IDS 566- Final Project

Due by 6:00pm CT on Thursday, May 9th

Dataset

For the final project we will use the "The 20 Newsgroups" data set available at:

http://qwone.com/~jason/20Newsgroups/

To simplify your project you are allowed to use the version of the "20 Newsgroups" data set that comes preloaded with scikit-learn library.

To load is you can use the following code:

```
#Loading the 20 Newsgroup data set. Example loading the training data.
from sklearn.datasets import fetch_20newsgroups
mydata_train = fetch_20newsgroups(subset='train', shuffle=True)
```

You can learn more from the scikit-learn documentation at:
https://scikit-learn.org/0.19/modules/generated/sklearn.datasets.fetch_20newsgroups.html#sklearn.datasets.fetch_20newsgroups

Project Description

The goal of the project is to train different classifier and test the outcomes trying to achieve as high accuracy as you can. Please make sure you are clear on which metrics you are using to measure how good is your classification.
You can use as many techniques as you have learned in class. Especially may be helpful preparing and transforming the data. Selecting the appropriate classification algorithm with the optimal parameters will give you different results.

Final deliverables:

You can work as a group but must submit each individually the following project deliverable:

1. Project report (Word or PDF) that describes in detail your project activities, what specific steps you did, what challenges you encountered and how you resolved them.
    a. Describe the different models you ran and the different results that you achieved with each model
    b. Describe the different types of data preparation and transformation you performed and how the different data steps affected the outcomes

    c. Finally describe your best model and outcome, show the results you achieve and comment on your insights and learnings from the project
2. Jupyter Notebook in Python 3.6 or higher that can be executed without errors
    a. The code must be very well documented explaining and self-explanatory from the notebook
    b. The code must run without errors
    c. The code must contain at least three (3) different models – trained and tested – showing the results
    d. Any interesting data transformations must be highlighted and their impact on the outcome must be described and explained
3. Configuration and setup file with user instruction
    a. The file must briefly describe how I should set up and execute your notebook. The easier it is for me to run and evaluate your code, but better your project will be accepted