BDS William Jiang

Prof. Bari 3/24/19

**Solutions**

Data Exploration

1. Top 20 hashtags are:

#followfriday 1861

#fb 1728

#squarespace 843

#FF 503

#seb-day 498

#FollowFriday 414

#1 403

#musicmonday 385

#iranelection 342

#ff 334

#fail 288

#BSB 273

#iremember 263

#myweakness 260

#asot400 254

#marsiscoming 245

#mcflyforgermany 230

#2 228

#andyhurleyday 187

#inaperfectworld 177

These are the top 20 hashtag mentions. Nothing interesting to note, other than that the first 2 hashtags dwarf that others.

1. The top 20 @handles are:

@mileycyrus 4331

@tommcfly 3853

@ddlovato 3383

@Jonasbrothers 1268

@DavidArchie 1239

@jordanknight 1109

@DonnieWahlberg 1086

@JonathanRKnight 1054

@mitchelmusso 1042

@taylorswift13 979

@jonasbrothers 958

@selenagomez 787

@dougiemcfly 786

@peterfacinelli 610

@aplusk 609

@joeymcintyre 564

@Dannymcfly 540

@gfalcone601 539

@shaundiviney 506

@YoungQ 502

Interestingly, all the twitter handles belong to celebrities, most of whom appeal to a younger crowd.

NGram Analysis

I was unable to compile using HPC Cluster’s batch system. However, I was able to compile and run on an interactive shell with a much smaller sample set (first 300 tweets) and these are my results:

----------------------------------NGrams------------------------------------

I = 145

to = 114

the = 111

i = 105

a = 65

and = 65

my = 58

it = 52

in = 41

for = 40

that = 38

but = 32

of = 32

Im = 31

you = 31

so = 31

me = 30

at = 29

have = 27

not = 25

----------------------------------END-----------------------------------

----------------------------------NGrams------------------------------------

in the = 12

going to = 10

I dont = 10

for the = 9

I mis = 8

have to = 8

to go = 8

of the = 7

I hate = 7

at the = 6

to sleep = 6

I have = 6

I wa = 6

Im so = 5

I cant = 5

I still = 5

I need = 5

want to = 5

for me = 5

go to = 5

----------------------------------END-----------------------------------

----------------------------------NGrams------------------------------------

I miss my = 4

I want to = 3

I need to = 3

I hate when = 3

to go to = 3

i have to = 3

to watch it = 2

I dont think = 2

is not going = 2

and now I = 2

I need a = 2

and I still = 2

one of my = 2

wednesday my bday = 2

dont know what = 2

know what do = 2

I cant see = 2

go to the = 2

Im sad that = 2

in the morning = 2

----------------------------------END-----------------------------------

----------------------------------NGrams------------------------------------

dont know what do = 2

so behind the time = 2

get shit done today = 2

shit done today im = 2

done today im so = 2

today im so screwed = 2

want to go to = 2

missed Brent at praise = 2

Brent at praise band = 2

at praise band No = 2

praise band No fun = 2

band No fun to = 2

No fun to not = 2

fun to not have = 2

your lead guitarist ltpoutgt = 2

is in the bathroom = 2

a bummer You shoulda = 1

Third Day to do = 1

httptwitpiccomyzl Awww thats a = 1

to do it D = 1

[1] Rule-based Methods

I was only able to complete the first part of this task (the POS tagger). If given more time, the rest should follow. My greatest obstacle to this assignment was to compile my code using HPC cluster’s batch server, even though the interactive server could compile it.