ML 512 Project Choice 2 - Explore Dataset(2-3)

In [2]:

```
import numpy as np
import pandas as pd

from sklearn import tree
from sklearn.pipeline import Pipeline
from sklearn import metrics
from sklearn.feat1ure_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.datasets import fetch_20newsgroups
```

In [3]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
categories = ['alt.atheism', 'talk.religion.misc','comp.graphics', 'sci.space']
newsgroups_train = fetch_20newsgroups(subset='train',categories=categories)
vectorizer = TfidfVectorizer()
vectors = vectorizer.fit_transform(newsgroups_train.data)
vectors.shape
```

Out[3]:

```
(2034, 34118)
```

In [4]:

```
vectors.nnz/ float(vectors.shape[0])
```

Out[4]:

```
159.0132743362832
```

In [5]:

```
sparsity = (vectors.nnz/ float(vectors.shape[0]))/ vectors.shape[1]
print("Sparsity : % 0.4f" %(100*(1-sparsity))+' %')
```

```
Sparsity :  99.5339 %
```

The extracted TF-IDF vectors are very sparse, with an average of 159 non-zero components by sample in a more than 30000-dimensional space (less than .5% non-zero features):

In [7]:

```
newsgroups_train = fetch_20newsgroups(subset='train')
newsgroups_test = fetch_20newsgroups(subset='test')
X_train = newsgroups_train.data
X_test = newsgroups_test.data
y_train = newsgroups_train.target
y_test = newsgroups_test.target

text_clf = Pipeline([('vect', CountVectorizer()),
                     ('tfidf', TfidfTransformer()),
                     ('clf', tree.DecisionTreeClassifier()),
                     ])

text_clf.fit(X_train, y_train)


predicted = text_clf.predict(X_test)

print(metrics.classification_report(y_test, predicted))
```

```
              precision    recall  f1-score   support

           0       0.49      0.48      0.49       319
           1       0.42      0.43      0.42       389
           2       0.51      0.56      0.53       394
           3       0.44      0.44      0.44       392
           4       0.54      0.57      0.55       385
           5       0.47      0.48      0.48       395
           6       0.69      0.73      0.71       390
           7       0.62      0.60      0.61       396
           8       0.73      0.77      0.75       398
           9       0.52      0.55      0.54       397
          10       0.68      0.67      0.68       399
          11       0.76      0.69      0.73       396
          12       0.35      0.33      0.34       393
          13       0.48      0.44      0.46       396
          14       0.67      0.64      0.65       394
          15       0.69      0.70      0.70       398
          16       0.50      0.61      0.55       364
          17       0.77      0.59      0.67       376
          18       0.40      0.39      0.39       310
          19       0.32      0.30      0.31       251

    accuracy                           0.56      7532
   macro avg       0.55      0.55      0.55      7532
weighted avg       0.56      0.56      0.56      7532
```