

HOW TO GET STARTED WITH TEXT ANALYSIS IN PYTHON

(And analyzing Donald
Trump's tweets)

2010-12-31

2011-12-31

2012-12-31

2013-12-31

2014-12-31

2015-12-31

2016-12-31

created per month

Please Options are my own
keep I'm not an expert on anything
in mind: It's a Python conference

DISCLAIMERS

OVERVIEW

01

Text Analysis
with TextBlob

02

Text
Classification
Basics

03

Machine
Learning with
Naïve Bayes

INTRO TO TEXTBLOB

A word cloud visualization where the size and color of words represent their frequency and context in political discourse. The most prominent words are 'great' (large, teal), 'democrat' (medium, teal), and 'republican' (large, green). Other significant words include 'president' (green), 'big news' (blue), 'people' (green), 'fake' (purple), 'country' (blue), 'tax' (blue), 'military' (blue), 'fbi' (teal), 'crime' (teal), 'hillary' (blue), 'obama' (green), 'russia' (purple), 'north' (orange), 'even' (purple), 'bad' (red), 'much' (orange), 'u' (orange), 'get' (purple), 'job' (blue), 'would' (purple), 'state' (purple), 'hunt' (purple), 'win' (purple), 'trump' (purple), 'make' (purple), 'house' (purple), 'thank' (purple), 'want' (purple), 'trade' (purple), and 'like' (purple).

TEXTBLOB BASICS

Term Counts

Parts of
Speech

Tokenization

Sentiment
(Subjectivity
/ Polarity)

Language
Detection /
Translation

Spelling

DEMO



DOWNLOAD CERTIFICATE PROBLEM

```
[nltk_data] Error loading brown: <urlopen error [SSL:  
[nltk_data]      CERTIFICATE_VERIFY_FAILED] certificate verify failed:  
[nltk_data]      unable to get local issuer certificate (_ssl.c:1045)>  
[nltk_data] Error loading punkt: <urlopen error [SSL:  
[nltk_data]      CERTIFICATE_VERIFY_FAILED] certificate verify failed:  
[nltk_data]      unable to get local issuer certificate (_ssl.c:1045)>  
[nltk_data] Error loading wordnet: <urlopen error [SSL:  
[nltk_data]      CERTIFICATE_VERIFY_FAILED] certificate verify failed:  
[nltk_data]      unable to get local issuer certificate (_ssl.c:1045)>  
[nltk_data] Error loading averaged_perceptron_tagger: <urlopen error  
[nltk_data]      [SSL: CERTIFICATE_VERIFY_FAILED] certificate verify  
[nltk_data]      failed: unable to get local issuer certificate  
[nltk_data]      (_ssl.c:1045)>  
[nltk_data] Error loading conll2000: <urlopen error [SSL:  
[nltk_data]      CERTIFICATE_VERIFY_FAILED] certificate verify failed:  
[nltk_data]      unable to get local issuer certificate (_ssl.c:1045)>  
[nltk_data] Error loading movie_reviews: <urlopen error [SSL:  
[nltk_data]      CERTIFICATE_VERIFY_FAILED] certificate verify failed:  
[nltk_data]      unable to get local issuer certificate (_ssl.c:1045)>  
Finished.
```

[/Applications/Python 3.X/Install Certificates.command](#)

MACHINE LEARNING



This Photo by Unknown Author is licensed under CC BY-NC-ND

MACHINE LEARNING

01

Problem

02

Text

03

Tools



Donald J. Trump

@realDonaldTrump

Congratulations to the 2016
#StanleyCup Champions,
Pittsburgh @penguins!

2:08pm · 13 Jun 2016 · Twitter for iPhone



Donald J. Trump

@realDonaldTrump

Wow, Twitter, Google and Facebook
are burying the FBI criminal
investigation of Clinton. Very
dishonest media!

10:26am · 30 Oct 2016 · Twitter for Android

KNOW YOUR PROBLEM |

KNOW YOUR PROBLEM

Donald J. Trump ✅
@realDonaldTrump

Wow! Thank you Louisville,
Kentucky!
#VoteTrump on 3/5/2016! Lets
#MakeAmericaGreatAgain!
facebook.com/DonaldTrump/po...
pic.twitter.com/xFwWRro3l



5:18pm · 1 Mar 2016 · Twitter for iPhone
© Louisville, KY, United States

871 REPLIES 3,787 RETWEETS 11,080 LIKES



Donald J. Trump ✅
@realDonaldTrump

Give the public a break - The FAKE NEWS media is trying to say that
large scale immigration in Sweden
is working out just beautifully. NOT!

9:15am · 20 Feb 2017 · Twitter for Android

48,090 REPLIES 33,143 RETWEETS
133,269 LIKES



David Robinson

*Chief Data Scientist at
DataCamp, works in R and
Python.*

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:



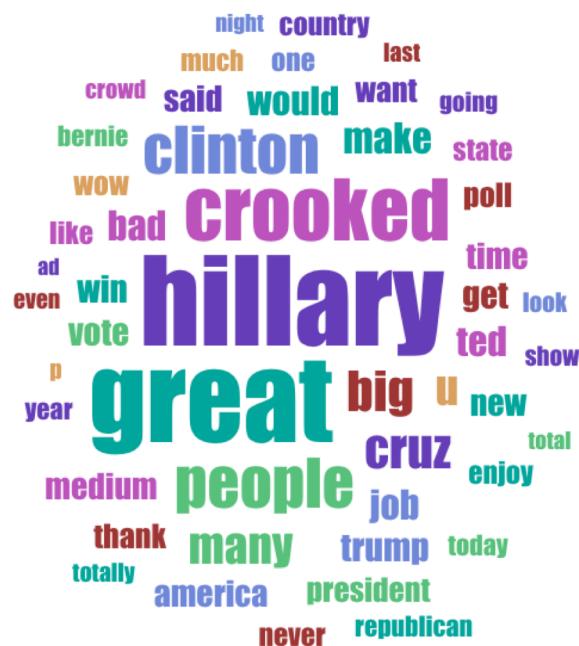
KNOW YOUR PROBLEM |

2016 TWITTER WORD COMPARISON

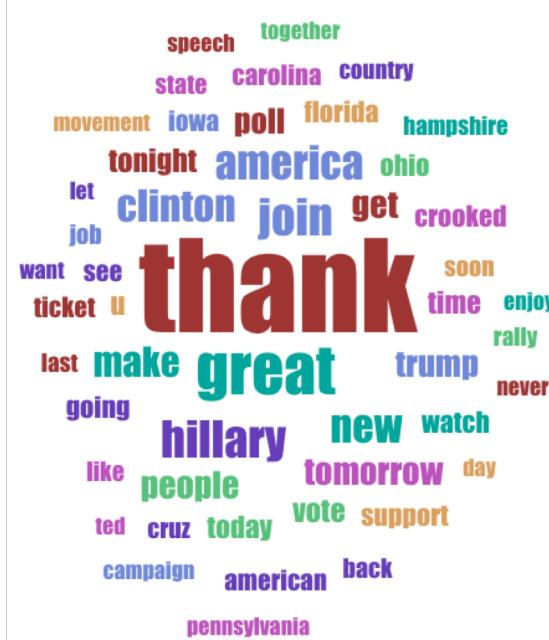
Top Words	Uses
hillary	230
great	223
crooked	156
people	125
clinton	120
big	99
cruz	97
many	90
u	85
make	75

Export: Raw Formatted

1 2 3 4 5 ...



Android



Non-Android

KNOW YOUR PROBLEM



Donald J. Trump ✅
@realDonaldTrump

There was No Collusion with Russia (except by the Democrats). When will this very expensive Witch Hunt Hoax ever end? So bad for our Country. Is the Special Counsel/Justice Department leaking my lawyers letters to the Fake News Media? Should be looking at Dems corruption instead?

1:43pm · 2 Jun 2018 · Twitter for iPhone



Donald J. Trump ✅
@realDonaldTrump

Landing in Las Vegas now for a Make America Great Again Rally supporting [@DeanHeller](#) and [@DannyTarkanian](#). Also doing interview there with [@seanhannity](#) live on [@FoxNews](#). Big crowd, long lines. Will be great! #MAGA

9:27pm · 20 Sep 2018 · Twitter for iPhone



Donald J. Trump ✅
@realDonaldTrump

The Failing New York Times wrote a story that made it seem like the White House Council had TURNED on the President, when in fact it is just the opposite - & the two Fake reporters knew this. This is why the Fake News Media has become the Enemy of the People. So bad for America!

8:06am · 19 Aug 2018 · Twitter for iPhone



Donald J. Trump ✅
@realDonaldTrump

Join me this Saturday in Wheeling, West Virginia at 7pmE! Tickets: donaldjtrump.com/rallies/wv-sep...



4:48pm · 26 Sep 2018 · Twitter for iPhone

RESEARCH AND REFERENCES

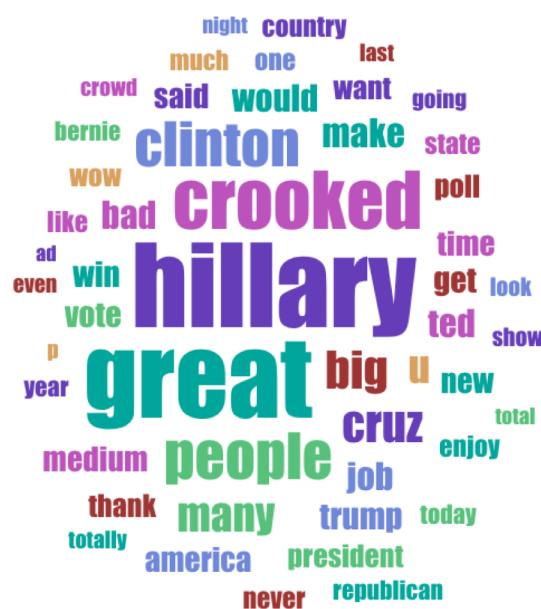
Source	Description
<u>Variance Explained, "Text analysis of Trump's tweets confirms he writes only the (angrier) Android half"</u>	David Robinson's original analysis of Android vs. iPhone tweets from @realdonaldtrump
<u>The New York Times Magazine, "The Man Behind the President's Tweets"</u>	Profile of Dan Scavino, Donald Trump's Director of Social Media, with discussion of others being able to tweet as @realdonaldtrump, by Robert Draper
<u>The New York Times, "A Homebody Finds the Ultimate Home Office"</u>	Analysis of Donald Trump's typical routine when arriving in the White House, including use of an Android phone, by Maggie Haberman
<u>Android Central, "Which Android Phone does Donald Trump use?"</u>	Analysis of Donald Trump's likely cell phone model based on public photos, by Alex Dobie
<u>Boston Globe, "Inside the Trump Tweet Machine"</u>	Background on White House staff members composing tweets to mimic Donald Trump's style, by Annie Linskey

Most Common Android Words

Top Words	Uses
hillary	230
great	223
crooked	156
people	125
clinton	120
big	99
cruz	97
many	90
u	85
make	75

Export: Raw Formatted

1 2 3 4 5 ...



Top iOS/Web Words

Top Words	Uses
thank	530
great	244
hillary	167
join	163
america	143
new	136
clinton	132
make	119
get	100
people	99

Export: Raw Formatted

1 2 3 4 5 ...

```
1 [  
2 {  
3     "text": "CNN is the worst - fortunately they have bad ratings because everyone  
4     "label": "Staff"  
5 },  
6 {  
7     "text": "Thank you to my great supporters in Wisconsin. I heard that the crowd  
8     "label": "Trump"  
9 },  
10 {  
11     "text": "Here is my statement. https://t.co/WAZiGoQqMQ",  
12     "label": "Staff"  
13 },  
14 {  
15     "text": "Mike Pence won big. We should all be proud of Mike!",  
16     "label": "Trump"  
17 },  
18 {  
19     "text": "Clinton's Top Aides Were Mired In Conflict Of Interest At The State  
20     "label": "Staff"  
21 },  
22 {  
23     "text": "Both are looking good! Now we begin!"  
24 }
```

GET SOME KNOWN TEXT

I

NAÏVE

Everything is
independent

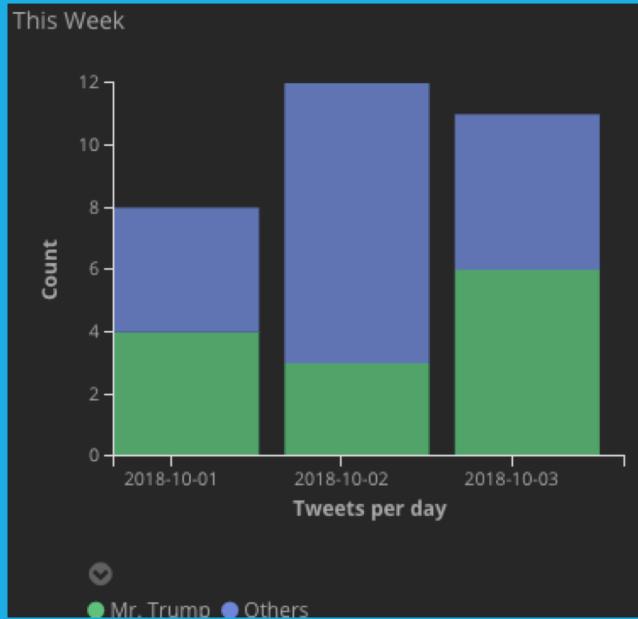
BAYES

Inputs (history) predict
outputs (future)

DEMO (2)



WHAT TO DO NOW?



Donald J. Trump @realDonaldTrump · 30 Sep 2017
...want everything to be done for them when it should be a community effort.
10,000 Federal workers now on Island doing a fantastic job.

31K 15K 68K

Trump Tweet Track @trumptweettrac

Following

Replies to @realDonaldTrump

Shhhhhh... your staff is still sleeping. Our bot gives it a 98% chance you wrote this yourself.
trumptweettrack.com

7:30 AM - 30 Sep 2017

3 Retweets 25 Likes

1 3 25

Trump or Not @trumpOrNotBot

Follow

This tweet was sent via Twitter for iPhone. I compute a 33% chance it was written by Trump himself.

The White House @WhiteHouse
"America will always be a nation of great builders, because in America, we honor work, we honor grit, we honor craftsmanship, we honor the men and women who turn dreams into a reality with their own two beautiful hands." — President @realDonaldTrump

1:39 2:24 PM - 2 Oct 2018

#aidetweet BOT @aidetweetbot

Follow

White House aides, like any other machine, are either a benefit or a hazard. If they're a benefit it's not my problem. #aidetweet Score:-85 Code:MLTHn Time:1826 #MAGA

Donald J. Trump @realDonaldTrump
GOD BLESS THE U.S.A! #MAGA

6:28 PM - 2 Oct 2018

Donald J. Trump @realDonaldTrump · Sep 29
NBC News incorrectly reported (as usual) that I was limiting the FBI investigation of Judge Kavanaugh, and witnesses, only to certain people. Actually, I want them to interview whoever they deem appropriate, at their discretion. Please correct your reporting!

29K 46K 159K

Trump Weather Report @realDonaldTrumpweather

Follow

Replying to @realDonaldTrump

Current: Partly cloudy with scattered tweets; 60% chance Donald Trump wrote this himself.

This is tweet number 277 mentioning NBC from Donald Trump -- 43 since inauguration.

7:50 PM - 29 Sep 2018

31 Retweets 245 Likes

27 31 245

Please
keep
in
mind:

No magic.

TextBlob isn't perfect

Slow
Overfit
Best practices

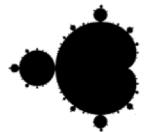
Naïve Bayes isn't perfect

It's all just probabilities

FINAL DISCLAIMERS

TextBlob Documentation

<https://textblob.readthedocs.io/>



TextBlob

Star 5,542

TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, translation, and more.

TextBlob: Simplified Text Processing

Release v0.15.1. [\(Changelog\)](#)

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

```
from textblob import TextBlob

text = """
The titular threat of The Blob has always struck me as the ultimate movie
monster: an insatiably hungry, amoeba-like mass able to penetrate
virtually any safeguard, capable of--as a doomed doctor chillingly
describes it—"assimilating flesh on contact.
Snide comparisons to gelatin be damned, it's a concept with the most
devastating of potential consequences, not unlike the grey goo scenario
```

Fork me on GitHub

NLTK Documentation

<https://www.nltk.org/>

NLTK 3.3 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

[Natural Language Processing with Python](#) provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_2ed.)

TABLE OF CONTENTS

[NLTK News](#)

[Installing NLTK](#)

[Installing NLTK Data](#)

[Contribute to NLTK](#)

[FAQ](#)

[Wiki](#)

[API](#)

[HOWTO](#)

SEARCH

Go

RESOURCES

POLITICAL TEXT ANALYSIS

Source	Description
@trumpornotbot	Twitter bot created by Andrew McGill of the Atlantic, using Naïve Bayes to determine the likely author of Donald Trump's tweets
TrumpTweetTrack.com @trumptweettrac	Full archive of Donald Trump's tweets analyzed with likely authors and various charts and graphs, including real-time analysis as tweets are posted using gradient boosting
TrumpTwitterArchive.com @realtrumptweet	Full archive of Donald Trump's tweets, updated regularly, in a fully searchable and mobile-friendly web page (along with several other politicians)
State of the Union	Online archive of State of the Union speeches with various text analysis and graphs
Factba.se @factbasefeed	Analysis of tweets, speeches, videos, depositions, and a wide variety of other topics regarding Donald Trump as well as Congress

You can call me Steve.

Sometimes I build [open source software](#).

Sometimes I scribble down [things I've learned](#).

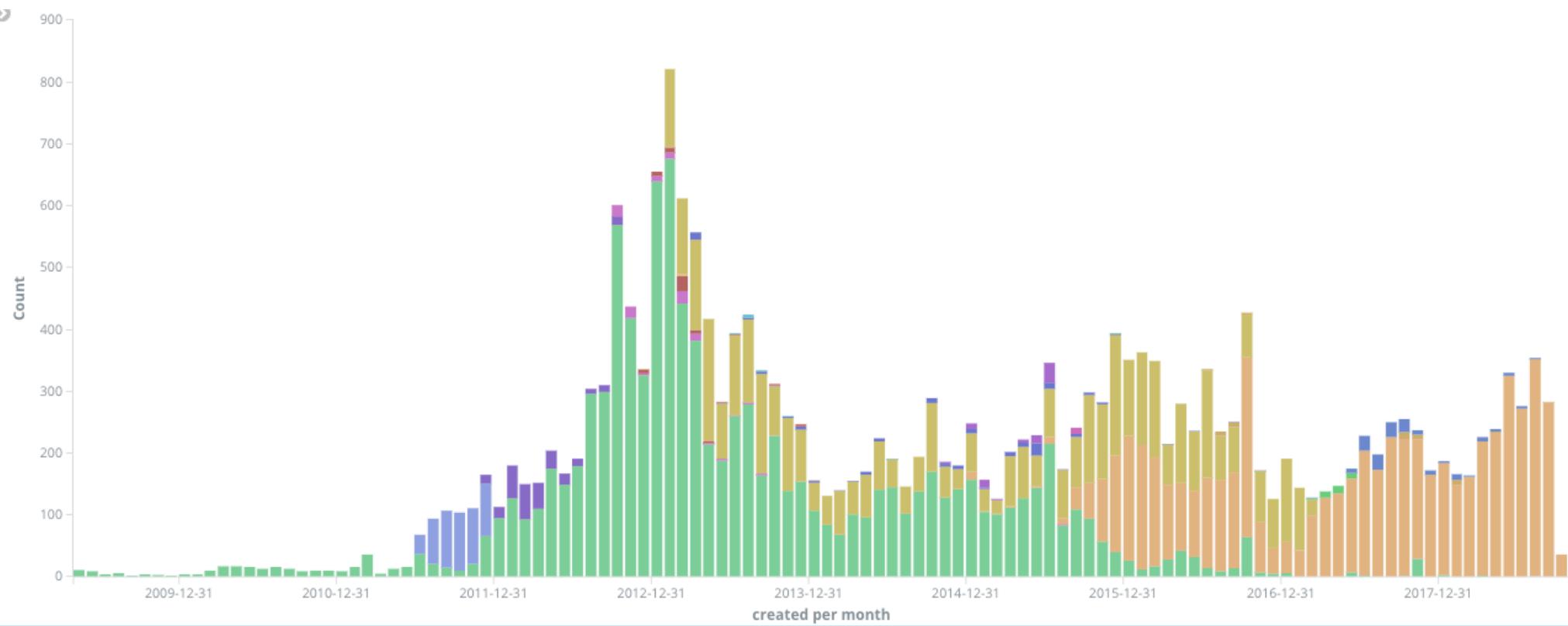
I'm currently living in Charlottesville, VA working for the [Center for Open Science](#).

I'm an amateur at everything in a constant struggle to [turn pro](#).

The easiest way to contact me is by [email](#).

THANKS TO STEVE LORIA

Maintainer of TextBlob and
several other projects
[@sloria1](#) | [stevenloria.com](#)



THANK YOU!



@daveklee