HW03 — STAT/CS 287  
NAME: Cecily Page   
DATE: October 17, 2018

## P1.1

The data was given on BlackBoard and I downloaded it to the folder “HW03\_crpage”

## P1.2

I had a total of 18000 errors. 10415 needed opening brackets “{“ and 7585 needed closing brackets “}”. These were JSON errors. After repairing the data from errors, I removed 2488 duplicate tweets. Leaving 76575 unique tweets.

## P1.3

To get rid of duplicate tweets, I used a set to keep track of the id\_str from the tweets. If the id is already in the set then it skips it an goes to the next tweet (with continue). At first I was using a list to keep the ids in and it was really slow. I realized after waiting for 5 minutes to clean the data I realized sets would be much more efficient. (line 20 and 38)I stored the cleaned tweets in a list. I have a list of all the tweets, a list of tweets that reference Obama, and a list of tweets that reference Romney. (lines 21,22,45,47). To calculate the number of tweets each day I used the Counter object for efficiency.

## P1.4

To split the tweets into tweets about Obama or Romney, if the key words ‘obama’, ‘barack’, ‘barry’ were in the lowercase text of the tweet, they were added to the Obama list of tweets. If ‘mitt’ or ‘romney’ were found in the lowercase text of the tweet they were added to the romney list of tweets

See lines 44-47

## P2

This plot is a time series of the number of tweets about Obama an Romney. You can see that around midnight on the October 24th, and then again around 7pm on October 25th. There’s also a large jump with the Obama tweet data on November 6th between 10pm and 11pm. There jumps from being as few as about 35 to being 221 which is more than any other tweet, so this is probably caused by an anomaly.

## P3.1

If a word appears 10 time in Obama and once in romney, it has the same weight as appearing 100000 times in Obama and 10000 times in romney. Althrough the proportion is the same, a word appearing 100000 times has significantly more weight than a word that appeared 10 times. I think this should be weighted differently.

## P3.2

Obama Romney

we’ve 0.986 ellenpage -0.991   
tix 0.985 reflect -0.986   
forward” 0.981 freed -0.984   
forall 0.977 sethmacfarlane -0.981   
confirm 0.976 sashay -0.979   
busabusss 0.974 90999 -0.978   
z 0.973 sandyshurricane -0.976   
submit 0.973 godspeed -0.973   
power” 0.972 considering -0.971   
katyperry 0.971 remembered -0.969   
expertise 0.970 dictated -0.968   
“let’s 0.968 defiance -0.966   
officials 0.966 1100 -0.956   
champion 0.965 puto -0.955   
magicjohnson 0.963 futures -0.955   
bbcbreaking 0.963 georgelopez -0.953   
jay 0.960 seize -0.952   
chapter 0.960 47percent -0.951   
instructions 0.959 occasion -0.951   
animator 0.958 closet -0.950   
seaboard 0.957 implying -0.949   
steven 0.955 shoq -0.949   
country’s 0.955 fairfax -0.947   
privatesector 0.955 bigot -0.947   
spy 0.953 bowl -0.944   
orleans 0.952 gym -0.944   
lenadunham 0.952 pretended -0.944   
women” 0.952 vagina -0.944   
bruce 0.952 metaquest -0.939   
huracán 0.951 romneylies -0.939   
performing 0.950 flashback -0.934   
madonnas 0.947 dictators -0.933   
ofaco 0.946 measured -0.933   
singlehandedly 0.945 ncgop -0.933   
ofava 0.942 electorals -0.931   
chant 0.940 jconason -0.931   
winds 0.939 rcmahoney -0.931   
destiny 0.938 latinas -0.926   
maya 0.938 amphitheater -0.926   
gottavote 0.935 beltwaybaca -0.923   
highlights 0.935 overflow -0.923   
ohvotesearly 0.935 thread -0.923   
proudofobama 0.934 zacharyquinto -0.920   
eric 0.933 expand -0.920   
lays 0.933 joshromney -0.920   
phyrefyter 0.931 evanaxelbank -0.917   
ofanh 0.930 attended -0.917   
commitment 0.929 rallys -0.917   
chipotle 0.929 jljacobson -0.913   
president’s 0.929 boca -0.913   
firefighter 0.926 taxreturns -0.913   
ofaia 0.926 solve -0.913   
campaign—thats 0.926 qualities -0.912   
craigatfema 0.926 marnus3 -0.909   
tammybaldwinwi 0.926 10moredays -0.909   
that’s 0.924 kissimmee -0.909   
madison 0.923 scottbrownma -0.909   
jm 0.923 visitation -0.908   
greeting 0.920 realstaceydash -0.905   
location 0.920 hilary -0.905   
what’s 0.919 prestoncnn -0.905   
yourfavwhiteguy 0.918 mrburlesk -0.905   
vota 0.917 evader -0.905   
shortgo 0.917 hurcanesandy -0.900   
formed 0.917 uniforms -0.900   
teaching 0.917 blasted -0.900   
peter 0.917 infomercial -0.900   
fights 0.914 otoolefan -0.900   
ofaoh 0.914 releasethereturns -0.900   
ofanc 0.913 httptcoek1v8qhs -0.898   
campus 0.913 robportman -0.897   
dadt 0.913 dgjackson -0.897   
54 0.911 repulsive -0.897   
go” 0.909 sayfie -0.895   
ucf 0.909 manchester -0.895   
promisekept 0.909 presidentelect -0.895   
theellenshow 0.909 p0tus -0.895   
booing 0.907 tommyforwi -0.895   
begun 0.907 badgering -0.895   
“you 0.906 basic -0.894   
tony 0.905 conviction -0.893   
obama’s 0.905 blackrepublican -0.892   
row 0.902 namely -0.892   
safer 0.902 morrowchris -0.890   
libertylynx 0.900 wrangler -0.889   
snoopdogg 0.900 i4 -0.889   
haven’t 0.900 lavellsays -0.889   
theblazetv 0.900 dirkz1 -0.889   
eastern 0.898 congressman -0.889   
toddkincannon 0.897 samesex -0.889   
tpbgirl 0.895 scumbag -0.889   
themick1962 0.895 believeinamerica -0.886   
aceofspadeshq 0.895 donating -0.883   
ofanv 0.895 rooting -0.883   
latinosforobama 0.895 1118 -0.882   
therealmix 0.895 kathrynjscags -0.882   
recruit 0.895 tom -0.882   
paulmccartney 0.895 7im -0.882   
commend 0.895 widely -0.882   
citing 0.894 azmoderate -0.882

\*\*This was all formatted perfectly in my output, but for some reason I didn’t copy with the correct spacing.

## P3.3

The long names were the people who were retweeted like katyperry and sethmacfarlane.There doesn’t seem to be any specific trend with words in Obama’s or romeny’s tweets. Obviously their names show up in their respective columns, but there doesn’t seem to be any big trend. ‘Commend’ ’president’ and ‘confirm’ seem like congratulatory words for obama’s win, but romney doesn’t have any themes.