Homework

# Text Mining Homework – Sentiment Analysis

**Step 1: Read in the positive and negative word files**

1. Find two files (one for positive words and one for negative words) from

<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html> (about halfway down the page, “A list of English positive and negative opinion words or sentiment words”).

2. Create two vectors of words, one for the positive words and one for the negative words.

3. Note that when reading in the word files, there might be lines at the start and/or the end that will need to be removed (i.e. you should clean you dataset).

**Step 2: Process in the MLK speech**

4. Read the text file from [www.analytictech.com/mb021/mlk.htm](http://www.analytictech.com/mb021/mlk.htm).

5. Parse the file using the XML package (or cut and paste into a text file). If you parse html file using the XML package, the following code might help:

# Read and parse HTML file

doc.html = htmlTreeParse('http://www.analytictech.com/mb021/mlk.htm', useInternal = TRUE)

# Extract all the paragraphs (HTML tag is p, starting at

# the root of the document). Unlist flattens the list to

# create a character vector.

doc.text = unlist(xpathApply(doc.html, '//p', xmlValue))

# Replace all \n by spaces

doc.text = gsub('\\n', ' ', doc.text)

# Replace all \r by spaces

doc.text = gsub('\\r', ' ', doc.text)

6. Create a term matrix.

7. Create a list of counts for each word.

**Step 3: Determine how many positive words were in the speech**

8. Scale the number based on the total number of words in the speech.

Hint: one way to do this is to use the “match” function and then “which” function.

**Step 4: Determine how many negative words were in the speech**

9. Scale the number based on the total number of words in the speech.

Hint: this is basically the same as step (3).

**Step 5: Redo the “positive” and “negative” calculations for each 25% of the speech**

10. Compare the results (e.g.. a simple barchart of the 4 numbers).