This assignment is based on the job descriptions obtained from stackoverflow.com

 obtained using a notebook like the one discussed in class several weeks ago

(Data\_Science\_Job\_Descriptions.ipynb, from May 16, in

**Some Jupyter Notebooks -- Data Science in Python**).

**Download your HW5 data from:**

*http://datamining.cs.ucla.edu/cs249/hw5/\_\_\_your\_nine\_digit\_UID\_\_\_.zip*

**Problem 1  (75 points):**

The Training data    train\_data.csv

has job classifications for about 800 job postings on careers.stackoverflow.com.

In the last column, about 20% of these are classified as "data science" jobs ('DS');

the remainder are "non-data science" ('nonDS').

Develop the most accurate classifier for data science jobs that you can.

Use your classifier to make predictions for the 1-column dataset

        test\_data.csv

and upload your predictions to CCLE in a 1-column dataset

        test\_predictions.csv

The file test\_data.csv has about 20% DS jobs, and 80% nonDS jobs.

Your score on this problem will be your classification accuracy for the test data ---

so if your classifier always predicts 'nonDS' you could get about 80%.

**Problem 2  (25 points):**

The scikit-learn documentation page

<http://scikit-learn.org/stable/modules/feature_selection.html>

describes a number of scikit-learn functions for feature selection.

Use the method **SelectKBest()** (and **chi2**) to find the top k = 10 features

for train\_data.csv, and upload them in a 1-column dataset

        top\_10\_features.csv

**Summary:**

for this (short) assignment, based on the data that you download

from the URL above,  you should upload two .csv files to CCLE:

test\_predictions.csv  and  top\_10\_features.csv .