JHU Data Science Capstone Project - Milestone Report

Ahmed Assal

Sunday, March 29, 2015

1 Synopsis

This report explains the explores the corpus, HC Corpora, from http://www.corpora.heliohost.org/ in preparation for the John Hopkins Data Science Capstone project. The project is a Shiny app that takes as input a phrase (multiple words) in a text box input and outputs a prediction of the next word. This document explains only the major features identified of the data including summary statistics about the data sets. It also reports any interesting findings that we amassed so far.

In addition, the report elaborates our goals for the eventual app and algorithm and briefly summarize our plans for creating the prediction algorithm and Shiny app. It also serves as a basis for collecting feedback on our plans for creating a prediction algorithm and Shiny app. #Data #Data Processing ##Loading packages

We used different R packages for this report. More specifically, we used: * For text manipulation: RWeka, stringi * For text mining: SnowballC, tm * For data visualization: stringr

1.1 Downloading data

```
## [1] "de_DE" "en_US" "fi_FI" "ru_RU"
```

```
## [1] "en_US.blogs.txt"   "en_US.news.txt"    "en_US.twitter.txt"
```

1.2 Loading data

```
dataCorporaPath = "../../data/Coursera-SwiftKey/final/en_US/"
newsCorpusFile = "en_US.news.txt"
blogsCorpusFile = "en_US.blogs.txt"
twitterCorpusFile = "en_US.twitter.txt"

sampleSize = 0.00001

readLines2=function(fname) {
 s = file.info( fname )$size
 buf = readChar( fname, s, useBytes=T)
 strsplit( buf,"\r\n",fixed=T,useBytes=T)[[1]]
}

# if (!file.exists("corpora.RData")){
  newsCorpus= readLines2(fname=paste0(dataCorporaPath, newsCorpusFile))
  blogsCorpus= readLines2(fname=paste0(dataCorporaPath, blogsCorpusFile))
  twitterCorpus= readLines2(fname=paste0(dataCorporaPath, twitterCorpusFile))
#   save.image("corpora.RData")
# } else{
#   load("corpora.RData")
# }
```

```
##              filename  size_MB total_words   lines mean_words
## 1   en_US.blogs.txt 200.4242    38154238  899288   42.42716
## 2    en_US.news.txt 196.2775    35010782 1010242   34.65584
## 3 en_US.twitter.txt 159.3641     2136398  167155   12.78094
```

```
## [1] 19
```

2 Results

3 Conclusion