**Homework 10 – Wielding the tm package and counting words**

Finding text, restructuring it, filtering it, counting particular words or phrases, and creating visualizations – these are critical skills for many data scientists. The tm package provides an essential toolbox for manipulating text data. In this homework, we build upon the activities described in Chapters 14 and 15 to count and plot the negative and positive words in a political speech.

In this homework, we will explore the free form text comments from the hotel survey responses (the same dataset that was used in HW8 and HW10).

**Part A: Load and condition the text file that contains the speech**

1. The data is available on blackboard, as a JSON file (see HW8 if you need a reminder on the dataset or how to load the dataset).
2. The key column to focus on is the ‘freeText’ column.

**Part B: Create a list of word counts from the speech**

1. Starting with the code at the bottom of page 180 in the text book, use a similar approach to transform the free text into a term document matrix, and then determine positive and negative word matches.
2. Calculate the percent positive words and negative words.
3. Write a block comment that summarizes what you learned from ratioPos and ratioNeg.

**Part D: Visualize the results**

1. Create a word cloud
2. Create a barplot of the positive and negative words that matched (at least twice)
3. Write a block comment on what you observe from these two barplots and the wordcloud.
4. Does these results make sense to you in terms of the kinds of emotions you see?  
   Which do you think is more informative – barplot or the wordcloud?

**Part E: Evaluate Happy and not Happy customer responses**

1. Create two subset of the text vectors: one for happy customers and one for not happy customers (based on overall customer satisfaction).
2. Redo Steps B, C & D, for these two subsets of the text strings.
3. Compare the positive and negative ratios for these two different group of customers