به نام خدا

تمرین دوم درس پردازش زبان طبیعی آماری

«آشنایی با انواع بازنمایی کلمات و کاربرد آن‌ها»

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

فروردین 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 ۲۰ اردیبهشت** می‌باشد.

۴- فایل های ارسالی خود شامل فایل‌های پیاده‌سازی و گزارش را فشرده کنید و با عنوان **«شماره دانشجویی_HW2»** مانند `HW2_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 بازنمایی متن را به چندین روش مختلف محاسبه‌کرده و سپس با استفاده از بازنمایی‌های به‌دست‌آمده، متن‌های موجود را خوشه‌بندی کنیم. به این منظور بخشی از مجموعه دادگان همشهری در قالب فایل `Hamshahri.txt` با ۸۵۹۹ سند در اختیار شما قرار گرفته‌است. در این دادگان هر خط یک داده را تشکیل می‌دهد. هر خط شامل یک سند و برچسب آن سند می‌باشد که با کاراکتر @ از هم جدا شده‌اند. مجموعه دادگان و کلیه فایل‌های مورد نیاز برای این تمرین از لینک[1] موجود در پاورقی قابل دانلود است.

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: بازنمایی

در این قسمت می‌خواهیم با استفاده از مجموعه دادگان همشهری بازنمایی اسناد موجود را به دست آوریم. برای این کار ۵ روش زیر را لحاظ کنید. برای تمام روش‌های بازنمایی زیر خروجی هر بردار متن ۳۰۰ بعد در نظر گرفته شود.

الف) آموزش بردار کلمات روی مجموعه داده با استفاده از `Word2Vec` مدل `Skip-gram`[2] و سپس استفاده از میانگین بازنمایی کلمات سند به منظور محاسبه بازنمایی هر سند.

ب) آموزش بردار کلمات روی مجموعه داده با استفاده از `Word2Vec` مدل `Skip-gram` و سپس استفاده از میانگین وزن‌دار بازنمایی کلمات سند با استفاده از `TF-IDF` هر یک از کلمات به منظور محاسبه بازنمایی هر سند. (منظور از میانگین وزن‌دار این است که `TF-IDF` هر کلمه به عنوان وزن بردار بازنمایی آن کلمه در نظر گرفته شود.)

ج) استفاده از بازنمایی کلمات موجود در مجموعه بردارهای از پیش آموزش داده شده با استفاده از `Word2Vec` بر روی حجم زیادی از مجموعه داده همشهری و سپس استفاده از میانگین بازنمایی کلمات سند به منظور محاسبه بازنمایی هر سند. (مجموعه

[1] https://drive.google.com/drive/folders/1r1tY9qhW8UJEaOt8ijMPkG2hRm-z3VuH?usp=sharing

[2] به عنوان نمونه می‌توانید از کتابخانه Gensim استفاده کنید.

بردارهای از پیش آموزش داده شده ذکر شده در فایل فشرده‌شده hamshahri.fa.text.300.vec.zip در لینک فایل‌های مورد نیاز تمرین موجود است.)

د) استفاده از بردارهای ذکرشده در بند (ج) و سپس استفاده از میانگین وزن‌دار بازنمایی کلمات سند با استفاده از TF-IDF هر یک از کلمات به منظور محاسبه بازنمایی هر سند.

ه) ساخت ماتریس سند-کلمه با استفاده از TF کلمات و سپس کاهش بعد ماتریس ایجادشده به ۳۰۰ از طریق روش SVD.

توجه:

در فایل hamshahri.fa.text.300.vec.zip فرمت محتوای فایل موجود در آن به این صورت است که در هر خط کلمه مورد نظر و بازنمایی متناظر با آن با یک فاصله از هم جدا شده‌اند.

بخش سوم: خوشه‌بندی

در این قسمت می‌خواهیم اسناد موجود در دادگان همشهری را با استفاده از بازنمایی‌های به‌دست‌آمده در بخش دوم و الگوریتم Kmeans خوشه‌بندی کنیم. پارامتر تعداد خوشه را برابر با ۵ در نظر بگیرید. برای ارزیابی خوشه‌بندی، پرتکرارترین برچسب اسناد موجود در هر خوشه را به عنوان برچسب حدس زده شده برای آن خوشه در نظر بگیرید. در این مرحله نیز با داشتن برچسب‌های واقعی داده‌های موجود در مجموعه داده و برچسب‌های حدس زده شده، خوشه‌بندی انجام‌شده را با استفاده از معیارهای Accuracy، F-Measure و NMI ارزیابی کنید.

بخش چهارم: مدل‌سازی موضوع

در پردازش زبان طبیعی مدل‌سازی موضوع[3] به منظور شناسایی موضوع‌های انتزاعی و کشف ساختارهای معنایی پنهان موجود در اسناد مورد استفاده قرار می‌گیرد. خروجی روش‌های موجود برای مدل‌سازی موضوع، یک توزیع احتمالی از موضوع‌های موجود در هر سند و یک توزیع احتمالی از کلمات مربوط به هر موضوع خواهد بود. در این بخش می‌خواهیم با استفاده از روش LDA[4] که یکی از روش‌های شناخته‌شده در مدل‌سازی موضوع است، موضوع‌ها و ساختارهای معنایی موجود در مجموعه دادگان همشهری را استخراج کنیم.

[3] Topic Modeling
[4] Latent Dirichlet Allocation

الف) با استفاده از مجموعه داده، مدل LDA را آموزش دهید. تعداد موضوع‌ها را مانند تعداد خوشه‌ها در بخش قبل برابر با ۵ در نظر بگیرید.

ب) برای ارزیابی مدل‌سازی موضوع، توزیع احتمالاتی موضوع‌ها بر روی اسناد موجود در مجموعه داده را به دست آورید. موضوع هر سند، موضوعی است که بیشترین احتمال را در توزیع احتمالاتی آن سند دارد. برای مقایسه موضوع حدس زده شده با برچسب‌های واقعی، پرتکرارترین برچسب اسناد موجود در هر موضوع را به عنوان برچسب حدس زده شده آن موضوع در نظر بگیرید. در این مرحله نیز مانند بخش قبل با داشتن برچسب‌های واقعی داده‌های موجود در مجموعه داده و برچسب‌های حدس زده شده برای آن‌ها، مدل‌سازی موضوع انجام‌شده را با استفاده از معیارهای Accuracy، F-Measure و NMI ارزیابی کنید.

بخش پنجم: تحلیل نتایج

در این قسمت نتایج خوشه‌بندی‌های مختلف انجام‌شده در بخش سوم و مدل‌سازی موضوع در بخش قبل را مقایسه و به صورت کامل تحلیل کنید.

موفق باشید