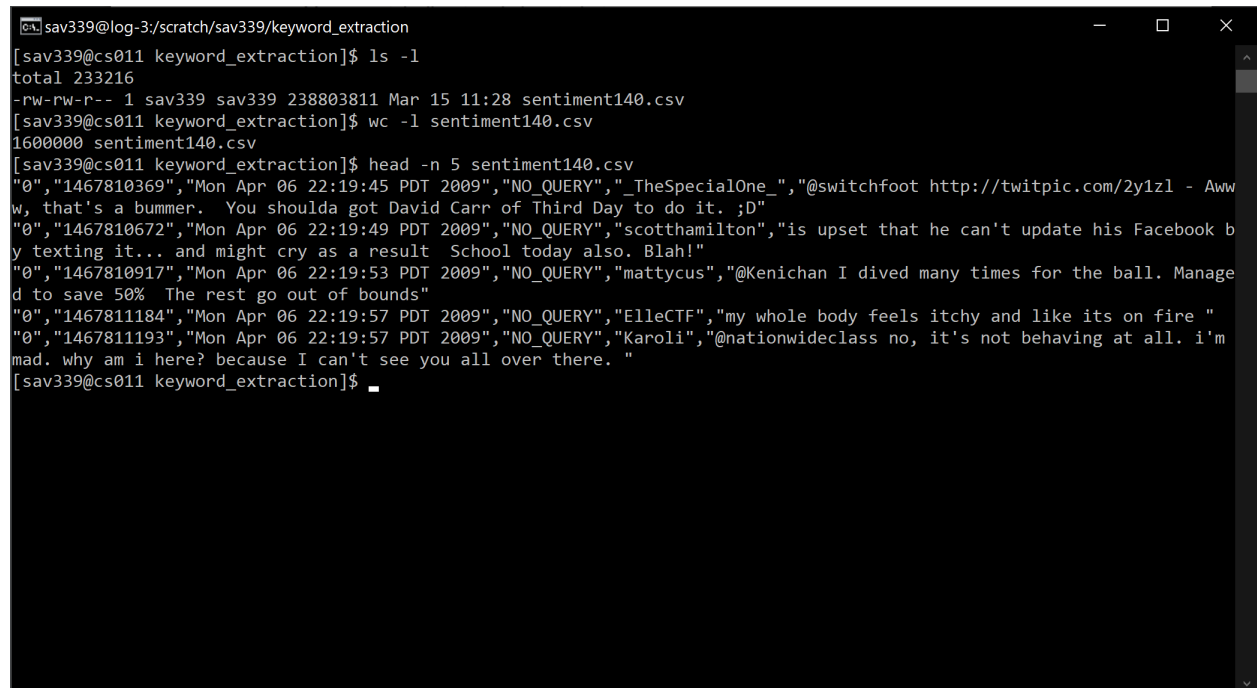


# Big Data Science Keyword Extraction

## Data exploration

### Interactive shell



```
sav339@log-3:/scratch/sav339/keyword_extraction
[sav339@cs011 keyword_extraction]$ ls -l
total 233216
-rw-rw-r-- 1 sav339 sav339 238803811 Mar 15 11:28 sentiment140.csv
[sav339@cs011 keyword_extraction]$ wc -l sentiment140.csv
1600000 sentiment140.csv
[sav339@cs011 keyword_extraction]$ head -n 5 sentiment140.csv
"0","1467810369","Mon Apr 06 22:19:45 PDT 2009","NO_QUERY","_TheSpecialOne_", "@switchfoot http://twitpic.com/2y1z1 - Aww
w, that's a bummer. You shoulda got David 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 b
y texting it... and might cry as a result School today also. Blah!"
"0","1467810917","Mon Apr 06 22:19:53 PDT 2009","NO_QUERY","mattycus", "@Kenichan I dived many times for the ball. Manage
d to save 50% The rest go out of bounds"
"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", "@nationwideclass no, it's not behaving at all. i'm
mad. why am i here? because I can't see you all over there. "
[sav339@cs011 keyword_extraction]$
```

### Hashtag and @-mention collection

Implemented own algorithm for hashtag and @-mention collection, runs in linear time (examines character by character)

```
sav339@log-2:/scratch/sav339/keyword_extraction/bds
[sav339@log-2 bds]$ [sav339@log-2 bds]$ ./DataExploration.sh
Compiling
srun: job 3796552 queued and waiting for resources
srun: job 3796552 has been allocated resources
Running
srun: job 3796553 queued and waiting for resources
srun: job 3796553 has been allocated resources
New parser from sentiment140.csv
Parsing CSV...
Collecting hashtags...
Collecting mentions...
{#fb=267, #asot400=113, #fail=86, #bgt=72, #followfriday=59, #2=58, #1=55, #myweakness=53, #marsiscoming=47, #ONTD=40, #f1=38, #BGT=38, #asylm=38, #bradiewebb=37, #mmwanted=32, #3turnoffwords=31, #fixreplies=26, #andyhurleyday=26, #ASOT400=24, #tag=24}
{@mileycyrus=508, @tommcfly=488, @ddlovato=331, @DavidArchie=167, @mitchelmusso=135, @jordanknight=124, @taylorswift13=116, @DonnieWahlberg=115, @JonathanRKNight=108, @selenagomez=106, @stephenfry=102, @doughiemcfly=90, @gfalcone601=81, @petewentz=80, @aplusk=77, @Jonasbrothers=77, @nick_carter=73, @iamdiddy=72, @joeymcintyre=65, @shaundiviney=63}
[sav339@log-2 bds]$
```

## Top hashtags

- #fb=267,
- #asot400=113,
- #fail=86,
- #bgt=72,
- #followfriday=59,
- #2=58,
- #1=55,
- #myweakness=53,
- #marsiscoming=47,
- #ONTD=40,
- #f1=38,
- #BGT=38,
- #asylm=38,
- #bradiewebb=37,
- #mmwanted=32,
- #3turnoffwords=31,
- #fixreplies=26,
- #andyhurleyday=26,
- #ASOT400=24,
- #tag=24

## Top mentions

- @mileycyrus=508,
- @tommcfly=488,

- @addlovato=331,
- @DavidArchie=167,
- @mitchelmusso=135,
- @jordanknight=124,
- @taylorswift13=116,
- @DonnieWahlberg=115,
- @JonathanRKnigh=108,
- @selenagomez=106,
- @stephenfry=102,
- @doughiemcfly=90,
- @gfalcone601=81,
- @petewentz=80,
- @aplusk=77,
- @Jonasbrothers=77,
- @nick\_carter=73,
- @iamdiddy=72,
- @joeymcintyre=65,
- @shaundiviney=63

## N-gram analysis

### No filtering

Implemented this myself, removed stop words and collected n-grams in linear time.

```

sav339@log-3:/scratch/sav339/keyword_extraction/bds
[sav339@log-3 bds]$ ./NgramAnalysis.sh
Compiling
srunk: job 3837191 queued and waiting for resources
srunk: job 3837191 has been allocated resources
Running
srunk: job 3837192 queued and waiting for resources
srunk: job 3837192 has been allocated resources
New parser from sentiment140.csv
Parsing CSV...
Collecting 1-grams...
Collecting 2-grams...
Collecting 3-grams...
Collecting 4-grams...
{I'm=16659, get=13033, go=12356, like=11644, work=11145, got=9056, going=8929, day=8816, don't=8420, back=8236, want=8200, miss=7995, really=7944, can't=7887, -=7779, still=7059, it's=6947, im=6907, today=6802, last=6468}
{last night=1584, feel like=1438, wish could=1291, want go=1157, don't know=1026, don't want=1018, I'm going=957, go back=936, wanna go=891, I'm sorry=773, looks like=701, . =653, looking forward=639, can't believe=631, don't think=624, didn't get=608, right now.=565, go work=558, last day=540, I'm gonna=536}
{. . =257, don't want go=225, wish could go=172, find good home.=158, I'm gonna miss=156, lost. Please help=156, Please help find=156, help find good=156, want go back=151, I'm going miss=141, I'm sorry hear=129, don't wanna go=129, wanna go back=119, go back sleep=114, hope feel better=107, don't feel like=105, really don't want=99, wanna go home=96, want go work=93, don't feel good=91}
{lost. Please help find=156, Please help find good=156, help find good home.=156, cant afford see Angels=86, afford see Angels Demons=86, see Angels Demons watched=86, Angels Demons watched free:=86, Demons watched free: http://tr.im/lvBu=86, im lonely keep company!=68, lonely keep company! 22=68, keep company! 22 female=68, . . . =55, @tommcfly plz say &quot;Happy=47, plz say &quot;Happy Birthday=47, say &quot;Happy Birthday Roni=47, &quot;Happy Birthday Roni &amp;=47, Birthday Roni &amp; Mickey!=47, Roni &amp; Mickey!=47, plz=47, &amp; Mickey!=47, plz plz=47, Mickey!=47, plz plz=47}
[sav339@log-3 bds]$

```

## 1-grams

- I'm=16659,
- get=13033,
- go=12356,
- like=11644,
- work=11145,
- got=9056,
- going=8929,
- day=8816,
- don't=8420,
- back=8236,
- want=8200,
- miss=7995,
- really=7944,
- can't=7887,
- -=7779,
- still=7059,
- it's=6947,
- im=6907,
- today=6802,
- last=6468

## 2-grams

- last night=1584,
- feel like=1438,
- wish could=1291,
- want go=1157,
- don't know=1026,
- don't want=1018,
- I'm going=957,
- go back=936,
- wanna go=891,
- I'm sorry=773,
- looks like=701,
- .=653,
- looking forward=639,
- can't believe=631,
- don't think=624,
- didn't get=608,
- right now.=565,
- go work=558,
- last day=540,
- I'm gonna=536

### 3-grams

- ...=257,
- don't want go=225,
- wish could go=172,
- find good home.=158,
- I'm gonna miss=156,
- lost. Please help=156,
- Please help find=156,
- help find good=156,
- want go back=151,
- I'm going miss=141,
- I'm sorry hear=129,
- don't wanna go=129,
- wanna go back=119,
- go back sleep=114,
- hope feel better=107,
- don't feel like=105,
- really don't want=99,
- wanna go home=96,
- want go work=93,
- don't feel good=91

### 4-grams

- lost. Please help find=156,
- Please help find good=156,
- help find good home.=156,
- cant afford see Angels=86,
- afford see Angels Demons=86,
- see Angels Demons watched=86,
- Angels Demons watched free:=86,
- Demons watched free: <http://tr.im/lvBu>=86,
- im lonely keep company!=68,
- lonely keep company! 22=68,
- keep company! 22 female=68,
- ...=55,
- @tommcfly plz say &quot;Happy=47,
- plz say &quot;Happy Birthday=47,
- say &quot;Happy Birthday Roni=47,
- &quot;Happy Birthday Roni &amp;=47,
- Birthday Roni &amp; Mickey!&quot;=47,
- Roni &amp; Mickey!&quot; plz=47,
- &amp; Mickey!&quot; plz plz=47,
- Mickey!&quot; plz plz plz=47

## Nouns only

```
sav339@log-2:/scratch/sav339/keyword_extraction/bds
[sav339@log-2 bds]$ ./POSTagger.sh
Compiling
srnun: job 3924072 queued and waiting for resources
srnun: job 3924072 has been allocated resources
Note: POSTagger.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
Running
srnun: job 3924073 queued and waiting for resources
srnun: job 3924073 has been allocated resources
New parser from sentiment140.csv
Parsing CSV...
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator tokenize
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator ssplit
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator pos
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS tagger from edu/stanford/nlp/models/pos-tagger/english-left3words-distisim.tagger ... done [2.5 sec].
Extracting nouns...
Analyzing n-grams on nouns...===== ]
Collecting 1-grams...
Collecting 2-grams...
Collecting 3-grams...
Collecting 4-grams...
{Im=18436, cant=11477, day=11154, today=10470, work=8984, im=8460, time=7706, night=6221, amp=5378, tomorrow=5351, sleep=3944, morning=3888, school=3828, bed=3809, tonight=3754, week=3607, weekend=3093, way=3043, days=2861, people=2704}
{cant sleep=1061, Im Im=425, school tomorrow=403, work today=403, work tomorrow=399, mothers day=363, day today=357, day work=357, Im work=354, ice cream=248, today day=244, Im bed=229, work day=229, Mothers Day=229, Cant sleep=226, swine flu=224, hours sleep=206, day tomorrow=200, today Im=196, Im cant=192}
{cant Angels Demons=86, plz quotHappy Birthday=47, quotHappy Birthday Roni=47, Birthday Roni amp=47, Roni amp Mickeyquot=47, amp Mickeyquot plz=47, Mickeyquot plz plz=47, plz plz plz=47, work work work=33, cant sleep Im=31, mothers day mom=30, Im cant sleep=28, bed work tomorrow=25, mom mothers day=25, bank holiday weekend=24, Jay Lenos show=24, email Miley cant=23, ow ow ow=22, Sarah Connor Chronicles=22, day work tomorrow=22}
{plz quotHappy Birthday Roni=47, quotHappy Birthday Roni amp=47, Birthday Roni amp Mickeyquot=47, Roni amp Mickeyquot plz=47, amp Mickeyquot plz plz=47, Mickeyquot plz plz plz=47, ACE prize FAN MADE=21, prize FAN MADE vid=21, Tonight Show Jay Leno=18, plz plz plz x=18, ow ow ow ow=16, yesterday quotsonny chancequot brazil=13, croissant bread rolls bacon=12, bread rolls bacon eggs=12, rolls bacon eggs sausages=12, quotsonny chancequot brazil lt33=12, Terminator Sarah Connor Chronicles=11, Jay Leno Tonight Show=9, tom google man man=8, cant upload pics reason=8}
```

### 1-grams

- Im=18436,
- cant=11477,
- day=11154,
- today=10470,
- work=8984,
- im=8460,
- time=7706,
- night=6221,
- amp=5378,
- tomorrow=5351,
- sleep=3944,
- morning=3888,
- school=3828,
- bed=3809,
- tonight=3754,
- week=3607,
- weekend=3093,
- way=3043,
- days=2861,

- people=2704

## 2-grams

- cant sleep=1061,
- Im Im=425,
- school tomorrow=403,
- work today=403,
- work tomorrow=399,
- mothers day=363,
- day today=357,
- day work=357,
- Im work=354,
- ice cream=248,
- today day=244,
- Im bed=229,
- work day=229,
- Mothers Day=229,
- Cant sleep=226,
- swine flu=224,
- hours sleep=206,
- day tomorrow=200,
- today Im=196,
- Im cant=192

## 3-grams

- cant Angels Demons=86,
- plz quotHappy Birthday=47,
- quotHappy Birthday Roni=47,
- Birthday Roni amp=47,
- Roni amp Mickeyquot=47,
- amp Mickeyquot plz=47,
- Mickeyquot plz plz=47,
- plz plz plz=47,
- work work work=33,
- cant sleep Im=31,
- mothers day mom=30,
- Im cant sleep=28,
- bed work tomorrow=25,
- mom mothers day=25,
- bank holiday weekend=24,
- Jay Lenos show=24,
- email Miley cant=23,
- ow ow ow=22,
- Sarah Connor Chronicles=22,

- day work tomorrow=22

#### 4-grams

- plz quotHappy Birthday Roni=47,
- quotHappy Birthday Roni amp=47,
- Birthday Roni amp Mickeyquot=47,
- Roni amp Mickeyquot plz=47,
- amp Mickeyquot plz plz=47,
- Mickeyquot plz plz plz=47,
- ACE prize FAN MADE=21,
- prize FAN MADE vid=21,
- Tonight Show Jay Leno=18,
- plz plz plz x=18,
- ow ow ow ow=16,
- yesterday quotsonny chancequot brazil=13,
- croissant bread rolls bacon=12,
- bread rolls bacon eggs=12,
- rolls bacon eggs sausages=12,
- quotsonny chancequot brazil lt33=12,
- Terminator Sarah Connor Chronicles=11,
- Jay Leno Tonight Show=9,
- tom google man man=8,
- cant upload pics reason=8

## POS patterns

```
sav339@log-2:/scratch/sav339/keyword_extraction/bds
Collecting POS patterns...
Collecting adjective noun
Collecting noun verb
Collecting adverb verb=====
{last night=2350, last day=566, next week=542, Good morning=312, next time=298, last time=296, sore throat=280, long day=277, last week=274, first time=268}
{Im going=910, cant believe=849, cant find=695, cant get=661, cant wait=564, Ill be=557, Im gon=550, cant go=536, im going=523, cant be=331}
{nt have=2178, nt know=1982, nt want=1943, nt get=1541, just got=1084, nt think=1074, nt wan=785, nt feel=734, not going=606, nt work=591}
[sav339@log-2 bds]$
```



## Adjective noun

- last night=2350,
- last day=566,
- next week=542,
- Good morning=312,
- next time=298,
- last time=296,
- sore throat=280,
- long day=277,
- last week=274,
- first time=268

## Noun verb

- Im going=910,
- cant believe=849,
- cant find=695,
- cant get=661,
- cant wait=564,
- Ill be=557,
- Im gon=550,
- cant go=536,
- im going=523,
- cant be=331

## Adverb verb

- nt have=2178,
- nt know=1982,
- nt want=1943,
- nt get=1541,
- just got=1084,
- nt think=1074,
- nt wan=785,
- nt feel=734,
- not going=606,
- nt work=591

## Dependency parser

{nsubj, root, dobj}

```

[Select sav339@log-2:/scratch/sav339/keyword_extraction/bds]
[sav339@log-2 bds]$ ./DependencyParser.sh
Compiling
srun: job 4120329 queued and waiting for resources
srun: job 4120329 has been allocated resources
Note: DependencyParser.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
Running
srun: job 4120330 queued and waiting for resources
srun: job 4120330 has been allocated resources
New parser from sentiment140.csv
Parsing CSV...
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator tokenize
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator ssplit
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator pos
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS tagger from edu/stanford/nlp/models/pos-tagger/english-left3words-distisim.tagger ... done [1.5 sec].
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator depparse
[main] INFO edu.stanford.nlp.parser.nnep.DependencyParser - Loading depparse model: edu/stanford/nlp/models/parser/nnep/english_UD.gz ... Time elapsed: 1.7 sec
Collecting dependency patterns...
Collecting nsubj root dobj
[main] INFO edu.stanford.nlp.parser.nnep.Classifier - PreComputed 20000 vectors, elapsed Time: 2.93 sec
[main] INFO edu.stanford.nlp.parser.nnep.DependencyParser - Initializing dependency parser ... done [4.7 sec].
{}===== ]
[sav339@log-2 bds]$
```

No results found (will try again with {nsubj, root, obj})

{nsubj, root, obj}

```

[sav339@log-2:/scratch/sav339/keyword_extraction/bds]
[sav339@log-2 bds]$ [sav339@log-2 bds]$ ./DependencyParser.sh
Compiling
srun: job 4121402 queued and waiting for resources
srun: job 4121402 has been allocated resources
Note: DependencyParser.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
Running
srun: job 4121403 queued and waiting for resources
srun: job 4121403 has been allocated resources
New parser from sentiment140.csv
Parsing CSV...
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator tokenize
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator ssplit
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator pos
[main] INFO edu.stanford.nlp.tagger.maxent.MaxentTagger - Loading POS tagger from edu/stanford/nlp/models/pos-tagger/english-left3words-distisim.tagger ... done [1.4 sec].
[main] INFO edu.stanford.nlp.pipeline.StanfordCoreNLP - Adding annotator depparse
[main] INFO edu.stanford.nlp.parser.nnep.DependencyParser - Loading depparse model: edu/stanford/nlp/models/parser/nnep/english_UD.gz ... Time elapsed: 1.6 sec
[main] INFO edu.stanford.nlp.parser.nnep.Classifier - PreComputed 20000 vectors, elapsed Time: 1.529 sec
[main] INFO edu.stanford.nlp.parser.nnep.DependencyParser - Initializing dependency parser ... done [3.1 sec].
Collecting dependency patterns...
Collecting nsubj root obj
{I miss you=220, i miss you=174, I hate it=112, I missed it=55, i love you=50, i hate it=46, i missed it=43, I love you=40, I miss him=37, i miss him=34}
[sav339@log-2 bds]$
```

- ## Named entity recognition

[illegible]

- Ive=570,
- David=230,
- Jay=199,
- Adam=194.

- Susan=174,
- Leno=152,
- Boyle=149,
- Danny=140,
- Shes=112,
- Cook=111

## Organizations

- Twitter=469,
- Hes=179,
- Facebook=155,
- Cavs=155,
- Google=138,
- House=108,
- BGT=98,
- Starbucks=82,
- Lakers=80,
- facebook=77

## Places

- Ill=2198,
- Miss=340,
- OH=235,
- NT=186,
- NY=180,
- New=114,
- DC=113,
- ME=109,
- California=102,
- OK=88

## Best method on positive and negative tweets

We claim that the best method is the Named Entity Recognition. We thus run it on the positive and negative subsets:

[illegible]

- Ive=570,
- David=230,
- Jay=199,
- Adam=194,
- Susan=174.

- Leno=152,
- Boyle=149,
- Danny=140,
- Shes=112,
- Cook=111

#### Organizations

- Twitter=469,
- Hes=179,
- Facebook=155,
- Cavs=155,
- Google=138,
- House=108,
- BGT=98,
- Starbucks=82,
- Lakers=80,
- facebook=77

#### Places

- Ill=2198,
- Miss=340,
- OH=235,
- NT=186,
- NY=180,
- New=114,
- DC=113,
- ME=109,
- California=102,
- OK=88