بتوانید نتایج را بهبود دهید نمره‌ای اضافه دریافت خواهید کرد.

**موفق باشید**

[9] Feed Forward Neural Network