**Laptop Specs**

Windows-8,8 gb ram,i7 proessor

I am new to Machine Learning and NLP field. Firstly I read both test and train csv files.

Then i removed---(**editing test and train.ipynb**)

* html tags
* everything inside <code> tags</code>
* stop words
* punctuations(except #,+) due to C#,c++
* Converted to lower form

The above part is done using dataframes and reglar expression.Finally i saved these files to test\_new and train\_new csv files

I found out that most of the tags are already present in title, making it more important than article.So i decided to only make a model on title.

Now I tried to train model on train\_new.csv but due to almost 8 lakh rows i got memory error while creating the Bag -of-Words model

So I finally removed duplicated based on the tags which left me with 5 lakhs 43 thousand rows in training data.

I again tried to train the model but still got memory error.

**final\_copy.ipynb**

I finally decided to create a tf-idf model using 50k rows of train set and predicitng all test\_set and finally taking the union of tags and removing repeated tags.

after tfidf model is created using multilabelbinarizer,onevsrest classifier,and logistic regression

So trainining data set of 5 lakhs 43 thousands rows became

[0-50000]/[50k to 1 lalk]/..........[5 lakh to 5 lakh 43 thousands]

Now this each small training set is trained and made predictions in

[0-5000.csv//.........5to5lk43.csv]

J**oining all the datasets through datframe.ipynb**

Finally i joined all these csv files tags through python coding skills.

such as set,join,list and it helped me landing at good rank

**Things i should have tried**

* **prediciton on title using BOW**
* **prediciton on article using tfidf**
* **prediciton on article using bow**
* **prediciton on union of article and title using bow and tfidf**