Assignment 04

Title: Stemming and feature selection techniques using vectors

Problem Statement : Consider a suitable text. Remove stop words, apply stemming and feature selection techniques to represent documents as vectors, Classify documents and evaluate precision and recall.

Learning Objective : • Implementation of problem using python.
• Remove stop words, applying stemming and feature selection.

Learning Outcome : • Understanding the stemming and feature selection process.
• Learn about precision and recall.

Theory:

1] STOP WORDS.

• In Computing, stop words are words which are filtered out before or after processing of natural. language data (text).

• Through "stop words" usually refers to most common values or words in a language, there is no universal list of stop words used by all natural processing tools, and indeed not all tools even use such a list.

• Any group of words can be chosen as the stop words for a given purpose. For some search engine following are the most common, short function words:- the, is, at, which, and, on, etc.

2] STEMMING :-

• Stemming is the process of reducing inflected words to their word stem, base or root form generally a written word form.

• The stem need not be identical to the morphological root of the word, it is usually sufficient that related words map to the same stem, even if the stem is not itself a valid root.

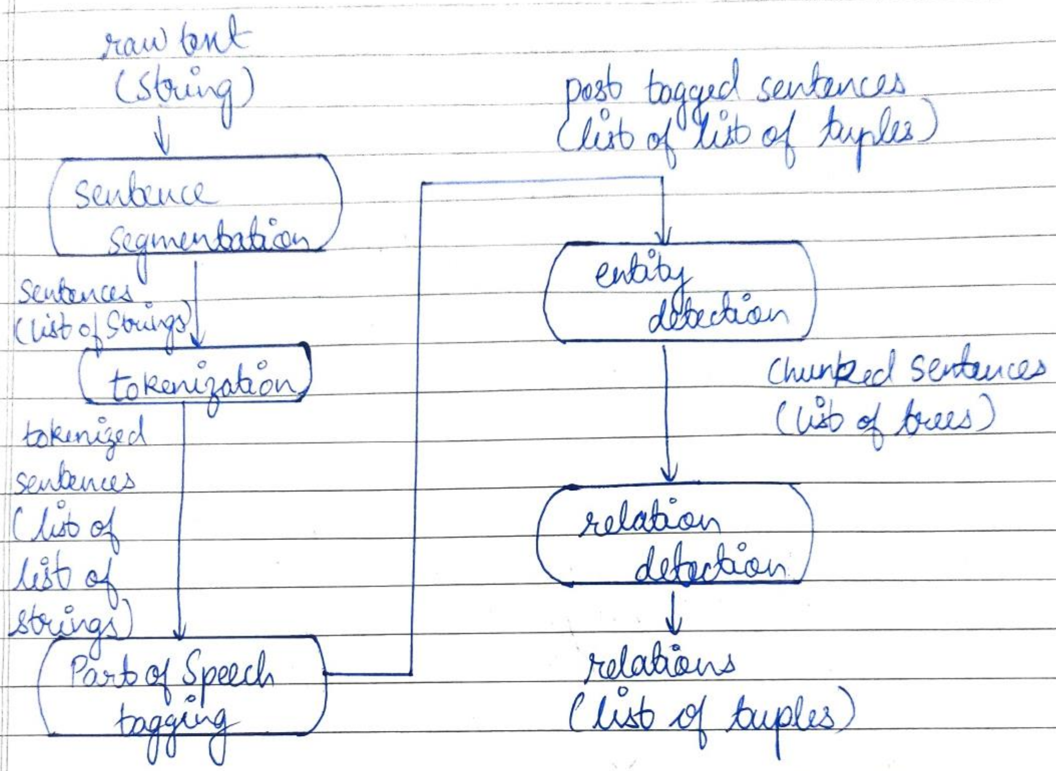
• Many search engines treat words with same stems as synonyms as a kind of query expansion, a process called conflation.

• Suffix - stripping algorithm is famous for stemming.

3] FEATURE SELECTION

• In machine learning and statistics, feature selection, also known as variable subset selection, attribute or variable subset selection, is a process of selecting a subset of relevant features (variables, predictions) for use in model construction.

• Feature extraction Architechture


raw text (string)
sentence segmentation
sentences (list of strings)
tokenization
tokenized sentences (list of list of strings)
Part of Speech tagging
post tagged sentences (list of list of tuples)
entity detection
chunked sentences (list of trees)
relation detection
relations (list of tuples)

→ Feature selection techniques are used for four reasons:

1. Simplification of models to make them easier to interpret by researchers / users.
2. Shorter training time.
3. To avoid the curse of dimentionality
4. Enhanced generalization by reducing lover fitting (formally, reduction of variance)

→ Stemming with nltk tool in python module.

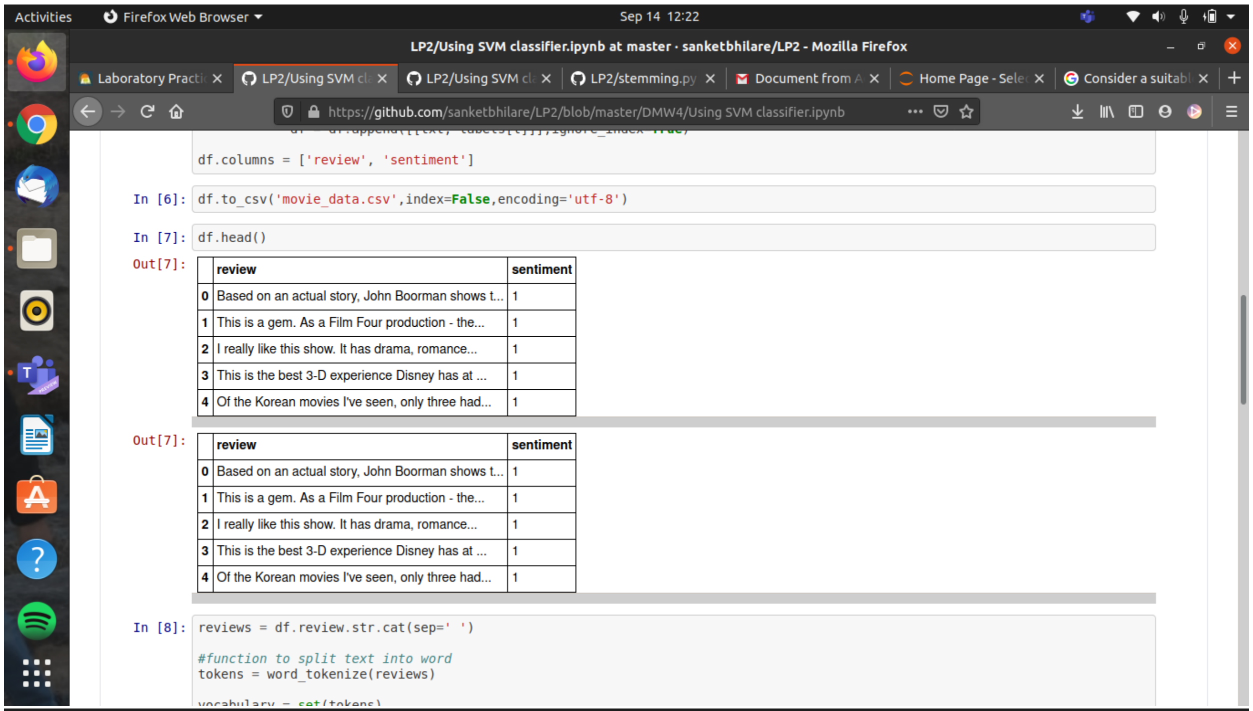
```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize,
                            word_tokenize.

ps = PorterStemmer().
example_words = ("python", "pythoner", "Pythoning",
          "Pythoned", "pythonly" ]
for w in example_words:
    print (ps.stem (w))
```

Test Cases:

Sr no	Description	Expected O/P	Actual O/P
1)	Import pandas, os and nltk libraries into jupyter notebook	success	success
2)	pre-process the data, and. append pos/neg to review and sentiment columns.	success	success
3)	Apply suffix-stripping stemming algorithm	success	success
4)	Divide data into train and test and obtain accuracy, precision and recall of the model built	success	success

Conclusion: Thus, we have studied to remove stop words, apply stemming and feature extraction techniques to represent documents as vectors.

Output:


Activities
Firefox Web Browser
Sep 14 12:22
LP2/Using SVM classifier.ipynb at master · sanketbhilare/LP2 - Mozilla Firefox
https://github.com/sanketbhilare/LP2/blob/master/DMW4/Using SVM classifier.ipynb

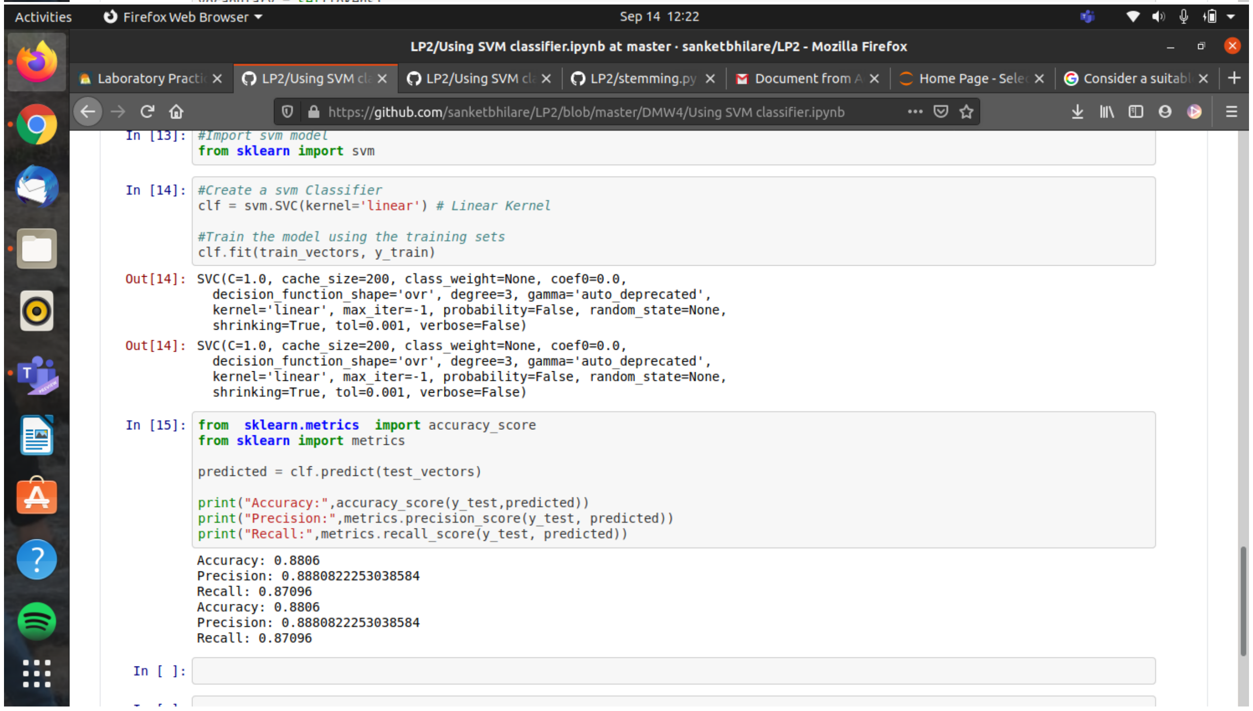
```
    df = df.append([[txt, labels[i]]],ignore_index=True)
    df.columns = ['review', 'sentiment']

In [6]: df.to_csv('movie_data.csv',index=False,encoding='utf-8')

In [7]: df.head()

Out[7]:
```

	review	sentiment
0	Based on an actual story, John Boorman shows t...	1
1	This is a gem. As a Film Four production - the...	1
2	I really like this show. It has drama, romance...	1
3	This is the best 3-D experience Disney has at ...	1
4	Of the Korean movies I've seen, only three had...	1

Out[7]:

	review	sentiment
0	Based on an actual story, John Boorman shows t...	1
1	This is a gem. As a Film Four production - the...	1
2	I really like this show. It has drama, romance...	1
3	This is the best 3-D experience Disney has at ...	1
4	Of the Korean movies I've seen, only three had...	1

```
In [8]: reviews = df.review.str.cat(sep=' ')

        #function to split text into word
        tokens = word_tokenize(reviews)

        vocabulary = set(tokens)
```


Activities
Firefox Web Browser
Sep 14 12:22
LP2/Using SVM classifier.ipynb at master · sanketbhilare/LP2 - Mozilla Firefox
https://github.com/sanketbhilare/LP2/blob/master/DMW4/Using SVM classifier.ipynb

```
In [13]: #Import svm model
         from sklearn import svm

In [14]: #Create a svm Classifier
         clf = svm.SVC(kernel='linear') # Linear Kernel

         #Train the model using the training sets
         clf.fit(train_vectors, y_train)

Out[14]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
           decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
           kernel='linear', max_iter=-1, probability=False, random_state=None,
           shrinking=True, tol=0.001, verbose=False)
```

```
Out[14]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
           decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
           kernel='linear', max_iter=-1, probability=False, random_state=None,
           shrinking=True, tol=0.001, verbose=False)
```

```
In [15]: from  sklearn.metrics  import accuracy_score
         from sklearn import metrics

         predicted = clf.predict(test_vectors)

         print("Accuracy:",accuracy_score(y_test,predicted))
         print("Precision:",metrics.precision_score(y_test, predicted))
         print("Recall:",metrics.recall_score(y_test, predicted))

         Accuracy: 0.8806
         Precision: 0.8880822253038584
         Recall: 0.87096
         Accuracy: 0.8806
         Precision: 0.8880822253038584
         Recall: 0.87096

In [ ]:
```