**Natural Language Processing (Spring 2017)**

**Homework # 2**

**DUE: 2/26/2017 11:59pm EST Total: 10 points**

Given a preprocessed document collection, please conduct document classification using LibSVM (you can download LibSVM at https://www.csie.ntu.edu.tw/~cjlin/libsvm/ ).

**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 first term in each line of the data file is the class label.) 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:**

1. You can use NLTK as we introduced previously in the class to generate the document-word matrix.
2. Once you download Libsvm, you can find .exe file in the "Windows" folder. You can also incorporate libsvm in Matlab, Python, and Java using the files in corresponding folders.
3. Change the document-word matrix into the following libsvm format:

   <label> <index>:<value> <index>:<value>
   for example, we can use label "1" represent "student", and one sample data can be changed as follows.
   1 5:1 7:1 14:1 19:1 39:1 40:1 51:1 63:1 67:1 73:1 74:1 76:1 78:1 83:1

4. Follow the instruction in README file to and choose linear kernel to perform training and testing using the given data set.

**Report:** write a report including the screenshots of generating document-word matrix, input data files for training and testing, command lines you use to train and test, and the output results (including accuracy) shown by libsvm.