Big Data Science
Homework 2
Keyword Extraction

Fang Han
fh643@nyu.edu

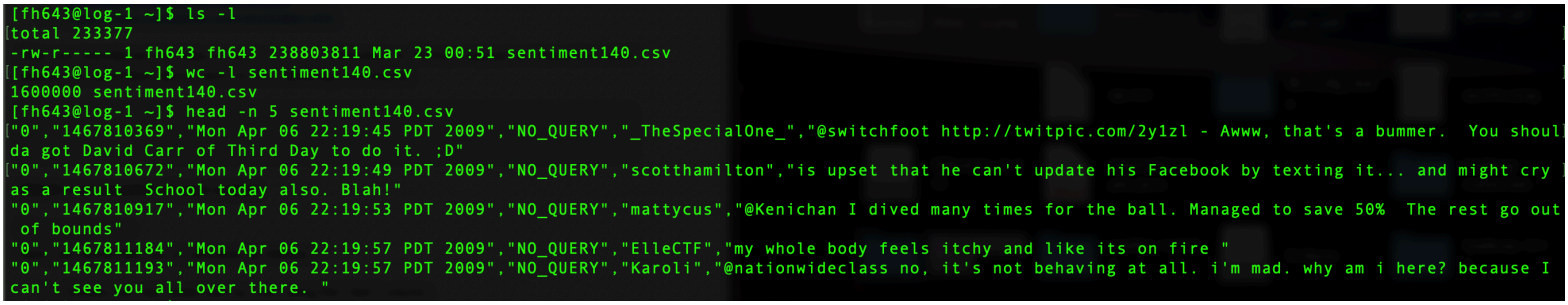
PART 0 — structure

— Code is stored in `/hw2_fh643/code`

— All output files (.txt) mentioned here below are in `/hw2_fh643/output`

— Detailed documentation as to how to run my projects on HPC is stored at README.md at `/hw2_fh643`

PART 1 — HPC setup



PART 2 — Data exploration

1. Extract raw tweet text: see text.txt

2. Top 20 most frequent hashtags: see topHashtags.txt

— Comment: manually combined identical entries out of the top 30 hashtags together due to my lack of clue as to how to smartly do so.

#followfriday / #followfriday!	1737
#fb"	1269
#ff	625
#squarespace	609
#seb-day	419

#iranelection	310
#fail	261
#bsb	232
#asot400	202
#mcflyforgermany	199
#iphone	170
#marsiscoming	169
#f1	164
#mileymonday	141
#delongeday	124
#musicmonday	111
#andyhurleyday	95
#e3	85
#bradiewebb	81
#spymaster	79

3. Top 20 most frequent @-mentions

— Comment: filtered out manually the empty @ symbol from the original output.

@jonasbrothers	522
@mileycyrus	499
@ddlovato	397
@tommcfly	197
@taylorswift13	173
@davidarchie	135
@jonathanrknight	135
@mitchelmusso	134
@donniewahlberg	131
@songzyuuup	112
@selenagomez	108
@nkotb	104
@jordanknight	102
@peterfacinelli	96
@dougiemcfly	94

@dannymcfly	93
@twitter	93
@reply	89
@replies	71
@jackalltimelow	69

PART 3 — N-gram analysis

1-grams:

love	9
the	8
this	7
with	7
and	7
for	7
lol	7
you	7
time	6
iphone	6
will	6
good	6
have	6
not	6
like	5
xxxx"	5
night	5
infamous	5
still	5
wish	5

2-grams:

the best	8
in the	8
my ipod	8
thank you	7
my new	7
love the	7
love you	7
the real	7
am so	7
is the	7
i love	7
he is	7
to the	7
and the	7
miss my	7
and you	7
it is	7
and i'm	7
on my	7
was the	7

3-grams:

i love the	10
i love you	10
but i did	9
and i love	9
is the best	9
but i love	9
but i am	8
the mtv movie	8

wish i was	8
i wish i	8
and i am	8
wish i could	8
have a great	8
mtv movie awards	8
but i have	8
i can't wait	8
"i love the	8
day to all	7
and i can't	7
i need a	7

4-grams:

the mtv movie awards	10
have a great day	9
i wish i was	9
happy mother's day to	8
i love the new	8
but i have to	8
and i have to	8
i wish i could	8
"happy fathers day to	8
now i have to	7
the new moon trailer	7
mother's day to all	7
" "watching the hills	7
thank you for the	7
"happy mother's day to	7
" "i love the	7
wish i could be	7

thank you so much	7
mothers day to all	7
you so much for	7

PART 4 — TextRank

1. TextRank algorithm is a graph-based ranking model that decides the importance of a vertex within a graph. The basic idea implemented by a graph-based ranking model is that of “voting” or “recommendation”. When one vertex links to another one, it is basically casting a vote for that other vertex. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex. Arbitrary values are assigned to each node in the graph initially, following iterations until convergence below a given threshold is achieved.

2. Run TextRank on the entire dataset: full output see textrank_output.txt

[java] 0.4 0.0 1.0 0.0	i
[java] 0.3 0.0 0.63 0.0	day
[java] 0.2 0.0 0.4 0.0	home
[java] 0.2 0.0 0.38 0.0	work
[java] 0.2 0.0 0.37 0.0	time
[java] 0.1 0.0 0.35 0.0	today
[java] 0.1 0.0 0.35 0.0	love
[java] 0.1 0.0 0.23 0.0	thanks
[java] 0.1 0.0 0.23 0.0	way
[java] 0.1 0.0 0.23 0.0	fun
[java] 0.1 0.0 0.21 0.0	week
[java] 0.1 0.0 0.2 0.0	friend
[java] 0.1 0.0 0.2 0.0	bed
[java] 0.1 0.0 0.19 0.0	morning
[java] 0.1 0.0 0.19 0.0	hour
[java] 0.1 0.0 0.17 0.0	guy

[java] 0.1 0.0 0.16 0.0	haha
[java] 0.1 0.0 0.16 0.0	everyone
[java] 0.1 0.0 0.16 0.0	people
[java] 0.1 0.0 0.16 0.0	twitter
[java] 0.1 0.0 0.16 0.0	tomorrow
[java] 0.1 0.0 0.15 0.0	lot
[java] 0.1 0.0 0.15 0.0	school
[java] 0.1 0.0 0.15 0.0	wish
[java] 0.1 0.0 0.15 0.0	weekend
[java] 0.1 0.0 0.14 0.0	life
[java] 0.1 0.0 0.14 0.0	lol
[java] 0.1 0.0 0.14 0.0	night
[java] 0.1 0.0 0.14 0.0	something
[java] 0.1 0.0 0.14 0.0	hope
[java] 0.1 0.0 0.13 0.0	need
[java] 0.1 0.0 0.13 0.0	thing
[java] 0.1 0.0 0.12 0.0	phone

3. Run TextRank on [“4”] tweets and [“0”] tweets respectively:

[“0”] group — full output see output_0tweet.txt

[java] 0.4 0.0 1.0 0.0	i
[java] 0.2 0.0 0.5 0.0	day
[java] 0.2 0.0 0.42 0.0	work
[java] 0.2 0.0 0.39 0.0	home
[java] 0.1 0.0 0.32 0.0	today
[java] 0.1 0.0 0.3 0.0	time
[java] 0.1 0.0 0.2 0.0	week
[java] 0.1 0.0 0.2 0.0	wish
[java] 0.1 0.0 0.2 0.0	hour
[java] 0.1 0.0 0.19 0.0	bed

[java] 0.1 0.0 0.18 0.0	way
[java] 0.1 0.0 0.17 0.0	morning
[java] 0.1 0.0 0.16 0.0	friend
[java] 0.1 0.0 0.15 0.0	school
[java] 0.1 0.0 0.14 0.0	love
[java] 0.1 0.0 0.14 0.0	phone
[java] 0.1 0.0 0.14 0.0	tomorrow
[java] 0.1 0.0 0.14 0.0	need
[java] 0.1 0.0 0.14 0.0	people
[java] 0.1 0.0 0.14 0.0	hate
[java] 0.1 0.0 0.14 0.0	fun
[java] 0.1 0.0 0.13 0.0	weekend
[java] 0.1 0.0 0.13 0.0	someone

["4"] group — full output see output_4tweet.txt

[java] 0.4 0.0 1.0 0.0	i
[java] 0.3 0.0 0.84 0.0	days
[java] 0.3 0.0 0.68 0.0	love
[java] 0.2 0.0 0.52 0.0	thanks
[java] 0.2 0.0 0.49 0.0	time
[java] 0.2 0.0 0.42 0.0	home
[java] 0.2 0.0 0.41 0.0	today
[java] 0.2 0.0 0.37 0.0	fun
[java] 0.1 0.0 0.31 0.0	work
[java] 0.1 0.0 0.3 0.0	way
[java] 0.1 0.0 0.3 0.0	haha
[java] 0.1 0.0 0.27 0.0	friend
[java] 0.1 0.0 0.26 0.0	everyone
[java] 0.1 0.0 0.25 0.0	guys

[java] 0.1 0.0 0.23 0.0	twitter
[java] 0.1 0.0 0.23 0.0	lol
[java] 0.1 0.0 0.23 0.0	morning
[java] 0.1 0.0 0.22 0.0	week
[java] 0.1 0.0 0.22 0.0	lot
[java] 0.1 0.0 0.21 0.0	bed
[java] 0.1 0.0 0.19 0.0	life
[java] 0.1 0.0 0.19 0.0	people
[java] 0.1 0.0 0.19 0.0	night
[java] 0.1 0.0 0.18 0.0	hope
[java] 0.1 0.0 0.18 0.0	tomorrow
[java] 0.1 0.0 0.17 0.0	good morning
[java] 0.1 0.0 0.17 0.0	weekend
[java] 0.1 0.0 0.17 0.0	hours
[java] 0.1 0.0 0.16 0.0	songs
[java] 0.1 0.0 0.16 0.0	something
[java] 0.1 0.0 0.16 0.0	movie
[java] 0.1 0.0 0.16 0.0	things
[java] 0.1 0.0 0.15 0.0	girl
[java] 0.1 0.0 0.15 0.0	school
[java] 0.1 0.0 0.14 0.0	sun
[java] 0.1 0.0 0.14 0.0	followers
[java] 0.1 0.0 0.14 0.0	years
[java] 0.1 0.0 0.13 0.0	mom
[java] 0.1 0.0 0.13 0.0	'tweets
[java] 0.1 0.0 0.12 0.0	house