به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان طبیعی

تمرین ۲

اسفند ماه ۱۴۰۲

فهرست

سوال اول

در این سوال هدف حل مسئله تشخیص احساسات (Sentiment Analysis) است. تجزیه و تحلیل احساسات شاخه‌ای از پردازش زبان طبیعی (NLP) است که شامل استفاده از ابزارهای پردازش زبان، جهت شناسایی و استخراج خودکار اطلاعات موجود در متن است. هدف از تجزیه و تحلیل احساسات، تعیین احساسات یا عواطف پشت یک متن است که در حالت پایه آن می‌تواند به شکل‌های مثبت، منفی و یا خنثی باشد. در حالت‌های دیگر این نوع از تحلیل، می‌توان کلاس‌هایی با برچسب‌های «بسیار مثبت»، «بسیار منفی» و غیره را نیز تعبیه کرد.

مدلی که برای تشخیص احساسات در این تمرین آموزش خواهید داد، یک مدل Naïve Bayes است. بنابراین، جهت آموزش این مدل نیاز به استخراج ویژگی از متن داشته و در این سوال می‌بایست ۳ نوع بردار جانمایی[1] مختلف را برای کلمات تولید کنید.

مجموعه داده

مجموعه داده‌ای که برای این سوال انتخاب شده است، مجموعه داده‌ی **Sentiment140** است. این مجموعه داده شامل ۱.۶ میلیون توییت[2] است که در سه کلاس منفی، خنثی و مثبت قرار دارند.

بخش اول – پیش پردازش مجموعه داده

در این بخش، ابتدا به صورت تصادفی، از هر کلاس 5000 نمونه را انتخاب نموده و سپس پیش‌پردازش‌های موردنیاز (همچون Normalization، Tokenization و موارد دیگر) را انجام دهید. 20 درصد داده‌ها را به عنوان مجموعه داده ارزیابی جدا کنید. جهت انجام این بخش می‌توانید از کتابخانه‌های آماده نیز استفاده کنید. لازم به ذکر است که در گزارش مربوط به این قسمت، می‌بایست توضیح مختصری درباره‌ی هر کدام از روش‌های پیش‌پردازش اعمال‌شده داده و علت استفاده از آن را ذکر کنید.

بخش دوم – ساخت بردار جانمایی اول TERM FREQUENCY

در این بخش لازم است تا برای هر نمونه بردار تعداد کلمات متناظر آن را بسازید و سپس در یک ماتریس ذخیره کنید.

[1] Word Embedding
[2] Tweet

مثال:

"the quick brown fox"

"jumped over the lazy dog"

"the quick dog"

Document	the	quick	brown	fox	jumped	over	lazy	dog
1	1	1	1	1	0	0	0	0
2	1	0	0	0	1	1	1	1
3	1	1	0	0	0	0	0	1

بخش سوم – ساخت بردار جانمایی دوم – TF-IDF[3]

در این بخش برای هر نمونه با روش Tf-Idf بردار جانمایی را بسازید و سپس در یک ماتریس ذخیره نمایید. دقت کنید که لازم است توابع Tf-Idf را خودتان پیاده‌سازی کنید.

بخش چهارم – ساخت بردار جانمایی سوم – PPMI[4]

در این بخش برای هر نمونه با روش PPMI بردار جانمایی را ساخته و سپس در یک ماتریس ذخیره نمایید. همچون بخش قبل، دقت کنید که برای پیاده‌سازی این قسمت نیز مجاز به استفاده از کتابخانه‌های آماده نیستید.

بخش پنجم – آموزش مدل

یک طبقه‌بند Naïve Bayes را با استفاده از هر یک از بردارهای جانمایی به‌دست‌آمده، جهت تحلیل احساسات آموزش دهید و پس از آن، مدل خود را بر اساس معیارهای F1-score، Precision و Recall ارزیابی نمایید. درنهایت نتایج را بررسی کرده و تحلیل خود را بنویسید.

[3] Term frequency–inverse document frequency
[4] Positive point-wise mutual information

سوال دوم

در این سوال هدف حل مسئله تشخیص کنایه[5] در متن است. امروزه کنایه به یک ویژگی فراگیر در بسیاری از متون تبدیل شده است و افراد برای بیان منظور خود به شکلی متفاوت‌تر از کنایه استفاده می‌کنند. تشخیص کنایه نیز، همچون تجزیه و تحلیل احساسات، یکی از شاخه‌های پردازش زبان طبیعی است که در آن مدل آموزش می‌بیند عباراتی را که در آن کنایه وجود دارند شناسایی کند.

مجموعه داده

برای این سوال، لازم است تا از مجموعه داده ضمیمه شده با عنوان sarcasm.json استفاده کنید. در این مجموعه‌داده، عناوین خبری برای تشخیص کنایه از دو وب‌سایت خبری جمع‌آوری شده است. این مجموعه داده شامل 28619 سطر و سه ستون headline ، article_link و is_sarcastic است.

بخش اول – پیش پردازش مجموعه داده

بر روی داده‌ها، تمامی پیش‌پردازش‌های موردنیاز را انجام داده و آن‌ها را جهت استفاده در بخش‌های بعدی آماده نمایید. 20 درصد داده‌ها را به عنوان مجموعه داده ارزیابی جدا کنید.

بخش دوم – بارگذاری GLOVE

همان‌طور که می‌دانید، GloVe (Global Vectors) پروژه ایست که توسط دانشگاه Stanford برای مدل زبانی ارائه شده است. از فواید اصلی word2vec توانایی آن در رمزگذاری[6] معنای کلمات و به طور دقیق‌تر، درنظرگرفتن روابط بین کلمات مانند queen – woman = king – man است. هدف از ساختن GloVe، رمزگذاری اطلاعات معنایی (Encoding Semantic Information) در بردارها است و همچنین میزان ارتباط هر دوکلمه را با سایر کلمات زمینه‌ای (Contextual Word) متن اندازه‌گیری می‌کند. در مدل زبانی، به این کلمات زمینه‌ای، Probe words نیز گفته می‌شود.

حال در این بخش لازم است برای بازنمایی کلمات، از بردارهای معنای GloVe نسخه 6b که در این آدرس موجود است، استفاده نمایید و با تشکیل دیکشنری کلمات، ماتریس جانمایی را بسازید.

[5] Sarcasm Detection
[6] Encoding

بخش سوم – آموزش مدل

مدل Logistic Regression را از sklearn دریافت کرده و با استفاده از بردارهای جانمایی‌های Glove آن را آموزش دهید. برای آشنایی با مدل logistic regression میتواند این لینک را مطالعه کنید. پس از آن، مدل خود را بر اساس معیارهای F1-score ، Precision و Recall ارزیابی نمایید. درنهایت نتایج را بررسی کرده و تحلیل خود را بنویسید.

سوال سوم

در این سوال هدف تولید بردار های معنا (Vector Semantic) برای جانمایی کلمات یک مجموعه داده با روش مشابه word2vec است. مدلی که از آن برای تولید این بردارهای معنا استفاده خواهید کرد Skipgram است. برای آشنایی با این مدل می‌توانید این مقاله را مطالعه فرمایید.

مجموعه داده

در این سوال از مجموعه داده متنی موجود در این لینک استفاده کنید. این مجموعه داده شامل متن داستان های شرلوک هلمز است. پس از دریافت آن، پیش پردازش های لازم را انجام دهید و سپس مراحل بعدی را طی کنید.

بخش اول

بر روی این مجموعه داده، ابتدا پردازش‌های لازم را 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 بردارها برای هر دو ماتریس را برابر با 100 در نظر بگیرید. با استفاده از تکنیک Negative Sampling به ازای هر نمونه مثبت 4 نمونه منفی تولید کنید. مدل Skipgram را پیاده‌سازی کرده و آموزش دهید. پس از آموزش بردار ویژگی کلمات را از جمع ماتریس‌های جانمایی و زمینه بسازید.

بخش دوم

با استفاده از ضرب داخلی میزان شباهت بردار queen را با بردار king – man + woman به دست آوردید و نتیجه را تفسیر کنید.

بخش سوم

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

1) brother - sister
2) uncle - aunt

تحلیل خود را از نتایج به دست آمده بیان کنید.

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.
- تمرین تا یک هفته بعد از مهلت تعیین شده با تاخیر تحویل گرفته می‌شود. دقت کنید که شما جمعا برای تمام تکالیف، 14 روز زمان تحویل بدون جریمه دارید که تنها از 7 روز آن برای هر تمرین می‌توانید استفاده کنید، در صورتی که این 14 روز به اتمام رسیده باشد، به ازای هر روز تاخیر در ارسال تمرین، ده درصد جریمه می‌شوید.
- **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 تشابه 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:

nastaran.ashoori@ut.ac.ir
namazifard@ut.ac.ir

- این تمرین شامل دو بخش است. مهلت تحویل آنها به شرح زیر است.

مهلت تحویل بدون جریمه سوال اول: 26 اسفند 1402

مهلت تحویل با تأخیر، با جریمه 10 درصد سوال اول: 4 فروردین 1403

مهلت تحویل بدون جریمه سوال دوم و سوم: 14 فروردین 1403

مهلت تحویل با تأخیر، با جریمه 10 درصد سوال دوم و سوم: 21 فروردین 1403