تمرین دوم – درس پردازش زبان طبیعی آماری **دکتر ممتازی** **بهار ۹۸ – دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر**	

1. درجدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین ها آورده شده است:

تاخیر (روز)	میزان نمره منفی
از ۱ الی ۲	هر روز ۵٪
از ۳ الی ۶	هر روز ۱۰٪

توجه داشته باشید در صورت تاخیر بین ۷ تا ۱۴ روز، نمره تمرین از ۵۰٪ محاسبه شده و پس از آن نمره ای تعلق نمی گیرد.

2. هدف از انجام تمرین ها یادگیری عمیق تر مطالب درسی است. در نتیجه هرگونه کپی برداری موجب کسر نمره خواهد شد.

3. تا **ساعت ۲۳:۵۵ روز دوشنبه ۲۳ اردیبهشت** فرصت دارید تمرین را در مودل بارگذاری کنید. تمام فایل های پیاده سازی را به همراه فایل pdf مربوط به گزارش تمرین را در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود (برای مثال HW2_97131024) قرار دهید.

4. زبان برنامه نویسی برای انجام تمرین: پایتون، جاوا و یا متلب در نظر گرفته شده است.

5. برنامه های نوشته شده خوانا باشد و کامنت گذاری مناسب باشد (طوری که روند کار کاملا مشخص باشد).

6. در صورت وجود هر گونه سوال می توانید از طریق ایمیل با تدریس یاران درس در ارتباط باشید:
javadforough@gmail.com و amin.ghsm@aut.ac.ir

بازنمایی و خوشه بندی متن

در این تمرین می خواهیم بازنمایی متن را به چندین روش محاسبه نموده و سپس روی این بازنمایی ها با استفاده از الگوریتم kmeans به خوشه بندی متن بپردازیم. برای این منظور همانند تمرین اول بخشی از مجموعه دادگان همشهری در نظر گرفته شده است.
در این راستا دو بخش را بایستی انجام دهید.

بخش اول: بازنمایی

در این قسمت با استفاده از پیکره‌ی ارائه شده، بازنمایی متن‌ها را بدست می‌آوریم. برای این کار ۴ روش زیر را لحاظ کنید.

- آموزش بردار کلمات روی پیکره ارائه شده (با استفاده از Word2Vec مدل Skip-gram) و سپس استفاده از میانگین بازنمایی کلمات متن به منظور محاسبه بازنمایی متن
- آموزش بردار کلمات روی پیکره ارائه شده (با استفاده از Word2Vec مدل Skip-gram) و سپس استفاده از میانگین وزن دار بازنمایی کلمات متن(با به کار بستن tf-idf هر یک از کلمات) به منظور محاسبه بازنمایی متن
- آموزش بردار متون با استفاده از مدل doc2vec روی پیکره ی ارائه شده
- ساخت ماتریس doc-word با استفاده از اطلاعات tf کلمات و سپس کاهش بعد ماتریس از طریق روش SVD

بخش دوم: خوشه‌بندی

در این بخش با استفاده از بازنمایی های بدست آمده در بخش قبل الگوریتم خوشه بندی Kmeans را اجرا کرده و پارامتر تعداد خوشه را برابر ۶ در نظر بگیرید. بدیهی است که هنگام خوشه بندی هیچ گونه استفاده از از برچسب اسناد صورت نخواهد گرفت.
حال برای ارزیابی خوشه های ایجاد شده دو سناریوی ارزیابی زیر را اجرا نمایید. یکی ارزیابی بر روی مجموعه داده آموزش و دیگری ارزیابی بر روی داده آزمون که توضیح آن به شرح زیر است:

❖ ارزیابی آموزش

بر چسب هر خوشه را برابر پرتکرارترین برچسب آن خوشه در نظر بگیرید. در نتیجه تمامی نمونه ها در یک خوشه، پرتکرارترین برچسب را می گیرند که به عنوان برچسب حدس زده شده توسط مدل در نظر گرفته می‌شود. همچنین برچسب‌های واقعی نیز همان برچسب‌های مشخص شده با @ در مجموعه آموزشی می‌باشد. با داشتن برچسب گلد و برچسب حدس زده شده کیفیت خوشه‌بندی را با استفاده از ۴ معیار NMI، Accuracy، F-Measure و V-Measure بدست آورید.

❖ ارزیابی آزمون

هر یک از نمونه های آزمون را با مراکز خوشه مقایسه کرده و هر کدام که نزدیک تر بود برچسب آن خوشه (که در ارزیابی آموزش توضیح داده شد) به عنوان برچسب حدس زده شده توسط مدل خوشه‌بندی در نظر گرفته می شود. مجددا کیفیت خوشه‌بندی را با استفاده از ۴ معیار NMI، Accuracy، F-Measure و V-Measureبدست آورده و نتایج را بررسی نمایید.

توجه:

- برای انجام تمرین می توانید از کتابخانه های آماده نظیر Gensim، NLTK، Scikit و ... استفاده نمایید.
- برای تمام روش‌های بازنمایی خروجی بردار هر متن ۳۰۰ بعد باشد
- لازم به ذکر است که لینک پیکره همشهری (شامل داده آموزش و داده آزمون) در زیر آورده شده است:

https://www.dropbox.com/s/zkmozli0mmentpy/HAM-Train-Test.zip?dl=0