بازیابی هوشمند اطلاعات

تمرین پنجم

مینا فریدی
810100430

بخش یک

سوال ۱)

ابتدا داده ها را با استفاده از تابع آماده read_csv از فایل می‌خوانیم.

	review	sentiment
0	This film is absolutely awful, but nevertheles...	0
1	Well since seeing part's 1 through 3 I can hon...	0
2	I got to see this film at a preview and was da...	1
3	This adaptation positively butchers a classic ...	0
4	Råzone is an awful movie! It is so simple. It ...	0

با استفاده از تابع train_test_split موجود در sklearn، داده‌های آموزش و داده را به نسبت ۸۰ درصد و ۲۰ درصد تقسیم می کنیم. با استفاده از تابع countVectorizer، برای هر یک از داده های تست و آموزش آرایه ای میسازیم که شامل تعداد تکرار کلمات در هر سند (کامنت) است. عملیات حذف stopword ها با دادن پارامتری به این تابع انجام می‌شود.

در مرحله بعد مدل دسته بندی Logistic Regression را میسازیم و داده‌های آموزش را به همراه خروجی آن‌ها در این مدل fit می کنیم. سپس برای داده های تست، با استفاده از این مدل مقادیر خروجی را پیش بینی می‌کنیم. در نهایت بین مقادیر پیش‌بینی شده و مقادیر خروجی واقعی داده های تست مقایسه انجام می‌دهیم. نتایج مقایسه به صورت زیر است:

روش Logistic Regression:

Accuracy score: 0.87

Precision:

0: 0.87

1: 0.87

Recall:

0: 0.87

1: 0.87

F1_score:

0: 0.87

1: 0.87

روش SVM:

Accuracy score: 0.85

Precision:

0: 0.85

1: 0.85

Recall:

0: 0.86

1: 0.84

F1_score:

0: 0.85

1: 0.85

روش Naïve Bayes:

Accuracy score: 0.85

Precision:

0: 0.85

1: 0.85

Recall:

0: 0.86

1: 0.84

F1_score:

0: 0.85

1: 0.85

سوال ۴) استفاده از tfidf:

در این روش به جای درنظرگرفتن تعداد کلمات منتخب، از حاصلضرب tf و idf آن ها استفاده می کنیم.

برای این کار از تابع tfidfVectorizer که مربوط به کلاس sklearn.feature_extraction.text است

استفاده می کنیم و به ورودی آن مقادیر پارامترهای ماکسیسمم ویژگی های ۱۰۰۰۰ (max_features)

و stopwords = 'english' می‌دهیم تا فیچرهای بیشتر از ۱۰۰۰۰ تا انتخاب نکند و همچنین کلمات stopword مثل حروف اضافه را حذف کند.

سپس دوباره این داده ها را به ورودی دسته بندی کننده های مختلف می دهیم و معیارهای مختلف را اندازه می‌گیریم.

روش Logistic Regression:

Accuracy score: 0.89

Precision:

0: 0.89

1: 0.89

Recall:

0: 0.88

1: 0.89

F1_score:

0: 0.89

1: 0.89

روش SVM:

Accuracy score: 0.85

Precision:

0: 0.85

1: 0.85

Recall:

0: 0.86

1: 0.84

F1_score:

0: 0.85

1: 0.85

روش Naïve Bayes:

Accuracy score: 0.85

Precision:

0: 0.85

1: 0.85

Recall:

0: 0.86

1: 0.84

F1_score:

0: 0.85

1: 0.85

سوال ۵) برای این قسمت تعداد ویژگی‌ها را که موقع countVectorize کردن ۱۰۰۰۰ گذاشته بودیم بیشتر می کنیم. نتایج تفاوت چندانی حاصل نشد که نشان می دهد که ویژگی های ذکر شده به اندازه کافی اثر گذار بوده اند. در مقایسه و تحلیل روش های مختلف می بینیم که استفاده از tfid در کل بهتر از استفاده از count است که مخصوصا در روش logistic regression این اثر مشاهده می‌شود. در کل هم روش logistic regression از دو روش دیگر بهتر عمل می کند.

بخش دو

در روش PLSA اجزای متن متغیرهای تصادفی چندجمله‌ای هستند که می‌توانند به عنوان موضوعات بازنمایی شوند. هر کلمه از موضوع واحد تولید می‌شود. کلمات مختلف در یک مدرک می‌توانند از موضوعات مختلف ایجاد شوند. تجزیه و تحلیل معنایی پنهان احتمالاتی مسئله چندمعنایی بودن کلمات را تا حدودی حل می‌کند. PLSA برای مسائلی که به یک موضوع پرداخته اند قابل استفاده است اما برخلاف LSA، این روش در کشف موضوعات و مضامین کلی متن کاربرد دارد.

سوال الف)

$$argmax_{i\in d}(argmax_{j\in w}\left\{\lambda p(d_{i,j}=w\mid D)+(1-\lambda)\sum_{k=1}^{K}p(z_{i,j}=k\mid \pi_i)p(d_{i,j}=w\mid \theta_k)\right\})$$

سوال ب) در کد داده شده قسمت هایی اضافه شد. از جمله ماتریس دو بعدی X که نشان دهنده تعداد رخداد هر کلمه در هر سند است، M نشان دهنده تعداد کلمات متمایز است.

کلمات به دست آمده به صورت زیر است:

```
said rating saudi mrs waste people fbi government new one
said fire roberts year would police people years monday jewish
said new would california states air gas york year north
said central state much snow northern school plant atlantic southern
said year percent soviet barry oil polish peres prices officers
said new greyhound year union president company children settlements one
percent said new dukakis administration people farmer rose government rate
said noriega bush also magellan scientists panama spacecraft president would
said bank new company police bush duracell city would two
said soviet percent gorbachev economy officials businesses black owned economic
```

سوال ج) تابع loglikelihoodBG به کد اضافه شد که احتمال زمینه را حساب می کند. برای محاسبه احتمال زمینه احتمال رخداد کلمه را بر تعداد کل کلمات اسناد تقسیم می کنیم و مقدار زمینه را در کد به متغیر p که در کد قبلی loglikelihood است مقداردهی می کنیم.

مقدار لاندا را در این قسمت ۰.۹ دادیم و کلمات زیر بدست آمد:

```
said new california farmer people year administration percent states government
said gorbachev soviet noriega barry central fbi economic jewish congress
said bank new company also greyhound peres man would oil
said bush rating waste former president would school last dukakis
said year dukakis roberts two black last national people percent
said soviet officers polish children plant world union nikolais years
percent said rate prices year month economy report rose index
said state soviet one government pope new president time church
said fire year bush north people officials monday two would
said new company duracell magellan spacecraft million mrs florio york
```

سوال د) با قراردادن ۰.۳ برای لاندا مقدار زیر بدست آمد:

```
said bank state year noriega central new would southern last
said rating program gorbachev would north fbi warming global new
percent said prices rose rate year month since report oil
said pope church may new people receptor man government could
said bush fire barry roberts police two people night thursday
said soviet duracell company polish union black owned plant officers
said soviet saudi congress united iraq two company military forces
said dukakis people bush administration farmer campaign magellan one spacecraft
said new york california year oil would florio gas exxon
said year new soviet like official peres greyhound israel school
```

سوال چ) مقدار شباهت ماکسیمم با استفاده از مدل بک گراند بهتر می شود پس به نظر می رسد استفاده از مدل زمینه ای در فرمول نتیجه را بهبود می بخشد.