**Document Categorization**

The task is categorize an article to an appropriate topic.

**Dataset**

The dataset was collected from <https://scdnlab.com/corpus/> . The dataset contains articles of 12 categories in 12 different folders. The number of articles each folder contains is given below.

1. Accident :: 6350 articles
2. Art :: 2669 articles
3. Crime :: 8840 articles
4. Economics :: 5351 articles
5. Education :: 12389 articles
6. Entertainment :: 10139 articles
7. Environment :: 6852 articles
8. International :: 5922 articles
9. Opinion :: 8116 articles
10. Politics :: 20479 articles
11. Science\_tec :: 2906 articles
12. Sports :: 12086 articles

**Feature Extraction**

We used TF-IDF values of each unique words in a document as a feature vector for a single document.

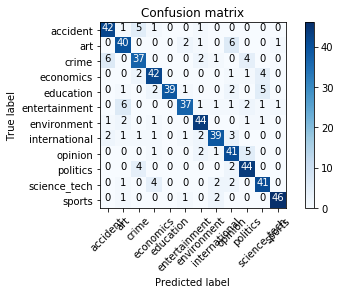
**Experiment Procedure**

The task is mainly classification. For each categories we took 500 articles in our train dataset and 50 articles in our test dataset. We fed the TF-IDF vectors of train dataset in a classifier and checked how the classifer worked on test dataset. We used Different Classifiers for this task. The results of our experiments is given below.

**Experiment Results**

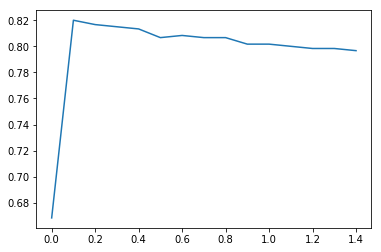
**Naïve Bayes:**

We achieved accuracy rate of 82.0 using Naïve Bayes Classifier. We used ***sklearn***’s MultinomialNB as classifier. The confusion matrix is given below.



*Figure: Confusion matrix of Naïve Bayes Classifier.*

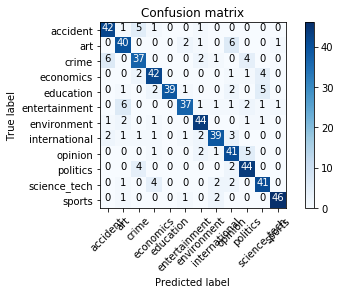
Multinomial Naive Bayes has a parameter ***alpha,*** which acts as a additive smoothing parameter. We tuned this value and got different accuracy rates.



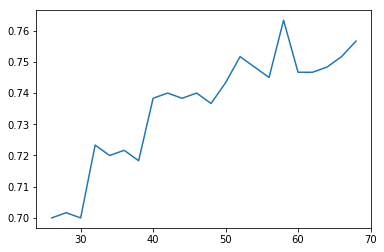
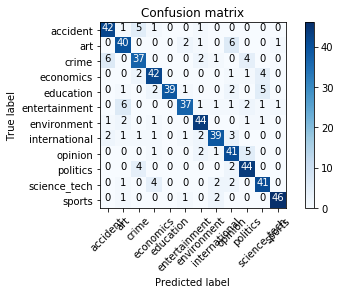
*Figure: Accuracy rate vs* ***alpha*** *value*

**Passive Aggressive Classifier:**

Using passive aggressive classifier, we got an accuracy rate of 77.3%. Below is the confusion matrix.



**Random Forest Classifier:**

We achieved an accuracy rate of 76.3% using Random Forest Classifier setting number of estimator trees to 58.  

*Fig: Accuracy rate vs number of trees Fig: Confusion matrix Random Forest*

**XGBoost:**

We achieved accuracy rate of 75.2% using XGBoost classifier. The confusion matrix is given below.

