Lab

# Lab 11 – Text Mining

For the association rules mining homework, you used a “termDocMatrix.” Now that we are doing text mining, we will be creating our own termDocMatrix.

This was also done in class, when we analyzed the structure of the “I have a dream” speech – in terms of the use of positive and negative words. However, in that effort, we treated all positive words the same (e.g. good is the same as great). This might not be appropriate – maybe we should count more positive (and negative) words than other words. For example, “I loved the movie” might be stronger than “I liked the movie.”

There is a different word file that ranks each word on a scale of −5 to 5 (negative to positive). It is known as the AFINN word list.

Your task for this homework is to adapt the lab that we did in class to compute the score for the MLK speech using the AFINN word list (as opposed to the positive and negative word lists).

1. First read in the AFINN word list (e.g., from <http://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-111.txt>).

Note that each line is both a word and a score (between −5 and 5). You will need to split the line and create two vectors (one for words and one for scores).

2. Read in the MLK “I have a dream” speech.

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

b. 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)

c. Create a term matrix.

d. Create a list of counts for each word.

3. Compute the overall score for the MLK speech using the AFINN word list (as opposed to the positive and negative word lists).