Machine Learning Online Assignment

Assignment 8 - Topic Modelling using SVM

Problem Statement - In this assignment, we work on 20 online newsgroups dataset. A newsgroup in a place on the Internet where people can ask questions related to certain topic. The data contains apprx 20k examples across 20 classes as shown below. The data is split into training at test sets as shown below. The dataset is available at http://qwone.com/~jason/20Newsgroups/ and is also available in sci-kit learn.

This notebook is present in "../Datasets/Newsgroups/TopicModellingDataset.ipynb" directory.

```
from sklearn.datasets import fetch_20newsgroups
import numpy as np
```

```
obj = fetch_20newsgroups()
#The dataset has the following classes, each class is mapped with an integer
id.
for i,j in enumerate(obj['target_names']):
    print(i,j)
```

```
0 alt.atheism
1 comp.graphics
2 comp.os.ms-windows.misc
3 comp.sys.ibm.pc.hardware
4 comp.sys.mac.hardware
5 comp.windows.x
6 misc.forsale
7 rec.autos
8 rec.motorcycles
9 rec.sport.baseball
10 rec.sport.hockey
11 sci.crypt
12 sci.electronics
13 sci.med
14 sci.space
15 soc.religion.christian
16 talk.politics.guns
17 talk.politics.mideast
18 talk.politics.misc
19 talk.religion.misc
```

```
data_train = fetch_20newsgroups(subset='train',random_state=10)
data_test = fetch_20newsgroups(subset='test',random_state=10)

X_train = data_train.data
Y_train = data_train.target

X_test = data_train.data
Y_test = data_train.target
```

```
print(X_train[0])
```

```
From: rj@ri.cadre.com (Rob deFriesse)
Subject: Can DES code be shipped to Canada?
Article-I.D.: fripp.1993Apr22.125402.27561
Reply-To: rj@ri.cadre.com
Organization: Cadre Technologies Inc.
Lines: 13
Nntp-Posting-Host: 192.9.200.19

Someone in Canada asked me to send him some public domain DES file
encryption code I have.  Is it legal for me to send it?

Thanx.
--
Eschew Obfuscation

Rob deFriesse                    Mail:  rj@ri.cadre.com
Cadre Technologies Inc.          Phone:  (401) 351-5950
222 Richmond St.                 Fax:    (401) 351-7380
Providence, RI  02903

I don't speak for my employer.
```

```
print(Y_train[0])
```

```
11
```

Clearly we can see the class 11 is sci.crypt, which is evident from words like "DES","file","encryption","code" in the document.

Tasks

1. Visualise a bar-chart chart, each bar showing the number of documents in each class.
2. Data Cleaning
   Perform Lemmatization and Stopword Removal on the dataset. Use WordNet Lemmatizer. Remove all numbers and non-english words like "94","a86". Filter the top 500 words based upon the frequency to use as features.
3. Use Count Vectorizer, or TF-IDF Vectorizer class to construct features.
4. Use the following approaches(you can use the sci-kit version)

- SVM with Linear Kernel and Grid Search over C = [1,2,5,10]
- Report best score(Training and Testing Set) and best value of C. Use Cross Validation count cv=3.

Bonus
Use K-Means(Unsupervised learning) to cluster similar documents together, all documents belonging to one topic can be clustered together.