Natural Language Processing - Mini-Project

Aarthi Reddy

September 22, 2017

DEADLINE: Oct 20^{th}, 2017

1 Introduction

In this assignment you will be asked to perform some tasks that are commonly used in Natural Language Processing. As in the previous project, you will be writing a shiny app to display your visualizations/work. Again as in the previous project, you may use any R package you like, however your work should be original.

2 Description of Dataset

There are two datasets that you will be using. First if the 'Reviews.csv' dataset which is a set of amazon reviews downloaded from Kaggle. It has already been cleaned, organized and ready for use. The table consists of the following columns:

1. Id: This is the review id number. Basically row number
2. ProductId: The id number of the product being reviewed. The number is not unique. There are several products that have been reviewed multiple times.
3. UserId: The id number of user providing the reviews. The number is not unique. There are several users that have reviewed multiple products
4. ProfileName: The profile name of the user writing the reviews.
5. HelpfulnessNumerator: Number of users who found the review helpful.
6. HelpfulnessDenominator: Number of users who indicated that they found the review helpful.
7. Score: Rating of the product from 1 to 5
8. Time: Timestamp for the review
9. Summary: Summary of the review

10. Text: Actual review itself.

The second dataset is a very popular one commonly used for sentiment analysis. This is AFINN-111 which is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Arup Nielsen in 2009-2011. The file is tab-separated.

3 Objectives

1. For each review remove all punctuation, convert uppercase to lowercase, and remove stop words. The results of the new 'normalized' review should be in its own column. Display Id, ProductId, UserId, Review and normalized review table in your app.

2. Create and display a table with two columns. The first column is a list of all words and the second is the count of occurrence of the words in the reviews. This table is not for each individual review, but for all reviews. This table is the term count table for all the reviews.

3. Create and display a word cloud of the 200 most frequently occurring words from the term count table you generated in 2.

4. Create and display a table with two columns. The first column is a list of all words and the second is the count of occurrence of the words in the normalized reviews you generated in 1. This table is not for each individual normalized review, but for all normalized reviews. This table is the term count table for all the normalized reviews.

5. Create and display a word cloud of the 200 most frequently occurring words from the term count table you generated in 4.

6. What are your observations of the 2 word clouds you generated in 3 and 5?

7. Use the AFINN-111 to obtain a sentiment score for each review in the csv file. First replace every **phrase** or **word** in each review with its corresponding AFINN score. If there is not AFINN score for a word then ignore it. Sum up the scores per review. A positive score indicates a positive sentiment, a negative score indicates a negative sentiment. The magnitude of the score indicates the degree of sentiment.

8. Create and display a table with three columns: Product ID, Number of Users who reviewed the product and Average score of the product.

9. Display top 6 most reviewed products and their average score.

10. For each of the top 6 most reviewed products show scatter plots with sentiment score per review on the x-axis and the Rating as provided per reviewer on the y-axis. So you should end up with 6 different plots.

11. In reference to 10, can you see/show any correlation between the sentiment score and the product rating? What are your thoughts or recommendations on the sentiment analysis work that you have done. How can we improve sentiment analysis? What modifications to the text did you have to do (if any) before performing sentiment analysis?

12. Be creative

4 Checking your Code

Aim to complete the following set of tasks before you send me the code:

1. Restart R and try and run your code before sending it to me. Restarting R will unload and loaded libraries or packages, and you will have to 'require' or 'library' your packages for the code to run.

2. Comment your code. The more commenting there is, the more it will help me figure out your thought process.

3. If you have tried any other package(s), let me know in the comments