HW3 – Engineering Analytics II

Due on March, 7th (11:59pm)

The data is about the page language revision history. Notations of the data can be seen in the data file. Develop a vandalism detector to distinguish a vandalism from a valid edit.

1) How many cases of vandalism were detected?

2) Preprocessing the data includes creating corpus for added column, removing stop words in English, words stemming and document term matrix building. How many terms appear in the document term matrix?

3) Filter out sparse terms by keeping only terms that appear in 0.8% or more of the revisions, and call the new matrix sparseAdded. How many terms appear in sparseAdded?

4) Convert sparseAdded to a data frame called wordsAdded, and then prepend all the words with the letter A. Now repeat all of the steps we've done so far (create a corpus, remove stop words, stem the document, create a sparse document term matrix, and convert it to a data frame) to create a Removed bag-of-words dataframe, called wordsRemoved, except this time, prepend all of the words with the letter R. How many words are in the wordsRemoved data frame?

5) Combine the two data frames wordsAdded and wordsRemoved into a data frame called wikiWords. Then add the Vandal column. Set the random seed to 123 and then split the data set to put 70% in the training set. What is the accuracy on the test set of a baseline method that always predicts "not vandalism" (the most frequent outcome)?

6) Build a CART model to predict Vandal, using all of the other variables as independent variables. Use the training set to build the model and the default parameters. What is the accuracy of the model on the test set, using a threshold of 0.5?

7) Plot the CART tree. How many word stems does the CART model use?

Submission instructions:

Please submit both your solution description file in .pdf or .doc and your R coding file in .R. Please also paste your R codes in the end of your solution description file. Each question need to be addressed with answers, missing of answers will lead to no credit. You have one week to finish the work. You can choose any packages you are familiar with.