

# Data Intensive Computing

## CSE 587

### LAB – 2 Report

#### **Part1: Data Collection using Twitter, NYTimes, Common Crawler**

Topic: Sports

Subtopic: Football, Baseball, Golf, Basketball, Boxing

Step1: Dataset- I

	Twitter[20k]	NewYorktimes[500]	CommonCrawl[500]
Football	4k	100	150
Baseball	4k	100	150
Golf	4k	100	150
Basketball	4k	100	150
Boxing	4k	100	150

Step2: Dataset- II Larger data – Week Data

	Twitter[40k]	NewYorktimes[1000]	CommonCrawl[5000] appx.
Football	8k	200	>1000
Baseball	8k	200	>1000
Golf	8k	200	>1000
Basketball	8k	200	>1000
Boxing	8k	200	>1000

Step3: Run the following python scripts to collect and pre-process (data sets into words, remove stop words, stem the words (ex: running to run) and reduce will count the useful words) the Twitter, NYT, CommonCrawl data.

`/snehamupLab2/part1/Code/ Datacollection.ipynb`  
`/snehamupLab2/part1/Code/CommonnCrawl.ipynb`  
`/snehamupLab2/part1/Code/CommonCrawl_WeekData.ipynb`

Step4: Data collected, pre-processed and stored in following locations:

- `/snehamupLab2/part1/Data/Twitter/`  
- `/snehamupLab2/part1/Data/NYT/`  
- `/snehamupLab2/part1/Data/Commoncrawl/`

#### **Part 2: Running Hadoop Infrastructure, mapper and reducer on docker:**

Step 1: We have installed docker using the documentation provided in lab2 other materials.

Step 2: After installing docker, we have pulled the cloudera image.

Step 3: After pulling cloudera image, we have configured docker to run the provided mapper and reducer.

Step 4: Create required directories in HDFS to put the files.

Step 5: Now download the required files to input “`/user/arpitrai/MR/input/`” using the following command: `curl https://cse.buffalo.edu/~bina/cse487/spring2019/Lectures/Lab2/data/pg345.txt > pg345.txt`

```

File Edit View Search Terminal Help
[root@quickstart data]# curl https://cse.buffalo.edu/~bina/cse487/spring2019/Lectures/Lab2/data/pg345.txt > pg345.txt
  % Total    % Received % Xferd  Average Speed   Time     Time   Current
          Dload  Upload Total Spent   Left Speed
100  862k  100  862k    0      0  1898k      0 --:--:-- --:--:-- 3382k
[root@quickstart data]#

```

```

[root@quickstart data]# hadoop fs -put pg345.txt /user/arpitrai/MR/input/
[root@quickstart data]#

```

Step 6: Now make sure to delete the existing output directory from “/user/arpitrai/MR/output/” else the script won’t run.

```

[root@quickstart data]# hadoop fs -rm -r /user/arpitrai/MR/output
Deleted /user/arpitrai/MR/output

```

Step 7: Now run the script: hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/mapper.py -mapper /src/mapper.py -file /src/reducer.py -reducer /src/reducer.py -input /user/arpitrai/MR/input/pg345.txt -output /user/arpitrai/MR/output

```

[root@quickstart data]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/mapper.py -mapper /src/mapper.py -file /src/reducer.py -reducer /src/reducer.py -input /user/arpitrai/MR/input/pg345.txt -output /user/arpitrai/MR/output
19/04/21 19:12:24 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar(/src/mapper.py /src/reducer.py [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob6455503261494490782.jar tmpDir=null
19/04/21 19:12:25 INFO Client.RMProxy: Connecting to resourceManager at /0.0.0.0:8082
19/04/21 19:12:25 INFO Client.RMProxy: Connecting to resourceManager at /0.0.0.0:8082
19/04/21 19:12:25 INFO mapred.FileInputFormat: Total input paths to process : 1
19/04/21 19:12:25 INFO mapred.JobSubmitter: number of splits:2
19/04/21 19:12:26 INFO mapreduce.Job: Submitting tokens for job: job_1555871009350_0012
19/04/21 19:12:26 INFO impl.YarnClientImpl: Submitted application application_1555871009350_0012
19/04/21 19:12:26 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1555871009350_0012/
19/04/21 19:12:26 INFO mapreduce.Job: Running job: job_1555871009350_0012
19/04/21 19:12:30 INFO mapreduce.Job: Job job_1555871009350_0012 running in uber mode : false
19/04/21 19:12:30 INFO mapreduce.Job: map 0% reduce 0%
19/04/21 19:12:34 INFO mapreduce.Job: map 50% reduce 0%
19/04/21 19:12:35 INFO mapreduce.Job: map 100% reduce 0%
19/04/21 19:12:49 INFO mapreduce.Job: map 100% reduce 100%
19/04/21 19:12:51 INFO mapreduce.Job: Job job_1555871009350_0012 completed successfully
19/04/21 19:12:51 INFO mapreduce.Job: Counters:
  File System Counters:
    FILE: Number of bytes read=1515580
    FILE: Number of bytes written=3382645
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=887488
    HDFS: Number of bytes written=198227
    HDFS: Number of read operations=9
    HDFS: Number of large read operations=2
    HDFS: Number of write operations=2
  Job Counters:
    Launched map tasks=2
    Launched reduce tasks=1
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=4114
    Total time spent by all reduces in occupied slots (ms)=3021
    Total time spent by all map tasks (ms)=4114
    Total time spent by all reduce tasks (ms)=3021
    Total vcore-seconds taken by all map tasks=4114
    Total vcore-seconds taken by all reduce tasks=3021
    Total auxiliary-seconds taken by all map tasks=4247726

```

Step 8: Now transfer the files, to local directory:

```

19/04/21 19:12:41 INFO streaming.StreamJob: Output directory: /user/arpitrai/MR/output
[root@quickstart data]# hadoop fs -cat /user/arpitrai/MR/output/part-00000 > pg345_output.txt
[root@quickstart data]#

```

-snehamupLab2/part2/pg345\_prepro\_output.csv

### Part3: Word Count

Step1: Run mapper.py and reducer.py similar to part2 from

- snehamupLab2/part3/Twitter/Code/reducer.py
- snehamupLab2/part3/Twitter/Code/mapper.py

- snehamupLab2/part3/NYT/Code/reducer.py  
- snehamupLab2/part3/NYT/Code/mapper.py

- snehamupLab2/part3/Commoncrawl/Code/reducer.py  
- snehamupLab2/part3/ Commoncrawl/Code/mapper.py

Step2: Output of word count is stored in following

```
/snehamupLab2/part3/Twitter/Code/WordCount/*  
/snehamupLab2/part3/ NYT /Code/WordCount/*  
/snehamupLab2/part3/ Commoncrawl /Code/WordCount/*
```

### **Part4: Word Count- larger data**

Step1: Now, we run mapper.py and reducer.py similar to part2 on the larger data present in

```
/snehamupLab2/part1/Data/Twitter/TwitterIweek_Full/*  
/snehamupLab2/part1/Data/NYT/NYtimes_Iweekfull/*  
/snehamupLab2/part1/Data/Commoncrawl/CC_Iweekfull/*
```

*Step2: Word count output is stored in*

```
/snehamupLab2/part3/Twitter/Code/WordCountIweek_full/*  
/snehamupLab2/part3/NYT/Code/WordCountIweek/*  
/snehamupLab2/part3/Commoncrawl/Code/WordCountIweek/*
```

### **Part 5: Word co-occurrence**

Step1: Get the top 10 words from word count output using sort.ipynb.

```
/snehamupLab2/part1/Code/sort.ipynb
```

Step 2: Run the mapper.py and reducer.py for word co-occurrence from following locations

```
/snehamupLab2/part3/Twitter/Code/co_baseball_mapper.py  
/snehamupLab2/part3/Twitter/Code/co_basketball_mapper.py  
/snehamupLab2/part3/Twitter/Code/co_boxing_mapper.py  
/snehamupLab2/part3/Twitter/Code/co_football_mapper.py  
/snehamupLab2/part3/Twitter/Code/co_golf_mapper.py
```

```
/snehamupLab2/part3/NYT/Code/co_baseball_nytimes_mapper.py  
/snehamupLab2/part3/NYT/Code/co_basketball_nytimes_mapper.py  
/snehamupLab2/part3/NYT/Code/co_boxing_nytimes_mapper.py  
/snehamupLab2/part3/NYT/Code/co_football_nytimes_mapper.py  
/snehamupLab2/part3/NYT/Code/co_golf_nytimes_mapper.py
```

```
/snehamupLab2/part3/Commoncrawl/Code/co_baseball_cc_mapper.py  
/snehamupLab2/part3/Commoncrawl/Code/co_basketball_cc_mapper.py  
/snehamupLab2/part3/Commoncrawl/Code/co_boxing_cc_mapper.py  
/snehamupLab2/part3/Commoncrawl/Code/co_football_cc_mapper.py  
/snehamupLab2/part3/Commoncrawl/Code/co_golf_cc_mapper.py
```

Step 3: Outputs are stored in

```
/snehamupLab2/part3/Twitter/Code/WordCo_occurrence/*  
/snehamupLab2/part3/NYT/Code/WordCo-occurrence/*  
/snehamupLab2/part3/Commoncrawl/Code/*
```

## **Part 6: Analyse and visualize using Tableau:**

Step 1: File-> open (txt or csv)-> chose the F1 and F2 and visualize and save the image.

Step 2: Images and .twb files are stored in image folders in respective data folders.

Step 3: snehamupLab2/Dashboard.twb contains visual display all 45 word clouds.

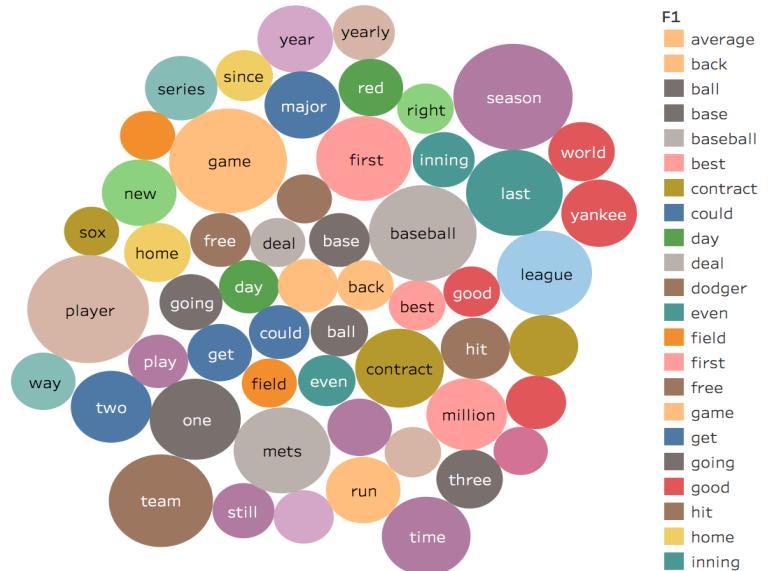
*/snehamupLab2/part3/Twitter /Images/  
/snehamupLab2/part3/NYT /Images/  
/snehamupLab2/part3/Commoncrawl/Images/*

Analysis:

1. The top word count across different sources is nearly same respective to each sub-groups.
2. Word co-occurrence for different data sources is also nearly the same for each of the sub-groups.
3. In golf and boxing, trump is coming as the top ten words for few sources.

Word Count [Output Images]:

## BaseballINYtimesdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 123.

F1
average
back
ball
base
baseball
best
contract
could
day
deal
dodger
even
field
first
free
game
get
going
good
hit
home
inning
last
league
major
manager
mets
million
national
new
one
pitcher
play
player
red
right
robinson
run
salary
season
seaver
series
since
sox
still
team
think
three
time
two
way
world
yankee
year
yearly

# BaseballINYtimes-data

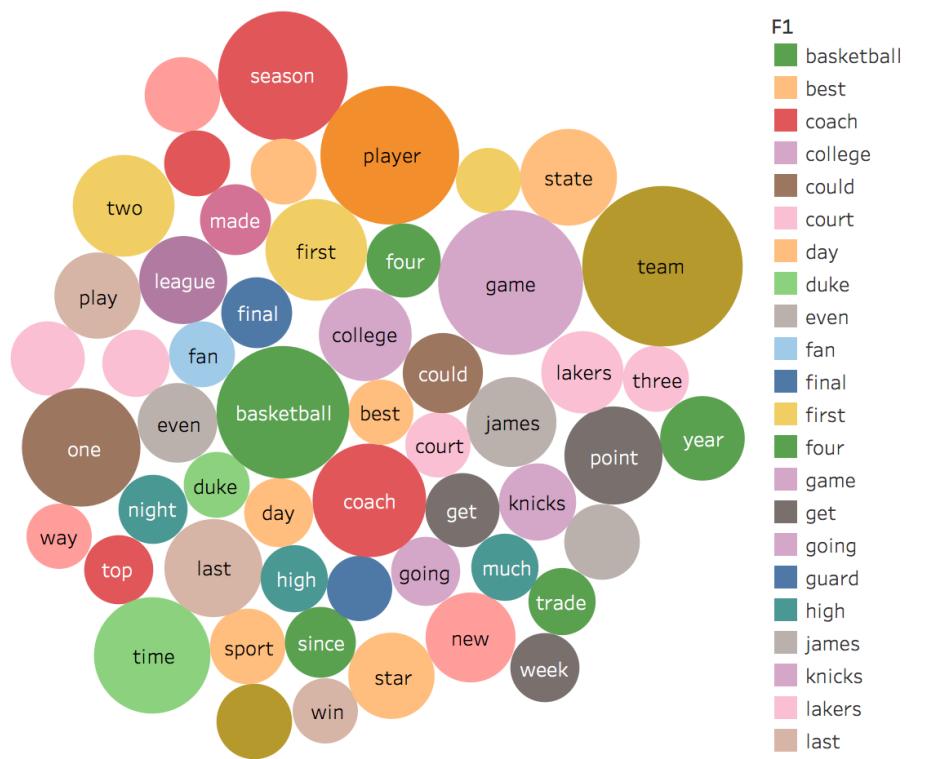
hit home yearly think three way field base best  
deal robinson one million seaver world  
back league manager yankee pitcher game st  
contract free season last sox player  
even first team going baseball mets t  
daytime national average major series re  
get good new dodger salary since right could two  
play

F1. Color shows details about F1.  
Size shows sum of F2. The view is  
filtered on sum of F2, which  
includes values greater than or  
equal to 123.

## F1

- average
- back
- ball
- base
- baseball
- best
- contract
- could
- day
- deal
- dodger
- even
- field
- first
- free
- game
- get
- going
- good
- hit
- home
- inning
- last
- league
- major
- manager
- mets
- million
- national
- new
- one
- pitcher
- play
- player
- red
- right
- robinson

## BasketballNYtimesdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 96.

F1
basketball
best
coach
college
could
court
day
duke
even
fan
final
first
four
game
get
going
guard
high
james
knicks
lakers
last
league
made
much
new
night
one
people
play
player
playoff
point
school
season
second
since
sport
star
state
team
three
time
top
tournament
trade
two
university
way
week
williamson
win
woman
year

## GolfNYtimesdata



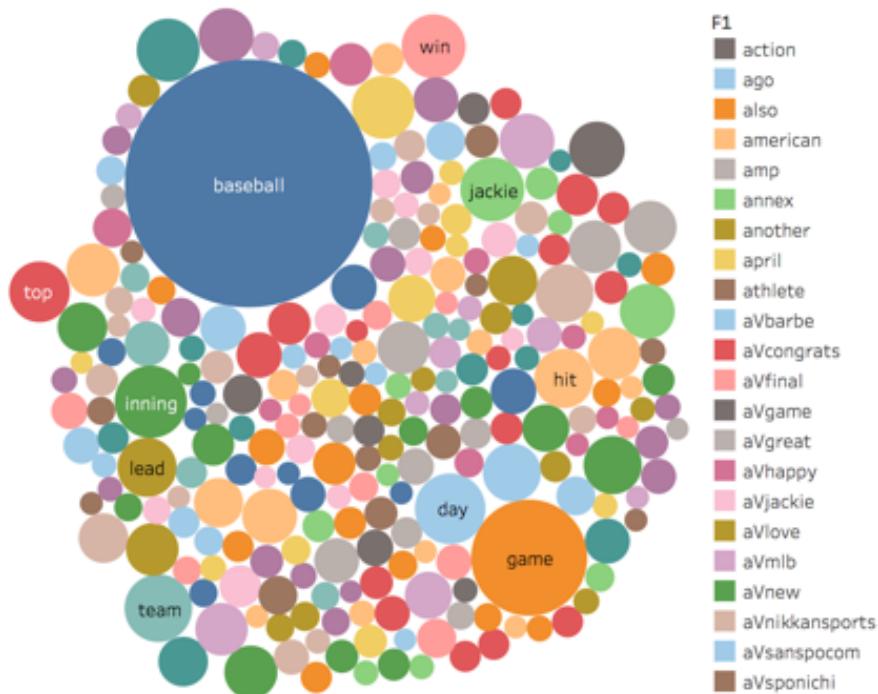
## Football|NYtimesdata

later think  
goingmade day title end giant much still right make  
pro case people two owner england another second head  
hit played patriot new coach kraft million former quarterback  
**game**team run football fan player ram season  
city super kaepernick sport one could year back league last  
field brady bowl time state defensive three many college even world  
way month long week including yard since school get woman soccer  
high



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 61.

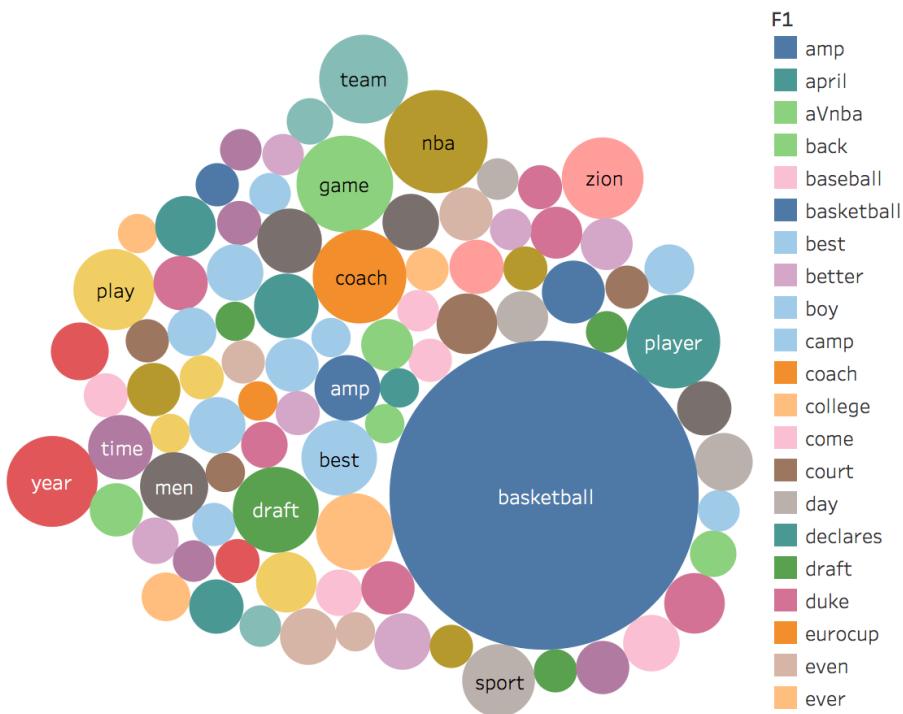
## BaseballTweetdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 17.

- F1
- action
  - ago
  - also
  - american
  - amp
  - annex
  - another
  - april
  - athlete
  - aVbarbe
  - aVcongrats
  - aVfinal
  - aVgame
  - aVgreat
  - aVhappy
  - aVjackie
  - aVlove
  - aVmib
  - aVnew
  - aVnikkansports
  - aVsanspocom
  - aVsponichi
  - aVthank
  - aVtoday
  - aVvarsity
  - away
  - back
  - bad
  - ball
  - barrier
  - base
  - baseball
  - basketball
  - bat
  - batting
  - beat
  - best
  - better
  - big
  - black
  - blue
  - bottom
  - boy
  - break
  - bring
  - cap
  - card
  - cardinal
  - center
  - central
  - change
  - changed
  - cheech
  - chrisba
  - christian
  - city
  - coach
  - college
  - color
  - come
  - coming
  - complete
  - contract
  - county
  - cub

## BasketballTweetdata

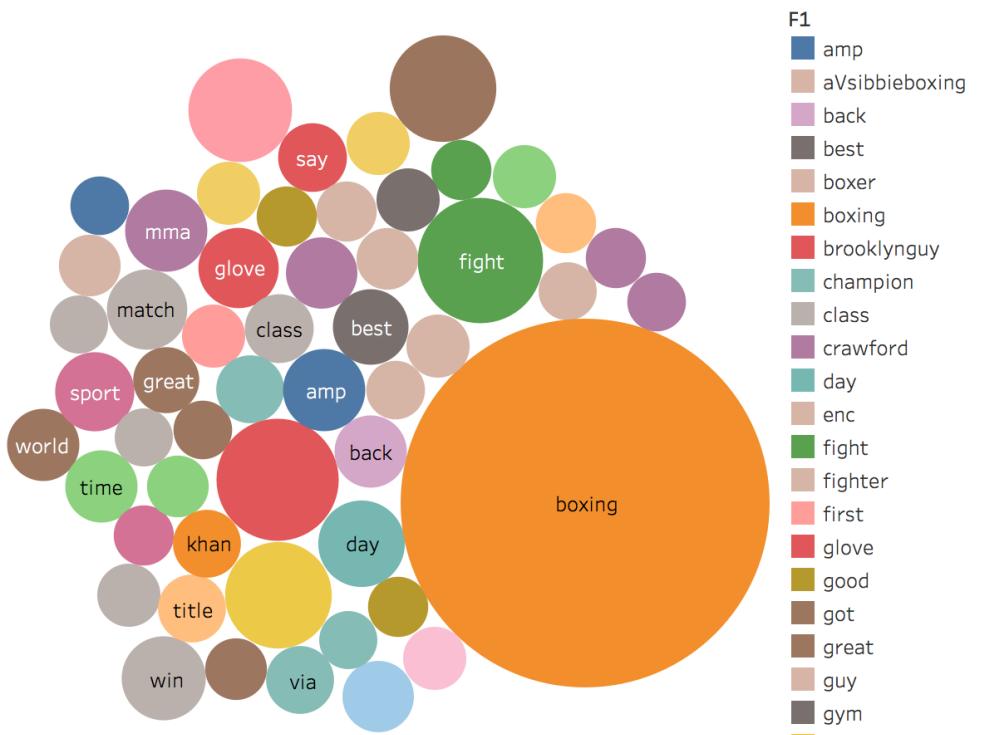


F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 43.

### F1

- amp
- april
- aVnba
- back
- baseball
- basketball
- best
- better
- boy
- camp
- coach
- college
- come
- court
- day
- declares
- draft
- duke
- eurocup
- even
- ever
- fan
- final
- first
- football
- game
- get
- girl
- going
- good
- got
- great
- guard
- high
- know
- let
- life
- look
- love
- make
- man
- men
- much
- national
- nba
- ncaa
- need
- net
- new
- news
- next
- people
- pick
- play
- player
- playing
- playoff
- really

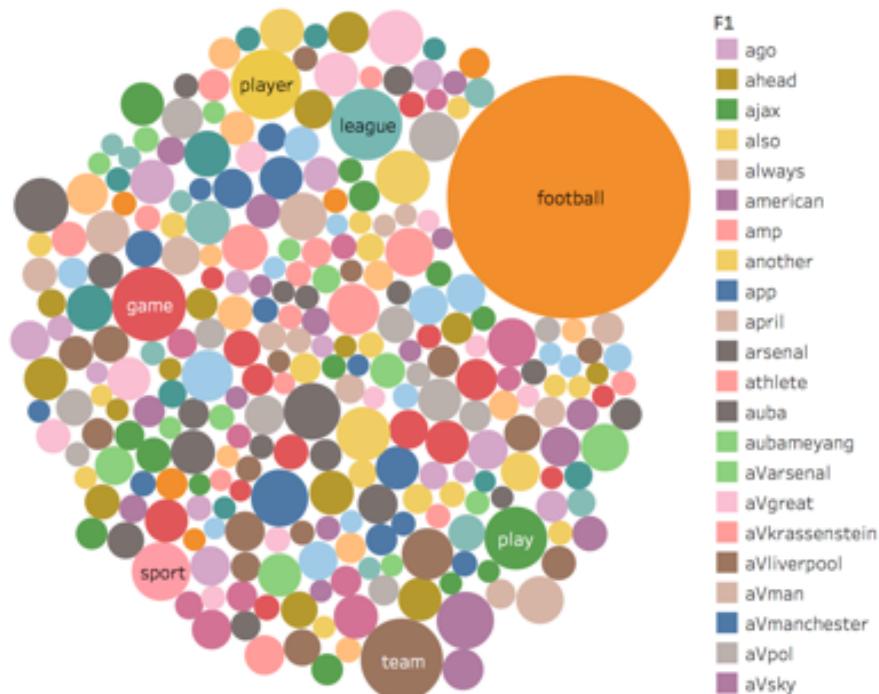
## BoxingTweetdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 53.

F1
amp
aVsibbieboxing
back
best
boxer
boxing
brooklynguy
champion
class
crawford
day
enc
fighter
first
glove
good
got
great
guy
gym
jackalterr
khan
kick
know
knstuff
make
match
mma
move
new
night
really
round
say
see
sibbieboxing
smokinyngbruno
sport
take
think
time
title
today
training
typhilly
via
week
win
woman
world
would
year

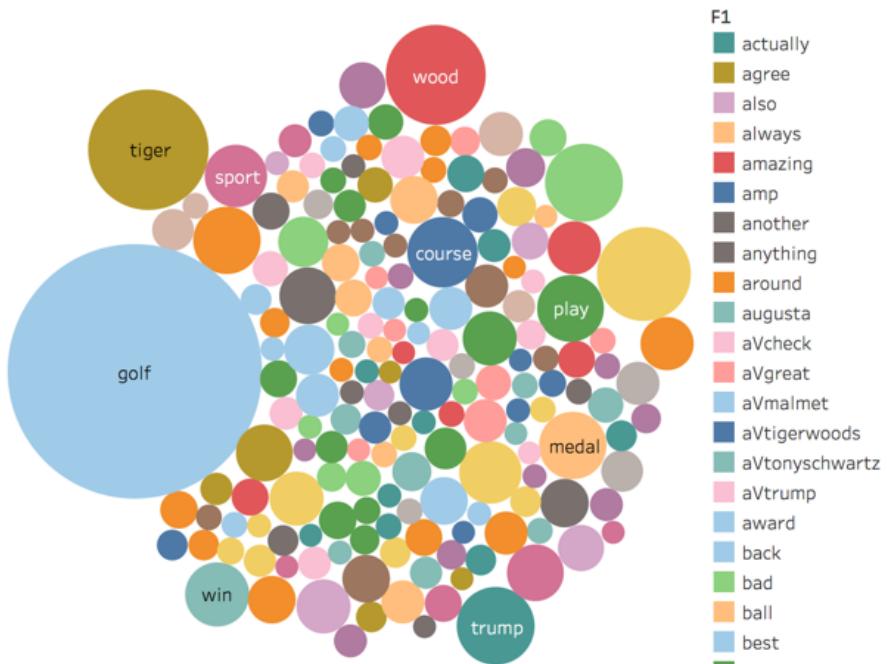
## FootballTweetdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 17.

F1
ago
ahead
ajax
also
always
american
amp
another
app
april
arsenal
athlete
auba
aubameyang
aVarsenal
aVgreat
aVkrassenstein
aVliverpool
aVman
aVmanchester
aVpol
aVsky
aVtroy
aVwatford
away
back
bad
ball
barcelona
baseball
basketball
beat
ben
beso
best
better
big
boy
call
camp
card
cardiff
care
cathedral
champion
check
chelsea
chrisporterfcn
city
class
clear
club
coach
college
come
could
cristiano
cup
dame
day
deeney
del
draft
draw
emerick

## GolfTweetdata



F1. Color shows details about F1. Size shows sum of F2. The marks are labeled by F1. The view is filtered on sum of F2, which includes values greater than or equal to 23.

- F1
- actually
  - agree
  - also
  - always
  - amazing
  - amp
  - another
  - anything
  - around
  - augusta
  - aVcheck
  - aVgreat
  - aVmalmet
  - aVtigerwoods
  - aVtonyschwartz
  - aVtrump
  - award
  - back
  - bad
  - ball
  - best
  - bet
  - better
  - big
  - black
  - boy
  - business
  - care
  - cart
  - championship
  - cheat
  - cheated
  - cheating
  - check
  - club
  - come
  - comeback
  - could
  - country
  - course
  - day
  - del
  - done
  - even
  - event
  - ever
  - every
  - everyone
  - fan
  - first
  - freedom
  - friend
  - game
  - get
  - getting
  - give
  - going
  - golf
  - golfer
  - good
  - got
  - great
  - greatest
  - green
  - guy
  - happy

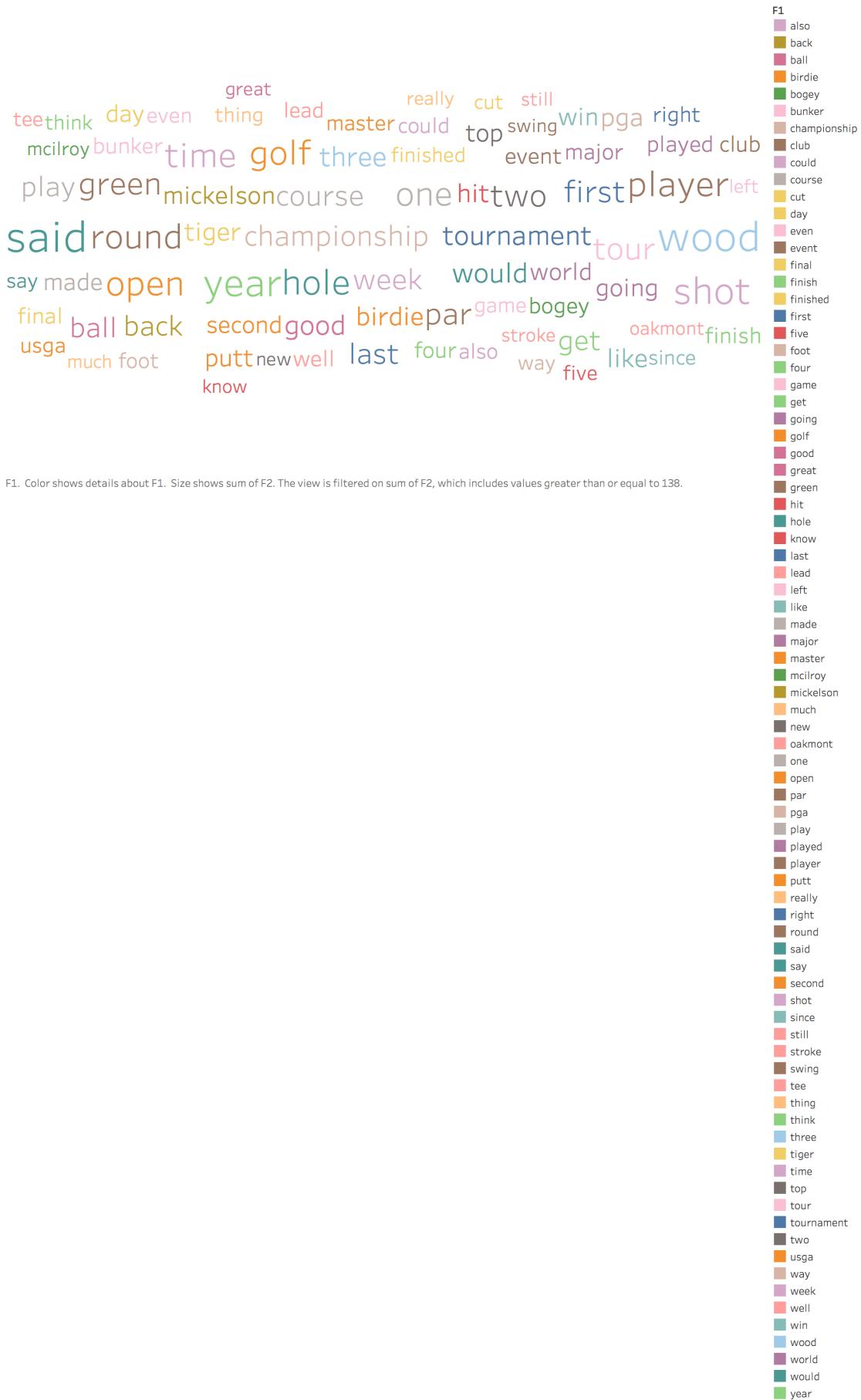
## Football\_CommonCrawl|WC

yard  
 going three night steelers jackson first free million rookie agent  
 cut released receiver four tackle round play right roster  
 last game preseason running training injury linebacker year  
 associated team season wide defensive nfl player  
 time quarterback practice would offensive coach said two  
 saturday back contract football thursday second  
 pick week camp injured veteran also brown raider draft  
 get end among new one

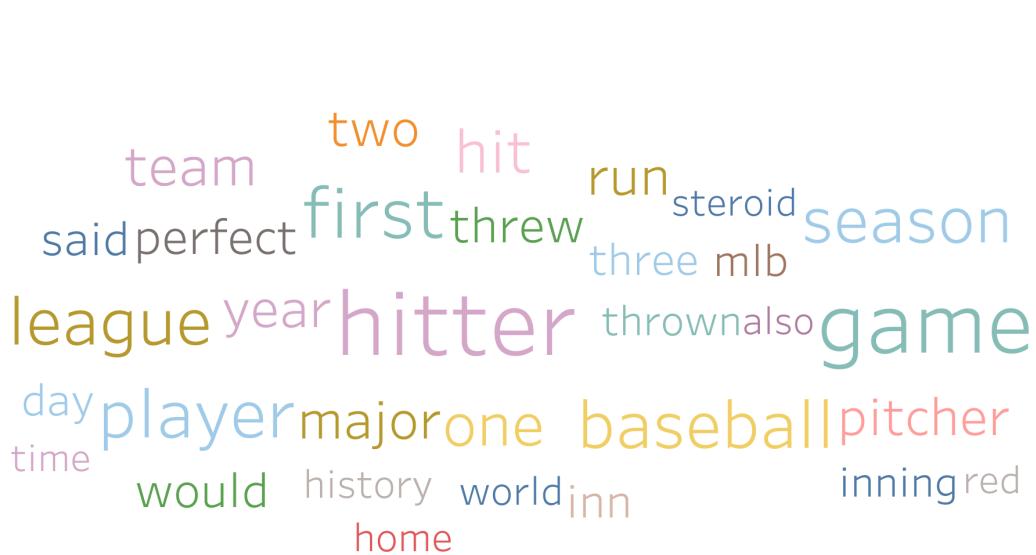


F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 21.

## Golf\_CommonCrawlWC



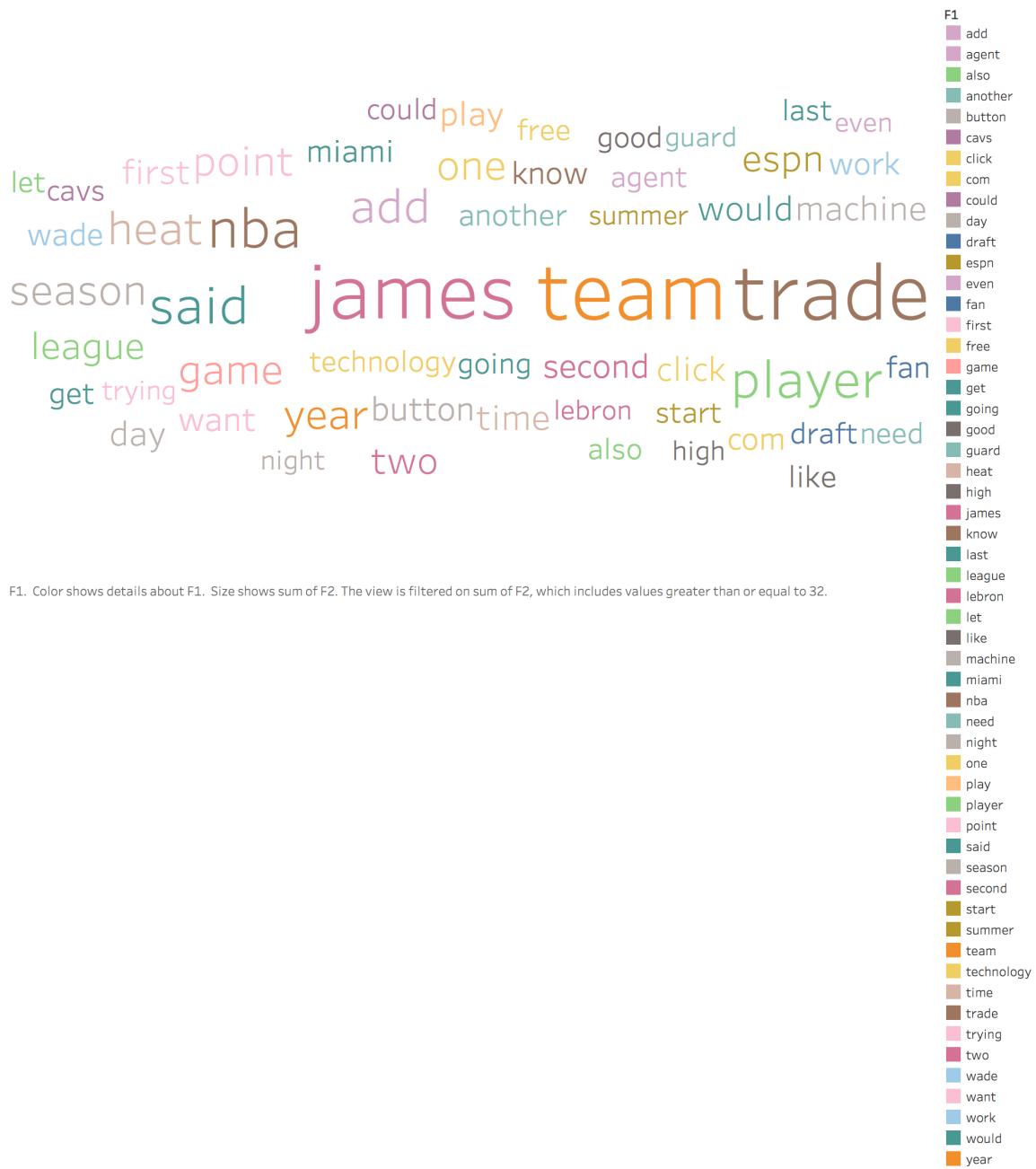
Baseball\_CommonCrawlIWC



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 31.

F1
also
baseball
day
first
game
history
hit
hitter
home
inn
inning
league
major
mlb
one
perfect
pitcher
player
red
run
said
season
steroid
team
three
threw
thrown
time
two
world
would
year

## Basketball\_CommonCrawlWC



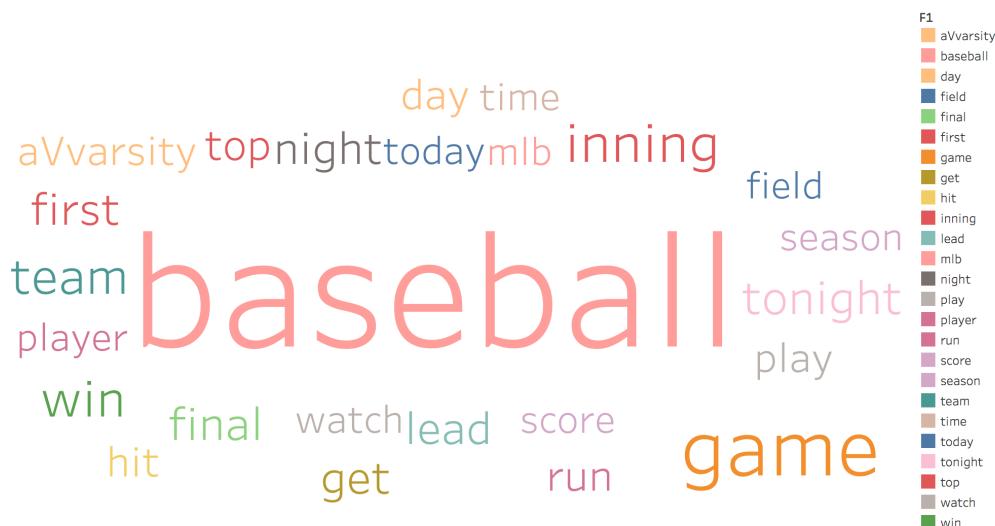
## Boxing\_CommonCrawlWC



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 235.

## Word Count larger data [Output Images]:

### BaseballTweet\_1week



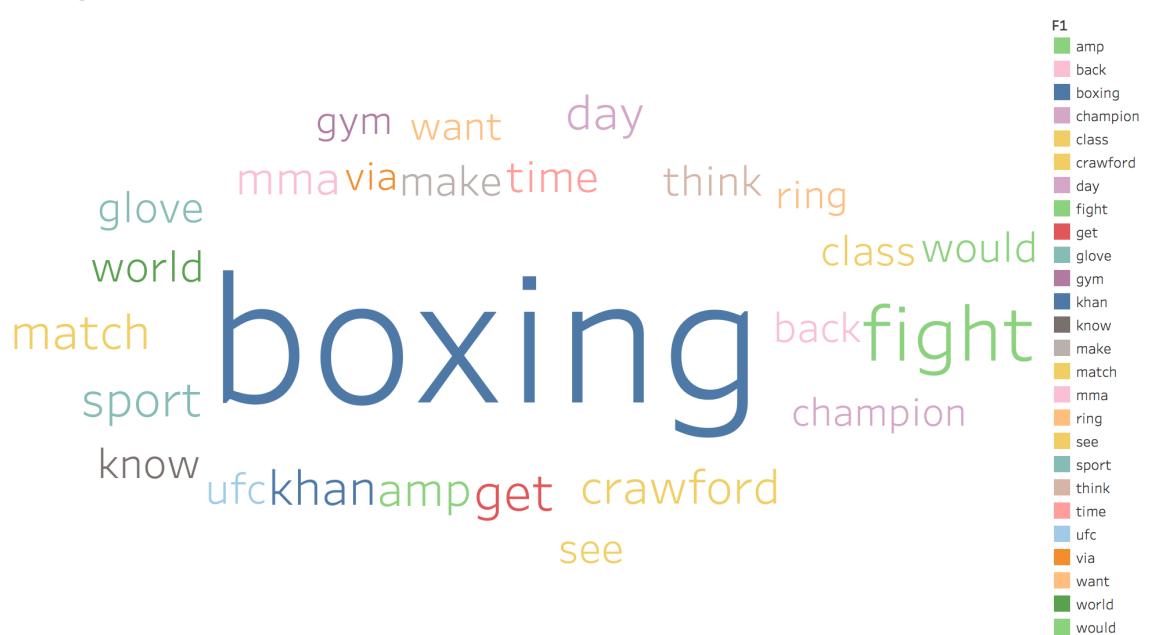
F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 165.

## BasketballTweet\_1week

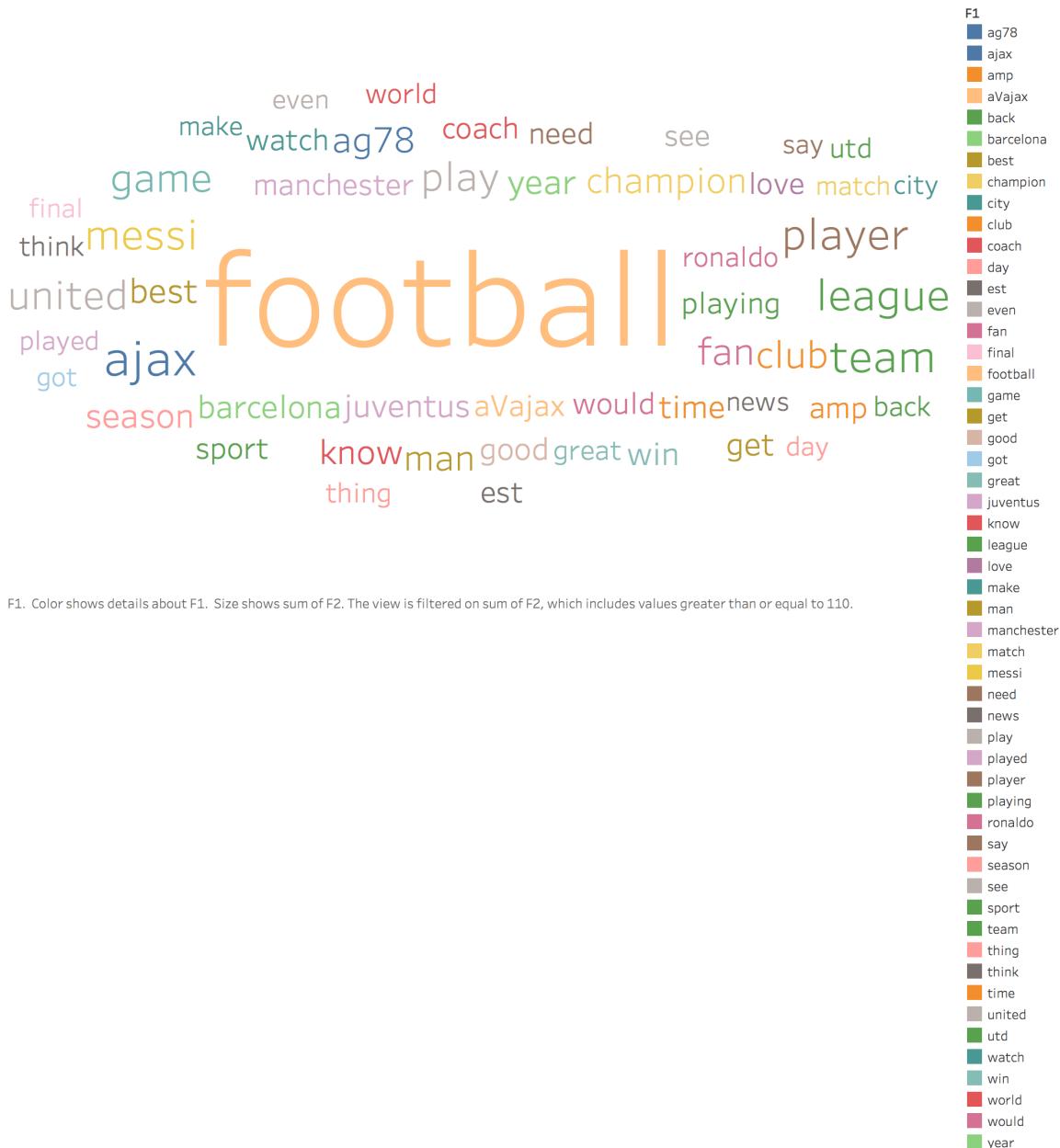


F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 202.

## BoxingTweet\_1week



## FootballTweet\_1week



GolfTweet\_1week

year look  
trump today game get timemaster  
club tiger  
tournament wood  
play tigerwoods team amp  
day

F1
amp
aVrealdonaldtrump
championship
club
course
day
game
get
golf
great
look
master
play
round
team
tiger
tigerwoods
time
today
tournament
trump
win
wood
year

F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 107.

## Baseball\_NYT1week



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 139.

F1
allowed
associated
avbyby
back
ball
base
baseball
career
day
double
eighth
era
field
fifth
first
five
four
fourth
game
gave
get
going
good
got
hit
home
homer
inning
last
lead
league
left
made
major
make
manager
mets
new
night
one
pitch
pitched
play
player
rbi
red
rhp
right
run
season
second
series
seven
single
six
sox
start
starter
straight
streak
struck
team
third
three
time
two
walk
went
win
yankee

Basketball NYT1week



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 58.

## Golf\_NYT1week

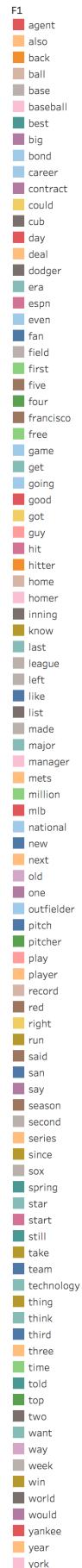
see take ever pga come hit nine still right fire little never bogey say  
 home play get state new made people winner right fire little never bogey say  
 putt record world president one old koepka sport lot lead way woman open could last playing  
 city day best time player first green six three hole sunday round  
**major national back wood tiger championship even master**  
 four shot victory tournament golf a v byby second week augusta since  
 par game going two year trump million played win another cup birdie course  
 many long good place great tour jacket make much title johnson really think behind know tee  
 left ball group got top club next five



BaseballCommonCrawl1week

told next free win a deal even take week ball got  
star four best also start know sox hitter third homer san  
record back career right time major good could spring three think fan  
pitch manager said home second cub series hit million team  
player league day season run baseball last game  
going technology first inning two top yankee would  
five like pitcher contract one dodger francisco made world year new  
want espn thing mlb list big left since old york national agent play mets way get base bond  
red  
say still

F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 2,261.



BasketCommonCrawl1week



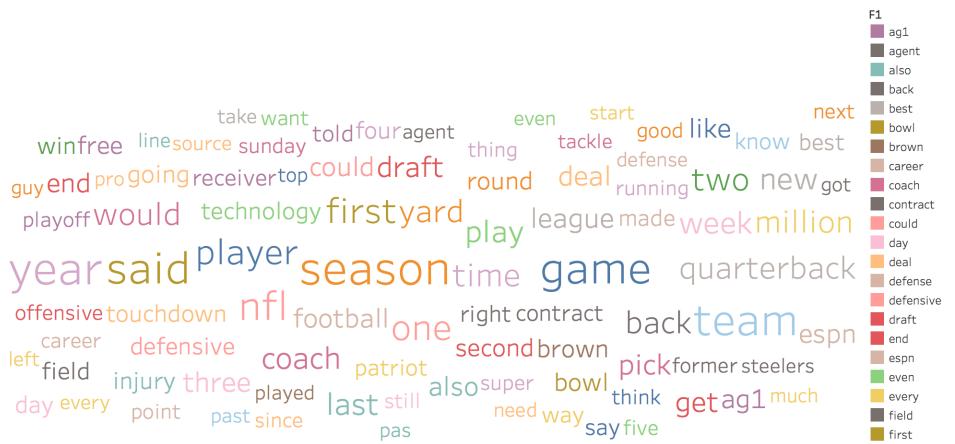
F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 2,415.

BoxingCommonCrawl1week



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 609.

## FootballCommonCrawl1week



F1. Color shows details about F1. Size shows sum of F2. The view is filtered on sum of F2, which includes values greater than or equal to 2,457.

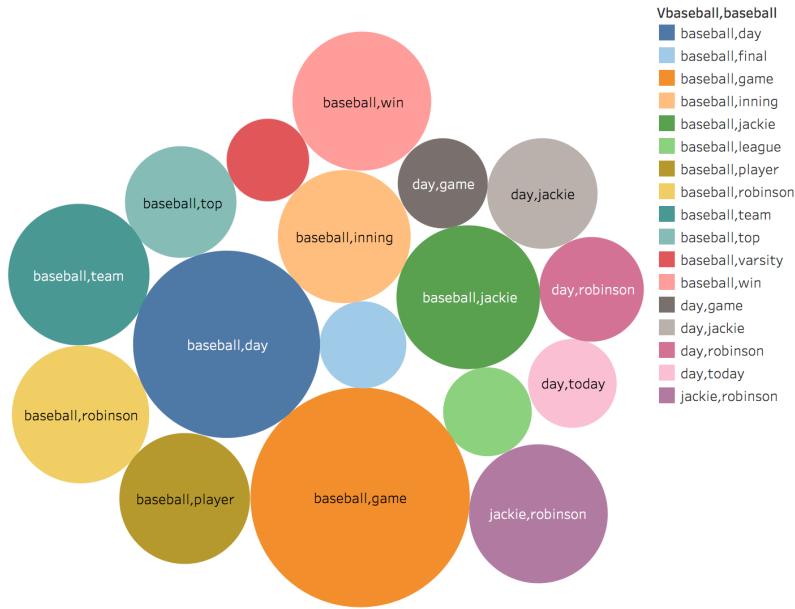
## GolfCommonCrawl1week

lot old want com ryder might yard tied went away long ago  
 next team take thing match well winner still foot way also behind career winning augusta national great  
 third game missed season event really another johnson since stroke good putt got  
 field saturday lead green mickelson cut made world playing major tiger  
 nine win open golf time course technology birdie mcilroy second shot  
 year player said championship pga wood last tournament  
 back top hole tour like first week champion one cup two round ag1  
 get bogey left would three five master going par tee final played start play tiger six  
 right think day victory finished know sunday finish ball could even spieh four thursday playoff  
 people golfer chance month club friday point best see took say hit come new never past around much title  
 place



## Word Co-occurrence [Output Images]::

Baseball\_Co-occurrence



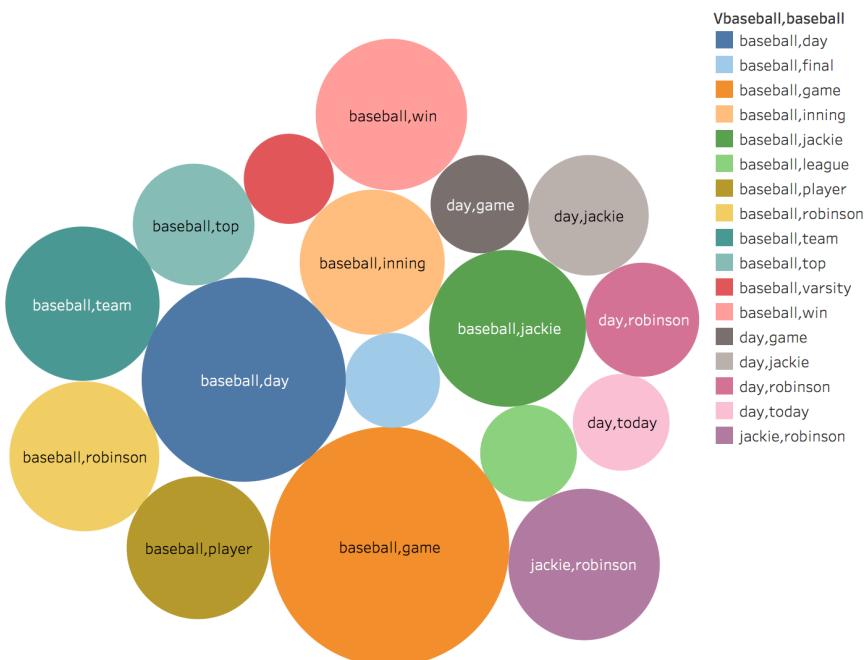
V<sub>baseball,baseball</sub>. Color shows details about V<sub>baseball,baseball</sub>. Size shows sum of 1. The marks are labeled by V<sub>baseball,baseball</sub>. The view is filtered on sum of 1, which includes values greater than or equal to 81.

Baseball\_Co-occurrence



V<sub>baseball,baseball</sub>. Color shows details about V<sub>baseball,baseball</sub>. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 81.

## Baseball\_Co



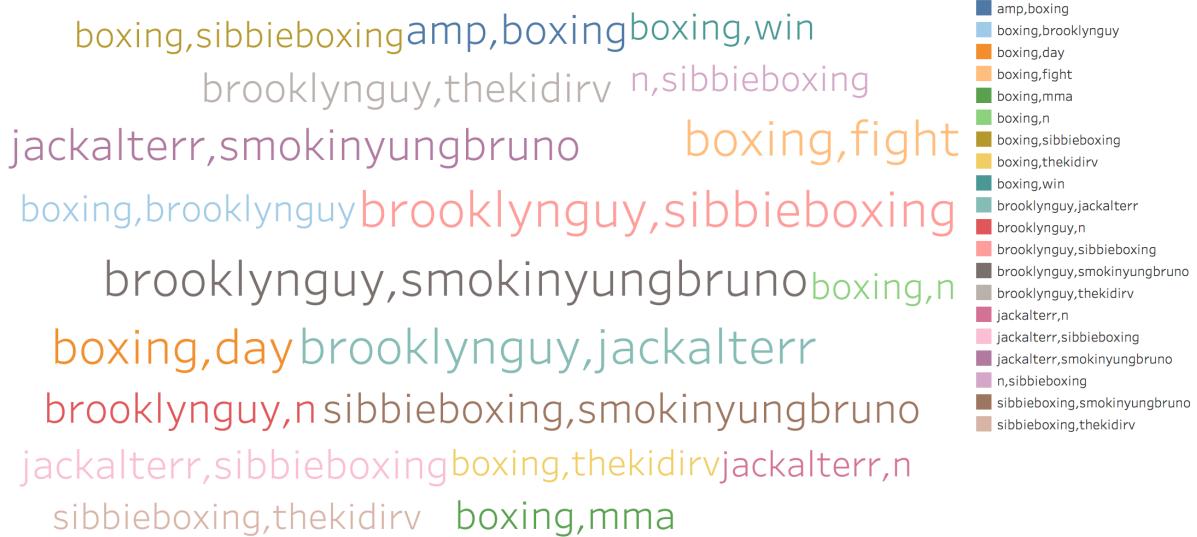
Vbaseball,baseball. Color shows details about Vbaseball,baseball. Size shows sum of 1. The marks are labeled by Vbaseball,baseball. The view is filtered on sum of 1, which includes values greater than or equal to 81.

## Basketball\_Co-occurrence



Vdetroitbadboys,basketball. Color shows details about Vdetroitbadboys,basketball. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 86.

BoxingTwitter\_Co-occurrence



A Vaaand,boxing. Color shows details about a Vaaand,boxing. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 167.

FootballTwitter\_Co-occurrence



Vshopping,football. Color shows details about Vshopping,football. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 59.

## GolfTwitter\_Co-occurrence



A<sub>Vabc,golf</sub>. Color shows details about a<sub>Vabc,golf</sub>. Size shows sum of 3. The view is filtered on sum of 3, which includes values greater than or equal to 264.

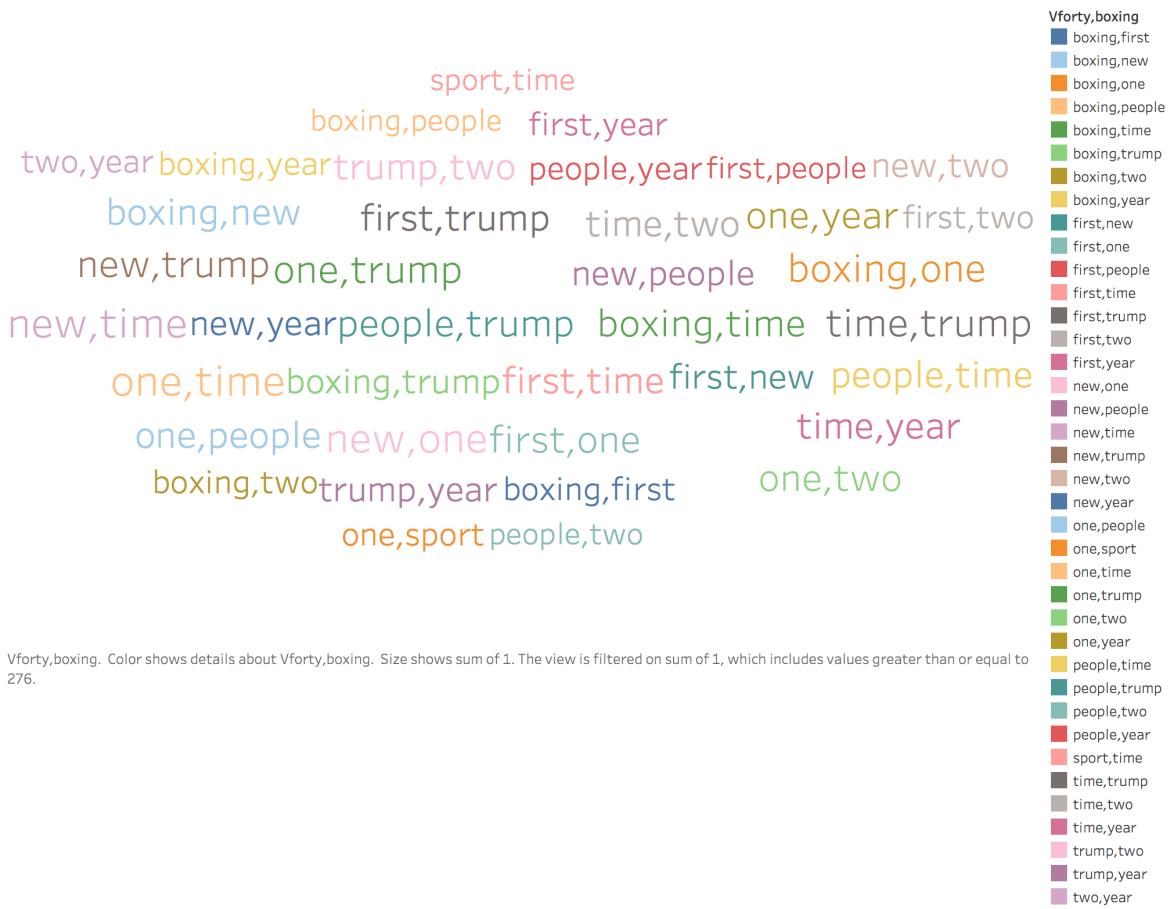
## BaseballNYTimes\_Co-occurrence



## BasketballNYTimes\_Co-occurrence



## BoxingNYTimes\_Co-occurrence



## FootballINYTimes\_Co-occurrence



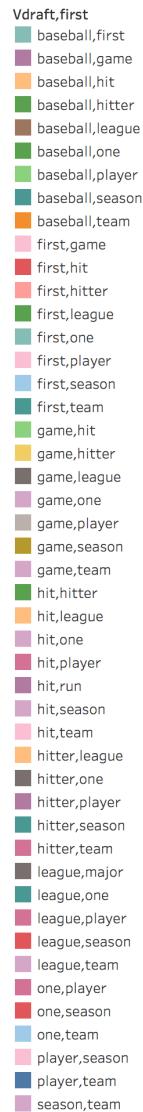
## GolfNYTimes\_Co-occurrence



## Baseball\_CommonCrawlCo-occ

one,season baseball,team league,one hitter,player baseball,hit  
 first,player hitter,teamfirst,league one,player baseball,one  
 hit,hitter game,league league,season game,team hit,one  
 game,season baseball,player baseball,game hitter,league  
 first,hitter hit,season game,hitter baseball,season hitter,season  
 first,game baseball,hitter baseball,first hit,run baseball,league  
 hit,league first,season league,player game,hit hitter,one  
 league,major league,team game,one game,player player,teamfirst,team  
 season,team player,season first,hit hit,player one,team  
 hit,team first,one

Vdraft,first. Color shows details about Vdraft,first. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 60.



## Basketball\_CommonCrawlCo-occ

baseball,hit baseball,one first,hit one,season hit,player one,team  
 league,major league,team first,league game,player league,one first,team  
 hit,hitter game,league hitter,team one,player league,season hit,one  
 game,season baseball,player baseball,game hitter,league  
 first,hitter hit,season game,hitter league,player hitter,season  
 first,game baseball,hitter hitter,one baseball,league  
 baseball,first first,season baseball,team baseball,season  
 first,player game,one game,team game,hit player,team first,one  
 season,team hit,team player,season hit,league

Vdraft,first
baseball,first
baseball,game
baseball,hit
baseball,league
baseball,one
baseball,player
baseball,season
baseball,team
first,game
first,hit
first,hitter
first,league
first,one
first,player
first,season
first,team
game,hit
game,hitter
game,league
game,one
game,player
game,season
game,team
hit,hitter
hit,league
hit,one
hit,player
hit,season
hit,team
hitter,league
hitter,one
hitter,player
hitter,season
hitter,team
league,major
league,one
league,player
league,season
league,team
one,player
one,season
one,team
player,season
player,team
season,team

Vdraft,first. Color shows details about Vdraft,first. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 63.

Boxing\_CommonCrawlCo-occ

one,round ko,said  
bradley,fight round,two  
round,said fight,first boxing,fight one,said  
fight,two fight,round fight,said  
fight,one round,title  
boxing,said first,round said,title said,two  
ko,round first,said

Vkubrat,boxing
boxing,fight
boxing,said
bradley,fight
fight,first
fight,ko
fight,one
fight,round
fight,said
fight,title
fight,two
first,round
first,said
ko,round
ko,said
one,round
one,said
round,said
round,title
round,two
said,title
said,two

Vkubrat,boxing. Color shows details about Vkubrat,boxing. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 966.

## Football\_Co-occurrence



Golf\_CommonCrawlCo-occ

one,open golf,said hole,one golf,one  
 hole,time hole,said time,year open,shot one,shot golf,time  
 golf,open player,shot open,year said,shot golf,shot shot,year  
 open,playertime,wood hole,player open,said said,year  
 player,said hole,wood player,year player,wood said,wood  
 open,wood player,time shot,wood wood,year  
 golf,player one,player one,said said,time one,wood  
 open,time golf,wood golf,year hole,shot hole,year  
 one,time one,year hole,open shot,time  
 golf,hole



AS'accepts,golf. Color shows details about aS'accepts,golf. Size shows sum of 1. The view is filtered on sum of 1, which includes values greater than or equal to 439.