JHU Data Science Capstone Project - Milestone Report

Sunday, March 29, 2015

1 Synopsis

This report 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.

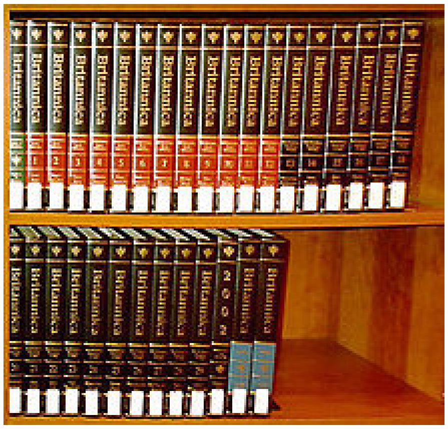

Figure 1: 15th edition of the Britannica. *Wikipedia.*

2.4 Data Features and 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
```

2.5 Data Sampling

Beacause the data size is too large which has resulted in many slowdowns and freezes in the computer while exploring the data, I have taken a sample that constitutes 10% of the original data. 10% of each of the files downloaded for the english corpora, the number of lines of the sample is around 200,000 lines verses 2,000,000 lines for the original corpora.

2.6 Data Cleaning

Before delving into tokenization of the corpora, i.e. extracting the unique n-gram words or phrases, we will clean the corpora through the following transofrmations:
1- Converting the corpora text into UTF-8 encoding.
2- Removing numbers.
3- Removing unnecessary punctuation.
4- Optionally removing English stopwords, we might cancel this step later as we progress as it may affect the perfromance of our predictive model.
5- Optionally removing profanity words, we might cancel this step later as we progress as it may affect the perfromance of our predictive model. 6- Converting all alphabetical characters into lowercase.
7- Removing extra white space such as tab, etc…
8- Stemming the different words, i.e. removing any suffixes such ing, er, s, etc…
9- Converting the text into plain text document.

2.7 Corpora Tokenization

2.7.1 Most Frequent Unigrams (1 word)

2.7.1.1 Bigrams Word Cloud The word cloud shows the most common unigrams larger in scale.

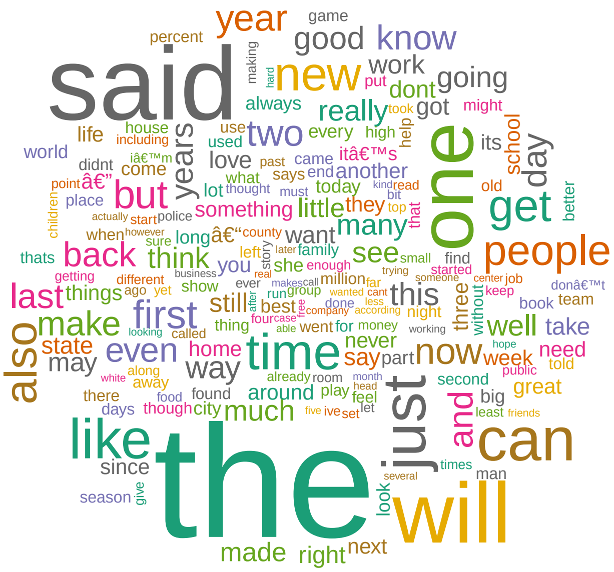

Figure 2:

2.7.1.2 Unigrams Histogram The histogram shows the different frequencies of the top unigrams.

2.7.2 Most Frequent Bigrams (2 words)

2.7.2.1 Bigrams Word Cloud The word cloud shows the most common bigrams larger in scale.

2.7.2.2 Bigrams Histogram The histogram shows the different frequencies of the top bigrams.

2.7.3 Most Frequent Trigrams (3 words)

2.7.3.1 Trigrams Word Cloud The word cloud shows the most common trigrams larger in scale.

2.7.3.2 Trigrams Histogram The histogram shows the different frequencies of the top trigrams.

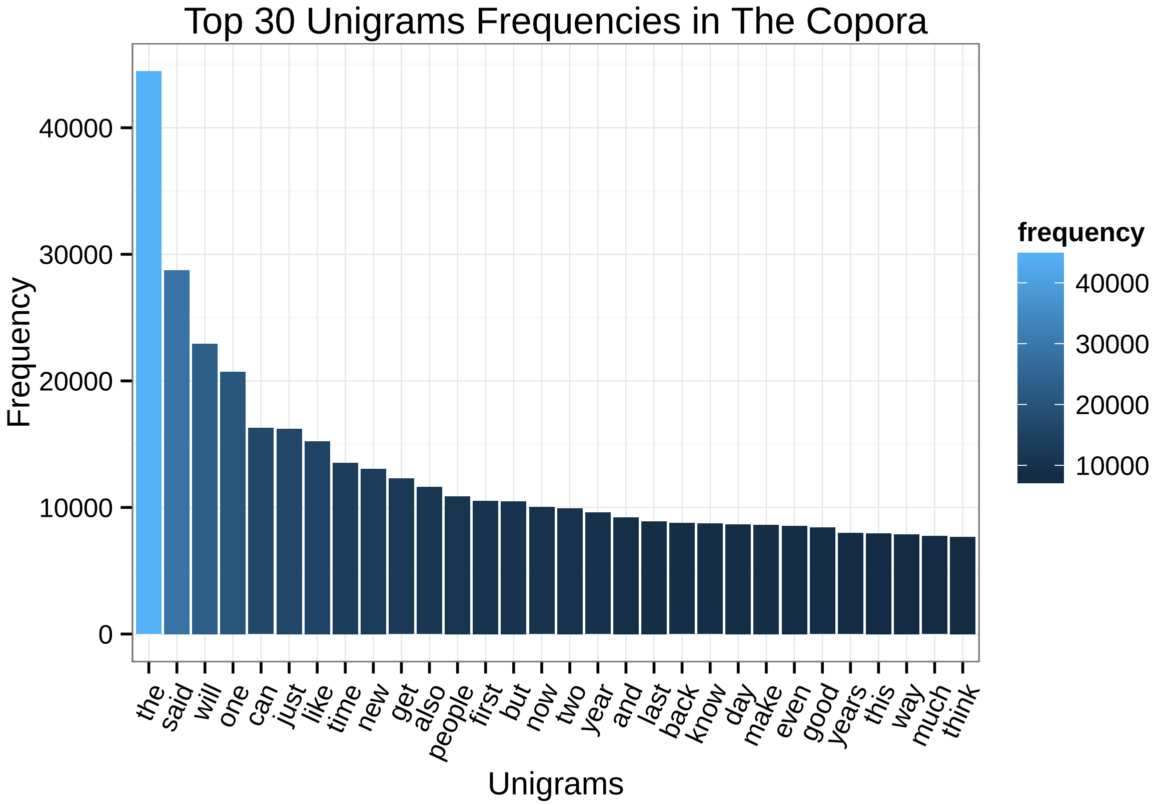
Top 30 Unigrams Frequencies in The Copora
Frequency
40000
30000
20000
10000
0
the
said
will
one
can
just
like
time
new
get
also
people
first
but
now
two
year
and
last
back
know
day
make
even
good
years
this
way
much
think
Unigrams
frequency
40000
30000
20000
10000

Figure 3:

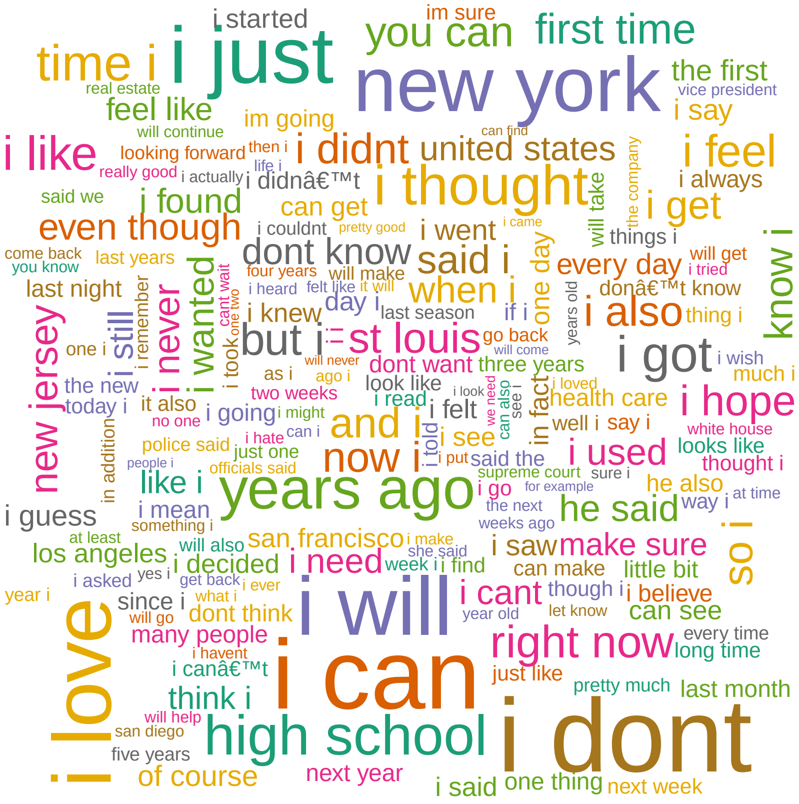
i started
im sure
time i
i just
you can
first time
new york
the first
real estate
vice president
feel like
im going
i say
will continue
can find
i like
looking forward
then i
i didnt
united states
i feel
really good
life i
i actually
i didnâ€™t
i thought
i always
said we
i found
can get
the company
will take
i get
even though
i couldnt
pretty good
i came
i went
things i
come back
you know
last years
dont know
said i
every day
will get
i tried
know i
four years
will make
last night
i remember
i wanted
cant wait
one two
i heard
felt like
it will
when i
one day
donâ€™t know
i never
i still
i knew
day i
last season
if i
years old
i also
thing i
new jersey
one i
i took
but i
st louis
go back
will come
i got
i wish
i i
will never
ago i
dont want
three years
much i
the new
as i
two weeks
look like
i look
we need
can also
see i
i loved
in fact
health care
i hope
today i
it also
i read
no one
i going
i might
and i
i felt
well i
say i
white house
can i
police said
i hate
i see
i told
said the
i used
looks like
just one
i put
people i
officials said
supreme court
sure i
thought i
in addition
like i
years ago
he also
i go
for example
at time
the next
way i
i mean
he said
i guess
something i
weeks ago
so i
at least
san francisco
i make
i saw
make sure
will also
she said
los angeles
i decided
i need
week i
i find
can make
little bit
yes i
get back
i ever
i asked
i cant
though i
i believe
year i
since i
what i
year old
let know
can see
will go
dont think
i will
i love
many people
every time
right now
i havent
long time
i canâ€™t
just like
i can
pretty much
last month
think i
will help
san diego
high school
i dont
five years
of course
next year
i said
one thing
next week

Figure 4:

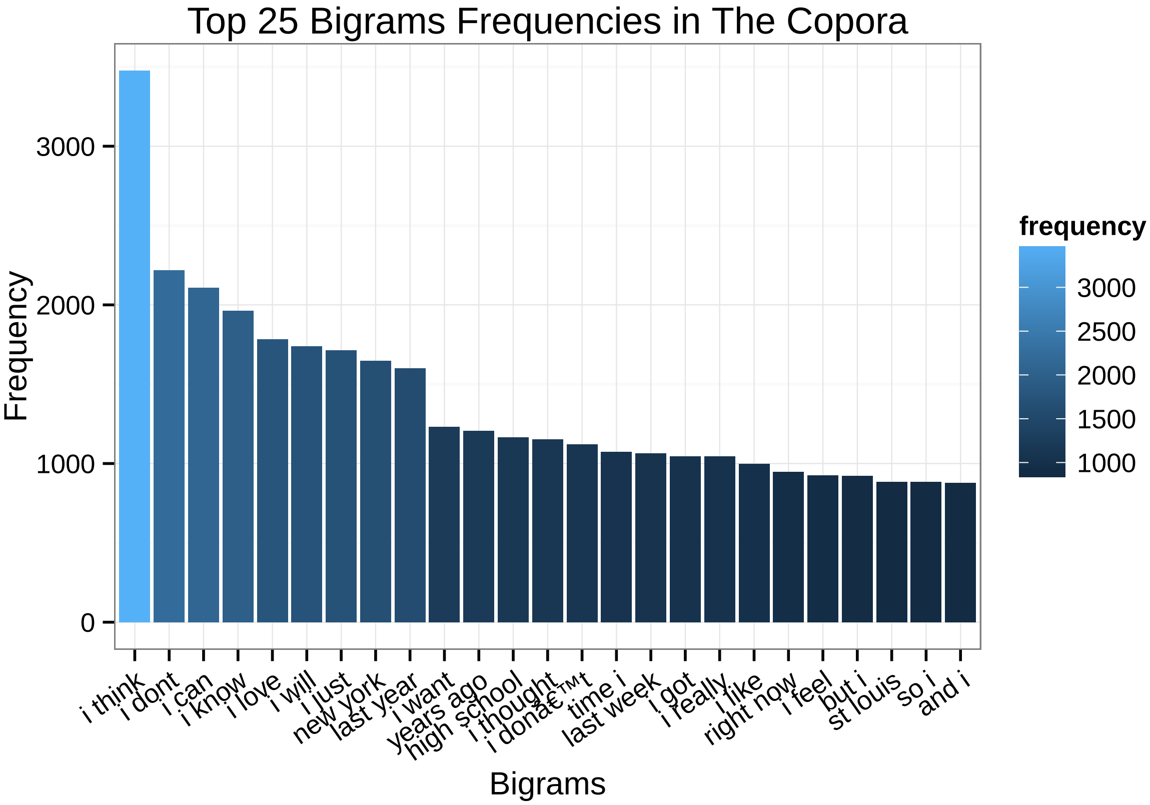
Top 25 Bigrams Frequencies in The Copora
Frequency
3000
2000
1000
0
Bigrams
i think
i dont
i can
i know
i love
i will
i just
new york
last year
i want
years ago
high school
i thought
i donâ€™t
time i
last week
i got
i really
i like
right now
i feel
but i
st louis
so i
and i
frequency
3000
2500
2000
1500
1000

Figure 5:

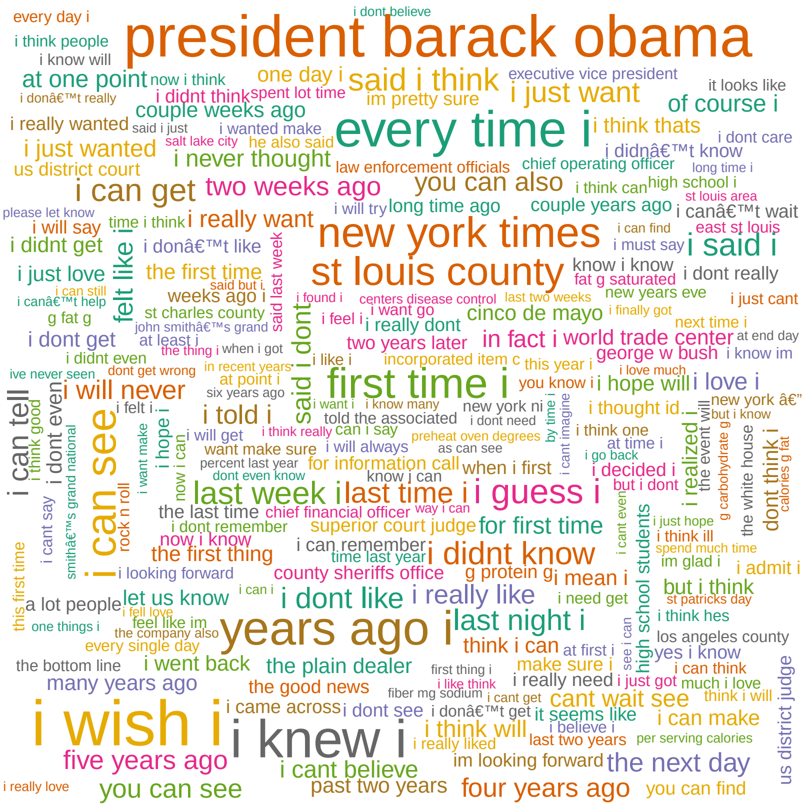
every day i
i dont believe
president barack obama
i think people
i know will
at one point
now i think
one day i
said i think
executive vice president
it looks like
i didnt think
spent lot time
im pretty sure
i just want
of course i
i donâ€™t really
couple weeks ago
i really wanted
said i just
i wanted make
every time i
i think thats
i dont care
i just wanted
salt lake city
he also said
i didnâ€™t know
i never thought
us district court
law enforcement officials
chief operating officer
long time i
i can get
two weeks ago
you can also
i think can
high school i
st louis area
please let know
i will try
long time ago
couple years ago
i canâ€™t wait
time i think
i really want
i will say
new york times
i can find
east st louis
i didnt get
i donâ€™t like
i must say
i said i
i just love
felt like i
the first time
st louis county
know i know
i dont really
said last week
i can still
said but i
fat g saturated
i canâ€™t help
weeks ago i
i found i
centers disease control
last two weeks
new years eve
i just cant
g fat g
st charles county
i want go
cinco de mayo
i finally got
i feel i
i really dont
next time i
john smithâ€™s grand
i dont get
at least i
two years later
in fact i
world trade center
at end day
the thing i
when i got
said i dont
george w bush
i know im
i didnt even
i like i
incorporated item c
this year i
ive never seen
dont get wrong
in recent years
i love much
at point i
first time i
you know i
i hope will
i love i
i can tell
i will never
six years ago
new york âe”
i dont even
i felt i
i want i
i know many
new york ni
i thought id
i think good
i told i
told the associated
i dont need
by time i
but i know
i can see
i want make
i hope i
can i say
i think one
i will get
i think really
preheat oven degrees
i cant imagine
the event will
smithâ€™s grand national
now i can
want make sure
i will always
as can see
at time i
i realized i
g carbohydrate g
the white house
percent last year
for information call
i go back
dont think i
dont even know
know i can
when i first
i decided i
calories g fat
rock n roll
last week i
last time i
i guess i
but i dont
i cant say
the last time
chief financial officer
way i can
i cant even
i dont remember
superior court judge
for first time
i just hope
now i know
i think ill
i can remember
high school students
this first time
the first thing
time last year
i didnt know
spend much time
im glad i
i looking forward
county sheriffs office
g protein g
i admit i
i mean i
but i think
a lot people
let us know
i can i
i dont like
i really like
i need get
st patricks day
i fell love
i think hes
feel like im
one things i
years ago i
last night i
the company also
see i can
los angeles county
every single day
think i can
at first i
yes i know
the bottom line
i went back
the plain dealer
first thing i
make sure i
i can think
many years ago
i really need
i just got
much i love
the good news
i like think
fiber mg sodium
i cant get
cant wait see
think i will
i wish i
i came across
i dont see
i donâ€™t get
it seems like
i can make
i knew i
i think will
i believe i
us district judge
i really liked
last two years
per serving calories
five years ago
im looking forward
the next day
i cant believe
i really love
you can see
past two years
four years ago
you can find

Figure 6:

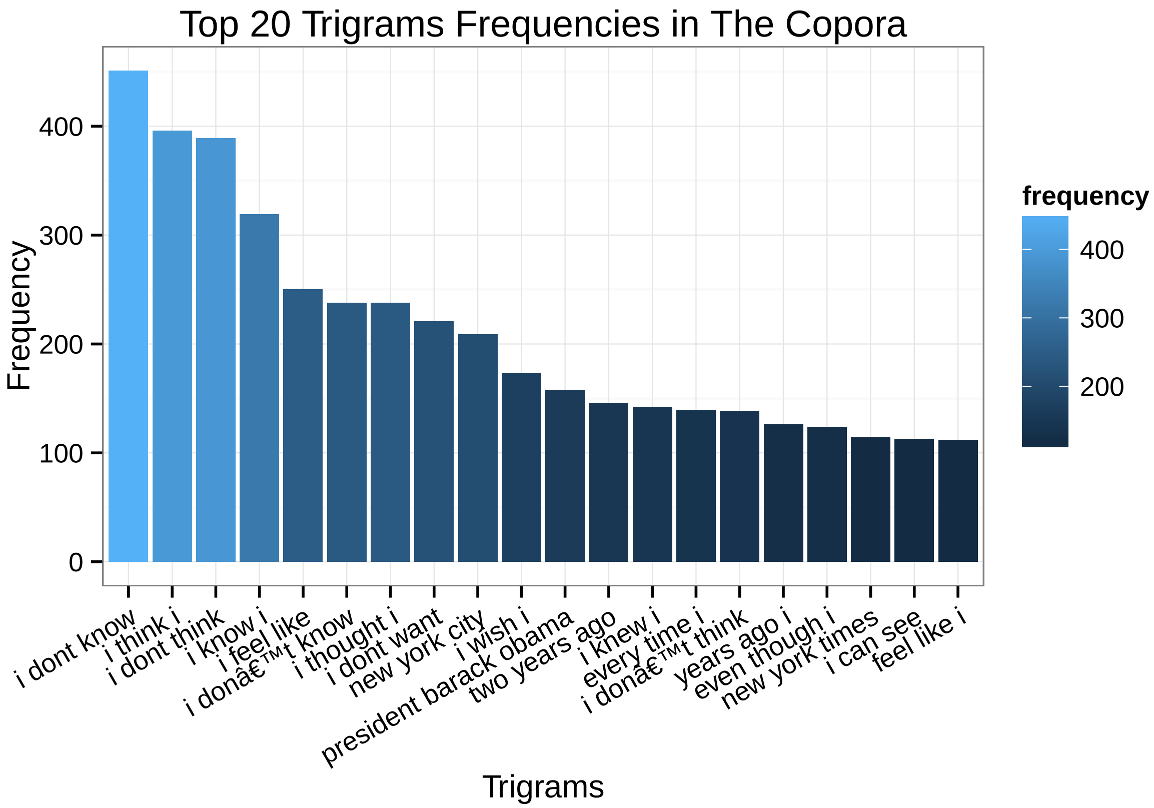
Top 20 Trigrams Frequencies in The Copora
Frequency
400
300
200
100
0
i dont know
i think i
i dont think
i know i
i feel like
i donâ€™t know
i thought i
i dont want
new york city
i wish i
president barack obama
two years ago
i knew i
every time i
i donâ€™t think
years ago i
even though i
new york times
i can see
feel like i
Trigrams
frequency
400
300
200

Figure 7:

2.7.4 Most Frequent Quadgrams (4 words)

2.7.4.1 Quadgrams Word Cloud The word cloud shows the most common quadgrams larger in scale.

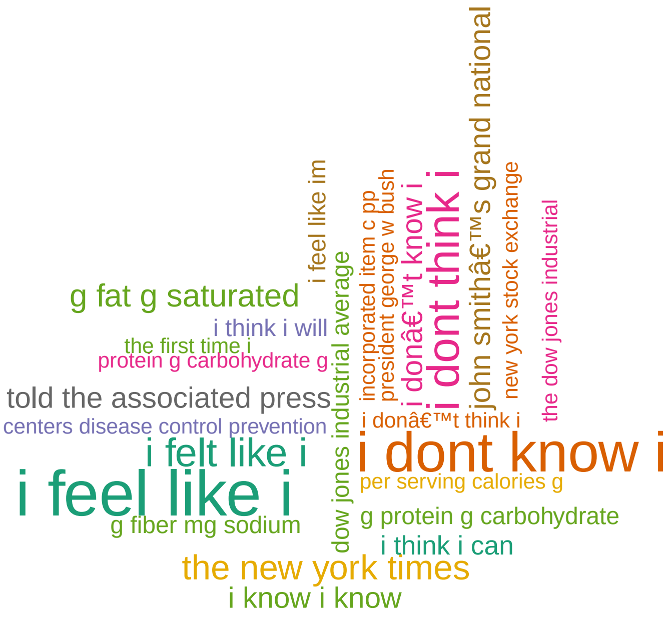

Figure 8:

2.7.4.2 Quadgrams Histogram The histogram shows the different frequencies of the top quadgrams.

3 Shiny App

3.1 Text Prediction Strategy

- Preprocessing a specific language corpora using the HC copora or augmenting it with other corpora.
- Building ngram models from 1 and upto maybe 5-7 words depending on the richness of the corpora.
- Building a table for the frequencies of each different ngram model.
- Merging the different tables into one consolidated frequency table for all the ngram models.
- Given a text string of n words, the prediction algorithm will match these words in any of the ngram tokens in the frequencies table, starting with the n+1 gram tokens, if it does not match any n+1 gram token, the last n-1 words of the input string is matched in the n gram tokens this time and the process

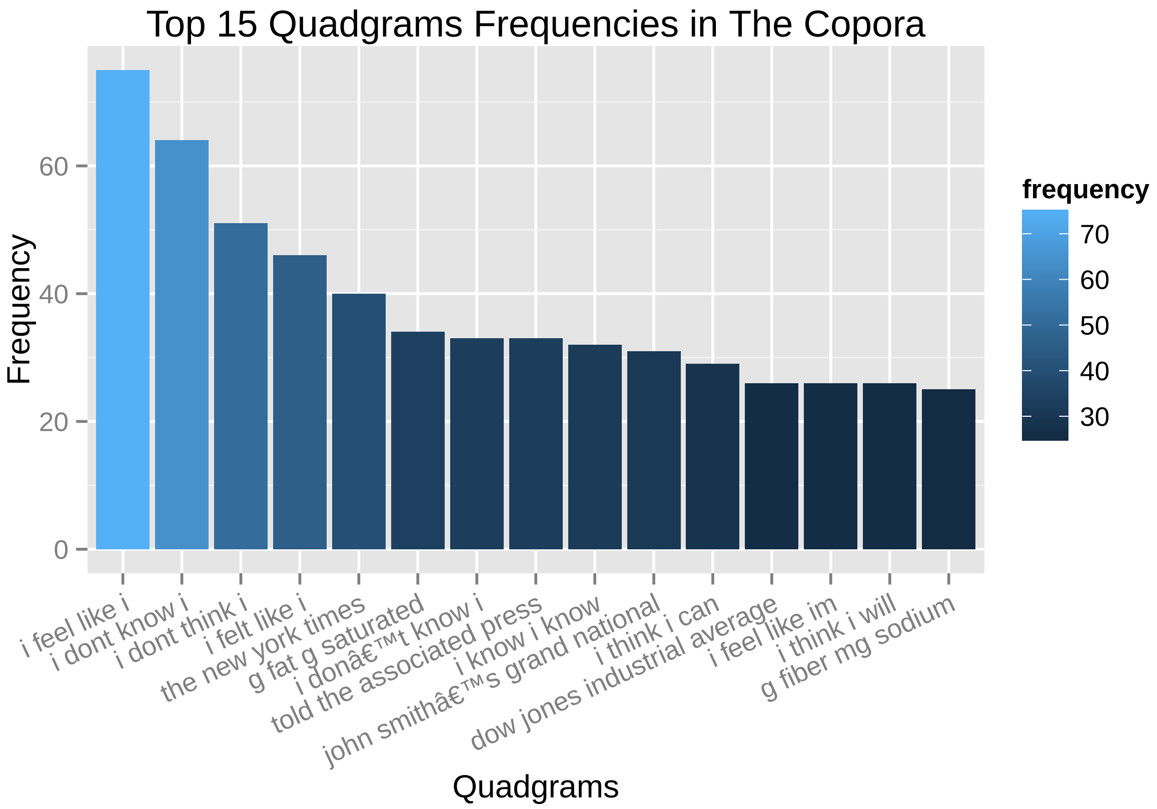
Top 15 Quadgrams Frequencies in The Copora
Frequency
0
20
40
60
frequency
70
60
50
40
30
i feel like i
i dont know i
i dont think i
i felt like i
the new york times
g fat g saturated
i donâ€™t know i
told the associated press
i know i know
john smithâ€™s grand national
i think i can
dow jones industrial average
i feel like im
i think i will
g fiber mg sodium
Quadgrams

Figure 9:

continues until a match is found or the last word, i.e. 1 word of the input text is matched in one of unigram tokens.
- If input is matched, the last word of the n+1 gram tokens with the highest frequencies are recommended for the user to complete his/her input text.

3.2 Shiny App Plan

- Deciding on the optional cleaning steps to include in the final prediction model. For example, will the final model include stopwords and profanity words?
- Looking for augmentary english corpora that can be used to enhance the prediction to be built.
- Completing the n-gram model, I have already built the unigram, bigram, trigram, and quadgram models. However, I might be considering ngram models based on 5-7 words as well if the dataset provided is rich enough to enable me to extract these models.
- Deisgning and prototyping the user interface for the shiny app.
- Implementing the Shiny App.
- Documenting the App.
- Developing a short pitch for the Shiny app.

4 Conclusion

This document reports the main data features of the HC Copora downloaded from http://www.corpora.heliohost.org/. For each text file in the English copora, the report shows the size in addition to other summary statistics such as, wordcount, number of lines, and the mean number of words per line. After building the unigram, bigram, trigram, and quadgram models, the report shows the most frequent n-gram words/tokens along with their frequency in the English Corpora. Finally, the report elaborates the text prediction strategy and the plan for the development of the Shiny App, the final product.