**Homework 13: Text Analytics**

**Due Dec 10 11:59 PM (22 points)**

IST 3420 – Fall 2017, Chen

**Name**: \_\_\_\_\_\_\_Adam Forestier\_\_\_\_\_\_\_\_\_\_\_\_\_

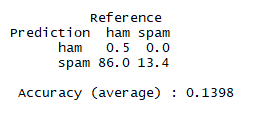
**Programming Task: Short Message Classification**

The dataset is a collection of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam. In total, there are 4,827 legitimate messages and 747 mobile spam messages.

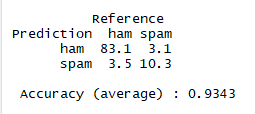
***Tasks***

1. Download the data file from <http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>. Click “Link 1” to download the data file in plain text. The data is stored in the file “SMSSpamCollection.txt”.
2. Read the plain text to create a data frame in R. The data frame should contain the following three columns. (2 points)
   1. ID: ID of the message;
   2. Text: the content of the message;
   3. Class: the category of the message (either ham or spam)
3. Create a corpus to represent the SMS. Each SMS is a document in the corpus. (2 points)
4. Apply text preprocessing methods such as strip whitespace, convert case, remove unimportant symbols, filter stop words, and stem. (2 points)
5. Create a word cloud to show the frequent terms occur in the corpus. (2 points)
6. Create a term-document matrix to represent the textual dataset. (2 points)
7. Use term features to classify the category of the message (documents as rows, term features as columns) by using the following detailed methods:
   1. Use a 30-70% data partitioning strategy (1 point)
   2. Use naïve Bayes classifier and logistic regression model
   3. Show confusion matrix for each classifier (2 points)
   4. Calculate accuracy, recall, precision, kappa, and AUC for each classifier (2 points)
8. Compare the performance of the naïve Bayes classifier and logistic regression model. Type the confusion matrix of each classifier below. Fill in the information in the following table. (2 points)

Type confusion matrix for naïve Bayes here:



Type confusion matrix for logistic regression here:



|  |  |  |
| --- | --- | --- |
| **Measures** | **Naïve Bayes** | **Logistic Regression** |
| Accuracy | .253 | .853 |
| Recall | .139 | .952 |
| Precision | .992 | .966 |
| Kappa | .392 | .707 |
| AUC | .771 | .913 |

***Question 1***: What conclusion can you reach based on the comparison in step 7? Please note that since this is an unbalanced dataset, the accuracy measure may not provide a good assessment of the classification performance. (3 points)

I can conclude that logistic regression is the more accurate method and has a more accurate effect on the dependent variable. Naïve Bayes is more precise however.

***Question 2***: What challenges and problems did you encounter in doing the text analytics assignment? How did you solve those problems? (2 points)

I encountered strange errors throughout that made me adjust code. I thought this was the most difficult assignment of the semester. I had many issues that I resolved by referring to examples provided, PowerPoints, and resources online.

***Submission:*** Submit this document with answers and the R Markdown file (.Rmd) that shows your analysis.