**CAP 6776 Information Retrieval (Fall 2016)**

**Homework # 2**

**DUE: 11/06/2016 11:59pm EST Total: 20 points**

Given a preprocessed document collection, please conduct document classification using Weka (you can download Weka at <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>).

**Data set:** WebKB containing 2803 training text data and 1396 test data. This data set contains WWW-pages collected from computer science departments of various universities. These web pages are classified into 4 categories: student, faculty, project, and course. The data set has been preprocessed with removing stop words and stemming. So you only need to count the word frequency to generate a document-word matrix before you start classification.

**Method:** please use Naïve Bayes and SVM in Weka to conduct text classification and return the classification accuracy. You can use either Weka to calculate the word frequency or you can use NLTK to generate the matrix as we did in homework #1 and load the matrix into Weka for classification.

**Report:** please write a report including the screenshots of generating document-word matrix, loading the given dataset into Weka, conducting classification using naïve bayes and svm. Please specify the parameters you choose if applicable and show the classification accuracy.