JHU Data Science Capstone Project - Milestone Report

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.

2 Data Processing

2.1 Loading packages

We used different R packages for producing report. More specifically:
For text manipulation, we used: RWeka, stringi
For text mining, we used: SnowballC, tm
For data visualization, we used: ggplot2

2.2 Downloading data

The HC Corpora where downloaded from www.corpora.heliohost.org . See the readme file at http://www.corpora.heliohost.org/aboutcorpus.html for details on the corpora available. The files have been language filtered but may still contain some foreign text.

The file downloaded had a large size of 548MB. After unzipping the file, we find the following directories:

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

Each of them represent Corpora for a specific language. We will use English Corpora here for our proficiency with the English language. If we list the files in the english corpora directory, we see that it contains files for news, blogs and twitter.

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

2.3 Loading data

When loading all of the english Corpora in the data file, we are actually loading around 151 million words in 2 million lines. That's about four times the size of the Encyclopedia Britannica.

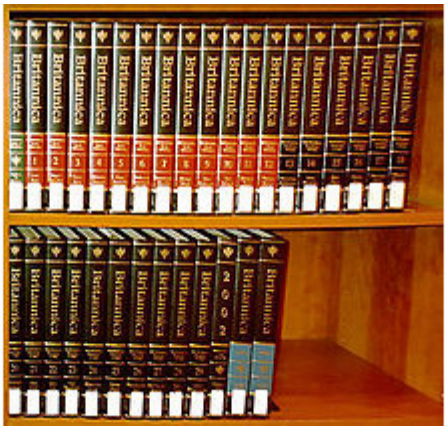

Figure 1: 15th edition of the Britannica. *Wikipedia.*

2.4 Data Features

2.4.1 Summary Statistics

To get a better sense of the size of this data, the following summary statistics elaborates the main features of the corpora.

```
##            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
## 4             Total 556.0658    75301418 2076685   34.65582
```

2.4.2 Word Cloud of Most Frequent Unigrams (1 word)

2.4.3 Word Cloud of Most Frequent Bigrams (2 words)

2.4.4 Word Cloud of Most Frequent Trigrams (3 words)

2.4.5 Word Cloud of Most Frequent Quadgrams (4 words)

3 Shiny App Strategy

4 Conclusion


can
rosa
review
reason
big
will
even
get
time
â€”
work
just
good

Figure 2:


mfr inc
factori littl
busi invest
like batman
polit charg
front camera
east interst
reason give
fill manner
fellow intellectu
anoth felt
even say
dirti joke
unabl use
back better

Figure 3:

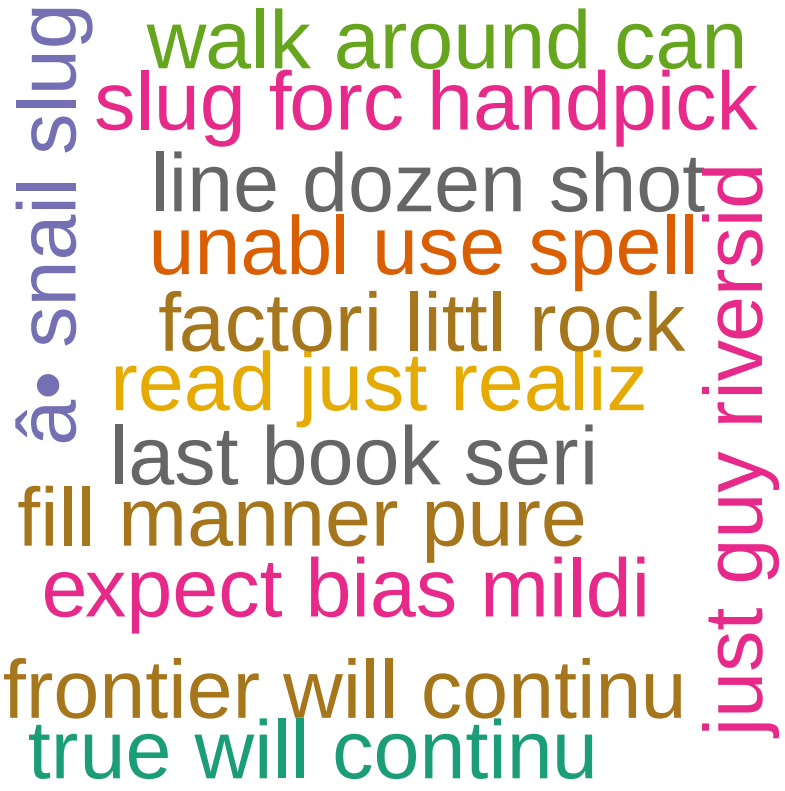
walk around can
slug forc handpick
line dozen shot
unabl use spell
factori littl rock
read just realiz
last book seri
fill manner pure
expect bias mildi
frontier will continu
true will continu
â• snail slug
just guy riversid

Figure 4:


front camera everi week
review write review last
teari backstori hes just
guy riversid reason weep
mfr inc new york
dirti joke apolog friend
bob big boy hair
fangirl contain best proof
oswego parent also told
shot pier â€" two
proxi busi invest plan
lake oswego parent also
flagstaff go east interst
can get fun shot
bias mildi funni somewhat
read just realiz mistaken
beechcraft employ peopl rough
wrote mani interest detail
lunaci place go realli
nondefens capit good exclud
photo bike complet origin
meanwhil financi diminish polit
urban grime shamen shenanigan
rather irrelev piec inform
play game â€" well
left bit encourag christina
call host passionetti women
actual discuss book pleas
anoth felt need tell
one feel reason give
best proof inde lunat
daniel rosa loud yellow
place go realli hurt
game â€" well beach
entertain fellow intellectu friend
like batman supervillain cross
peopl rough work wichita
batman supervillain cross diner
waiter rosa teari backstori
parti place neighbor call
septemb two straight declin
chief us economist mfr

Figure 5: