**Team Members:**

1. فارس احمد عبد التواب عثمان
2. عبد الرحمن عمرو محمد احمد
3. عبد الرحمن علاء حموده
4. عبد العزيز جمال على محمد
5. يوسف سامح رشدي عبد الحميد

[NewsGroup Classifcaiton]

NEWSGROUP CLASSIFCAITON

Introduction:

The main idea of project is to classify articles to their news group .

The objectives of this project is to preprocessing on data ,cleaning and train it with proper model .

Methodology:

We read files data from folders ['alt.atheism','comp.graphics','comp.os.ms-windows.misc','comp.sys.ibm.pc.hardware','comp.sys.mac.hardware','comp.windows.x','misc.forsale','rec.autos','rec.motorcycles','rec.sport.baseball','rec.sport.hockey','sci.crypt','sci.electronics','sci.med','sci.space','soc.religion.christian','talk.politics.guns','talk.politics.mideast','talk.politics.misc','talk.religion.misc']

And then make preprocessing , cleaning it with these steps and save it at pickle file:

* Tokenizing all text data at files
* clear all stop words
* clear all words that have length less than 2
* clear all punctuation and the words with numbers
* do lemmatize on words to get the source meaning of them

after that we read that file pickle and split it to train and test sets , do vectorizing for them and then we use support vector machine classifier to train on data and we continually change at the hyper parameters until we get best test accuracy.

Data Set Summary:

1-What is the data set used?

Twenty Newsgroups = > 20news-18828 contains files 18828

2- What is the summary of the dataset columns?

Two columns:

* The first one for targets
* The second one for features after cleaning it

RangeIndex: 18828 entries, 0 to 18827

Data columns (total 2 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 0 18828 non-null int32

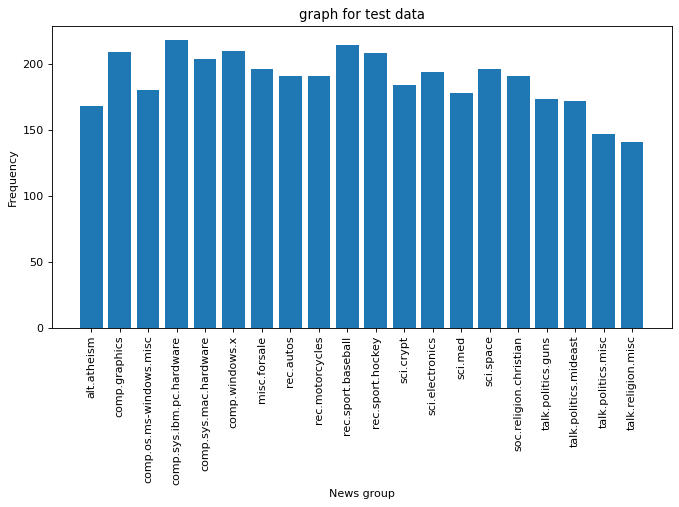
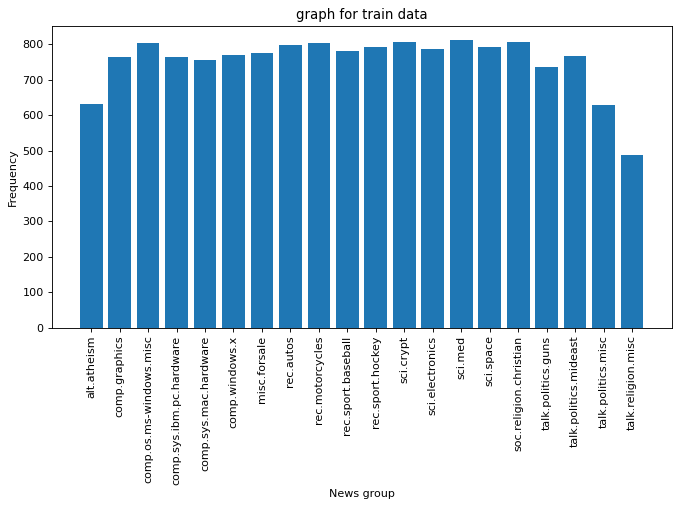
1 1 18828 non-null object

dtypes: int32(1), object(1)

memory usage: 220.8+ KB

None

3- Visualize the dataset statistics



Results:

SVC Model Test Score is: 0.8950863213811421

SVC Model Train Score is: 0.9996016729735112

صورة تحتوي على نص, شاشة عرض

تم إنشاء الوصف تلقائياً