**Twitter Sentiment Analysis**

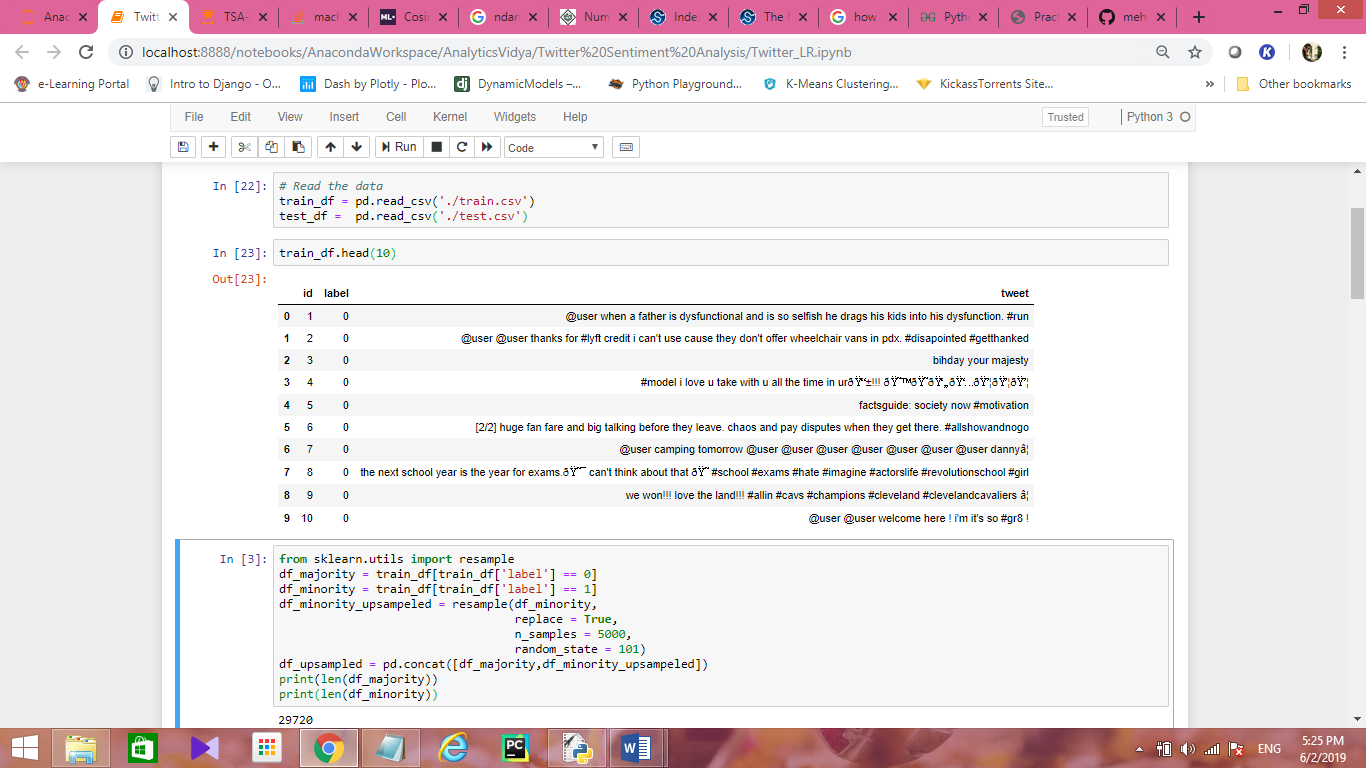
Problem Statement

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets. Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, your objective is to predict the labels on the test dataset.

**Data Description**

**Features**

1. **Label**- Indicates wheather tweet is positive or negative
2. **Tweet**- User tweet



**Observations**

Tweet contains

1. @user tags
2. hash tags
3. special symbols
4. no emoji’s
5. slang words

Data Preparation

1. Removing the @user from tweet
2. Removal of html tags
3. Converting the slang words into normal words

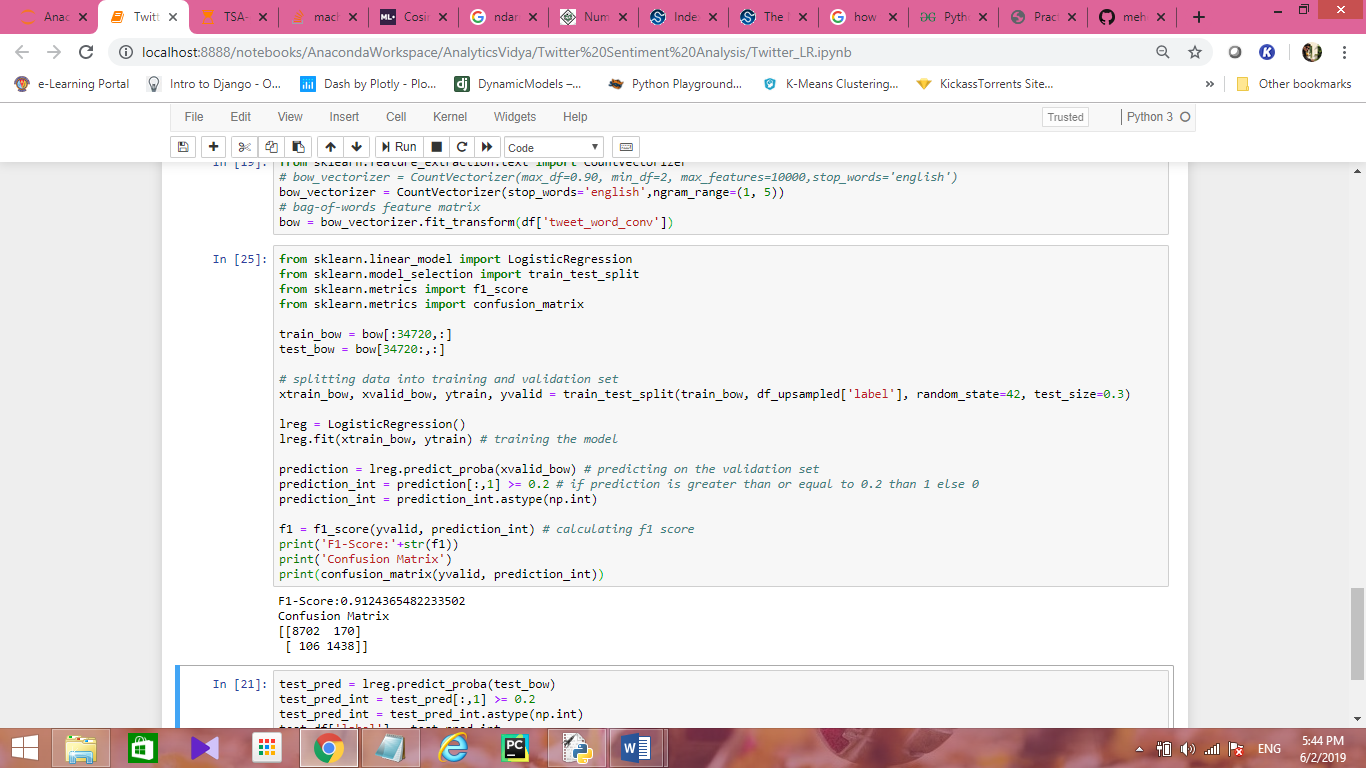
Ex: Gr8 to Great, OMG to oh my god

1. Removing the special characters and numbers
2. Removal of rare words from tweets
3. Removing the words whose length is <3 Ex: is, an ,am,he,it
4. Lemmatization
5. Preparation of Bag Of Words

Model

Selected Model: Logistic Regression

Evaluation Metric: F1-Score



Here we got f1-score as 0.91 on our training data

On test set we got public score as 0.72 which is good score.