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

**Homework # 1**

**DUE: 09/25/2016 11:59pm EST Total: 20 points**

Given a collection of documents, conduct text preprocessing including tokenization, stop words removal, stemming, tf-idf calculation, and pairwise cosine similarity calculation using NLTK (similar to the lecture on 09/08 or you can choose other text processing tools you prefer). The following steps should be completed:

1. Install Python 2 and NLTK (3 points)
2. Tokenize the documents into words, remove stop words, and conduct stemming (5 points)
3. Calculate tf-idf for each word in each document and generate document-word matrix (each element in the matrix is the tf-idf score for a word in a document) (7 points)
4. Calculate pairwise cosine similarity for the documents (5 points)

Please include your screen shots for each of the above steps and also the final results of the pairwise cosine similarity scores in your report.

Note: the detailed NLTK tutorial and datasets for homework 1 have been uploaded in Blackboard.