Project Report 24

Text Classification

NAME: DHIVAKAR.R

COURSE: AI and ML

Batch : Aug 2020

Question:

Using vector semantics, we can easily convert a given text into its corresponding vector form. Given any text, first pre process the text and convert it into a vector using BoW methods. Given this vector, implement your own classifier to classify the vector is pre-defined categories. You may use of these datasets for training and for defining the categories: 14 Best Text classification Datasets for Machine Learning

Prerequisites

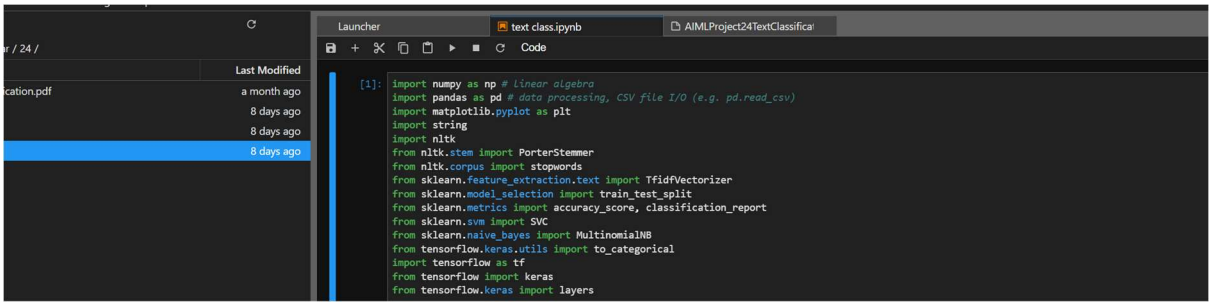
What things you need to install the software and how to install them:

Python 3.6 This setup requires that your machine has latest version of python. The following url https://www.python.org/downloads/ can be referred to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-asan-internal-or-externalcommand/ . Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.

Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url https://www.anaconda.com/download/ You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.6

Implementation

Importing the libraries and dataset



Loading the dataset :

```
[2]: train = pd.read_csv('Corona_NLP_train.csv')
     test = pd.read_csv('Corona_NLP_test.csv')
     train.head(10)
```

[2]:

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative
5	3804	48756	ÃœT: 36.319708,-82.363649	16-03-2020	As news of the regionÂ's first confirmed COVID...	Positive
6	3805	48757	35.926541,-78.753267	16-03-2020	Cashier at grocery store was sharing his insig...	Positive
7	3806	48758	Austria	16-03-2020	Was at the supermarket today. Didn't buy toile...	Neutral
8	3807	48759	Atlanta, GA USA	16-03-2020	Due to COVID-19 our retail store and classroom...	Positive
9	3808	48760	BHAVNAGAR,GUJRAT	16-03-2020	For corona prevention,we should stop to buy th...	Negative

Removing the punctuation and storing the punctuation free text

```
[3]: punctuation = ["'","@","#",",",".","/",":",";",'"',"(",")","_"]
     def remove_punctuation(text):
         punctuationfree="".join([i for i in text if i not in punctuation])
         return punctuationfree

[4]: #storing the puntuation free text
     train['clean_msg']= train['OriginalTweet'].apply(lambda x:remove_punctuation(x))
     train.head(10)
```

[4]:

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment	clean_msg
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	MeNyrbie PhilGahan Chrisitv httpstcoiFz9FAn2Pa...
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive	advice Talk to your neighbours family to excha...
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia Woolworths to give elder...
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive	My food stock is not the only one which is emp...
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative	Me ready to go at supermarket during the COVID...
5	3804	48756	ÃœT: 36.319708,-82.363649	16-03-2020	As news of the regionÂ's first confirmed COVID...	Positive	As news of the regionÂ's first confirmed COVID...
6	3805	48757	35.926541,-78.753267	16-03-2020	Cashier at grocery store was sharing his insig...	Positive	Cashier at grocery store was sharing his insig...
7	3806	48758	Austria	16-03-2020	Was at the supermarket today. Didn't buy toile...	Neutral	Was at the supermarket today Didnt buy toilet ...
8	3807	48759	Atlanta, GA USA	16-03-2020	Due to COVID-19 our retail store and classroom...	Positive	Due to COVID-19 our retail store and classroom...
9	3808	48760	BHAVNAGAR,GUJRAT	16-03-2020	For corona prevention,we should stop to buy th...	Negative	For corona preventionwe should stop to buy thi...

```
[5]: test['clean_msg']= test['OriginalTweet'].apply(lambda x:remove_punctuation(x))
     test.head(10)
```

[5]:

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment	clean_msg
0	1	44953	NYC	02-03-2020	TRENDING: New Yorkers encounter empty supermar...	Extremely Negative	TRENDING New Yorkers encounter empty supermark...
1	2	44954	Seattle, WA	02-03-2020	When I couldn't find hand sanitizer at Fred Me...	Positive	When I couldnt find hand sanitizer at Fred Mey...
2	3	44955	NaN	02-03-2020	Find out how you can protect yourself and love...	Extremely Positive	Find out how you can protect yourself and love...
3	4	44956	Chicagoland	02-03-2020	#Panic buying hits #NewYork City as anxious sh...	Negative	Panic buying hits NewYork City as anxious shop...
4	5	44957	Melbourne, Victoria	03-03-2020	#toiletpaper #dunnypaper #coronavirus #coronav...	Neutral	toiletpaper dunnypaper coronavirus coronavirus...
5	6	44958	Los Angeles	03-03-2020	Do you remember the last time you paid $2.99 a...	Neutral	Do you remember the last time you paid $299 a ...
6	7	44959	NaN	03-03-2020	Voting in the age of #coronavirus = hand sanit...	Positive	Voting in the age of coronavirus = hand saniti...
7	8	44960	Geneva, Switzerland	03-03-2020	@DrTedros "We can□t stop #COVID19 without prot...	Neutral	DrTedros We can□t stop COVID19 without protect...
8	9	44961	NaN	04-03-2020	HI TWITTER! I am a pharmacist. I sell hand san...	Extremely Negative	HI TWITTER! I am a pharmacist I sell hand sani...
9	10	44962	Dublin, Ireland	04-03-2020	Anyone been in a supermarket over the last few...	Extremely Positive	Anyone been in a supermarket over the last few...

Removing the unwanted Columns

```
[7]: unwanted_cols = ['UserName', 'ScreenName', 'Location', 'TweetAt', 'OriginalTweet', 'clean_msg','Sentiment']
```

```
[8]: train_X = train.drop(unwanted_cols, axis=1)
     train_Y = train['Sentiment']
     test_X = test.drop(unwanted_cols, axis=1)
     test_Y = test['Sentiment']
     stop_words = nltk.corpus.stopwords.words('english')
     stop_words[:5]
```

```
[8]: ['i', 'me', 'my', 'myself', 'we']
```

```
[9]: train_X['length'] = train_X['msg_lower'].apply(len)
     train_X.head(10)
```

[9]:

	msg_lower	length
0	menyrbie philgahan chrisitv httpstcoifz9fan2pa...	92
1	advice talk to your neighbours family to excha...	237
2	coronavirus australia woolworths to give elder...	124
3	my food stock is not the only one which is emp...	284
4	me ready to go at supermarket during the covid...	287
5	as news of the regionâ's first confirmed covid...	238
6	cashier at grocery store was sharing his insig...	168
7	was at the supermarket today didnt buy toilet ...	107
8	due to covid-19 our retail store and classroom...	272
9	for corona preventionwe should stop to buy thi...	260

```
[12]: def pre_process(text):
          text = [word for word in text.split() if word.lower() not in stop_words]
          words = ""
          for i in text:
                  words += (ps.stem(i))+" "
          return words
```

```
[13]: textFeatures_train = train_X['msg_lower'].copy()
      textFeatures_train = textFeatures_train.apply(pre_process)
      textFeatures_train[:5]
```

```
[13]: 0    menyrbi philgahan chrisitv httpstcoifz9fan2pa ...
      1    advic talk neighbour famili exchang phone numb...
      2    coronaviru australia woolworth give elderli di...
      3    food stock one empti pleas dont panic enough f...
      4    readi go supermarket covid19 outbreak im paran...
      Name: msg_lower, dtype: object
```

[13]:
```
textFeatures_train = train_X['msg_lower'].copy()
textFeatures_train = textFeatures_train.apply(pre_process)
textFeatures_train[:5]
```

[13]:
```
0    menyrbi philgahan chrisitv httpstcoifz9fan2pa ...
1    advic talk neighbour famili exchang phone numb...
2    coronaviru australia woolworth give elderli di...
3    food stock one empti pleas dont panic enough f...
4    readi go supermarket covid19 outbreak im paran...
Name: msg_lower, dtype: object
```

[14]:
```
textFeatures_test = test_X['msg_lower'].copy()
textFeatures_test = textFeatures_test.apply(pre_process)
textFeatures_test[:5]
```

[14]:
```
0    trend new yorker encount empti supermarket she...
1    couldnt find hand sanit fred meyer turn amazon...
2                   find protect love one coronaviru ?
3    panic buy hit newyork citi anxiou shopper stoc...
4    toiletpap dunnypap coronaviru coronavirusaustr...
Name: msg_lower, dtype: object
```

[15]:
```
textFeatures_combined = pd.concat([textFeatures_train, textFeatures_test], axis = 0)
len(textFeatures_combined)
```

[15]:
```
44955
```

[16]:
```
(44955, 546515)
```

[17]:
```
train_X = features_combined[:len(textFeatures_train)]
test_X = features_combined[len(textFeatures_train):]
train_X.shape, test_X.shape
```

[17]:
```
((41157, 546515), (3798, 546515))
```

[18]:
```
len(train_Y), len(test_Y)
```

[18]:
```
(41157, 3798)
```

[19]:
```
target_names = train_Y.unique()
target_names
```

[19]:
```
array(['Neutral', 'Positive', 'Extremely Negative', 'Negative',
       'Extremely Positive'], dtype=object)
```

[20]:
```
# Prediction using Support Vector Machine
svc = SVC(kernel='sigmoid', gamma=1.0)
```

[21]:
```
svc.fit(train_X, train_Y)
```

[21]:
```
SVC(gamma=1.0, kernel='sigmoid')
```

[23]:
```python
# accuracy_score(labels_test,prediction)
print(classification_report(test_Y, prediction, target_names = target_names))
```

```
                    precision    recall  f1-score   support

           Neutral       0.72      0.39      0.50       592
          Positive       0.77      0.47      0.58       599
Extremely Negative       0.50      0.56      0.53      1041
          Negative       0.61      0.63      0.62       619
Extremely Positive       0.48      0.67      0.56       947

          accuracy                           0.56      3798
         macro avg       0.62      0.54      0.56      3798
      weighted avg       0.59      0.56      0.56      3798
```

[24]:
```python
# Prediction using Multinomial Naive Bayes Model
mnb = MultinomialNB(alpha=0.2)
mnb.fit(train_X, train_Y)
prediction = mnb.predict(test_X)
# accuracy_score(labels_test,prediction)
print(classification_report(test_Y, prediction, target_names = target_names))
```

```
                    precision    recall  f1-score   support

           Neutral       0.81      0.07      0.13       592
          Positive       0.77      0.14      0.24       599
Extremely Negative       0.39      0.50      0.43      1041
          Negative       0.64      0.18      0.28       619
Extremely Positive       0.33      0.75      0.46       947

          accuracy                           0.38      3798
         macro avg       0.59      0.33      0.31      3798
      weighted avg       0.54      0.38      0.34      3798
```