**Text Analytics**

**Assignment 1**

Identifying hot topics in news articles, press release, social media, etc. is becoming an important task for many businesses. While many of them still rely on manual monitoring, some companies have turned to an automatic approach to explore hot topics using text mining. In general, words/phrases/concepts that appear most frequently can be considered as a “hot topic.”

This is an exercise to explore the hottest topics in news article. You are required to run 3 comparison studies. The website you will be collecting news articles from is <http://www-03.ibm.com/press/us/en/pressreleases/recent.wss>. Please collect at least 20 news articles from this website (crawling or manual download). You may use tools such as NLTK, Stanford parser, Weka or any tools you have access to determine the frequency of keywords in the collection. You may also write your own program. Please describe each step you take to **find and rank hot topics**. Compare the following three approaches:

1. Use a simple bag-of-words approach
2. Use a bag-of-words with stemming and stop word removal approach
3. Use POS and focus on NNP approach
4. Discuss other things you can do to improve the performance (e.g. IDF). Implementation of improved methods are not required. However, if you decide to implement other methods, you may get extra points.

You should submit your final work to Blackboard in a word file. Figures and tables are encouraged in the report. Try to limit your final report to 3 pages maximum.