Text Analytics Individual Assignment

This assignment involves building and testing classification models to predict salaries from the text contained in the job descriptions. The data for this assignment can be found at http://www.kaggle.com/c/job-salary-prediction

Randomly select **2500** data points from the **training dataset** ("Train_rev1.csv") for ease of analysis. Then split the 2500 data points into training (80%) and test (20%) sets.

You will create classification models to predict **high** (75th percentile and above) or **low** (below 75th percentile) salary from the text contained in the job descriptions. Use the **Naïve Bayes** classifier.

Hint: Check out http://www.nltk.org/book/ch06.html (esp. 1.3) for illustrations. Also, you can find plenty of relevant resources online.

1. Build a classification model with text (full job description) as the predictor. What is the accuracy of your model? Show the confusion matrix. Also show the top 10 words (excluding stopwords) that are most indicative of (i) high salary, and (ii) low salary.

2. If you wanted to increase the accuracy of the model above, how can you accomplish this using the dataset you have?

Deliverables

The deliverables for this assignment are python scripts, outputs (including plots & tables where applicable) and your answers to the questions.