Big Data: Data Wrangling Boot Camp
Python Sentiment Analysis

Chuck Cartledge, PhD

17 September 2016

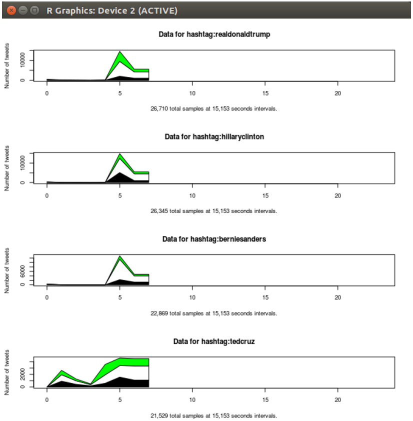
Table of contents I

What are we going to cover?

- Look to the future
- Talk briefly about sentiment analysis
- Address the polyglot of computer languages
- Talk about our sentiment analysis system
- Data wrangle tweets using Python

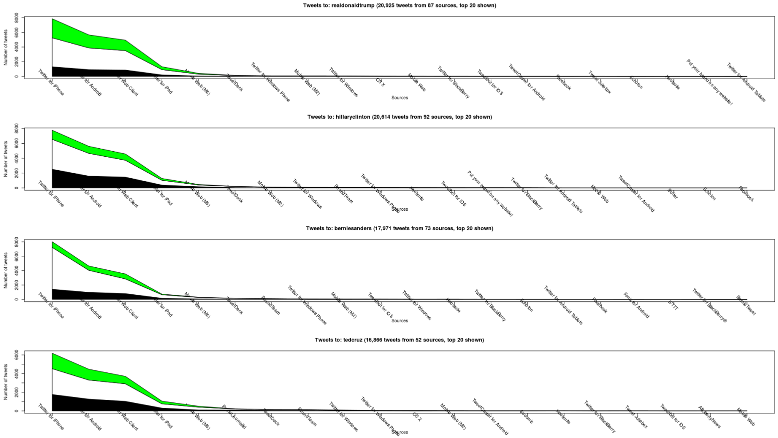
Things that we will be doing.

1. Data wrangle tweets using python
2. Conduct sentiment analysis on tweets
3. Look at the sentiments in different ways

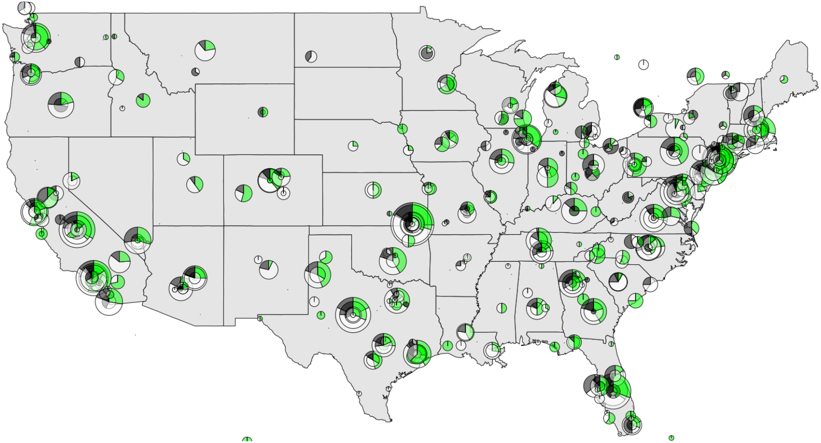


Sentiments over time.

Sentiment by sending device



Sentiment by geographic location



A working definition

> *“Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.”*
>
> *W. Staff [3]*

More formal definitions

> *"The field of opinion mining and sentiment analysis is well-suited to various types of intelligence applications. Indeed, business intelligence seems to be one of the main factors behind corporate interest in the field."*
>
> *Pang and Lee [2]*

> *"Sentiment analysis, also called opinion mining, is the field of study that analyzes peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes."*
>
> *Liu [1]*

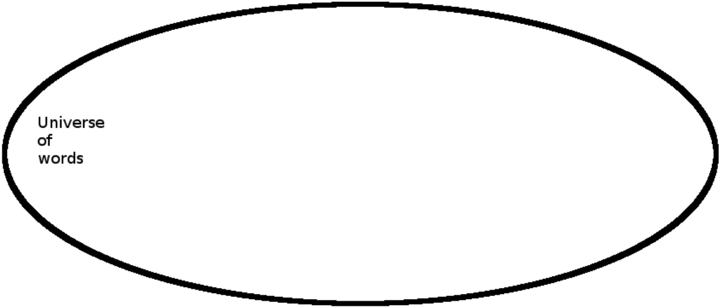
Our approach to sentiment analysis

We will:

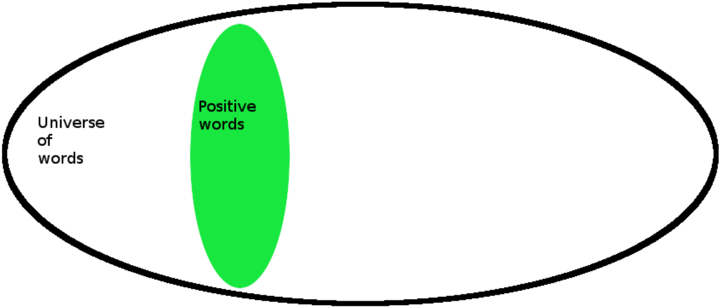
1. Search the “twitterverse” for tweets using specific hashtags
2. Tokenize each tweet
3. Data wrangle each token
4. Remove all stop words from the tokens
5. Count number of positive and negative tokens
6. Compute the positive, negative, or neutral sentiment for the tokens
7. Display the results

Our approach is language agnostic.

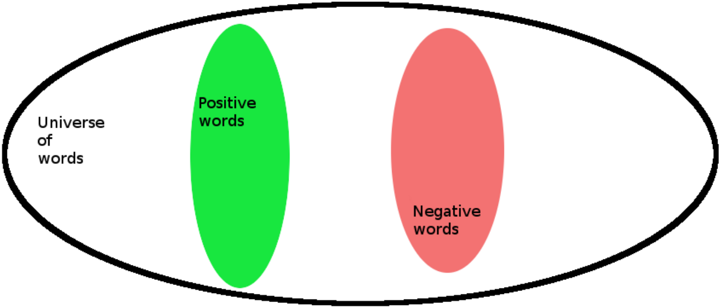
A “Universe” of words



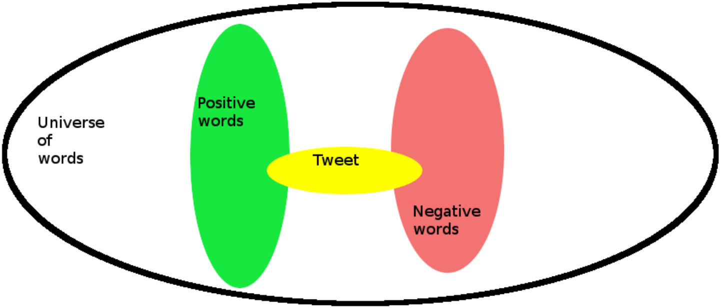
Some words are “Positive”



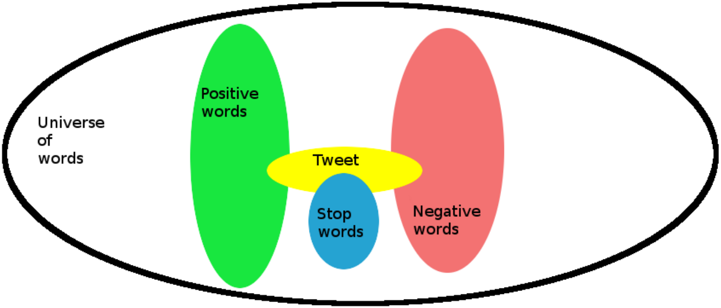
Some words are “Negative”



A tweet will/may have positive and negative words



Some words we don't care about



Mechanically this is what we are doing

The steps are:

1. Break the tweet into tokens
2. Remove stop words from the tokens
3. Compute the percentage of remaining tweet token that are positive
4. Compute the percentage of remaining tweet tokens that are negative
5. Classify the tokens as positive, negative, or neutral

Mathematically this is what we are doing

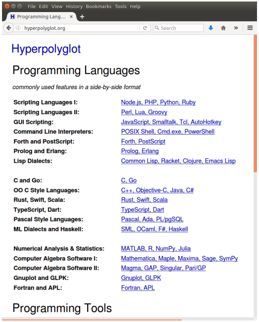
The steps are:

$$\begin{aligned} tokens &= \{\text{words in tweet}\} \\ tokensLessStop &= tokens - stopWords \\ positivePart &= positiveWords \cap tokensLessStop \\ negativePart &= negativeWords \cap tokensLessStop \end{aligned}$$

$$classification = \begin{cases} positive, & \begin{array}{l} \text{if} positiveThreshold \leq \frac{positivePart}{tokensLessStop} \\ \text{AND} \\ \frac{negativePart}{tokensLessStop} < negativeThreshold \end{array} \\ negative, & \begin{array}{l} \text{if} negativeThreshold \leq \frac{negativePart}{tokensLessStop} \\ \text{AND} \\ \frac{positivePart}{tokensLessStop} < positiveThreshold \end{array} \\ neutral, & \text{otherwise} \end{cases}$$

Comparing a known language with Python and R

A specific language was not a requirement for this boot camp. Python and R used in boot-camp. Too many languages to compare to Python and R.



Hyperpolyglot

Programming Languages

commonly used features in a side-by-side format

Scripting Languages I:	Node.js, PHP, Python, Ruby
Scripting Languages II:	Perl, Lua, Groovy
GUI Scripting:	JavaScript, Smalltalk, Tcl, AutoHotkey
Command Line Interpreters:	POSIX Shell, Cmd.exe, PowerShell
Forth and PostScript:	Forth, PostScript
Prolog and Erlang:	Prolog, Erlang
Lisp Dialects:	Common Lisp, Racket, Clojure, Emacs Lisp
C and Go:	C, Go
OO C Style Languages:	C++, Objective-C, Java, C#
Rust, Swift, Scala:	Rust, Swift, Scala
TypeScript, Dart:	TypeScript, Dart
Pascal Style Languages:	Pascal, Ada, PL/pgSQL
ML Dialects and Haskell:	SML, OCaml, F#, Haskell
Numerical Analysis & Statistics:	MATLAB, R, NumPy, Julia
Computer Algebra Software I:	Mathematica, Maple, Maxima, Sage, SymPy
Computer Algebra Software II:	Magma, GAP, Singular, Pari/GP
Gnuplot and GLPK:	Gnuplot, GLPK
Fortran and APL:	Fortran, APL

Programming Tools

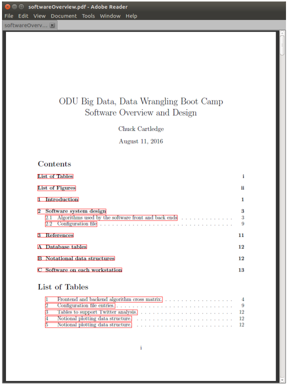
`http://hyperpolyglot.org/`

Use Hyperploygloт.org as a cross reference.

A software design document

The document contains:

1. Overall system design
2. Algorithms used through out the system
3. Details about the configuration file
4. Details about the database tables


softwareOverview.pdf - Adobe Reader
File Edit View Document Tools Window Help
softwareOverv...

ODU Big Data, Data Wrangling Boot Camp
Software Overview and Design

Chuck Cartledge

August 11, 2016

Contents

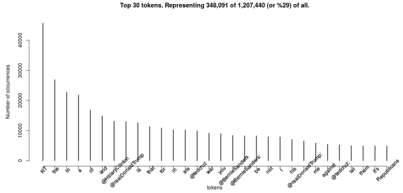
List of Tables

i

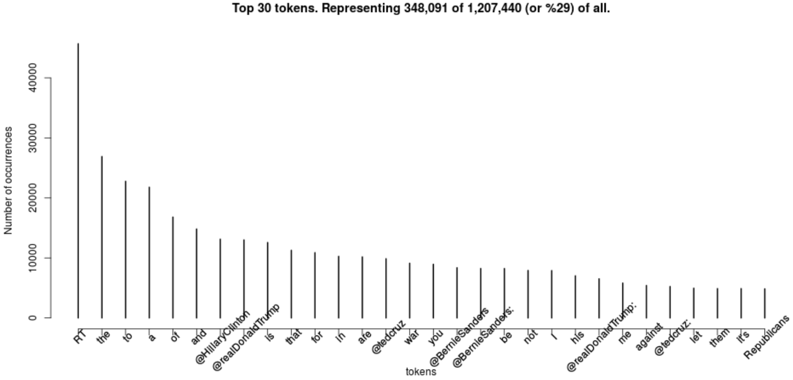
The file is attached.

What will happen to the tweets when we do nothing.

A histogram of how often a token occurs.

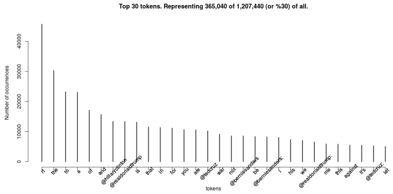


Same image.

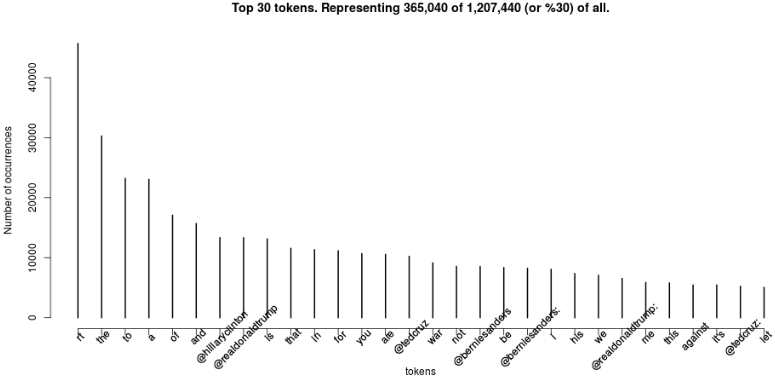

Top 30 tokens. Representing 348,091 of 1,207,440 (or %29) of all.
Number of occurrences
0
10000
20000
30000
40000
tokens
RT
the
to
a
of
and
@HillaryClinton
@realDonaldTrump
is
that
for
in
are
@tedcruz
war
you
@BernieSanders
@BernieSanders:
be
not
I
his
@realDonaldTrump:
me
against
@tedcruz:
let
them
it's
Republicans

What will happen to the tweets when we make everything the same case

Changing case is easy, unless they are emojois.



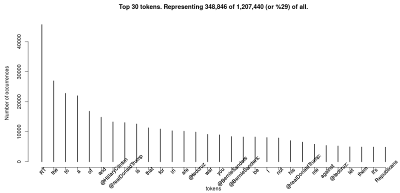
Same image.



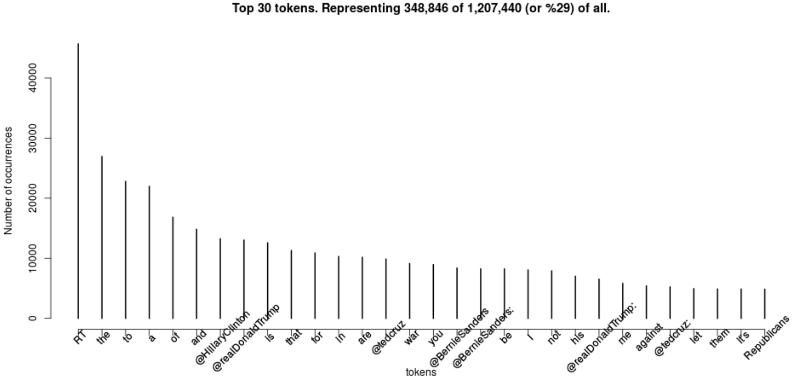
Top 30 tokens. Representing 365,040 of 1,207,440 (or %30) of all.

What will happen to the tweets we remove non-ASCII

We need to talk about ASCII and the rest of the character systems.



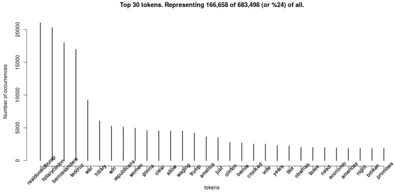
Same image.


Top 30 tokens. Representing 348,846 of 1,207,440 (or %29) of all.
Number of occurrences
0
10000
20000
30000
40000
RT
the
to
a
of
and
@HillaryClinton
@realDonaldTrump
is
that
for
in
are
@tedcruz
war
you
@BernieSanders
@BernieSanders:
be
I
not
his
@realDonaldTrump:
me
against
@tedcruz:
let
them
it's
Republicans
tokens

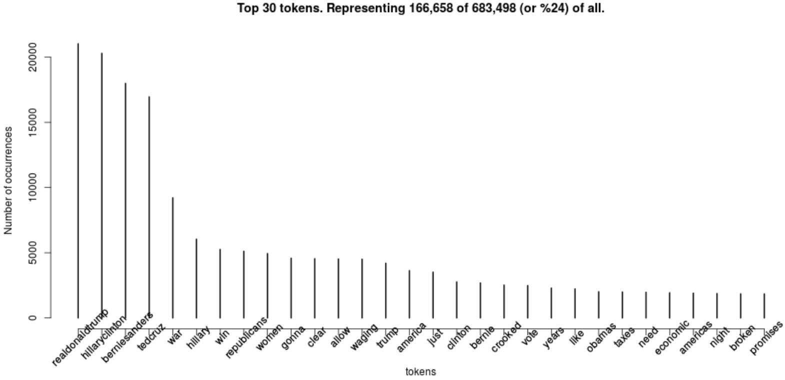
What will happen to the tweets when we remove “stop words”

Stop words are words/tokens that have no use in whatever we are doing. Stop words are domain specific.

And we change case, remove non-ASCII, remove punctuation, etc.



Same image.

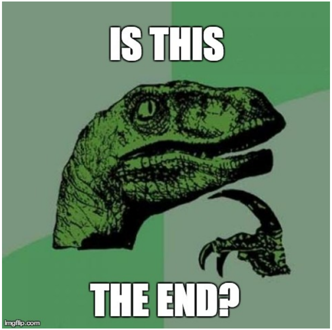

Top 30 tokens. Representing 166,658 of 683,498 (or %24) of all.
Number of occurrences
0
5000
10000
15000
20000
tokens
realdonaldtrump
hillaryclinton
berniesanders
tedcruz
war
hillary
win
republicans
women
gonna
clear
allow
waging
trump
america
just
clinton
bernie
crooked
vote
years
like
obamas
taxes
need
economic
americas
night
broken
promises

Q & A time.

“’The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,’ said Deep Thought, with infinite majesty and calm.”
Douglas Adams, The Hitchhiker’s Guide to the Galaxy

What have we covered?

- A preview of today's activities
- An overview of the sublime IDE



Next: Hands on analysing tweets with Python.

References I

[1] Bing Liu, Sentiment analysis and opinion mining, 2012.

[2] Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, 2008.

[3] Wikipedia Staff, Sentiment analysis, https://en.wikipedia.org/wiki/Sentiment_analysis, 2016.

Files of interest

1. softwareOverview.pdf