## LP1 -> DA4

* Title: Twitter Data Analysis

* Problem Statement:
Use Twitter data for sentiment analysis. The data is 3MB in size and has 31,962 tweets. Identify the tweets which are hate and which are not.

* Objective:
To classify the tweets as hate ~~and~~ or not.

* S/w and H/w requirements:

i) 64 bit processor
ii) RAM
iii) Linux OS
iv) Python 3

* Theory:
Natural Language Processing (NLP) is a subfield of linguistics, CS and AI, concerned with interactions between computer and human language in particular, and how to program computers to process and also analyze large amounts of data.

Stop words are the words that are filtered out before or after the language data is processed.

Stemming for grammatical reasons, text can use different forms of a word.

There are also families of derivationally related words with similar meaning.
Stemming reduces inflectional forms and sometimes derivationally linked forms to its common base form.

Feature selection is the process of selection of a subset of the terms occuring in the training set and using only this subset of features in text classification.

This makes the classifier more efficient, as well as more accurate because it eliminates noise.

Vectorization is the process of converting the text data into machine readdble form.
IF-IDF vectors are related to one-hat encoding, but instead of featuring a count, they feature numerical representations where words aren't just binary. Instedd, they're represented by their term frequency multiplied by their Inverse doc frequency.

For this problem, the classification used were: Multinomial Naive Bayes, Random Forest and Linear support vector classifier.

* Accuracy of 95% was achieved. Tweets were pre-processed to convert them lower case, remove @ mentions, removed numbers and punctuations.

* Conclusion:

From this assignment, I was able to understand the basics of Natural Language Processing (NLP) and hence classify the tweets as hate and non-hate.