HW03 — STAT/CS 287
NAME: Galen Byrd
DATE: 10/15/18

P1.1

I downloaded the zipped JSON objects from blackboard and saved it to my Stat 287 folder on my desktop, inside a HW3 folder. This is opened using gzip and relative paths.

P1.2

Immediately I got a JSONDecodeError: Extra data. Some of the JSON objects were missing squiggly braces at the beginning/end (or both) of the object, so we had to add those to 18,000 records. After repairs we see 76,575 tweets. I also get a weird character in the word table that I think is an encoding issue but I am not sure how to fix this. For interpretability I made all characters lower case and removed punctuation.

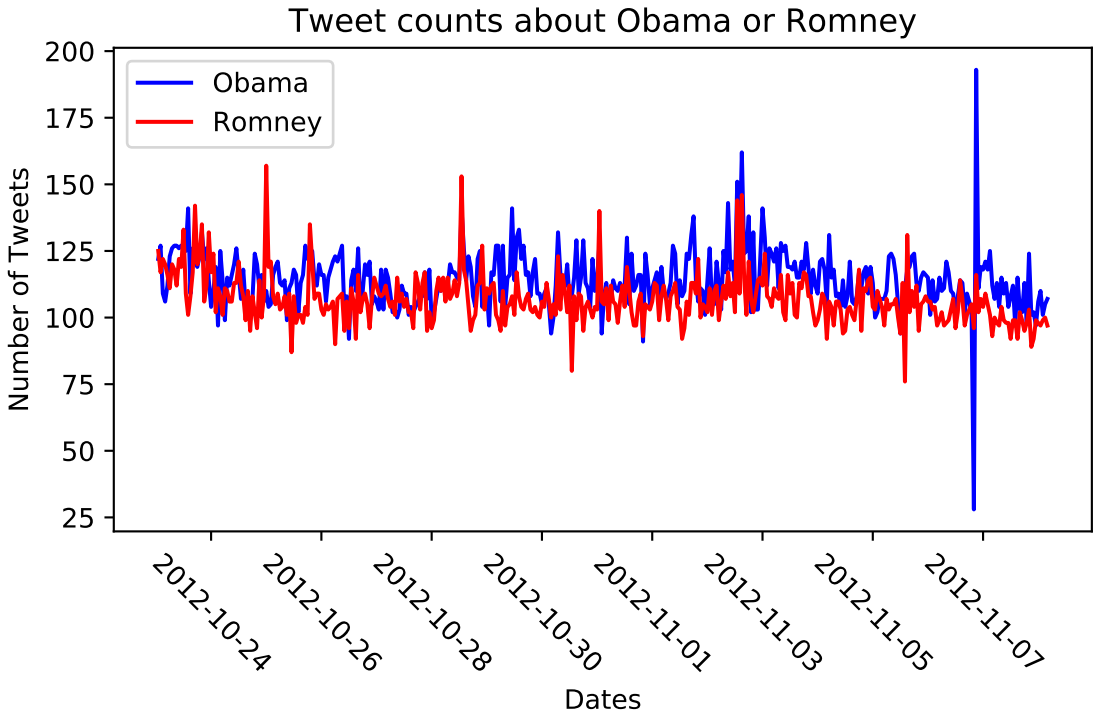
P1.3

I used a list to store each tweet dictionary containing all the information for that tweet. This is done on lines 57 and 69 for uncorrupted/corrupted data respectively.

P1.4

I made a list of possible names people might refer to either Obama or Romney with (lines 15/17) and checked if any of those names show up in the text of any of the tweets, appending the tweet to the corresponding corpus (lines 79-82). I didn't add Hussein for Obama as I thought it would give more false positives, because anyone who would call Obama by his middle name would also use either his first or last or both too.

P2



P3.1

If a word shows up in one corpus many times and the other only once it will be heavily weighted toward the corpus in which it shows up many times. This gives us a few odd words that are less meaningful. A better statistic would be TFIDF as it can take in more than just two corpuses.

P3.2

cScores

tix	0.984848	ellenpage	-0.991031
magicjohnson	0.981308	rickgorka	-0.987013
forward”	0.979381	reflect	-0.985714
katyperry	0.978182	freed	-0.982906
confirm	0.975709	dean	-0.982143
busabusss	0.973684	sethmacfarlane	-0.981481
submit	0.973333	sashay	-0.978947
z	0.972973	sandyshurricane	-0.97561
we’ve	0.972789	godspeed	-0.973333
power”	0.971429	considering	-0.971429
ofanc	0.970149	remembered	-0.969231
expertise	0.969697	90999	-0.967568
“let’s	0.966667	dictated	-0.966942
ofava	0.966102	defiance	-0.966102
officials	0.965812	tedcruz	-0.964072
bbcbreaking	0.962617	puto	-0.954545
ofanh	0.961538	fresh	-0.953488
jay	0.959596	seize	-0.95122
instructions	0.958763	futures	-0.95122
—president	0.958159	closet	-0.95
country’s	0.954545	47percent	-0.95
steven	0.954545	occasion	-0.95
privatesector	0.953488	bigot	-0.94702
lenadunham	0.952381	bowl	-0.944444
women”	0.952381	gym	-0.944444
orleans	0.952381	pretended	-0.944444
spy	0.952381	vagina	-0.943662
forall	0.951807	canned	-0.941606
bruce	0.951807	stale	-0.94
huracán	0.95122	realstaceydash	-0.938053
performing	0.95	politicspr	-0.937984
champion	0.948718	georgelopez	-0.937984
booing	0.948718	dgjackson	-0.935484

madonnas	0.947368	barackobama’s	-0.934959
proudofobama	0.947368	flashback	-0.934426
ofaco	0.945946	unraveling	-0.933333
singlehandedly	0.945205	measured	-0.933333
themick1962	0.942857	ncgop	-0.932584
chant	0.940299	delivering	-0.932203
winds	0.939394	rcmahoney	-0.931034
what’s	0.938272	electorals	-0.931034
maya	0.9375	jconason	-0.931034
you”	0.9375	ميت	-0.928571
eastern	0.9375	amphitheater	-0.925926
highlights	0.935484	latinas	-0.925926
gottavote	0.934066	qualities	-0.924528
that’s	0.933702	overflow	-0.923077
ohvotesearly	0.931034	thread	-0.923077
phyrefyter	0.931034	expand	-0.92
you’ll	0.931034	zacharyquinto	-0.92
lays	0.928571	kattwillliams	-0.92
location	0.928021	beltwaybaca	-0.92
president’s	0.927928	goods	-0.918367
eric	0.926606	dictators	-0.918033
craigatfema	0.925926	rallys	-0.916667
tammybaldwinwi	0.925926	1100	-0.913043
firefighter	0.925926	attended	-0.913043
jm	0.923077	culture	-0.913043
nvdecides	0.921569	ayeee	-0.913043
chapter	0.921569	marnus3	-0.909091
madison	0.92	kissimmee	-0.909091
fri	0.918919	davidshepardson	-0.909091
rousing	0.918367	taxreturns	-0.909091
yourfavwhiteguy	0.918367	enquirer	-0.906977
formed	0.916667	joss	-0.904762
shortgo	0.916667	evader	-0.904762
vota	0.916667	mrburlesk	-0.904762
dadt	0.913043	prestoncnn	-0.904762
condition	0.913043	10moredays	-0.904762

blocking	0.913043	implying	-0.902439
begun	0.911894	solve	-0.901639
fights	0.911765	infomercial	-0.9
campus	0.909091	hilary	-0.9
go”	0.909091	barakobama	-0.9
peter	0.909091	thepresobama	-0.9
destiny	0.909091	uniforms	-0.9
defendpaulryan	0.904762	blasted	-0.9
ofaoh	0.904762	presidentelect	-0.894737
cnns	0.904762	badgering	-0.894737
tony	0.904762	otoolefan	-0.894737
“you	0.904762	robportman	-0.894737
casino	0.903226	releasethereturns	-0.894737
obama’s	0.901316	blackrepublican	-0.891892
theblazetv	0.9	conviction	-0.890909
haven’t	0.9	morrowchris	-0.889908
ofaia	0.9	p0tus	-0.889299
row	0.898734	jljacobson	-0.888889
toddkincannon	0.897436	nominee	-0.888889
ofanv	0.895522	congressman	-0.888889
safer	0.894737	i4	-0.888889
recruit	0.894737	dirkz1	-0.888889
tpbgirl	0.894737	rooting	-0.888
commend	0.894737	romneylies	-0.885714
paulmccartney	0.894737	reppaulryan	-0.883721
libertylynx	0.894737	dailykos	-0.882353
citing	0.893617	1118	-0.882353
basketball	0.892857	mittromneyisapathologicalliar	-0.882353
sacrificed	0.891892	wut	-0.882353
americaforward	0.88961	azmoderate	-0.878788
tracyjeffords	0.888889	gut	-0.878788

P3.3

Some of the words make sense. Many words are just here because they show up many times in one corpus and only a few in the other. There are a lot of famous people that I think is because they either endorsed the candidate or said something polarizing about them. We also see many

words with negative connotations weighted towards Romney. This could be because people did not like him, hence why he lost.