Naive Bayes Classifier on Text data

CIS400/600 Fundamentals of Data and Knowledge Mining

February 23, 2017

Problem 1

Install and load packages `tm` and `e1071` to implement text classification using Naive Bayes classifier

Problem 2

1. Load `sms_spam.csv`. Build a CORPUS using text data and perform following text cleaning steps (Refer `tm_map` method)
   (a) Convert text to lower case
   (b) Remove additional white spaces
   (c) Remove numbers
   (d) Remove English stopwords
   (e) Perform stemming
2. Perform hold out method by splitting data into 80% training and 20% test set
3. Use terms with frequency 10 that are part of CORPUS to train Naive Bayes classifier
4. Test the above model using test set
5. Print confusion matrix

Problem 3

To improve classification accuracy, perform Laplacian smoothing using output of steps (1), (2) and (3) of problem 2.