Northeastern University
Khoury College of
Computer Sciences

Practice Problems 4

As always, complete the practice problems on time and review them with your peers. If you would like more detailed feedback come to office hours.

Problem 1 (50 Points)

Build an R Notebook of the SMS message filtering example in the textbook on pages 103 to 123 (data set). Show each step and add appropriate documentation. Note that the attached data set differs slightly from the one used on the book; the number of cases differ.

Problem 2 (50 Points)

Install the requisite packages to execute the following code that classifies the built-in ***iris*** data using Naive Bayes. Build an R Notebook and explain in detail what each step does. Be sure to look up each function to understand how it is used.

```
library(klaR)
data(iris)

nrow(iris)
summary(iris)
head(iris)

testidx <- which(1:length(iris[, 1]) %% 5 == 0)

# separate into training and testing datasets
iristrain <- iris[-testidx,]
iristest <- iris[testidx,]

# apply Naive Bayes
nbmodel <- NaiveBayes(Species~., data=iristrain)
```

Notes

- Producing a word cloud directly from raw text: might produce the error "invalid input 'â£1000' in 'utf8towcs'". There is a solution at stackoverflow. Basically you need to remove non-graphical characters from the raw text using regular expressions (explained in the link).
- The *Dictionary()* function from the **tm** package is deprecated and no longer available. Instead, store the result of *findFreqTerms* directly as a character vector with: *sms_dict <- findFreqTerms(sms_dtm_train, 5)*.
  (*thanks to Annie Bryant for these suggestions*)

Submission Details

- Practice Problems are for learning and practice and therefore are not graded and no submission is required. You are encourage to discuss and review them with your peers. Additionally, they are reviewed during weekly recitations. If you desire, you may ask for individual feedback from the instructional staff during office hours. Completing practice problems will prepare you for the graded practicums and their completion is critical to doing well on the practicums and the final project.

Useful Resources

- R Markdown Notebooks
- SMS Spam Collection Data Set (CSV)

Northeastern University
Khoury College of
Computer Sciences

Learning

Blackboard
Lynda.com
Data Camp

Northeastern University
Khoury College of
Computer Sciences

Search