Natural Language Processing - Mini-Project

Aarthi Reddy

March 6, 2018

DEADLINE: April 4^{th}, 2018; 8am

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. I have found the following package very useful: `https://www.tidytextmining.com`. You can use this or any other package of your choice.

2 Description of Dataset

You will be analyzing Toyota Camry car reviews for this assignment. The reviews are available online and you will be programmatically downloading them.
`https://www.cars.com/research/toyota-camry/`
We are specifically interested in the 2012 - 2017 Camry model. You will be using the 2012, 2013, 2014, 2015 and 2016 reviews for training and the 2017 reviews for testing purposes.

3 Objectives

1. (10 points) Download the training data (reviews) from the online link. Include the star rating given by the reviewer. Display this as a table where each row corresponds to a single review. The table should have three columns: year of the review, star rating given by the reviewer, text of review.

2. (5 points) Download the test data (reviews) from the online link. Include the star rating given by the reviewer. Display this as a table where each row corresponds to a single review. The table should have three columns: year of the review, star rating given by the reviewer, text of review.

3. (5 points) For each review in train and test set, remove all punctuation, convert uppercase to lowercase. The results of the new 'normalized' review should be in its own column (this will be the fourth column).

4. (5 points) Tag each review according to the presence of the following words: 'service', 'price', 'handling', 'interior'. The tag can be in the fifth column. It is possible that you will have multiple tags per review.

5. (25 points) There are several lexicons that help compute sentiment analysis (bing, afinn, nrc etc.). Use any sentiment analysis text to compute the sentiment of each review.

6. Give the following information:

   (a) (5 points) What is the average sentiment rating of the reviews in the training set of the Toyota Camry? Compare to average star rating provided by user.

   (b) (5 points) What is the average sentiment rating of the reviews corresponding to each of the 4 tags? Compare to average star rating provided by user and also to average star rating from 6a

7. (10 points)Build a model that can predict the star rating given the sentiment analysis you compute. The model should be built using just the training data. Use any type of model you choose

8. (10 points)Use the model you built in 7 to predict the star rating of the test reviews. What is the accuracy of your model. Remember this is not an exercise in optimizing your model, so don't spend too much time on this part of the assignment.

9. (20 points) For each tag in the training dataset compute the TF-IDF for every word in the set. Remove stop words. For each tag set visualize the TF-IDF scores for the top 10 words.

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