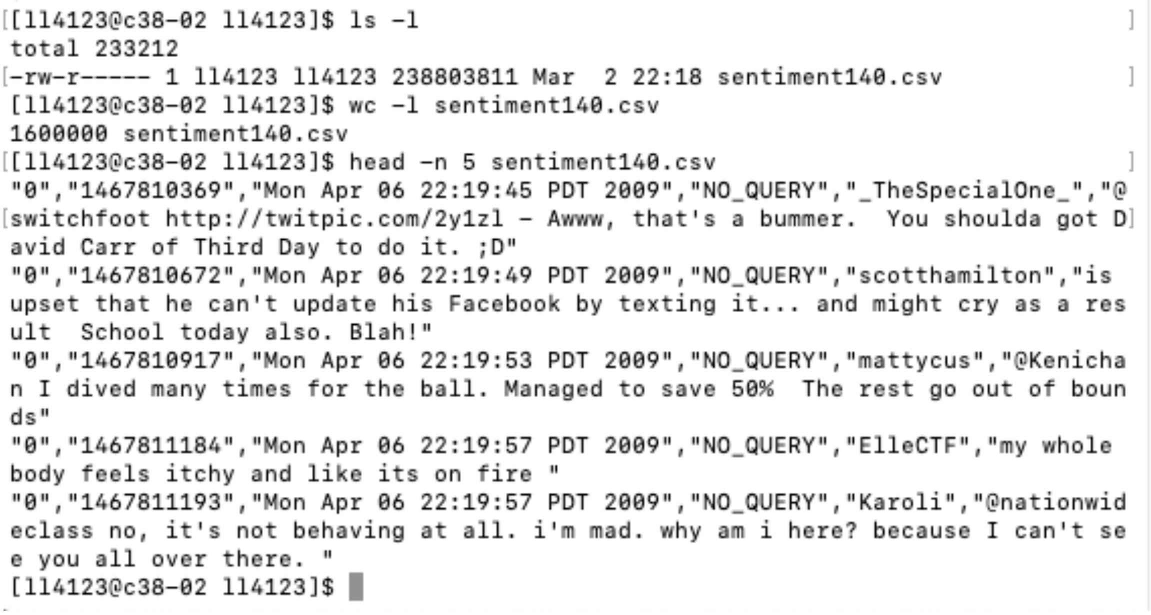
# Big Data Science

# HW3 — Twitter keyword extraction

Lingwei Luo (ll4123@nyu.edu)

## Part 1 HPC setup

Screenshot of required commands output.



```
[[ll4123@c38-02 ll4123]$ ls -l                                            ]
total 233212
[-rw-r----- 1 ll4123 ll4123 238803811 Mar  2 22:18 sentiment140.csv       ]
[ll4123@c38-02 ll4123]$ wc -l sentiment140.csv
1600000 sentiment140.csv
[[ll4123@c38-02 ll4123]$ head -n 5 sentiment140.csv                       ]
"0","1467810369","Mon Apr 06 22:19:45 PDT 2009","NO_QUERY","_TheSpecialOne_","@
[switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got D]
avid Carr of Third Day to do it. ;D"
"0","1467810672","Mon Apr 06 22:19:49 PDT 2009","NO_QUERY","scotthamilton","is
upset that he can't update his Facebook by texting it... and might cry as a res
ult  School today also. Blah!"
"0","1467810917","Mon Apr 06 22:19:53 PDT 2009","NO_QUERY","mattycus","@Kenicha
n I dived many times for the ball. Managed to save 50%  The rest go out of boun
ds"
"0","1467811184","Mon Apr 06 22:19:57 PDT 2009","NO_QUERY","ElleCTF","my whole
body feels itchy and like its on fire "
"0","1467811193","Mon Apr 06 22:19:57 PDT 2009","NO_QUERY","Karoli","@nationwid
eclass no, it's not behaving at all. i'm mad. why am i here? because I can't se
e you all over there. "
[ll4123@c38-02 ll4123]$ █
```

## Part 2 Data exploration

1 Top 20 most frequent hashtags: please see top#.txt

| **#fb** | **941** |
|---|---:|
| **#squarespace** | 515 |
| **#followfriday** | 493 |
| **#seb-day** | 276 |
| **#1** | 230 |
| **#iranelection** | 203 |
| **#musicmonday** | 178 |
| **#marsiscoming** | 164 |
| **#fail** | 161 |

| | |
|---|---:|
| **#BSB** | 161 |
| **#2** | 159 |
| **#andyhurleyday** | 146 |
| **#FF** | 145 |
| **#mcflyforgermany** | 144 |
| **#myweakness** | 142 |
| **#iremember** | 128 |
| **#bgt** | 117 |
| **#FollowFriday** | 113 |
| **#ff** | 111 |
| **#trackle** | 96 |

## 2 Top 20 most frequent @-mentions: please see top@.txt

| **@mileycyrus** | **2509** |
|---|---:|
| **@tommcfly** | 2182 |
| **@ddlovato** | 1686 |
| **@DavidArchie** | 730 |
| **@Jonasbrothers** | 726 |
| **@JonathanRKnight** | 667 |
| **@taylorswift13** | 579 |
| **@jordanknight** | 574 |
| **@DonnieWahlberg** | 565 |
| **@mitchelmusso** | 554 |
| **@jonasbrothers** | 509 |
| **@selenagomez** | 423 |
| **@dougiemcfly** | 393 |
| **@aplusk** | 351 |
| **@peterfacinelli** | 330 |
| **@joeymcintyre** | 304 |
| **@gfalcone601** | 289 |
| **@YoungQ** | 281 |
| **@Dannymcfly** | 257 |
| **@shaundiviney** | 256 |

# Part 3 N-gram analysis

1 Top 20 1-grams: please see top20_1gram.txt

| [to] | 293190 |
|---|---:|
| [I] | 271480 |
| [the] | 255501 |
| [a] | 191464 |
| [my] | 150309 |
| [and] | 147103 |
| [i] | 138317 |
| [you] | 124499 |
| [is] | 117076 |
| [it] | 111553 |
| [for] | 109012 |
| [in] | 108568 |
| [of] | 93969 |
| [on] | 83372 |
| [me] | 80755 |
| [have] | 70950 |
| [so] | 69666 |
| [that] | 69039 |
| [but] | 60988 |
| [with] | 58136 |

2 Top 20 2-grams: please see top20_2gram.txt

| [in, the] | 23039 |
|---|---:|
| [going, to] | 19456 |
| [for, the] | 17211 |
| [to, be] | 15782 |
| [to, go] | 15716 |
| [I, have] | 15505 |
| [on, the] | 15439 |
| [to, the] | 15190 |
| [have, to] | 14631 |
| [I, am] | 13362 |
| [of, the] | 12673 |
| [have, a] | 12552 |
| [to, get] | 11878 |

| | |
|---|---:|
| **[want, to]** | 11591 |
| **[go, to]** | 10366 |
| **[I, was]** | 10310 |
| **[at, the]** | 10162 |
| **[I, don't]** | 9959 |
| **[to, do]** | 9943 |
| **[for, a]** | 9878 |

3 Top 20 3-grams: please see top20_3gram.txt

| [to, go, to] | 5282 |
|---|---:|
| **[I, have, to]** | 3960 |
| **[I, want, to]** | 3089 |
| **[going, to, be]** | 3047 |
| **[I, wish, I]** | 2790 |
| **[I, have, a]** | 2650 |
| **[want, to, go]** | 2366 |
| **[is, going, to]** | 2337 |
| **[I'm, going, to]** | 2317 |
| **[i, have, to]** | 2204 |
| **[I, need, to]** | 2116 |
| **[looking, forward, to]** | 2102 |
| **[go, to, the]** | 2002 |
| **[a, lot, of]** | 1767 |
| **[have, to, go]** | 1708 |
| **[I, think, I]** | 1707 |
| **[to, be, a]** | 1671 |
| **[i, wish, i]** | 1667 |
| **[I, don't, know]** | 1630 |
| **[in, the, morning]** | 1619 |

4 Top 20 4-grams: please see top20_4gram.txt

| [I, wish, I, could] | 1050 |
|---|---:|
| **[to, go, to, the]** | 1023 |
| **[want, to, go, to]** | 940 |
| **[is, going, to, be]** | 895 |
| **[I, don't, want, to]** | 821 |
| **[to, go, to, work]** | 811 |
| **[you, are, on, the]** | 788 |
| **[on, the, train, or]** | 775 |

| | |
|---|---|
| **[are, on, the, train]** | 773 |
| **[a, day, using, www]** | 773 |
| **[add, everyone, you, are]** | 773 |
| **[followers, a, day, using]** | 773 |
| **[Once, you, add, everyone]** | 773 |
| **[the, train, or, pay]** | 773 |
| **[you, add, everyone, you]** | 773 |
| **[train, or, pay, vip]** | 773 |
| **[com, Once, you, add]** | 773 |
| **[100, followers, a, day]** | 773 |
| **[everyone, you, are, on]** | 773 |
| **[Get, 100, followers, a]** | 773 |

# Part 4 TextRank

1 Summary of TextRank:

TextRank is a general purpose, graph based ranking algorithm for NLP. It calculates the importance of a vertex in a graph. The core ideas in this model are “voting” and “recommendation”. If one vertex is linked to another vertex, a vote will be casted for that another vertex. The number of votes that a vertex obtained, represents the importance of the vertex. Initially, arbitrary values will be assigned to each node in the graph, consistent iterations will be conducted until convergence reaches below a given threshold.

2 Run the TextRank algorithm on the entire dataset: please see allTextRank.txt

| **1** | **I** |
|---:|---|
| **2** | I'm |
| **3** | day |
| **4** | good |
| **5** | today |
| **6** | work |
| **7** | time |
| **8** | love |
| **9** | u |
| **10** | 2 |
| **11** | night |

| | |
|---|---|
| **12** | The |
| **13** | home |
| **14** | My |
| **15** | lol |
| **16** | 3 |
| **17** | feel |
| **18** | tomorrow |
| **19** | great |
| **20** | Just |
| **21** | morning |
| **22** | bad |
| **23** | hope |
| **24** | sleep |
| **25** | gonna |
| **26** | But |
| **27** | sad |
| **28** | fun |
| **29** | tonight |
| **30** | It's |
| **31** | And |
| **32** | You |
| **33** | 4 |
| **34** | bed |
| **35** | wait |
| **36** | people |
| **37** | haha |
| **38** | I'll |
| **39** | It |
| **40** | week |
| **41** | twitter |
| **42** | nice |
| **43** | watching |
| **44** | So |
| **45** | I've |
| **46** | days |
| **47** | hate |
| **48** | school |
| **49** | watch |
| **50** | No |

| | |
|---|---|
| **51** | long |
| **52** | wanna |
| **53** | happy |
| **54** | show |
| **55** | A |
| **56** | Oh |
| **57** | find |
| **58** | weekend |
| **59** | tired |
| **60** | Good |
| **61** | working |
| **62** | ur |
| **63** | awesome |
| **64** | 1 |
| **65** | sick |
| **66** | friends |
| **67** | hours |
| **68** | ready |
| **69** | yeah |
| **70** | guys |
| **71** | phone |
| **72** | house |
| **73** | life |
| **74** | pretty |
| **75** | Now |
| **76** | n |
| **77** | feeling |
| **78** | left |
| **79** | We |
| **80** | Thanks |
| **81** | movie |
| **82** | Not |
| **83** | thought |
| **84** | bit |
| **85** | This |
| **86** | cool |
| **87** | man |
| **88** | early |
| **89** | LOL |

| | |
|---|---|
| **90** | 5 |
| **91** | lost |
| **92** | start |
| **93** | guess |
| **94** | year |
| **95** | friend |
| **96** | big |
| **97** | missed |
| **98** | Im |
| **99** | What |
| **100** | coming |

Analysis: The top keywords are mixed, there are both positive and negative words.

3 Run the TextRank algorithm on the positive [“4”]: please see poisitiveTextRank.txt

| 1 | I |
|---|---|
| **2** | I'm |
| **3** | good |
| **4** | day |
| **5** | love |
| **6** | u |
| **7** | today |
| **8** | time |
| **9** | great |
| **10** | 2 |
| **11** | night |
| **12** | The |
| **13** | lol |
| **14** | work |
| **15** | You |
| **16** | fun |
| **17** | 3 |
| **18** | Just |
| **19** | morning |
| **20** | home |
| **21** | nice |
| **22** | haha |
| **23** | hope |

| | |
|---|---|
| **24** | tomorrow |
| **25** | wait |
| **26** | My |
| **27** | I'll |
| **28** | It's |
| **29** | happy |
| **30** | watching |
| **31** | And |
| **32** | Good |
| **33** | twitter |
| **34** | tonight |
| **35** | gonna |
| **36** | awesome |
| **37** | Thanks |
| **38** | people |
| **39** | 4 |
| **40** | It |
| **41** | But |
| **42** | bed |
| **43** | A |
| **44** | watch |
| **45** | sleep |
| **46** | ur |
| **47** | week |
| **48** | I've |
| **49** | So |
| **50** | days |
| **51** | show |
| **52** | cool |
| **53** | feel |
| **54** | LOL |
| **55** | movie |
| **56** | amazing |
| **57** | ready |
| **58** | pretty |
| **59** | guys |
| **60** | yeah |
| **61** | friends |
| **62** | long |

| | |
|---|---|
| **63** | song |
| **64** | weekend |
| **65** | Oh |
| **66** | We |
| **67** | Have |
| **68** | school |
| **69** | www |
| **70** | 1 |
| **71** | life |
| **72** | Twitter |
| **73** | Now |
| **74** | bit |
| **75** | ya |
| **76** | birthday |
| **77** | excited |
| **78** | house |
| **79** | If |
| **80** | Hope |
| **81** | friend |
| **82** | party |
| **83** | start |
| **84** | hey |
| **85** | No |
| **86** | bad |
| **87** | follow |
| **88** | big |
| **89** | glad |
| **90** | check |
| **91** | tweet |
| **92** | music |
| **93** | finally |
| **94** | beautiful |
| **95** | n |
| **96** | find |
| **97** | working |
| **98** | thought |
| **99** | What |
| **100** | Hey |

Analysis: The top keywords are mainly positive words.

4 Run the TextRank algorithm on the negative [“0”]: please see negativeTextRank.txt

| 1 | I |
|---:|---|
| 2 | I'm |
| 3 | work |
| 4 | day |
| 5 | today |
| 6 | 2 |
| 7 | time |
| 8 | good |
| 9 | u |
| 10 | home |
| 11 | My |
| 12 | sad |
| 13 | feel |
| 14 | night |
| 15 | bad |
| 16 | The |
| 17 | tomorrow |
| 18 | sleep |
| 19 | love |
| 20 | 3 |
| 21 | hate |
| 22 | But |
| 23 | lol |
| 24 | gonna |
| 25 | Just |
| 26 | tonight |
| 27 | hope |
| 28 | morning |
| 29 | bed |
| 30 | sick |
| 31 | No |
| 32 | school |
| 33 | And |
| 34 | wanna |
| 35 | week |

| | |
|---|---|
| **36** | It's |
| **37** | 4 |
| **38** | people |
| **39** | It |
| **40** | I've |
| **41** | days |
| **42** | So |
| **43** | tired |
| **44** | find |
| **45** | fun |
| **46** | long |
| **47** | lost |
| **48** | Oh |
| **49** | twitter |
| **50** | phone |
| **51** | working |
| **52** | I'll |
| **53** | wait |
| **54** | left |
| **55** | hours |
| **56** | feeling |
| **57** | watch |
| **58** | great |
| **59** | show |
| **60** | watching |
| **61** | weekend |
| **62** | missed |
| **63** | Not |
| **64** | 1 |
| **65** | haha |
| **66** | house |
| **67** | A |
| **68** | n |
| **69** | sucks |
| **70** | early |
| **71** | Now |
| **72** | guess |
| **73** | thought |
| **74** | You |

| | |
|---|---|
| **75** | man |
| **76** | nice |
| **77** | 5 |
| **78** | friends |
| **79** | life |
| **80** | This |
| **81** | rain |
| **82** | year |
| **83** | car |
| **84** | ready |
| **85** | hard |
| **86** | Im |
| **87** | bit |
| **88** | missing |
| **89** | yeah |
| **90** | Why |
| **91** | yesterday |
| **92** | weather |
| **93** | wanted |
| **94** | guys |
| **95** | damn |
| **96** | ur |
| **97** | start |
| **98** | leave |
| **99** | late |
| **100** | hot |

Analysis: The top keywords are mainly negative words.