	تمرین سوم درس علوم داده در مهندسی نرم‌افزار	
دانشکده مهندسی کامپیوتر	تاریخ: ۱۴۰۰/۰۹/۲۹	موعد تحویل: ۱۴۰۰/۱۰/۱۷
	مدرس: دکتر عباس حیدرنوری	دستیاران آموزشی: مهسا مسعود (mah.masoud76@gmail.com) حامد طاهرخانی (th.hamed75@gmail.com)

نکات مهم در مورد تحویل تکلیف

- **پاسخ‌های خود را در یک فایل فشرده با فرمت HW3_fullname_studentNumber.zip در صفحه CW درس ارسال کنید.**
- **هر قسمتی از کد که خطا داشته باشد، بررسی نخواهد شد و نمره‌ای هم برای آن در نظر گرفته نمی‌شود.**
- **تمامی توابعی که نیاز دارید از آن‌ها استفاده کنید یا به شما تدریس شده‌اند و یا با یک جستجوی ساده آن‌ها را پیدا خواهید کرد.**

****در این تمرین استفاده از jupyter notebook و یا ابزارهای دیگر به منظور ارائه پاسخ در قالب ipynb الزامی می‌باشد.**

سوال اول:

دیتاست news شامل مجموعه‌ای از اخبار می‌باشد. که برخی از این اخبار Fake(label=1) و برخی real(label=0) می‌باشد. ستون text را به عنوان متن و ستون label را به عنوان target در نظر بگیرید.

A. ابتدا قرار است بر روی ستون text عملیات PreProcess را به صورت زیر انجام دهیم:
   a. تمامی کاراکترها lowercase شوند.
   b. تمامی ارقام حذف شوند.
   c. کلمات به طول ۱ حذف شوند.
   d. Punctuation ها حذف شوند.
   e. Stop word ها حذف شوند.
   f. عملیات stemming, lemmatization انجام شود.

B. با استفاده از CountVectorizer ماتریسی از داده‌های text که عملیات PreProcess بر روی آن‌ها انجام شده است بسازید.

C. با استفاده از عملیات train_test_split که عملیات shuffle را نیز انجام می‌دهد ۲۰ درصد از داده‌ها را به عنوان داده تست جدا کنید.

D. با استفاده از الگوریتم MultinominalNB مدل‌هایی بر روی داده train آموزش داده و سپس بر روی داده‌های تست عمل prediction انجام دهید.

E. Classification report و confusion matrix را برای نتایج هر دو مدل چاپ کنید.

سوال دوم:

A. با دستور read_json آدرس news group را لود کنید. این دیتاست حاوی ۱۱ هزار پست در موضوعات مختلف است. با head قسمتی از دیتای لود شده را نشان دهید.

B. قسمت محتوی(content) را انتخاب کنید و آن را به لیست تبدیل کنید.(با target و target_names کاری نداریم) سعی کنید که آن را را تمیز کنید. این کار میتواند شامل حذف کردن ایمیل ها و کاراکتر های خاص مثل نقل قول تک و دوگانه، خط جدید و.. باشد. استفاده از کتابخانه re توصیه میشود.

C. با استفاده از simple_preprocess پیش پردازش اولیه روی متن انجام دهید و پست ها را tokenize کنید. این تابع token ها را به حروف کوچک نیز تبدیل می کند. از پارامتر deacc=True استفاده کنید تا علائم نگارشی را حذف کند. خروجی این مرحله لیستی از پست هایی است که هر یک شامل لیستی از token های آن پست است. راه های دیگر tokenize کردن متن در پایتون را در این لینک میتوانید بیابید.

D. در مدل N-gram به دنبال پیدا کردن احتمال وقوع N کلمه پشت سرهم هستیم. در حال حاضر واژه ها به صورت unigram جدا شده اند در حالی که بسیاری از واژه ها در متن به طور مکرر به دنبال هم ظاهر میشوند. می خواهیم در این دیتاست با استفاده از خروجی قسمت قبل یک مدل bigram بسازیم. در مدل bigram سعی میکنیم که علاوه بر واژه های تکی، جفت کلماتی که در متن به فراوانی به دنبال هم تکرار شده اند را نیز شناسایی کنیم.
دو کلاس FrozenPhrases و Phrases را با دستور زیر import کنید.

```
From gensim.models.phrases import Phrases, FrozenPhrases
```

مدل bigram را با استفاده از کلاس Phrases آموزش دهید. مدل ایجاد شده را به FrozenPhrases پاس دهید و مدلی جدید با این کلاس بسازید. به این ترتیب مدل بدست آمده freeze می شود و میتوانیم با سرعت بیشتری از مدل استفاده کنیم.

E. میخواهیم عملیات زیر را بر روی توکن های بدست آمده از مرحله قبل اعمال کنیم
   a) واژه هایی را که در لیست stop_words هستند حذف کنید.
   b) توکن های unigram را با استفاده از مدل بدست آمده در قسمت D به توکن های bigram تبدیل کنید.

F. با استفاده gensim corpora یک دیکشنری از دیتای مرحله قبل بسازید. سپس با استفاده از doc2bow و دیکشنری ایجاد شده، شکل bag of words را برای دیتای مرحله قبل بسازید.

G. با استفاده از gensim_ldamodel مدل lda را با ۲۰ تاپیک بسازید. با استفاده از print_topics موضوعات بدست آمده و واژه های مهم در هر موضوع را نمایش دهید.