به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

پـزدازش زبـان هـای طبـیـعـی

CA2

فاطمه سلیقه

۸۱۰۱۹۸۳۰۶

اسفند ماه ۱۳۹۸

پیش پردازش

برای پیش پردازش در ابتدا stop word ها را حذف کرده :

```
#remove stop words
stopword = stopwords.words('english')
all_words_without_stopwords = [word.lower() for word in allWords if word not in stopword]
```

سپس punctuationها را حذف کرده:

```
#remove punctuation
punctuations = '''!()-[]--'{}‹;:'"\,<>./?@#$%^&*_~'''
all_words_without_punctuation1 = [word for word in all_words_without_stopwords if word not in string.punctuation]
all_words_without_punctuation = [word for word in all_words_without_punctuation1 if word not in punctuations]
```

سپس اعداد را حذف می کنیم :

```
#remove numbers
all_words_without_numbers = []
for word in all_words_without_punctuation:
    x = re.sub(r"\d+", "", word)
    if(x != ''):
        all_words_without_numbers.append(x)
```

و در اخر هم lemmatize می کنیم :

```
#lemmatization
lemmatizer = WordNetLemmatizer()
all_words_lemmatized = [lemmatizer.lemmatize(word) for word in all_words_without_numbers]
```

ورودی تابع preprocess ، tokenize شده هستند و اینکه متعلق به کدام کلاس هستند :

```
for category in movie_reviews.categories():
    for fileid in movie_reviews.fileids(category):
        reviews.append((preprocess(movie_reviews.words(fileid)), category))
```

با توجه به آنکه می خواهیم sentiment analysis انجام دهیم ، مفهوم کلمات مهم است و نه شکل کلمه و یا حتی stop word ها . بنابراین لازم است تا کلماتی که ضرورتی به وجود انها نیست را حذف کنیم . مثلا stop word ها و punctuation ها و هم چنین تنها بن کلمات را نگه داریم .هم چنین اعداد هم نمی توانند تاثیر زیادی داشته باشند و بیشتر دقت را پایین می آورند . بنابراین بهتر است پیش پردازش انجام دهیم .

استخراج ویژگی

در کل داده ها میزان تکرار هر کلمه را پیدا می کنیم و تعدادی از کلمات پر تکرار را به عنوان ویژگی انتخاب می کنیم :

```
words_frequency = FreqDist(words)
```

سپس بررسی کردیم که میزان تکرار کلمات تا چه حد است و در اخر ۱۵۰۰ ویژگی پرتکرار را انتخاب نمودیم . زیرا معمولا سایر ویژگی ها میزان تکرار ۱ تا ۵۰ بار داشتند که بهتر است حذف شوند .

آموزش رده بند

برای استفاده از naïve bayes از کتابخانه scikit learn استفاده می کنیم .

```
gnb = GaussianNB()
gnb.fit(feature_set,labels)
```

برای k-fold validation به این صورت عمل می کنیم که داده ها را به صورت رندوم جابه جا می کنیم و سپس آنها را به ۵ قسمت تقسیم می کنیم . و هر بار یک قسمت را به عنوان داده تست و بقیه قسمت ها را به عنوان داده آموزش انتخاب می کنیم . و در اخر میانگین می گیریم .

```
y_pred = gnb.predict(feature_set[1201:1600])
accuracy_score(labels[1201:1600], y_pred)
```

ارزیابی مدل آموزش دیده

```
0.8395989974937343
[[182  24]
 [ 40 153]]
```

$$precision = \frac{182}{222} = 0.81$$

$$recall = \frac{182}{206} = 0.88$$

$$F1 = 2 * \frac{(0.81 * 0.88)}{0.81 + 0.88} = 0.84$$

```
0.8421052631578947
[[181  23]
 [ 40 155]]
```

$$precision = \frac{181}{221} = 0.81$$

$$recall = \frac{181}{204} = 0.88$$

$$F1 = 0.84$$

```
0.8571428571428571
[[164  22]
 [ 35 178]]
```

$$precision = \frac{164}{199} = 0.82$$

$$recall = \frac{164}{186} = 0.88$$

$$F1 = 0.84$$

```
0.8646616541353384
[[188  14]
 [ 40 157]]
```

$$precision = \frac{188}{228} = 0.82$$

$$recall = \frac{188}{202} = 0.93$$

$$F1 = 0.87$$

```
0.8225
[[177  24]
 [ 47 152]]
```

$$precision = \frac{177}{224} = 0.79$$

$$recall = \frac{177}{201} = 0.88$$

$$F1 = 0.83$$

دقت میانگین رده بند برابر است با :

```
accuracy =  0.8452017543859649
```

F1 میانگین:

```
F1 =  0.844
```

با توجه به نتایج به دست آمده طبقه بند می تواندبا دقت ۸۴ درصد تشخیص دهد که هر نظر آیا نظری مثب است یا منفی . کلماتی که به عنوان ویژگی انتخاب شدند . کلمات خوبی هستند که می توانند دو جنبه مثبت یا منفی بودن را به خوبی نشان دهند . مثلا کلمه "خوب" کلمه ای است که در نظرات مثبت زیاد تکرار می شود و کلمه "بد" در منظرات منفی زیاد تکرار می شود .زمانی که ما کلماتی را به عنوان ویژگی انتخاب کردیم که تعداد بار تکرارشان زیاد است ، این کلمات می توانند معیار خوبی برای مقایسه باشند .