HW03 — STAT/CS 287  
NAME: Michael Arnold  
DATE: 2018-10-18

## P1.1

I retrieved the data by downloading it from blackboard, and placing it in a directory called data in the HW03 directory. Since the data is stored locally, it only needs to be downloaded once.

## P1.2

When loading the data in the data\_loader function, I encountered 10415 missing leading bracket errors, and 7585 missing trailing brackets. There were 76575 total “twitter records.” However after adding to a dictionary where the key was the unique id, only 74087 unique tweets remained.

## P1.3

I used a dictionary to store the data, where each key was the tweet id and each value was a list where the first element is the created\_at timestamp and the second element is the text of the tweet.

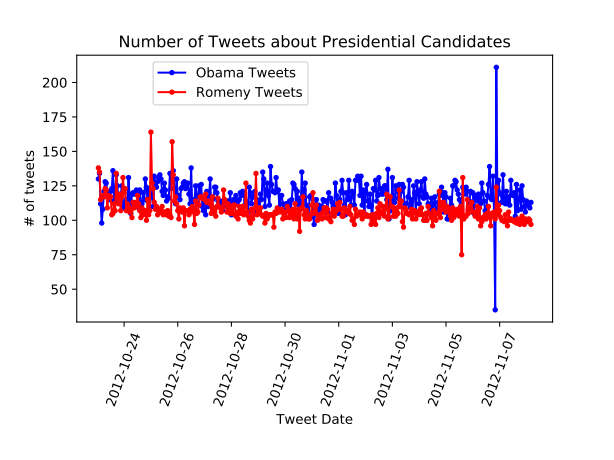
## P1.4

All of the operations to split the tweets into Obama and Romney lists were applied in the tweet\_categorizer function beginning on line 46 in the main homework file HW03\_mvarnold.py. I used lists of keywords to identify which tweets referred to each candidate, with the common names for Obama being ‘obama’, ‘barack’, and ‘barry,’ while common names for Romney were “mitt” and “romney.”

I performed a sanity check by making sure all the tweets were assigned to at list one category, and found there were no uncategorized tweets. Then, in the interest of completeness, I added a few derogatory nicknames from the internet for each candidate. For Obama there were "obummer", "bammy", and "odummy", and for Romney there was “robme” and “romnuts.” However, the addition of these key words only added 9 entries to the sum of the two lists.

If any one of these key words was found to be in a tweet’s text, the tweet date and text were added to that Candidate’s tweet list. After iterating over all tweets and all key words, the program returns the two lists.

## P2

  
Illustration 1: Time Series plot showing the number of tweets in each hour written about candidates Obama and Romney.

## P3.1

One potential issue is that this coefficient would rank a word which turned up twice in each corpus as the same coefficient as a word which turned up 1000 times in each. So one thing we could do to smooth everything out is add something to the denominator, so that things with less entries are pushed towards the middle. But this isn’t really better, and the since we have a lot of text, the top and bottom of our sorted list are really used a lot more in one corpus than in another.

## P3.2

Here are the formated lines

@MichelleObama: 0.990698 problems. -0.987578

let’s 0.990148 reflect -0.985714

sun 0.988304 THROUGH -0.985075

Confirm 0.982759 remembered -0.984496

officials 0.982301 freed -0.983607

President’s 0.981651 @reppaulryan -0.978947

together." 0.980952 REDCROSS -0.977778

instructions 0.979167 90999 -0.977778

safe. 0.978947 dictators. -0.977011

Jay 0.978947 qualities -0.971631

#progress 0.975610 considering -0.971014

5.4 0.974684 .@PaulRyanVP -0.970149

performing 0.974359 dictated -0.968254

forward: 0.972973 Victim -0.966102

power.” 0.971831 Homeless -0.964912

tix 0.971831 ENDORSEMENT -0.964912

forward." 0.970588 FLASHBACK: -0.964286

weekend: 0.970588 FEMA! -0.957447

#GottaVote 0.970149 movement. -0.957447

fights 0.969231 lives, -0.956522

#ForAll 0.969231 Text -0.956044

expertise 0.969231 listen. -0.954545

"major 0.967742 seize -0.952381

Milwaukee, 0.967213 rally" -0.950000

"four 0.966667 prosperity. -0.950000

We’ve 0.966102 support! -0.949008

reelect 0.966102 "disaster -0.948718

re-election: 0.964912 intended -0.948718

Office, 0.964912 ASAP! -0.947826

FL: 0.964286 "Some -0.945946

succeed 0.962963 pretended -0.944444

already!! 0.960784 @Rick\_Gorka -0.944444

"No 0.960784 Yo!! -0.944444

stretch: 0.960000 fact." -0.942857

chapter 0.960000 Together, -0.940299

try." 0.958333 okay. -0.939394

animator 0.958333 @metaquest: -0.937500

@OFA\_NH: 0.955556 #WTF -0.937500

up." 0.955556 adopt -0.937500

(like 0.955556 FAMILY -0.937500

@OFA\_OH: 0.955224 .@mittromney -0.937220

That’s 0.954887 believer -0.935484

private-sector 0.954545 thinking! -0.935484

stand, 0.954545 expanding -0.935484

location 0.952000 measured -0.933333

Bruce 0.950617 changes. -0.933333

vote." 0.950000 bad? -0.931818

Forward, 0.950000 #RomneyLies -0.931034

Orleans 0.950000 @PaulRyanVP. -0.931034

begun 0.947137 Rocks -0.929412

from, 0.945946 track. -0.929293

BRILLIANT: 0.945946 returns? -0.928571

believe. 0.944444 platform. -0.928571

Spread 0.942857 BRO. -0.928571

state: 0.942857 going. -0.927273

jobs." 0.942857 30k -0.925926

Iraq: 0.942857 @DMRegister, -0.925926

you.” 0.941176 profit -0.925926

sleep, 0.939394 @NCGOP -0.924051

days, 0.938596 Congressman -0.923077

#ProudOfObama 0.937853 @mittromney's -0.921569

chant 0.937500 Change" -0.921569

Madonna's 0.937500 'You -0.920000

Obama-Biden? 0.936508 OH! -0.920000

destiny 0.936508 earth. -0.918919

chief. 0.936000 @tedcruz -0.918699

highlights 0.935484 @MittROMney -0.916667

Maya 0.935484 rally's -0.916667

mean. 0.935484 Visitation -0.916084

voucher 0.934426 Cross. -0.913669

lays 0.933333 @REALStaceyDash -0.913043

democracy 0.933333 solve -0.913043

@themick1962: 0.933333 Incredible -0.913043

32 0.932203 #47percent -0.913043

2010, 0.931034 repulsive! -0.910714

Quit 0.931034 Boca -0.909091

visions 0.931034 "It’s -0.909091

$4 0.929825 attended -0.909091

“One 0.928571 @David\_Brody: -0.909091

Cmon 0.928571 @rcmahoney: -0.909091

@OFA\_VA 0.928571 nations -0.907692

commitment 0.928571 LIL -0.906977

basketball 0.927273 honesty, -0.904762

39 0.926829 Katrina? -0.904762

safer 0.925926 evader -0.904762

Richmond, 0.925926 Note -0.904762

helpful 0.925926 #Kissimmee -0.904762

@CraigatFEMA 0.925926 34% -0.904762

@UN 0.925926 Hospital -0.904110

Firefighter 0.925926 Endorsement: -0.902439

campaign—that's 0.925926 Same-Sex -0.901408

champion 0.923077 paper. -0.900000

stand. 0.923077 blasted -0.900000

voices. 0.923077 @HurcaneSandy: -0.900000

word, 0.921569 FLORIDA -0.900000

Grabbing 0.921053 #10moredays -0.900000

Obama: 0.920419 http://t.co/Ek1v8Qhs -0.897959

CONGRATULATIONS 0.920000 @P0TUS: -0.896825

reform, 0.920000 Couples -0.894737

best, 0.920000 Detroit. -0.894737

## 

## P3.3

Looking at the words, some things make sense. Typical Obama like slogans are more common in tweets that mention only him like “#progress”, “together”, “let’s”, and “#ForAll”. Also @MichelleObama is fitting for the Obama side. Similarly on the Romney side there is tagging of the running-mate, and words that reflect the campaign he was trying to run, questioning the success of Obama’s first term. So negative words like “problems” “victims,” and “disaster” show up more in tweets only talking about Romney,