



---

## LAB – 2

Data Intensive Computing (Spring 2019)

Paritosh Velalam, Gaurav Avula



**TABLE OF CONTENTS**

<b>OVERVIEW .....</b>	<b>3</b>
<b>HOW DID WE DO IT?.....</b>	<b>3</b>
<b>BIG DATA .....</b>	<b>4</b>
BIG DATA (WORD COUNT) .....	4
<i>Common Crawl</i> .....	4
<i>NY Times</i> .....	7
<i>Twitter</i> .....	10
BIG DATA (WORD CO-OCCURRENCE)	13
<i>Common Crawl</i> .....	13
<i>NY Times</i> .....	16
<i>Twitter</i> .....	19
 <b>SMALL DATA .....</b>	 <b>22</b>
SMALL DATA (WORD COUNT) .....	22
<i>Common Crawl</i> .....	22
<i>NY Times</i> .....	25
<i>Twitter</i> .....	28
SMALL DATA (WORD CO-OCCURRENCE) .....	31
<i>Common Crawl</i> .....	31
<i>NY Times</i> .....	34
<i>Twitter</i> .....	37

## Overview

An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov) have their data available to the public for various purposes. Social network applications such as Twitter and Facebook collect enormous amount of data contributed by their numerous and prolific user. For other businesses such as Amazon and NY Times data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make subset of their data available for use by registered users for free. Some of them as downloadable data files (.csv, .xlsx) as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as a web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrap the data from these web pages and extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products.

We had collected data about from three sources, one opinion-based social media in twitter, research data in New York Times, and the third is the common crawl data for the same topic or key phrase, and similar time periods. Processed the three data sets collected individually using classical big data methods. Compared the outcomes using popular visualization methods.

## How did we do it?

We have chosen our point of interest as sports with sub categories baseball, basketball, NFL, soccer and tennis. As sports being everyone's interest and there will be a great number of articles from NY Times also every person tweets about their best interest of sport in a while, we thought we could gather huge amount of data.

Twitter:

- We collected the data by querying the twitter developer account.
- Queries used are: baseball, basketball, NFL, soccer, tennis.
- We used twitter search API for 1,00,000 Tweets.

NY Times:

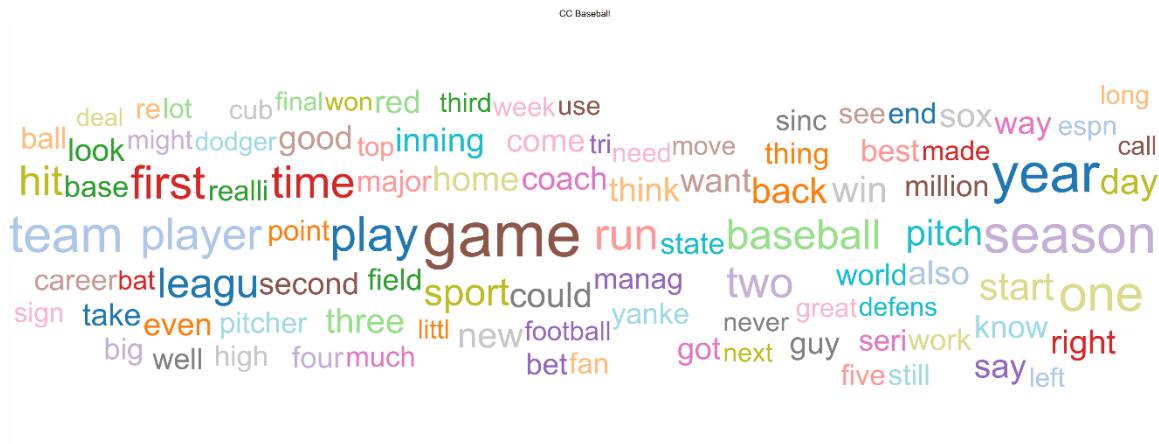
- We used the NY Times Article Search API to get URL's.
- The query words we sued are: baseball, basketball, NFL, soccer, tennis.
- Then the URL's are used to retrieve the HTML content.
- With python script we extracted the paragraph data of certain articles.
- Saved all the retrieved articles.

Common Crawl:

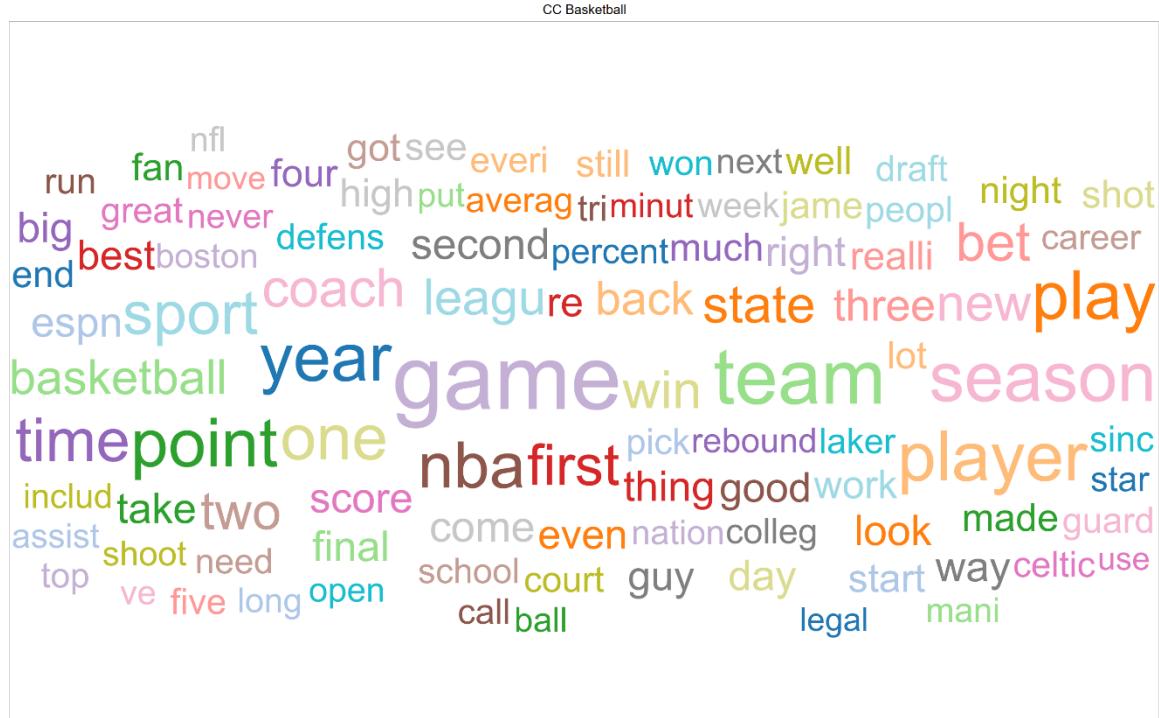
- We obtained the data from common crawl by passing the parameters url [www.espn.com](http://www.espn.com), match type: domain and got the output as JSON file. We selected espn as it covers all sports on daily basis.
- Used common index API to set the above parameters.
- Parsed the JSON file to get the articles by using offset, length for each WARC file.
- Downloaded the exact file using offset, length from Amazon AWS.
- Unzip the WARC file to get text
- Extract paragraph data and save the article

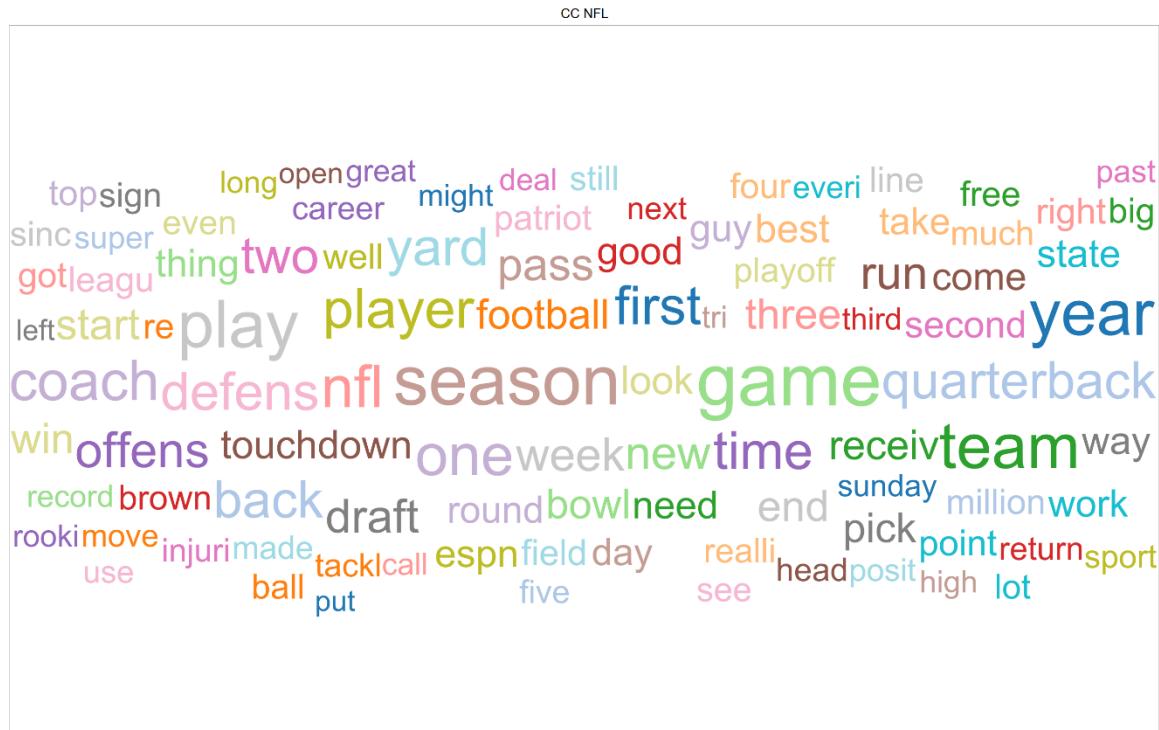
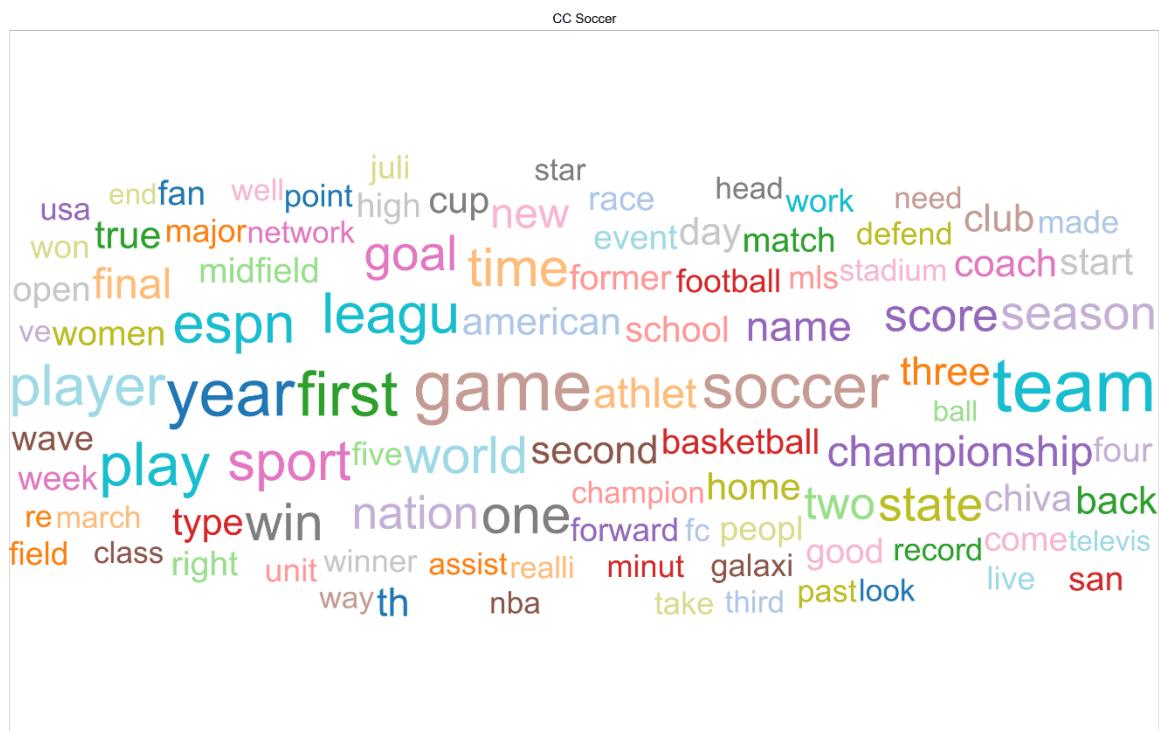
## Big Data

### Word Count of Common Crawl (Baseball)

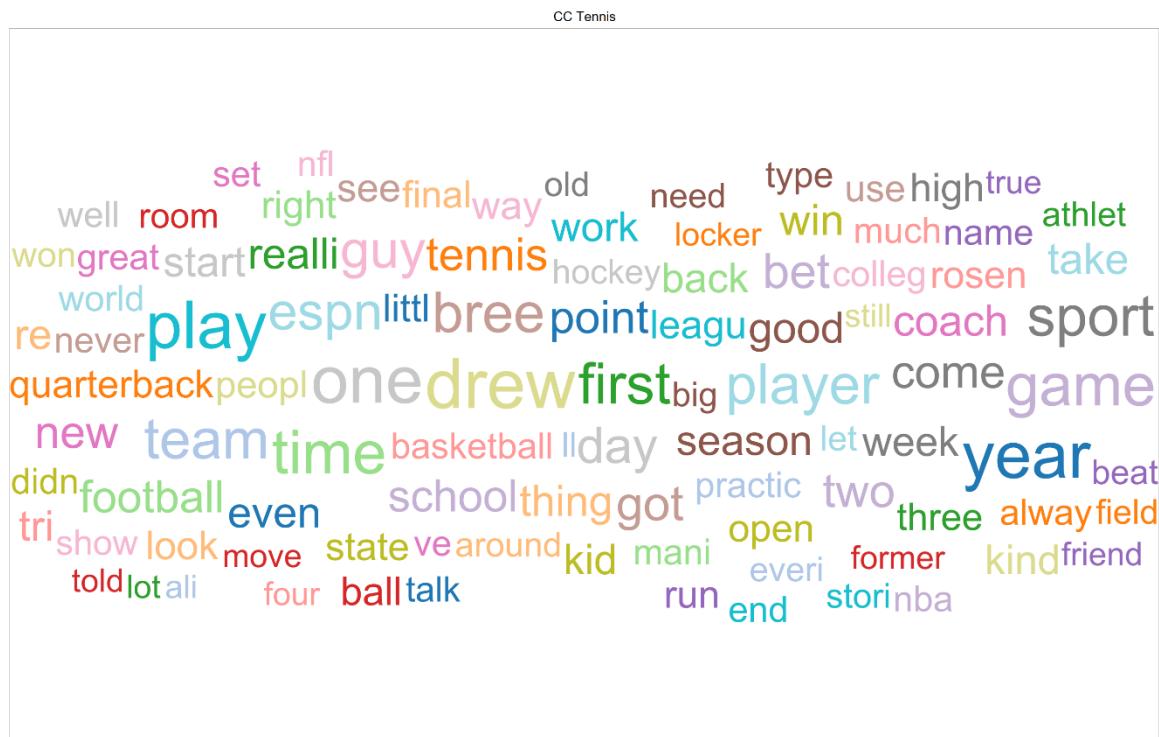


### Word Count of Common Crawl (Basketball)

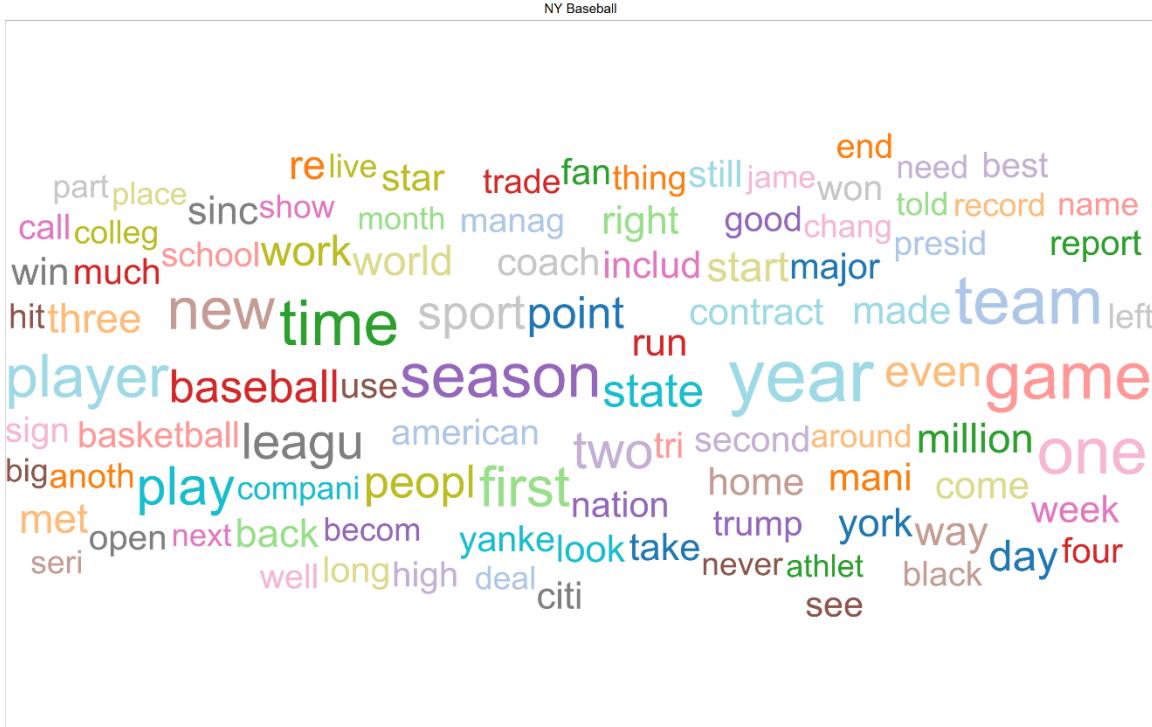


**Word Count of Common Crawl (NFL)****Word Count of Common Crawl (Soccer)**

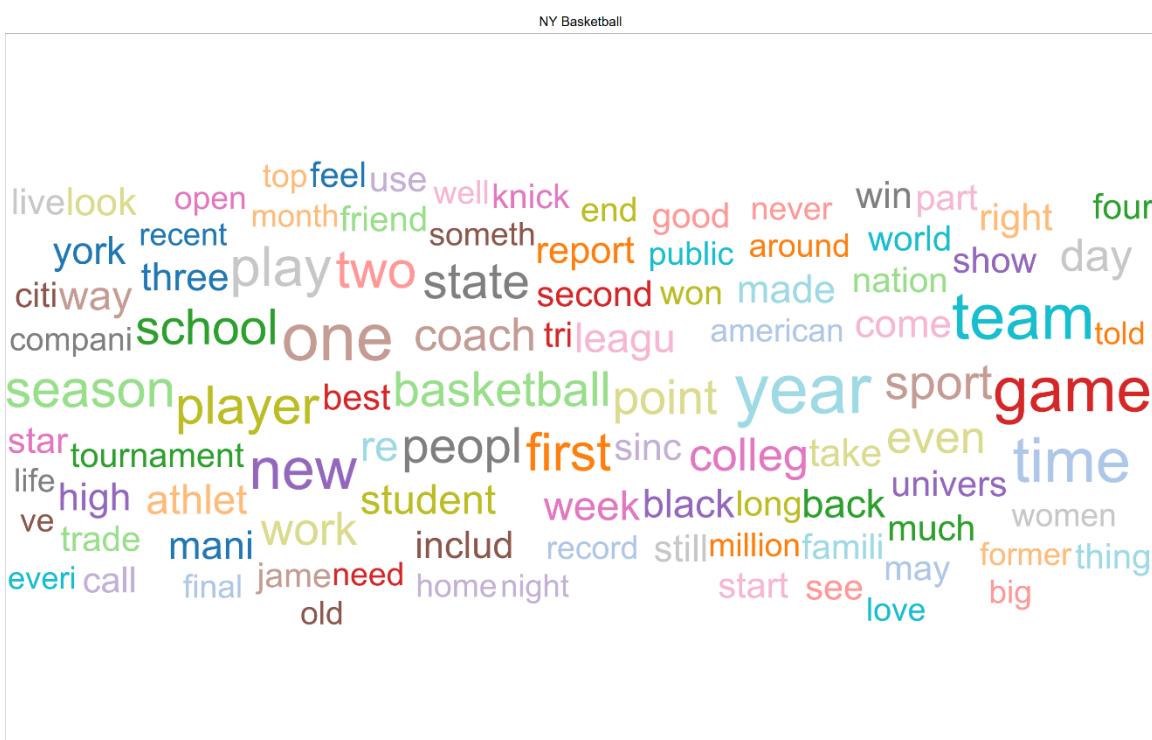
## Word Count of Common Crawl (Tennis)



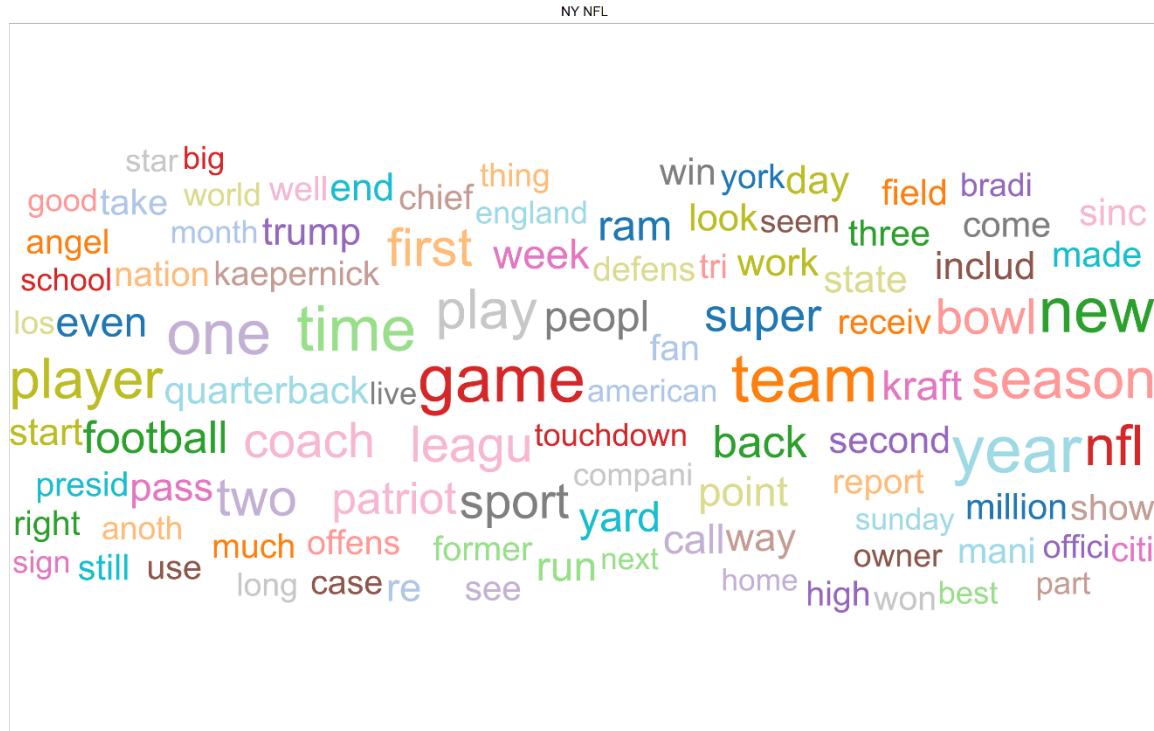
## Word Count of NY Times (Baseball)



## Word Count of NY Times (Basketball)



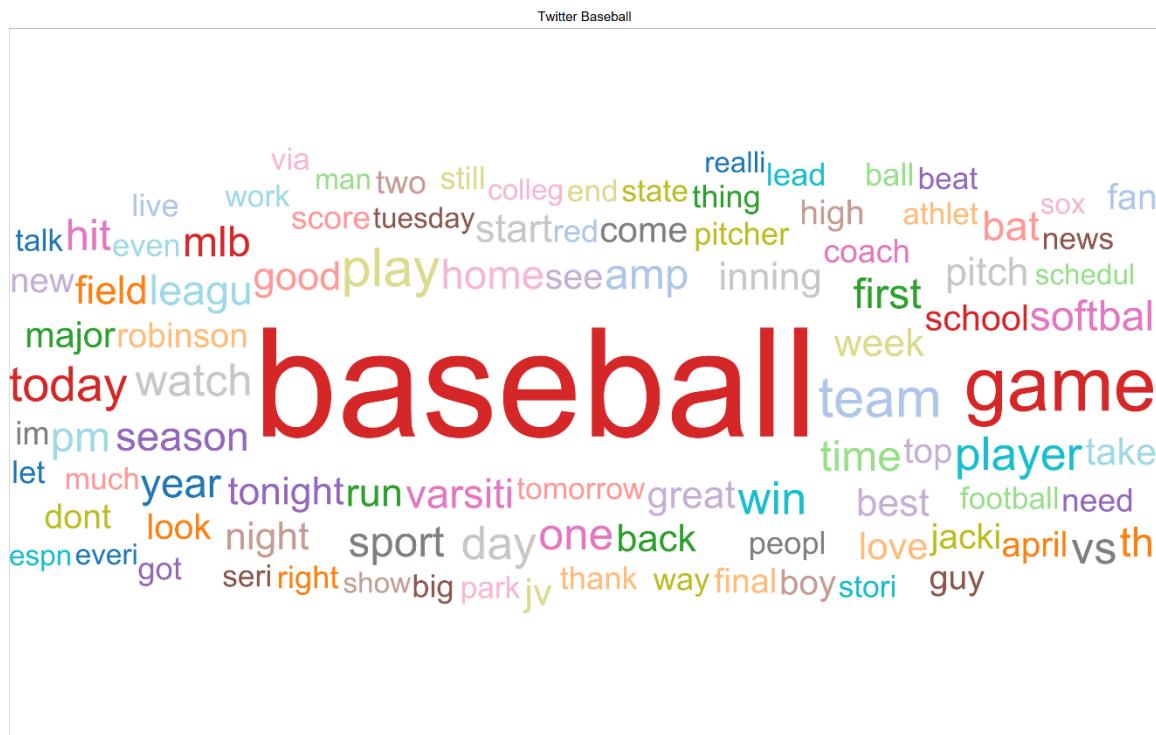
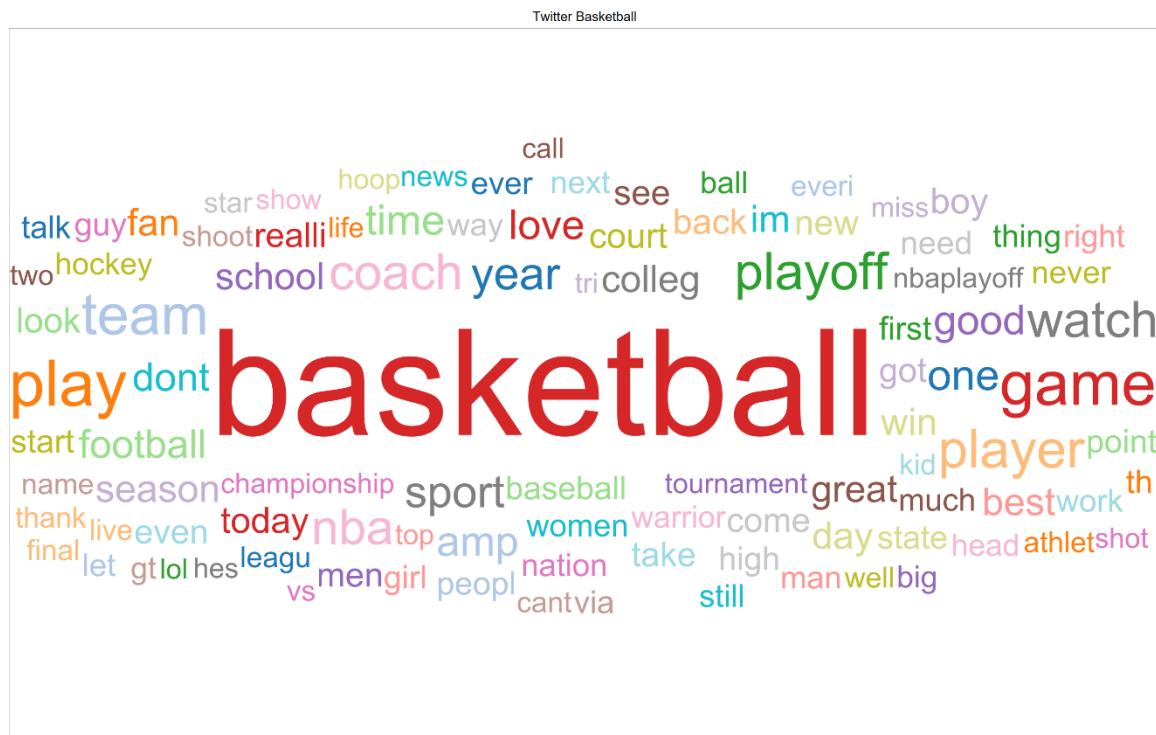
## Word Count of NY Times (NFL)



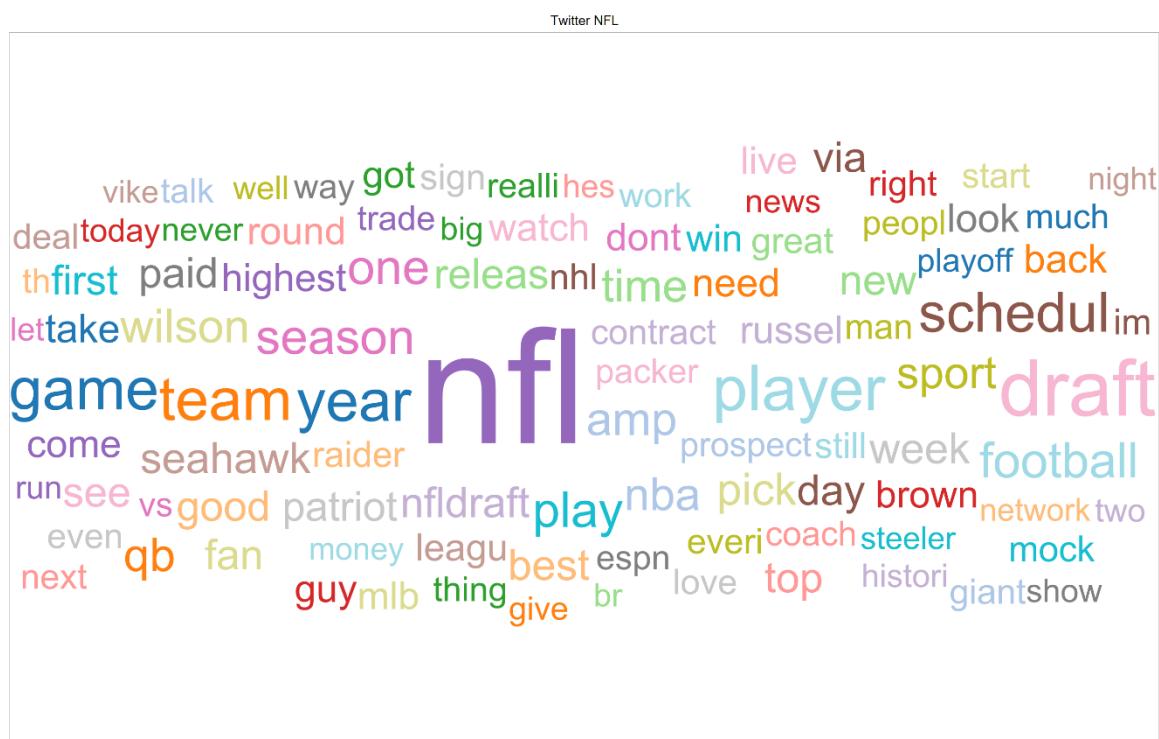
## Word Count of NY Times (Soccer)



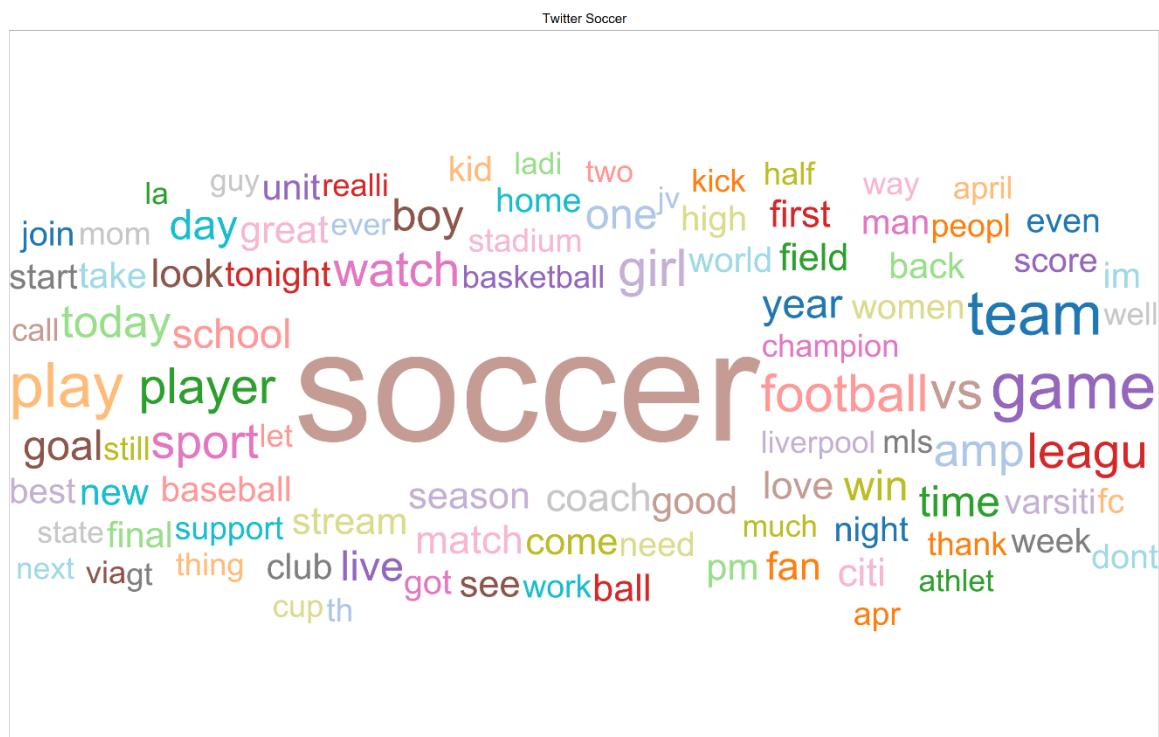
**Word Count of NY Times (Tennis)**

**Word Count of Twitter (Baseball)****Word Count of Twitter (Basketball)**

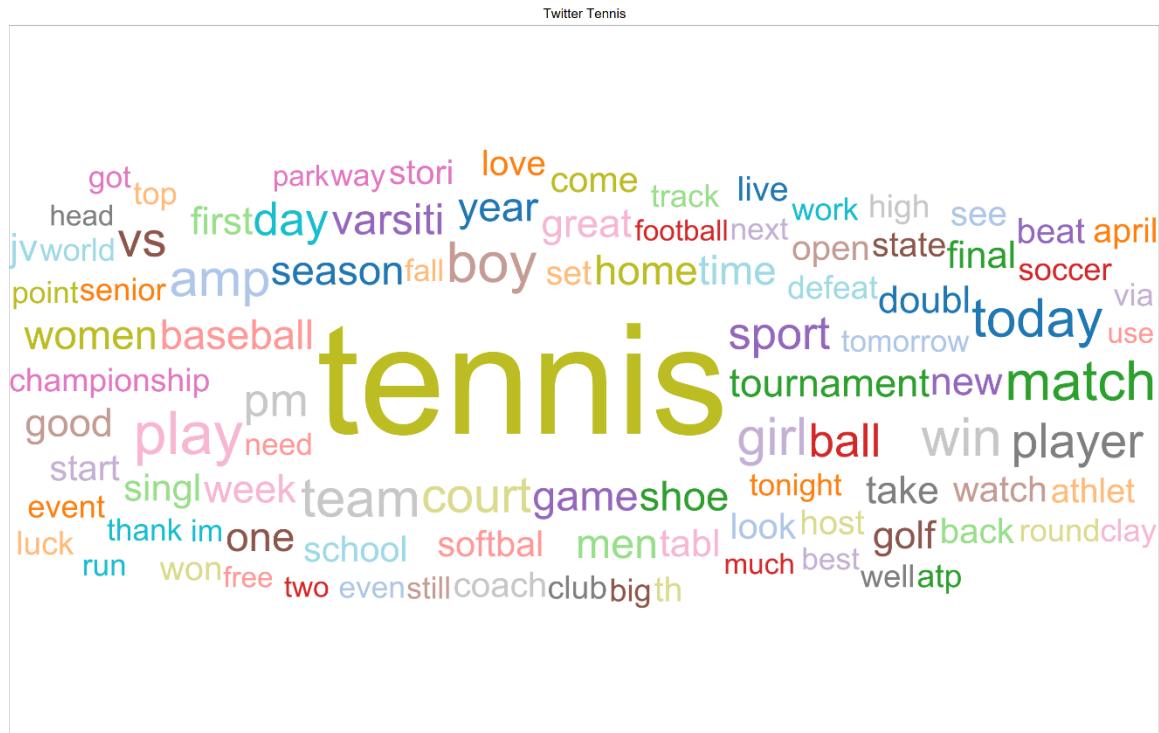
## Word Count of Twitter (NFL)



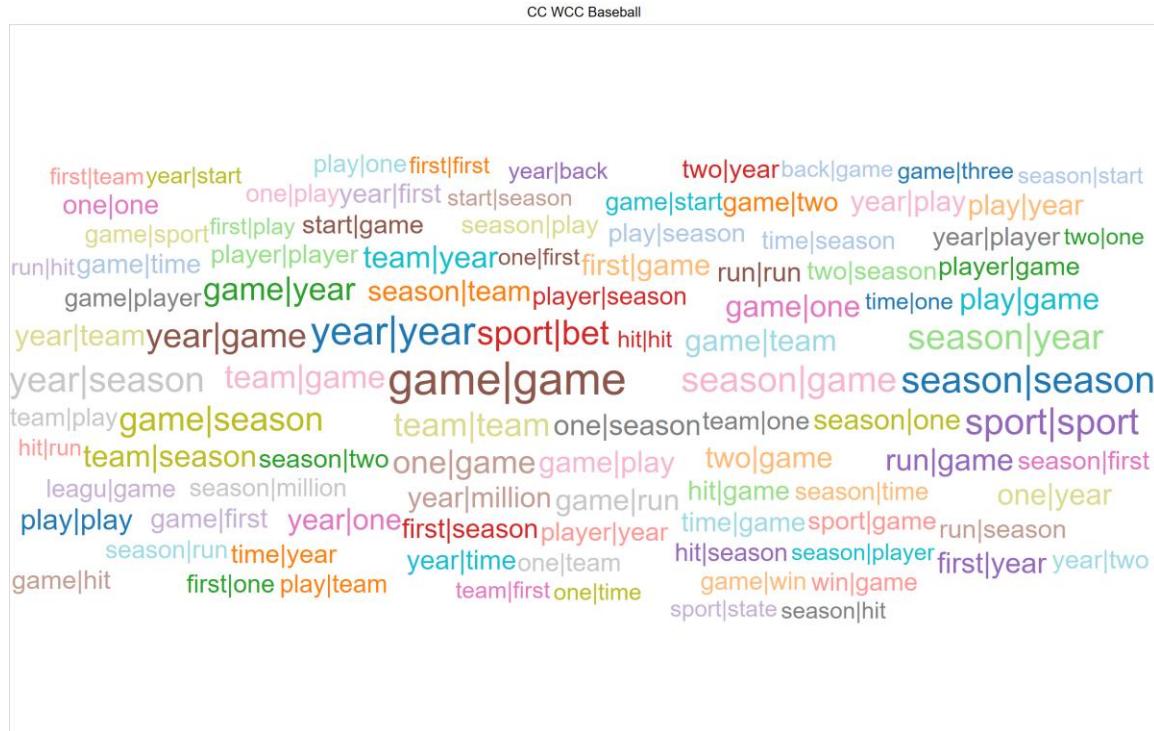
## Word Count of Twitter (Soccer)



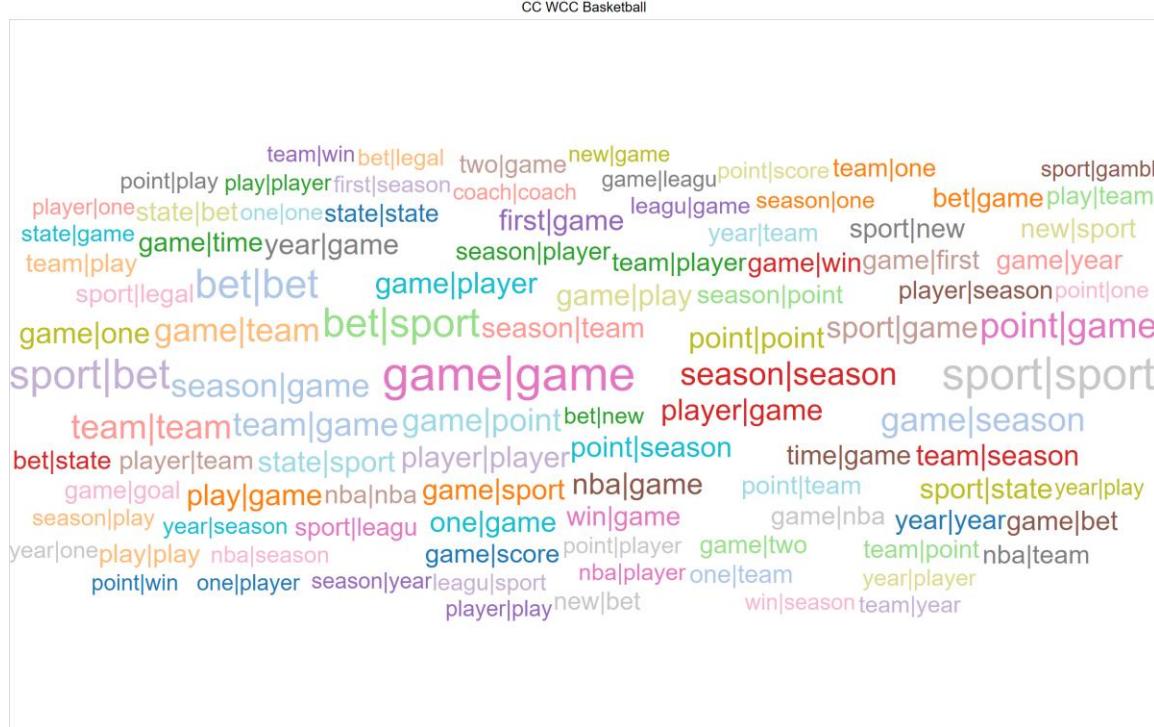
## Word Count of Twitter (Tennis)



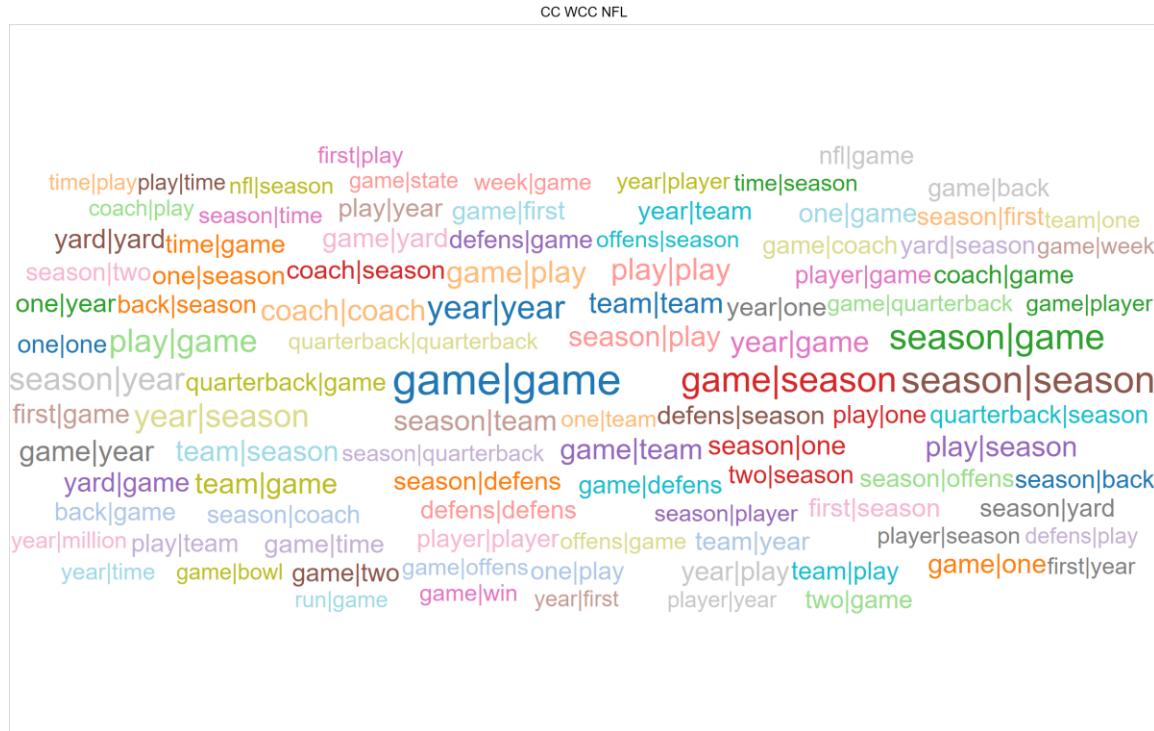
## Word Co-occurrence of Common Crawl (Baseball)



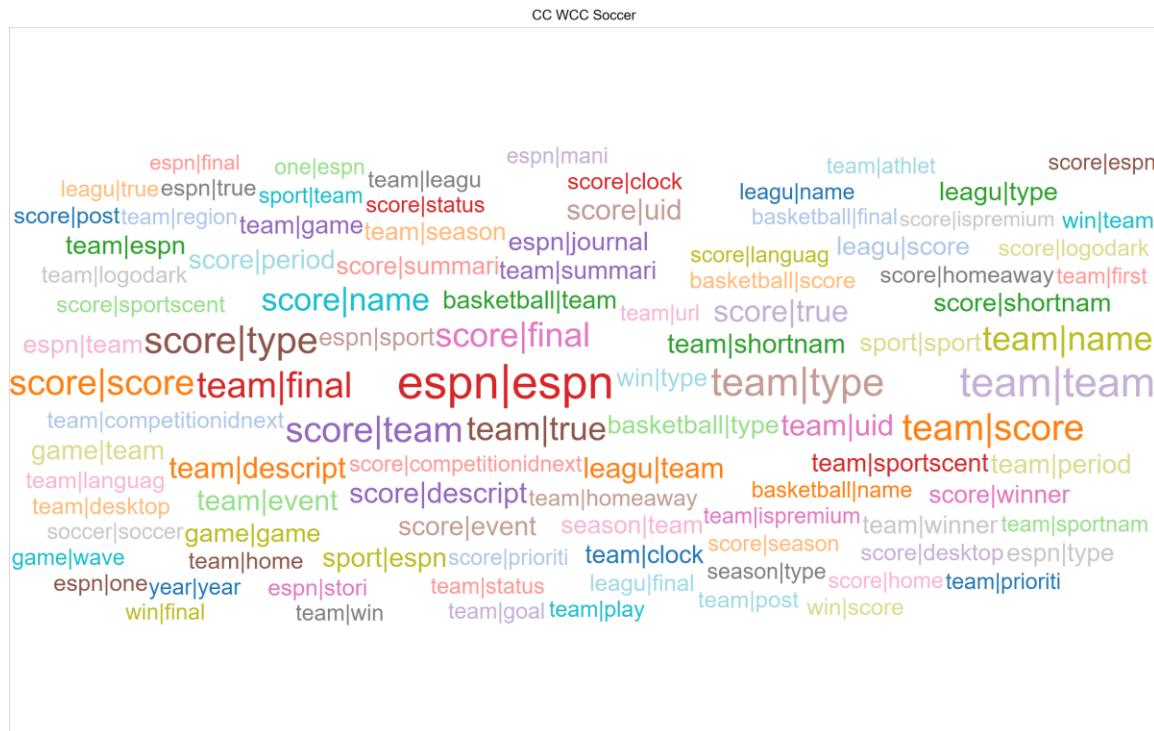
## Word Co-occurrence of Common Crawl (Basketball)



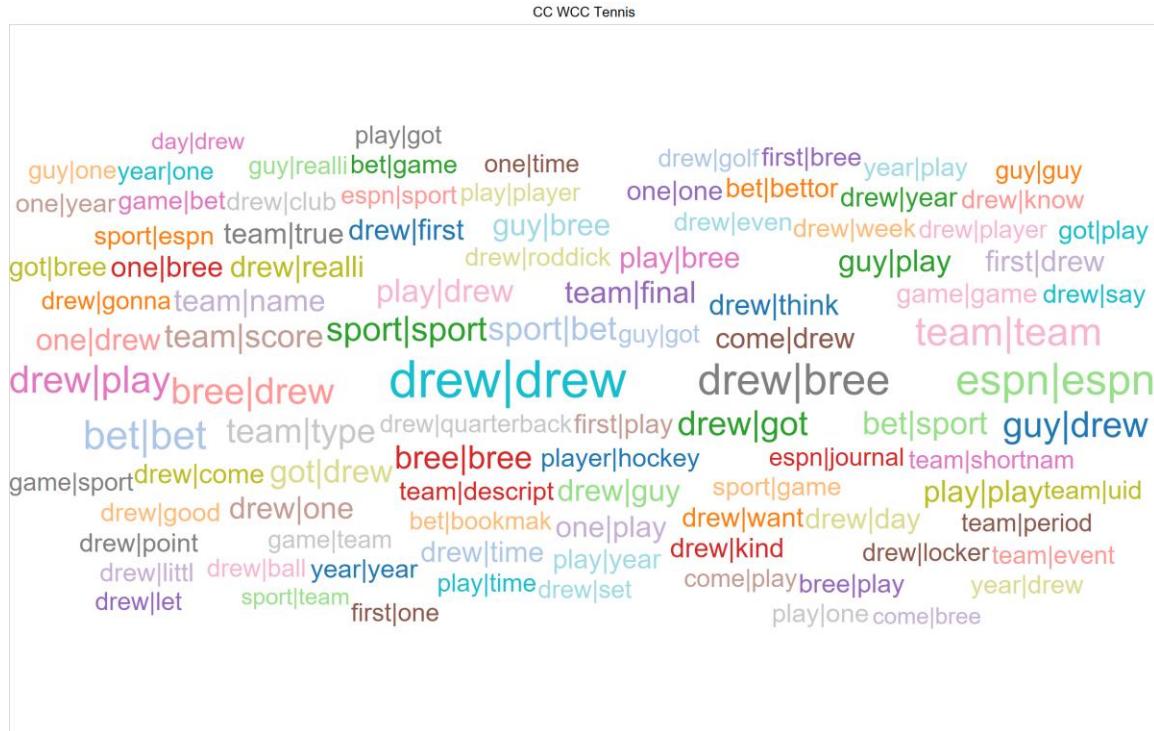
## Word Co-occurrence of Common Crawl (NFL)



## Word Co-occurrence of Common Crawl (Soccer)



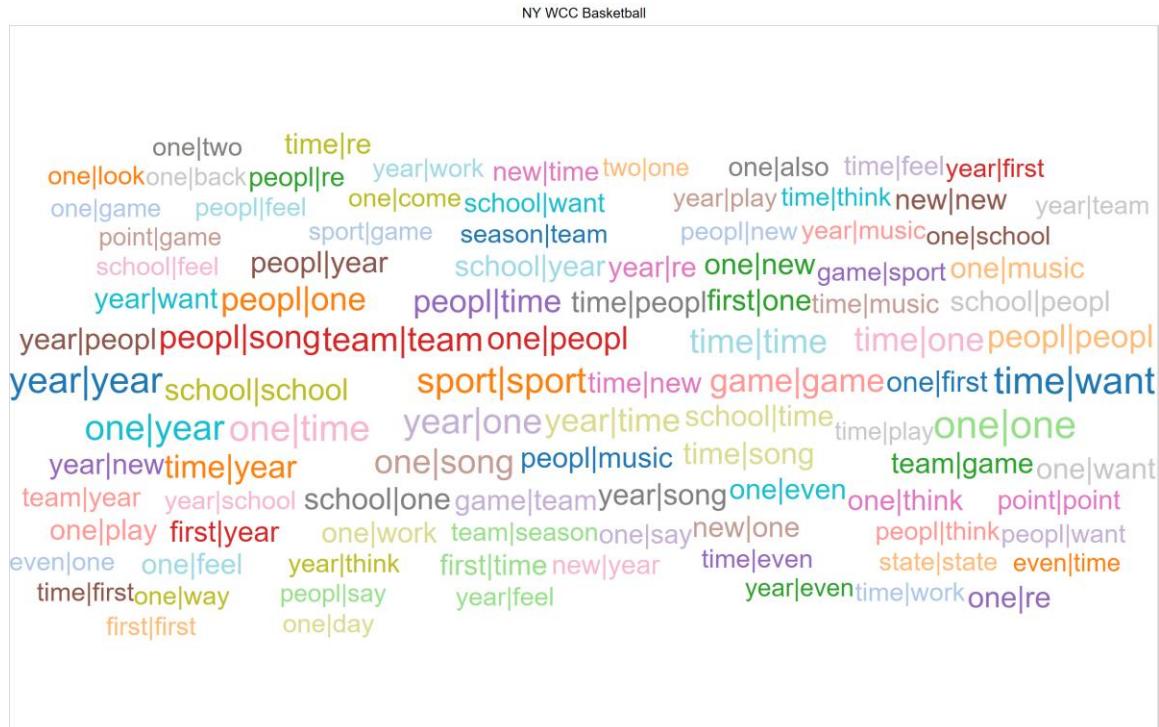
## Word Co-occurrence of Common Crawl (Tennis)



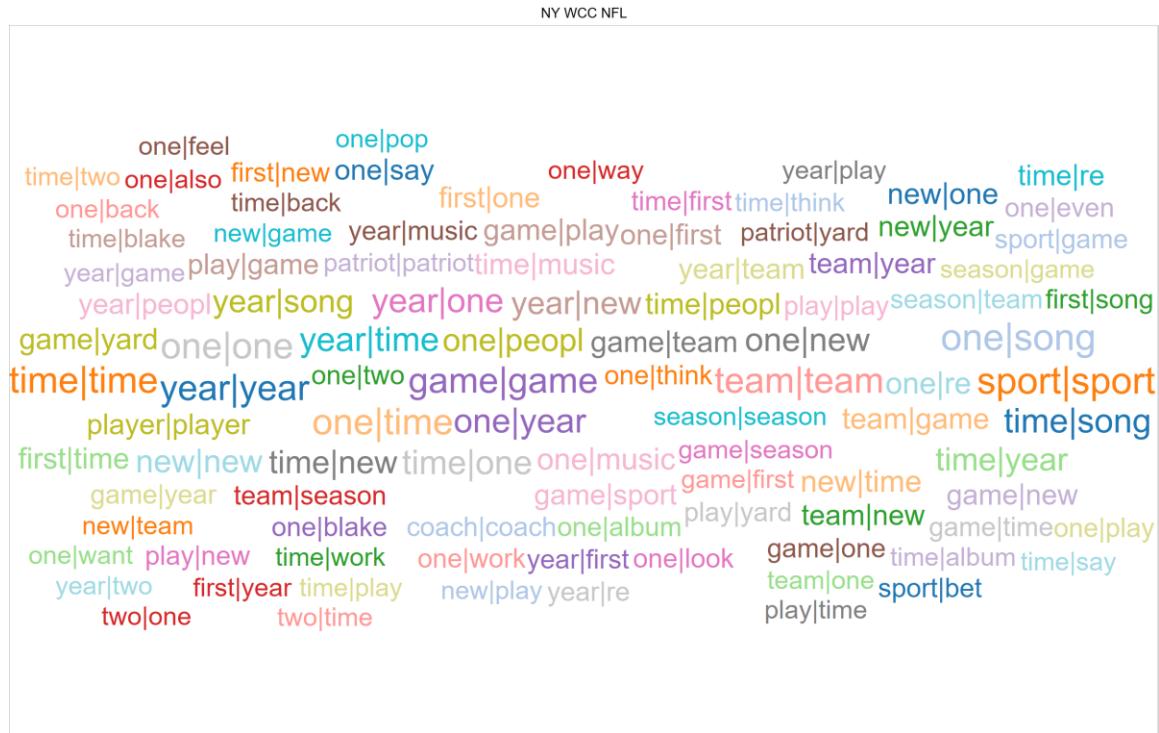
### Word Co-occurrence of NY Times (Baseball)



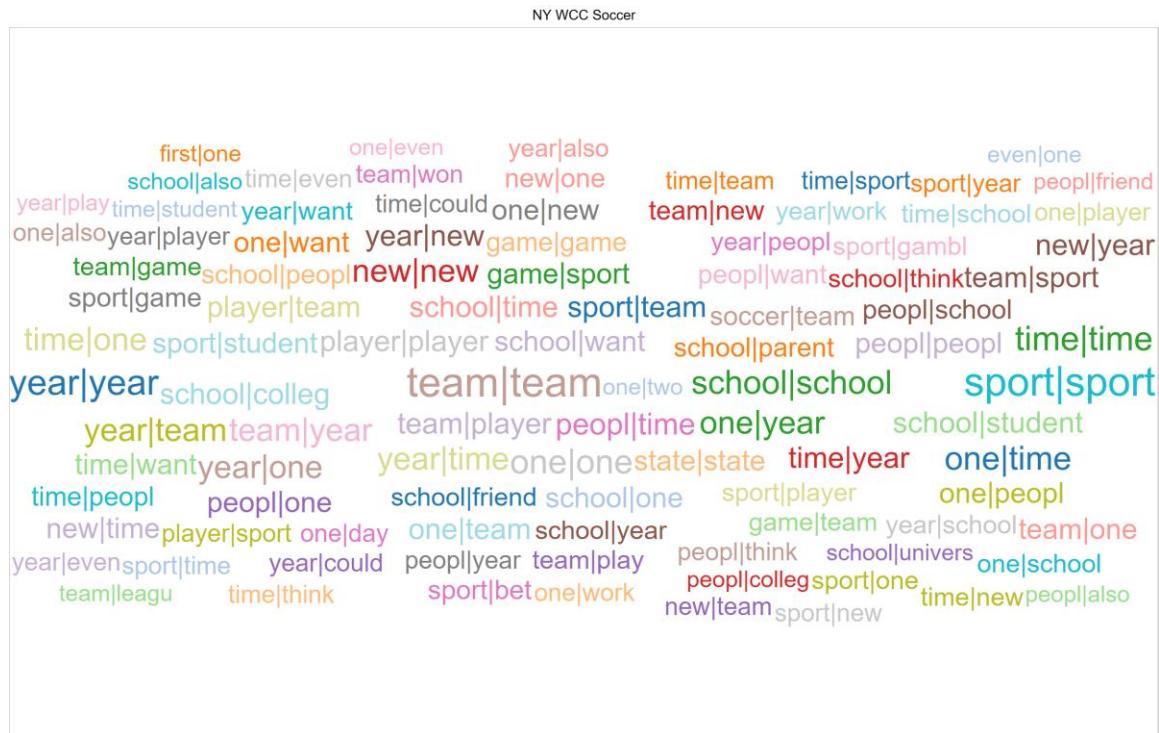
### Word Co-occurrence of NY Times (Basketball)



### Word Co-occurrence of NY Times (NFL)

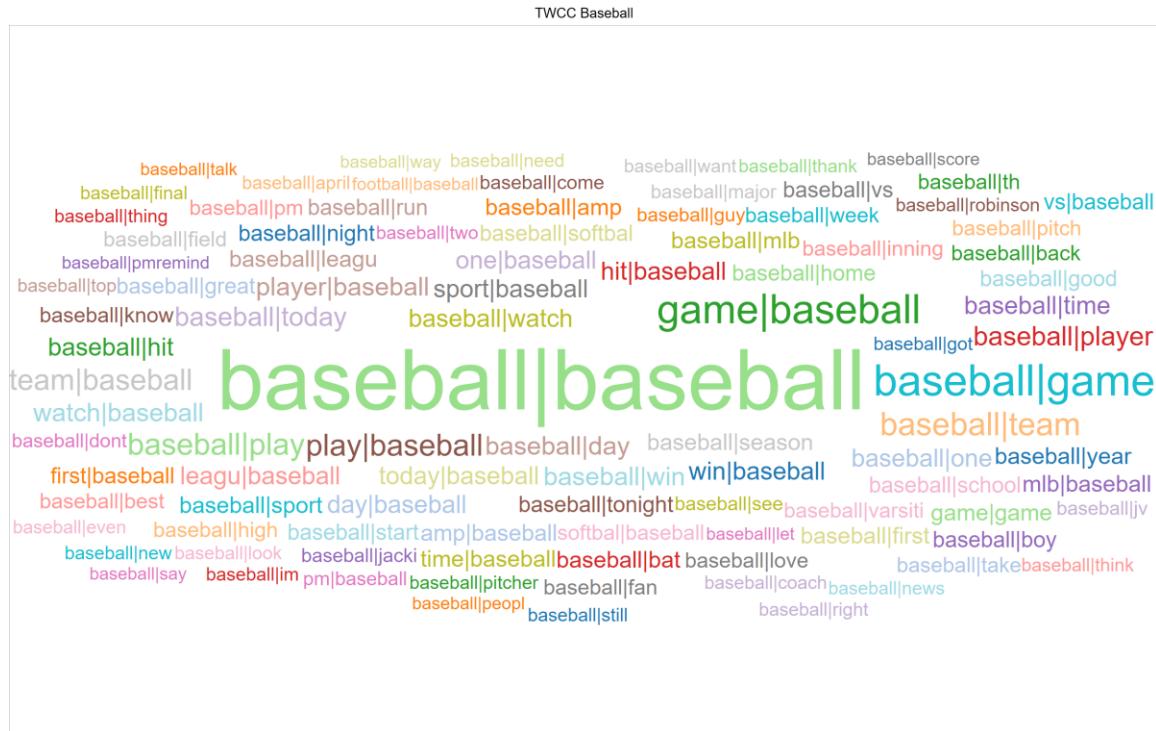


### Word Co-occurrence of NY Times (Soccer)



**Word Co-occurrence of NY Times (Tennis)**

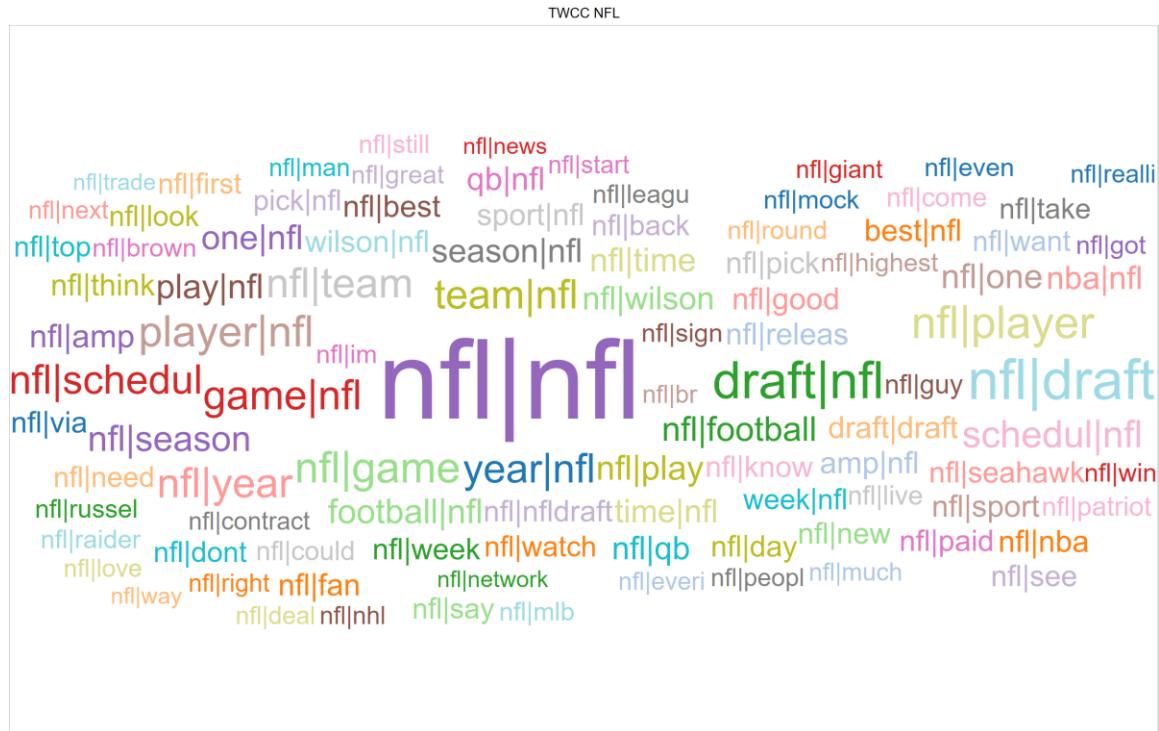
### Word Co-occurrence of Twitter (Baseball)



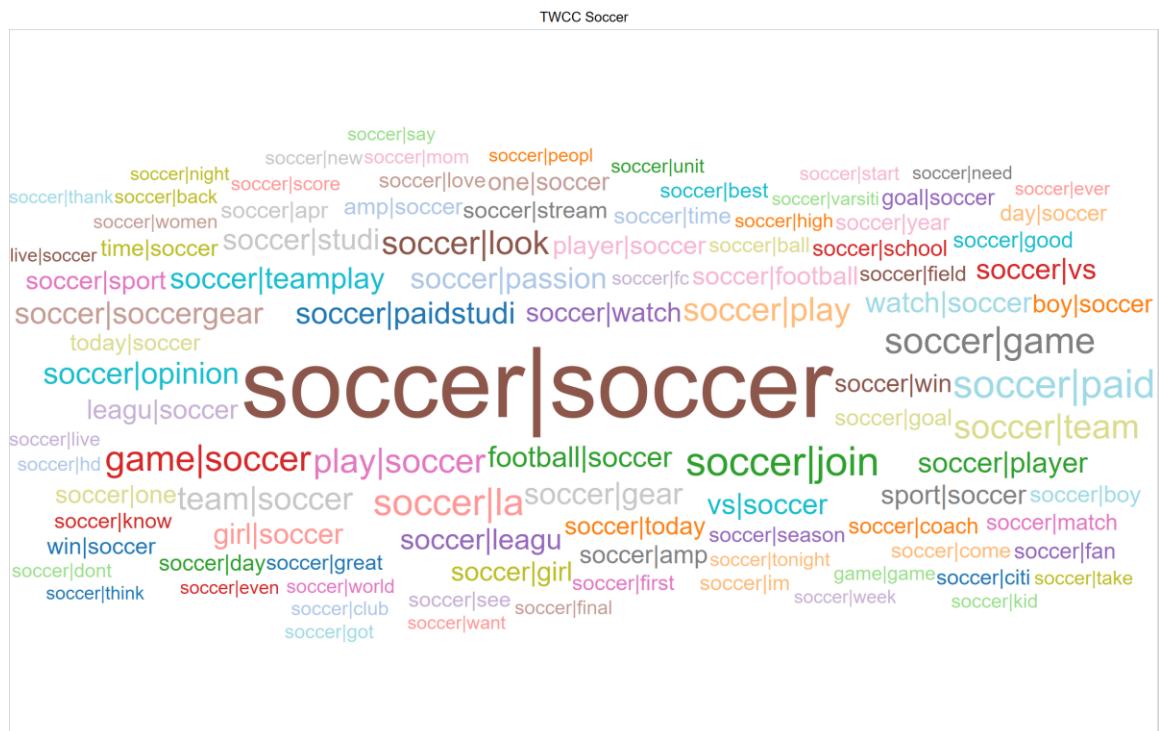
### Word Co-occurrence of Twitter (Basketball)

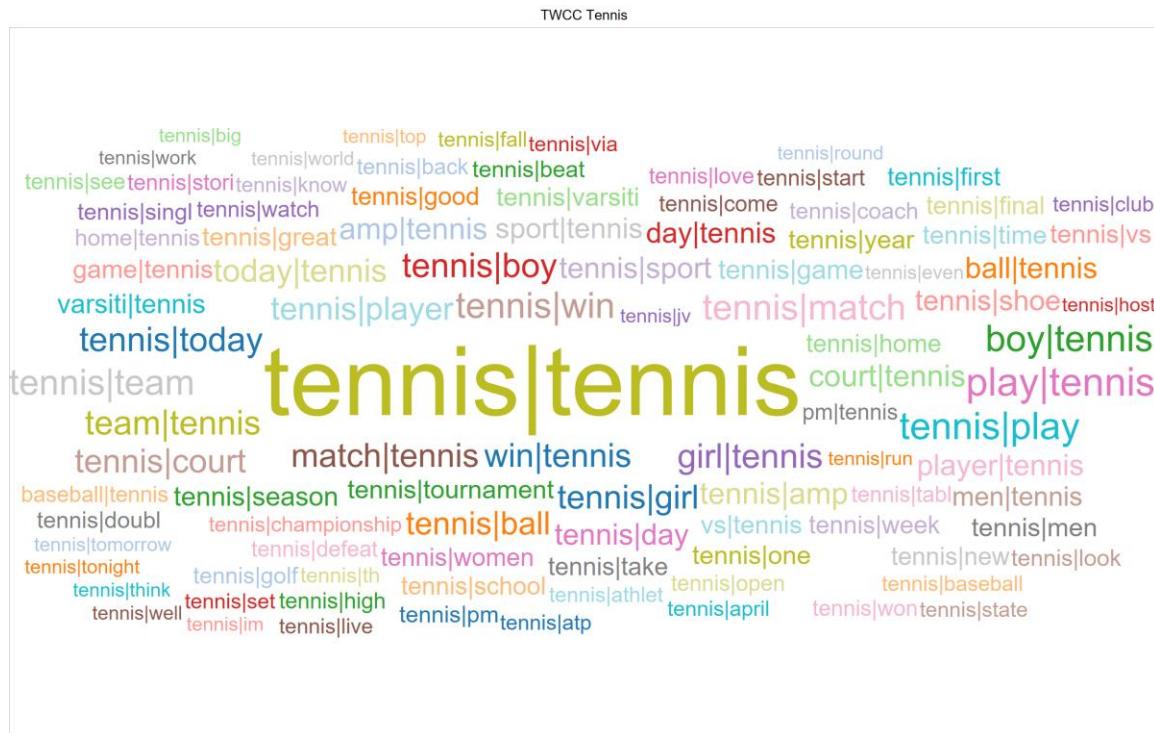


## Word Co-occurrence of Twitter (NFL)



## Word Co-occurrence of Twitter (Soccer)

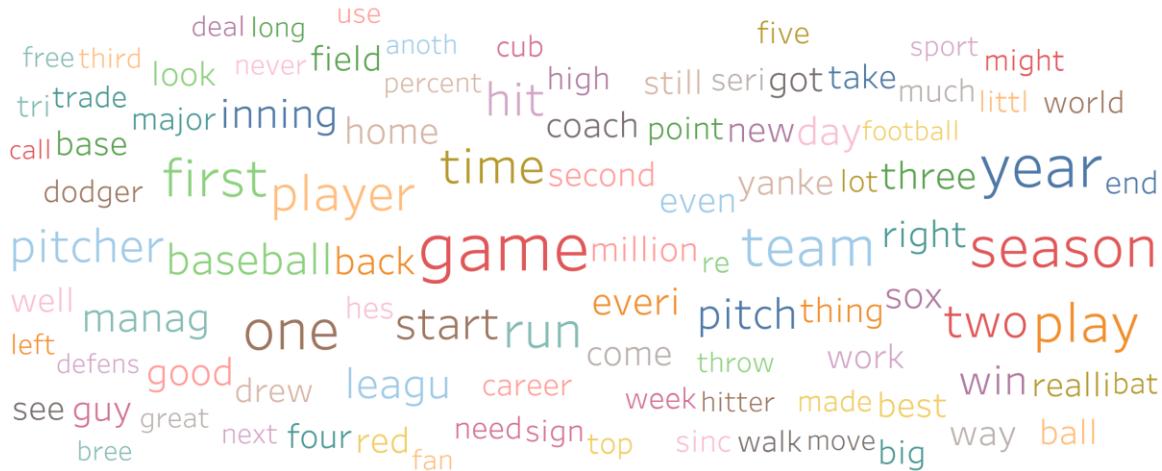


**Word Co-occurrence of Twitter (Tennis)**

# Small Data

## **Word Count of Common Crawl (Baseball)**

CC WC Baseball



## Word Count of Common Crawl (Basketball)

CC WC Basketball



### Word Count of Common Crawl (NFL)

CC WC NFL

A word cloud visualization for the Common Crawl (NFL) dataset. The words are arranged in a cluster, with larger words indicating higher frequency. The colors of the words represent different categories or themes:

- Red/Orange:** lot, put, call, big, long, take, bradisaint, field, everi, day, reali, look, free, injuri, much, right, made, win, record, need, start, football, sunday, good, part, three, hes, super, pass, cowboy, two, offens, touchdown, time, triplayoff, got, draft, work, yard, player.
- Blue/Cyan:** play, year, defens, season, round, game, quarterback, ball, new, coach, first, one, back, thing, week, patriot, team, run, still, end, rooki, nfl, second, receiv, come, best, guy, brown, point, bowl, practic, four, even, re, top, see, pick, posit, move, espn, sinc, great, way, career, leagu, head, past, deal, next, well, chief, five.
- Green/Yellow:** citibet, select, san, career, percent, winner, defend, ball, high, minutespn, stadium, score, world, week, open, award, former, day, see, athlet, leagu, state, first, coach, two, second, cupgatoradnation, mani, reali, past, take, work, divis, call, diego, includ, field, peopl, final, home, select, san, career, percent, winner, defend, ball, high, minutespn, stadium, score, world, week, open, award, former, day, see, athlet, leagu, state, first, coach, two, second, cupgatoradnation.

### Word Count of Common Crawl (Soccer)

CC WC Soccer

A word cloud visualization for the Common Crawl (Soccer) dataset. The words are arranged in a cluster, with larger words indicating higher frequency. The colors of the words represent different categories or themes:

- Red/Orange:** five, ve, citibet, select, san, career, percent, winner, defend, ball, high, minutespn, stadium, score, world, week, open, award, former, day, see, athlet, leagu, state, first, coach, two, second, cupgatoradnation, mani, reali, past, take, work, divis, call, diego, includ, field, peopl, final, home, select, san, career, percent, winner, defend, ball, high, minutespn, stadium, score, world, week, open, award, former, day, see, athlet, leagu, state, first, coach, two, second, cupgatoradnation.
- Blue/Cyan:** team, play, sport, soccer, chiva, fc, game, goal, player, assist, time, made, season, basketball, forward, american, wave, year, back, come, three, thmidfield, school, football, one, univers, portland, colleg, senior, club, mls, win, galaxi, match, start, fraser, need, new, won, way, usa, four, honor, right, star, look, fan, thing, name, point, good, major, well.

**Word Count of Common Crawl (Tennis)**

CC WC Tennis

A word cloud visualization showing the frequency of words from the CC WC Tennis dataset. The words are colored in various shades of green, blue, red, and orange, and are arranged in a roughly circular pattern. The most frequent words include 'player', 'year', 'time', 'drew', 'let', 'game', 'bree', 'littl', 'quarterback', 'show', 'first', 'big', 'football', 'team', 'realli', 'start', 'tennis', 'thing', 'play', 'back', 'much', 'walk', 'someth', 'come', 'got', 'wil', 'sport', 'teammat', 'espn', 'take', 'end', 'mani', 'room', 'nfl', 'throw', 'alway', 'high', 'three', 'never', 'two', 'leagu', 'right', 'use', 'nba', 'talk', 'old', 'peopl', 'stori', 'see', 'close', 'decid', 'well', 'colleg', 'run'.

**Word Count of NY Times (Baseball)**

NY WC Baseball

A word cloud visualization showing the frequency of words in NY Times Baseball articles. The size of each word indicates its frequency, with larger words being more prominent. The words are colored in various shades of red, orange, yellow, green, blue, and purple.

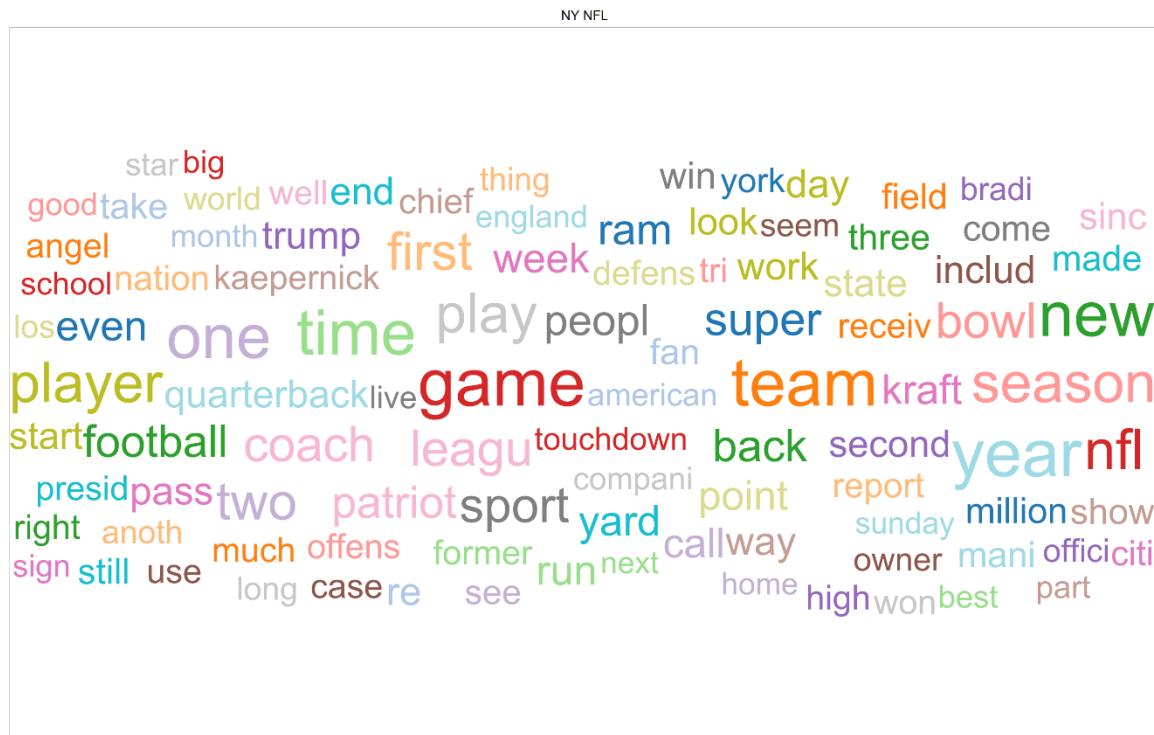
Key words include: time, baseball, team, new, year, two, ms, contract, work, season, three, sinc, leagu, sport, first, world, mani, met, manag, fan, game, week, won, averag, pitch, much, use, thing, high, next, best, state, look, win, got, york, includ, nation, former, back, trum, report, chang, black, good, open, see, left.

**Word Count of NY Times (Basketball)**

NY WC Basketball

A word cloud visualization showing the frequency of words in NY Times Basketball articles. The size of each word indicates its frequency, with larger words being more prominent. The words are colored in various shades of red, orange, yellow, green, blue, and purple.

Key words include: bet, told, live, need, four, news, re, take, ever, high, call, part, never, offici, davi, world, citi, show, laker, jame, work, even, look, york, friend, day, made, homeman, compani, million, second, confer, tri, includ, former, big, women, coach, school, colleg, two, trade, leagu, report, game, life, season, time, win, new, team, athlet, basketball, play, year, start, player, sport, state, student, right, first, come, univers, nation, three, american, peopl, black, point, presid, recent, week, one, month, name, famili, good, might, way, still, gambl, chang, back, around, knick, next, top, well, see, best, open, white, long, much, use, sinc, rule, star.

**Word Count of NY Times (NFL)****Word Count of NY Times (Soccer)**

NY WC Soccer



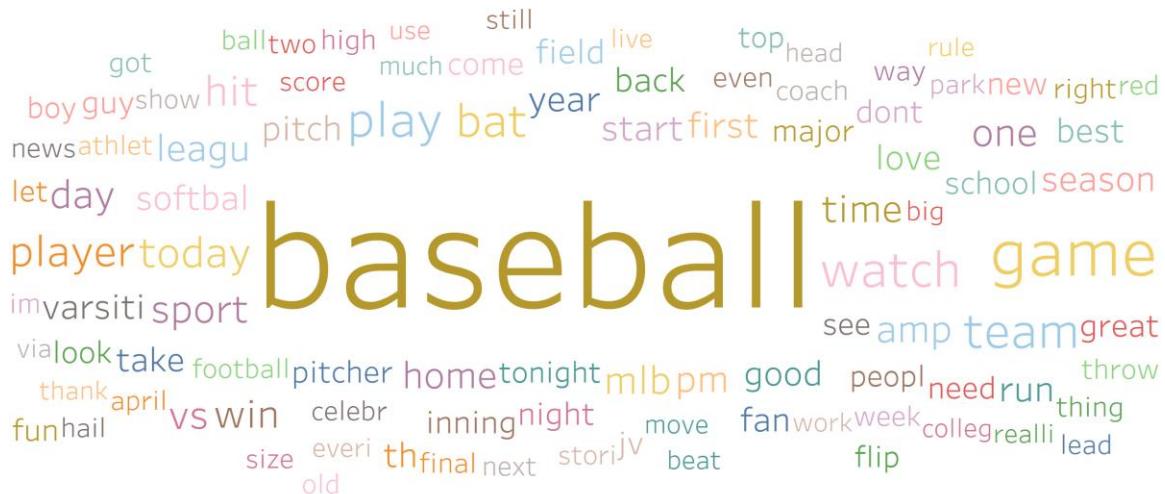
**Word Count of NY Times (Tennis)**

NY WC Tennis

slam seed hous well big final live movi still best april  
feel take month even realli round osaka highlook mani leftmade  
th includ thingmillion second work statefriend call  
women play player match american superhero day  
peopl william open tournament presid tennis year time  
win two australian first new point world week long school ms  
much follow trump game back coach come around country australia one  
talk feder show court three offici never good grand way team great next  
top hit use part facetri love reve sinc won see sport alway

## Word Count of Twitter (Baseball)

TWC Baseball



## Word Count of Twitter (Basketball)

TWC Basketball



## Word Count of Twitter (NFL)

TWC NFL



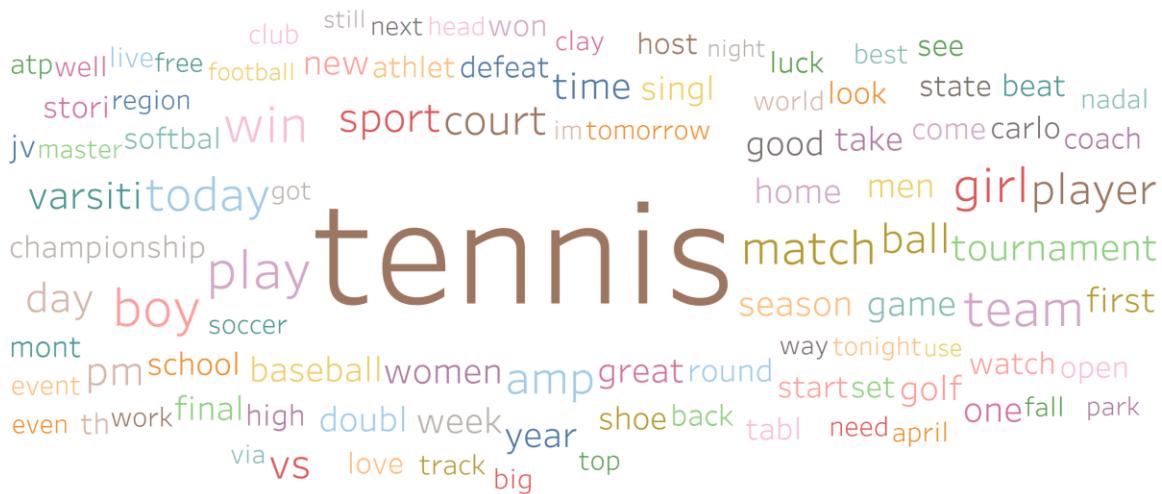
## Word Count of Twitter (Soccer)

TWC Soccer



## Word Count of Twitter (Tennis)

TWC Tennis



## Word Co-occurrence of Common Crawl (Baseball)

CC WCC Baseball



## Word Co-occurrence of Common Crawl (Basketball)

CC WCC Basketball



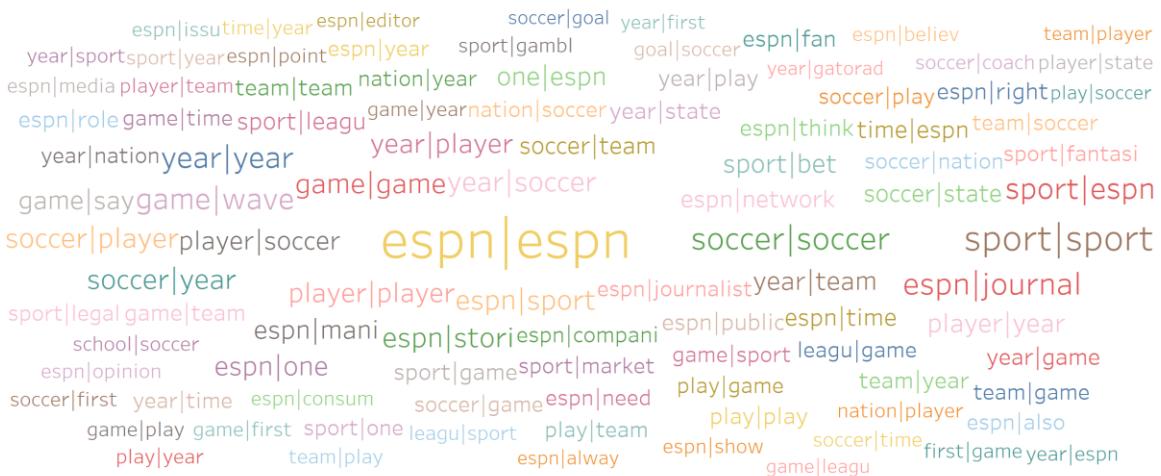
## Word Co-occurrence of Common Crawl (NFL)

CC WCC NFL



## Word Co-occurrence of Common Crawl (Soccer)

CC WCC Soccer



### Word Co-occurrence of Common Crawl (Tennis)

CC WCC Tennis

bree|one                one|year                play|time  
drew|kid    play|gotgot|play    play|year    drew|didn't    bree|got    drew|say    guy|year  
drew|two    drew|school    drew|golf    drew|take    drew|gonna    day|drew    drew|ball    come|bree  
bree|realli    one|bree    drew|first    drew|know    play|bree    rosen|rosen    got|bree    drew|locker  
drew|throw    drew|one    come|drew    kind|drew    guy|got    drew|even    drew|year    drew|littl  
drew|day    bree|bree    got|drew    drew|guy    drew|beat    drew|think    one|play    bree|drew  
guy|drew    drew|got    drew|drew    got|got    drew|play    drew|bree  
drew|tri    drew|realli    drew|quarterback    guy|bree    drew|week    drew|roddick    play|drew  
realli|drew    drew|point    drew|time    one|drew    drew|want    guy|play    drew|good    first|drew    bree|club  
            point|drew    drew|kind    drew|come    play|play    drew|club    team|team    drew|player    time|drew  
drew|former    drew|practic    good|drew    kind|bree    first|bree    first|play    play|realli    guy|realli    kind|play  
drew|saint    bree|time    drew|set    come|play    guy|guy    play|one    good|bree  
guy|think    year|bree    one|one    bree|day    year|play

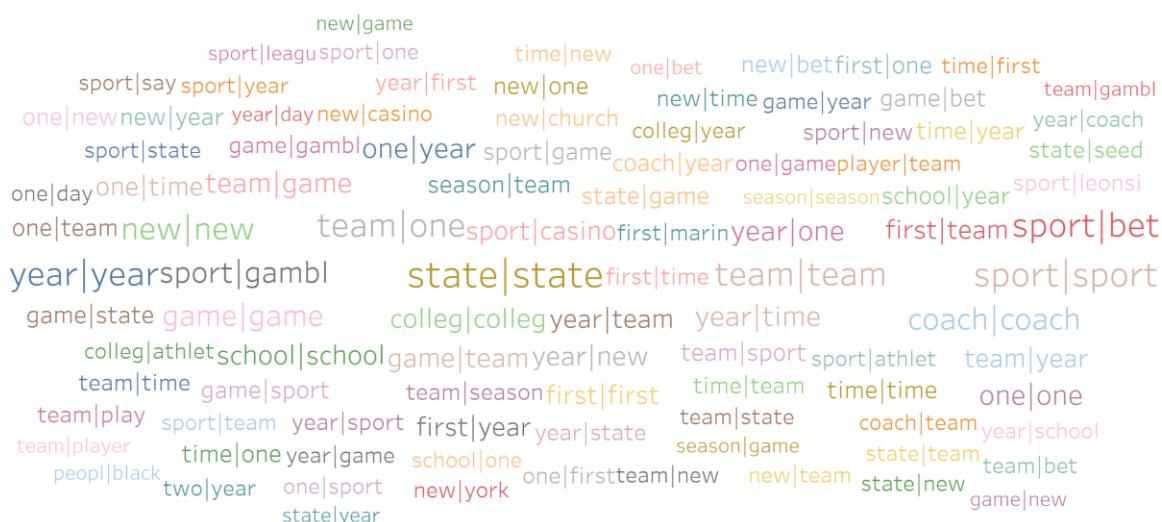
## Word Co-occurrence of NY Times (Baseball)

NY WCC Baseball



## Word Co-occurrence of NY Times (Basketball)

NY WCC Basketball



### Word Co-occurrence of NY Times (NFL)

NY WCC NFL

A 2D word co-occurrence matrix for the NY Times NFL dataset. The words are represented by colored squares, where the color indicates the frequency or strength of the co-occurrence between two words. The matrix is roughly square and symmetric, with a dense cluster of high-frequency words in the center.

### Word Co-occurrence of NY Times (Soccer)

NY WCC Soccer

A 2D word co-occurrence matrix for the NY Times Soccer dataset. The words are represented by colored squares, showing the frequency of co-occurrence between various soccer-related terms like team, year, and city.

**Word Co-occurrence of NY Times (Tennis)**

NY WCC Tennis

one|feel play|one day|point even|one first|year even|peopl  
first|play year|want peopl|even friend|hous week|time point|team peopl|come peopl|feel one|play  
new|one year|one tennis|tennis peopl|year friend|school time|think first|one one|one year|year  
first|first open|year time|want peopl|talk time|time game|game play|point peopl|re friend|invit  
time|left friend|feel point|state point|game first|point point|pointer peopl|know play|play  
point|rebound peopl|time peopl|want time|friend peopl|left friend|time peopl|friend  
william|william peopl|joke friend|friend peopl|peopl point|point  
friend|peopl peopl|think one|friend friend|want one|left friend|one one|peopl even|friend  
peopl|one friend|think game|point point|tournament time|peopl first|time friend|left time|feel  
even|time peopl|say peopl|thing one|first one|time year|time point|seed point|first time|one point|play  
one|think time|year first|friend peopl|school open|open friend|hang point|one point|second  
year|open friend|felt one|point new|time one|year friend|know one|want peopl|social  
year|peopl play|april year|friend game|state peopl|way

## Word Co-occurrence of Twitter (Baseball)

TWCC Baseball



## Word Co-occurrence of Twitter (Basketball)

TWCC Basketball



### Word Co-occurrence of Twitter (NFL)

TWCC NFL



### Word Co-occurrence of Twitter (Soccer)

TWCC Soccer



### Word Co-occurrence of Twitter (Tennis)

TWCC Tennis

