LAB – 2

Data Intensive Computing (Spring 2019)

Paritosh Velalam, Gaurav Avula

TABLE OF CONTENTS

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:

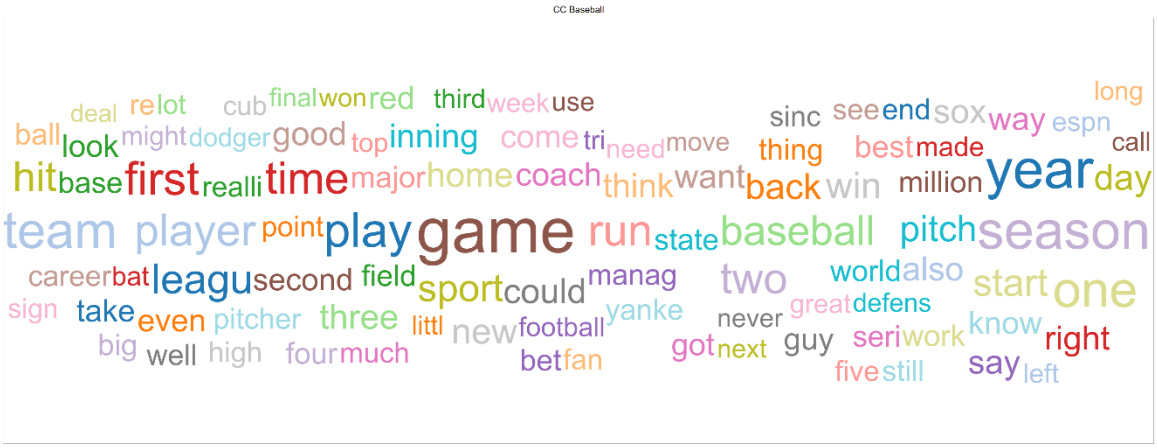
- We used the NY Times Article Search API to get URL's.
- The query words we sued are: baseball, basketball, NFL, soccer, tennis.
- Then the URL's are used to retrieve the HTML content.
- With python script we extracted the paragraph data of certain articles.
- Saved all the retrieved articles.

Common Crawl:

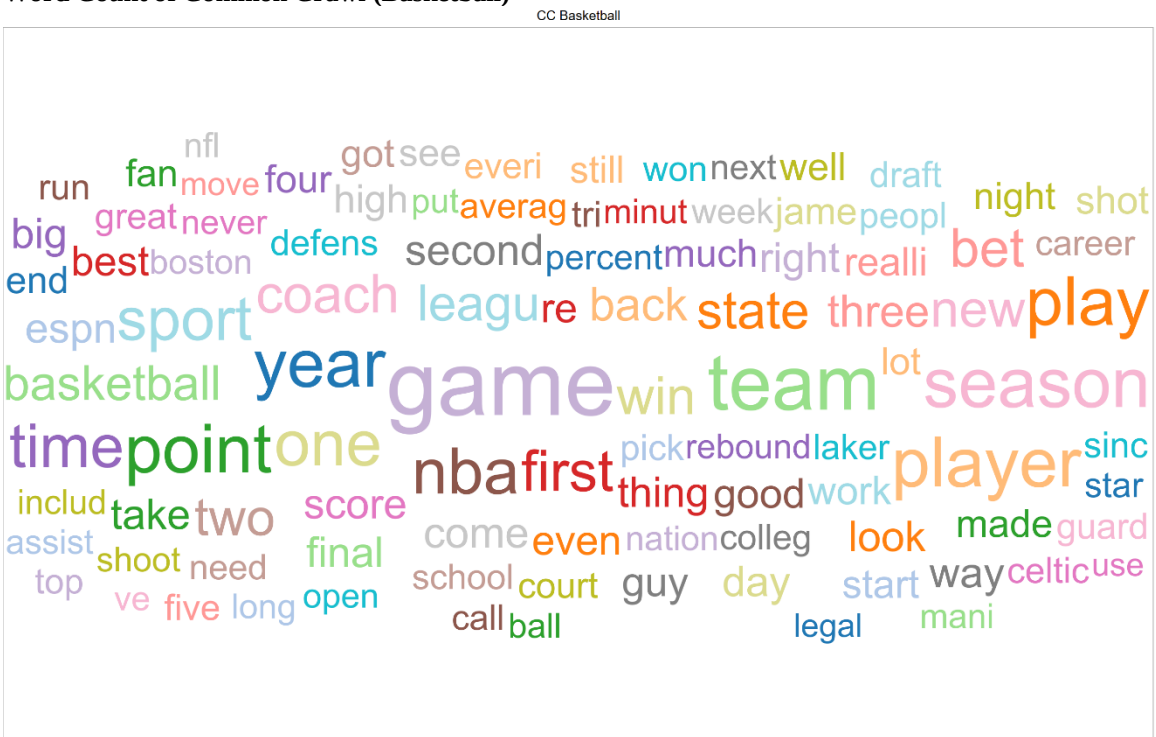
- We obtained the data from common crawl by passing the parameters url www.espn.com, match type: domain and got the output as JSON file. We selected espn as it covers all sports on daily basis.
- Used common index API to set the above parameters.
- Parsed the JSON file to get the articles by using offset, 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

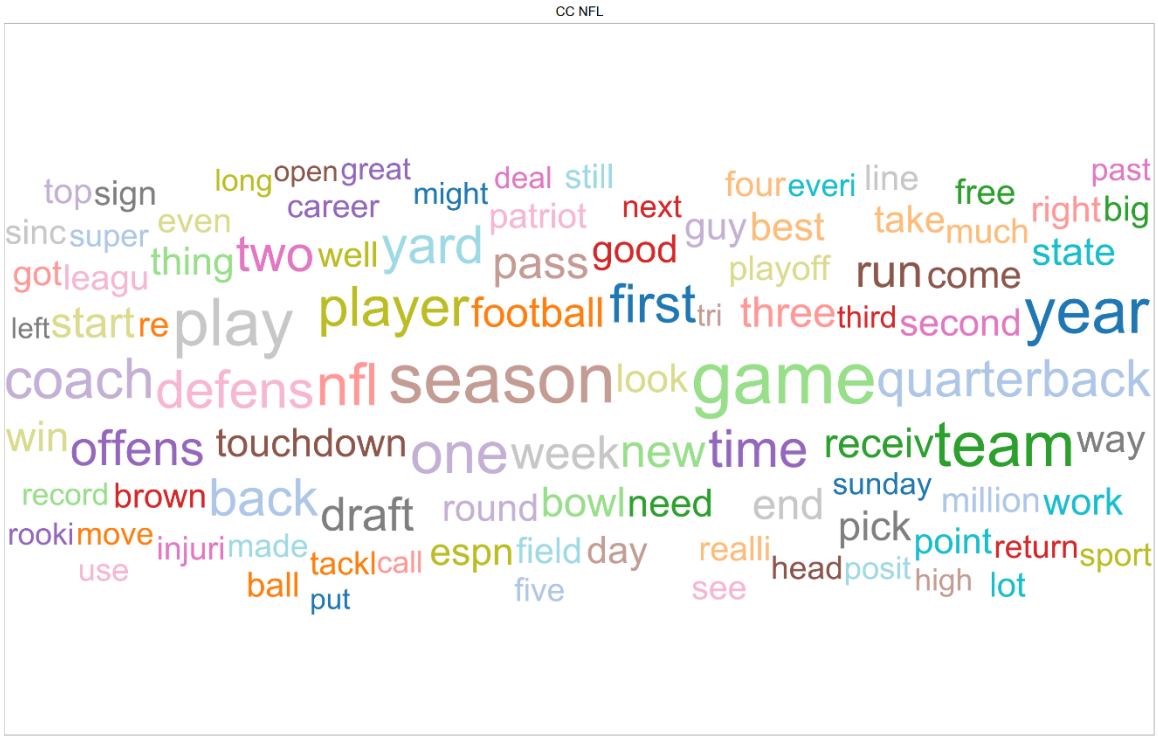
Word Count of Common Crawl (Baseball)



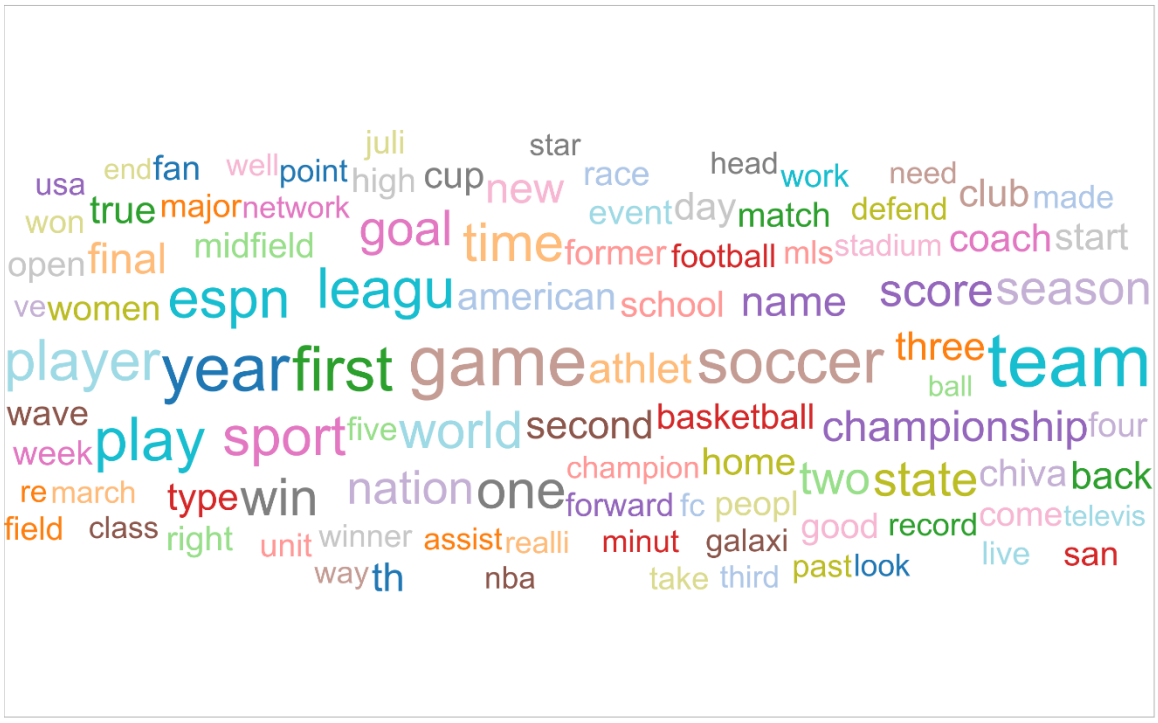
Word Count of Common Crawl (Basketball)



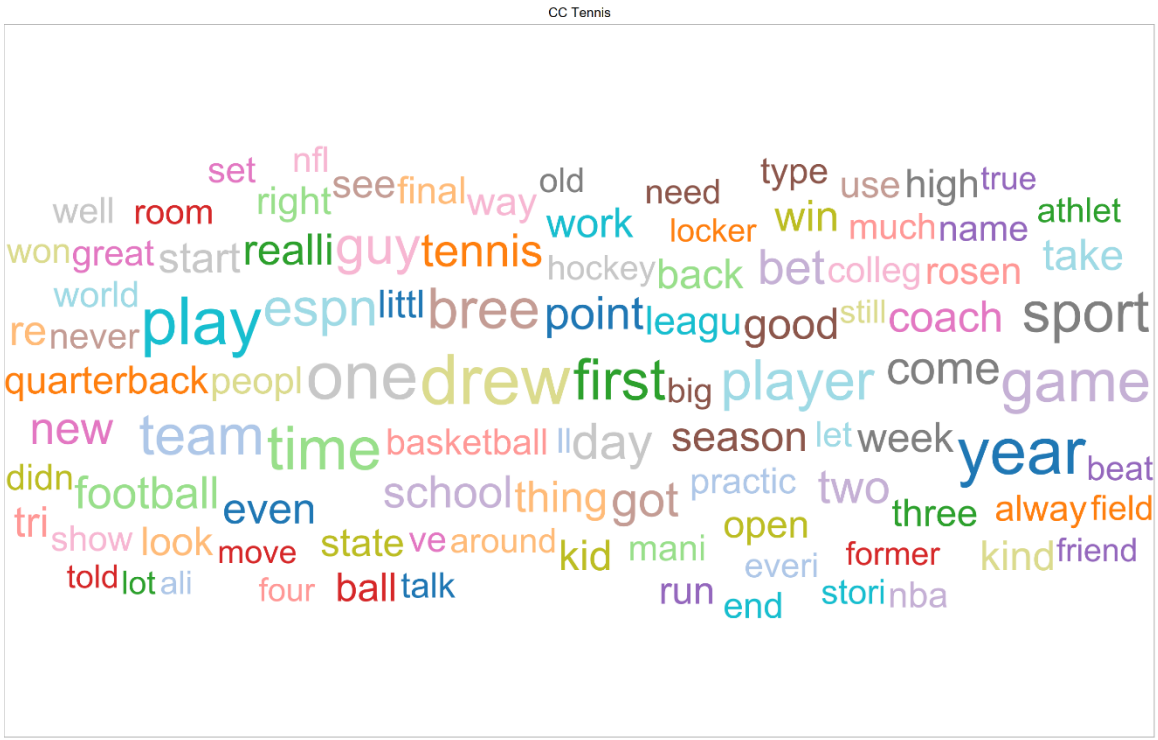
Word Count of Common Crawl (NFL)



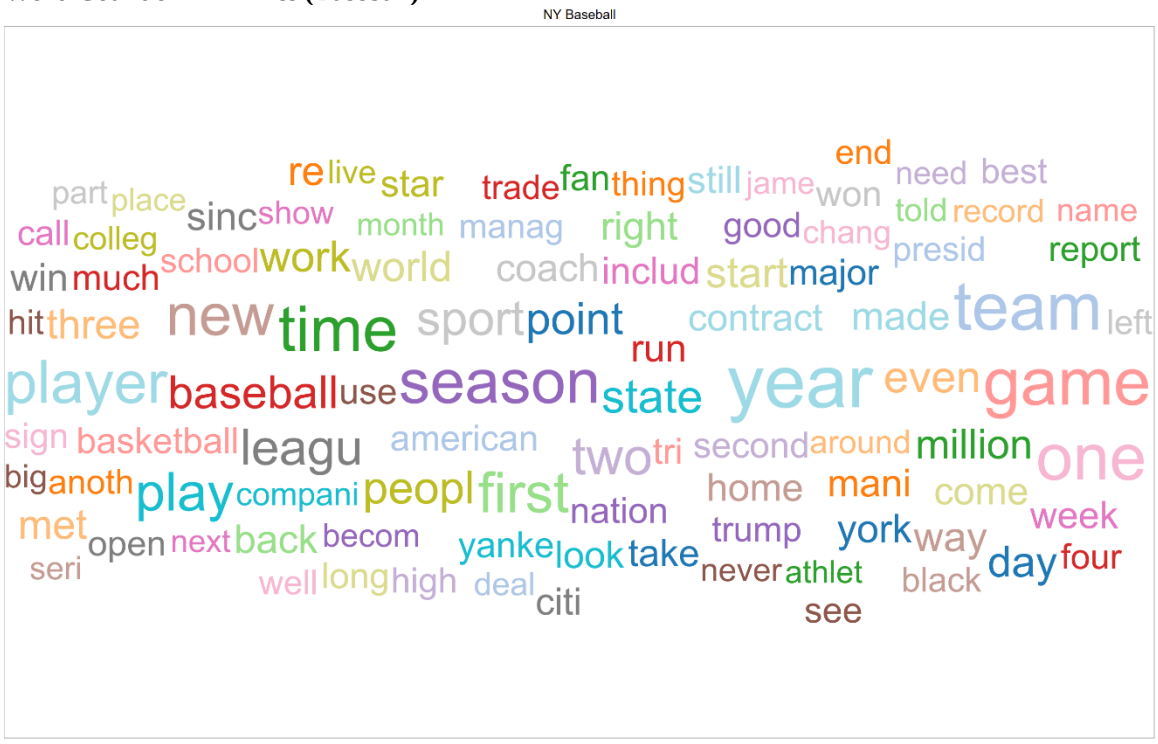
Word Count of Common Crawl (Soccer)



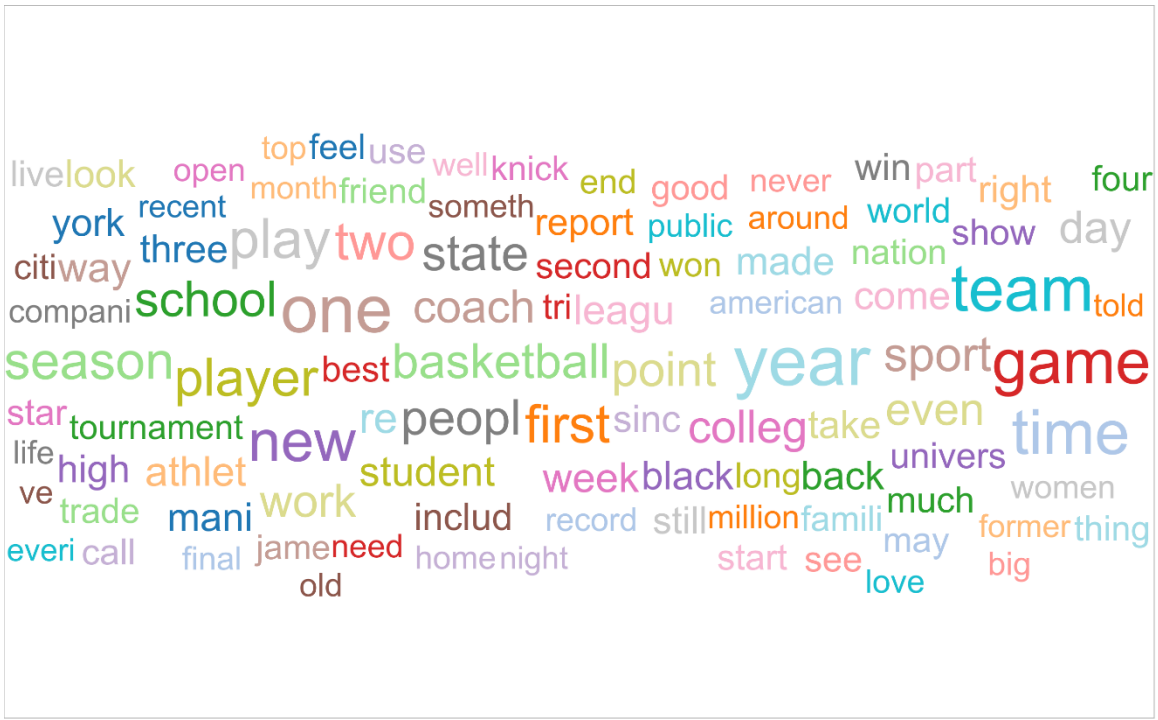
Word Count of Common Crawl (Tennis)



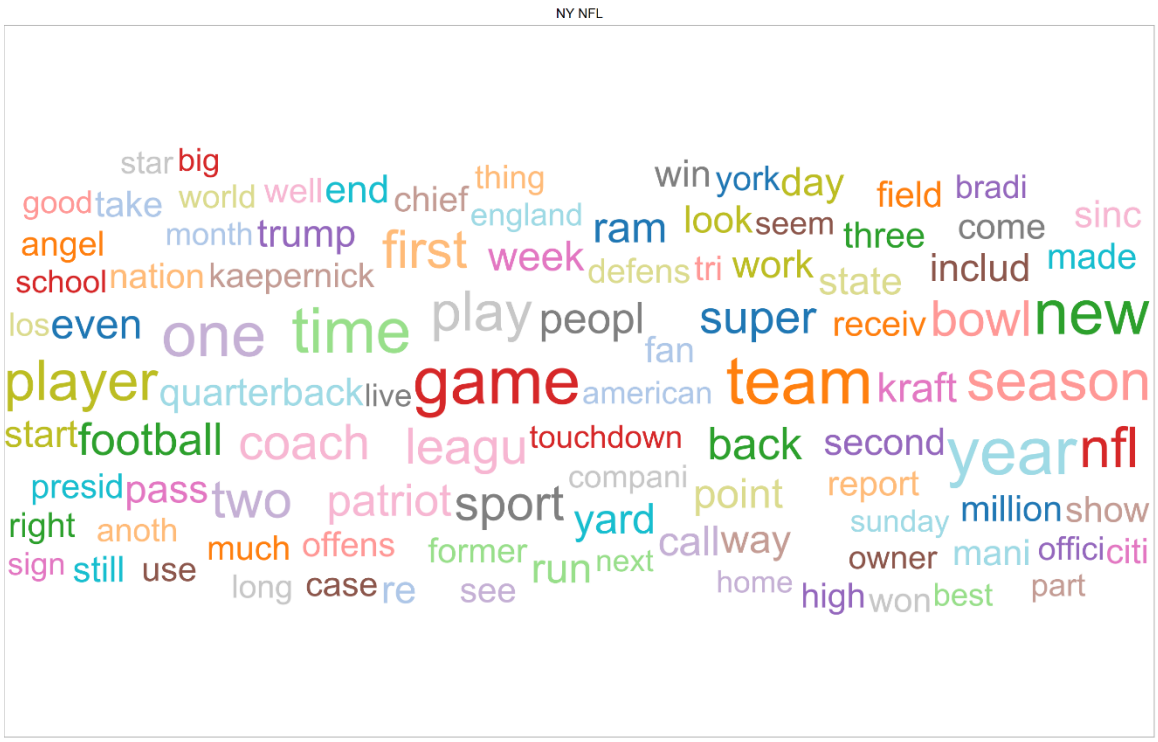
Word Count of NY Times (Baseball)



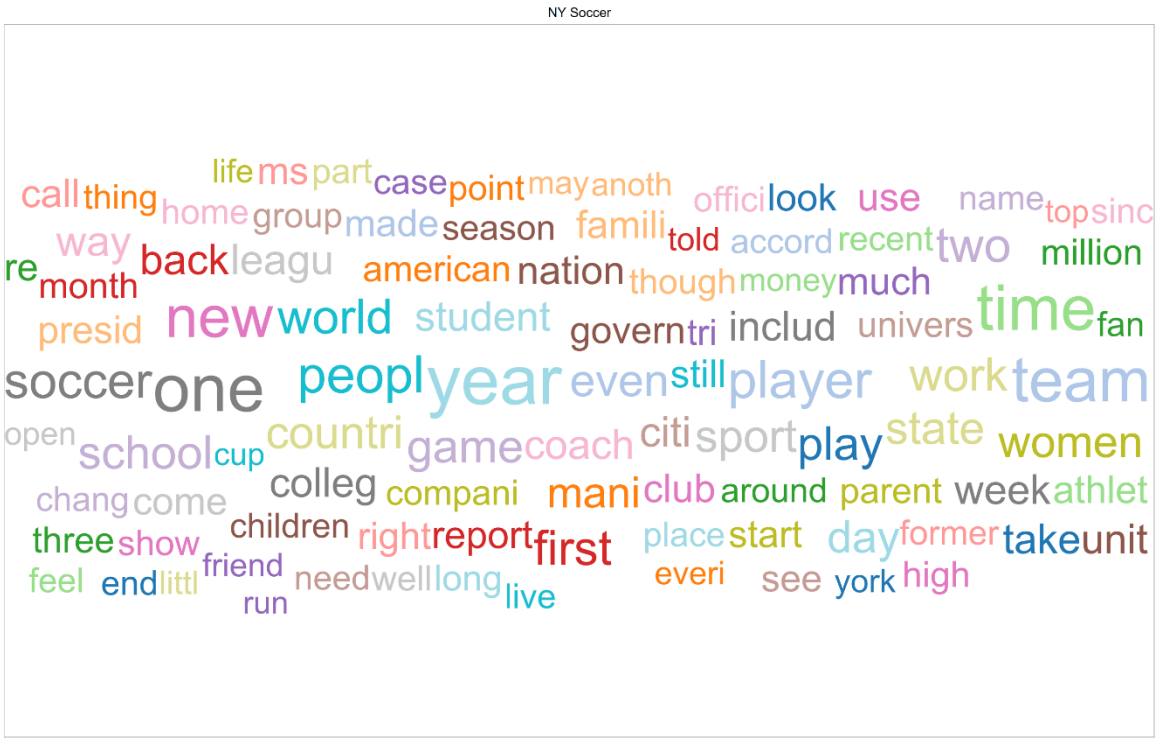
Word Count of NY Times (Basketball)



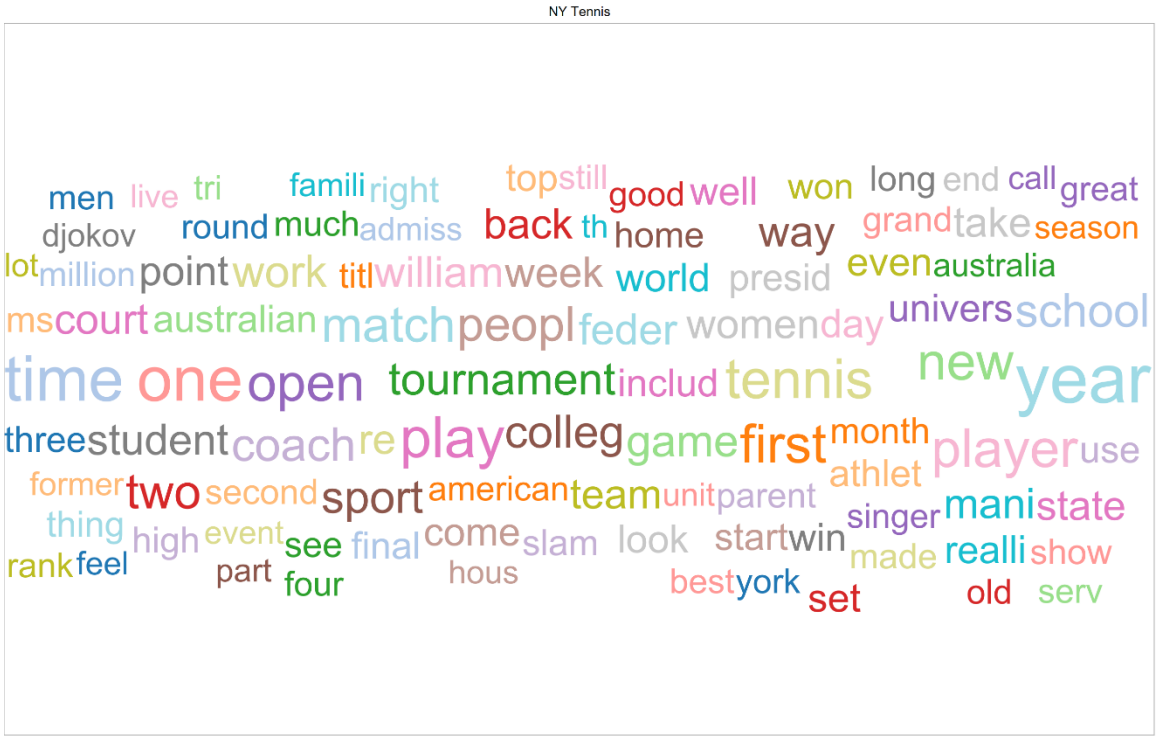
Word Count of NY Times (NFL)



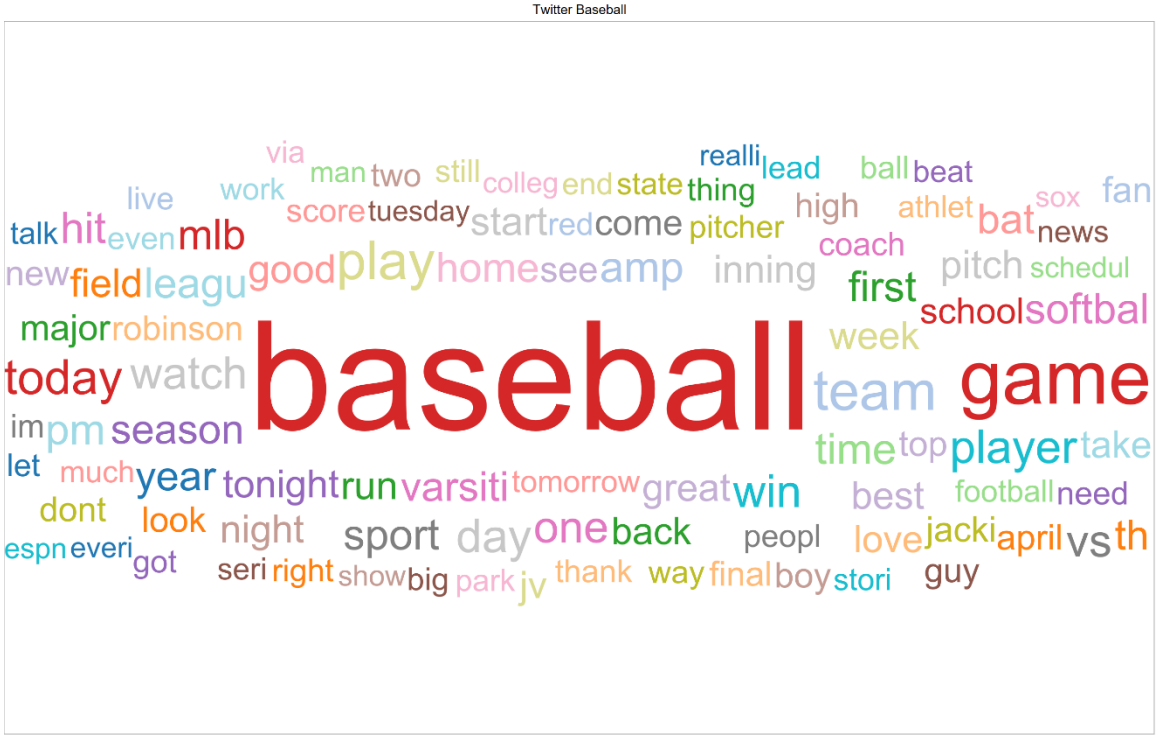
Word Count of NY Times (Soccer)



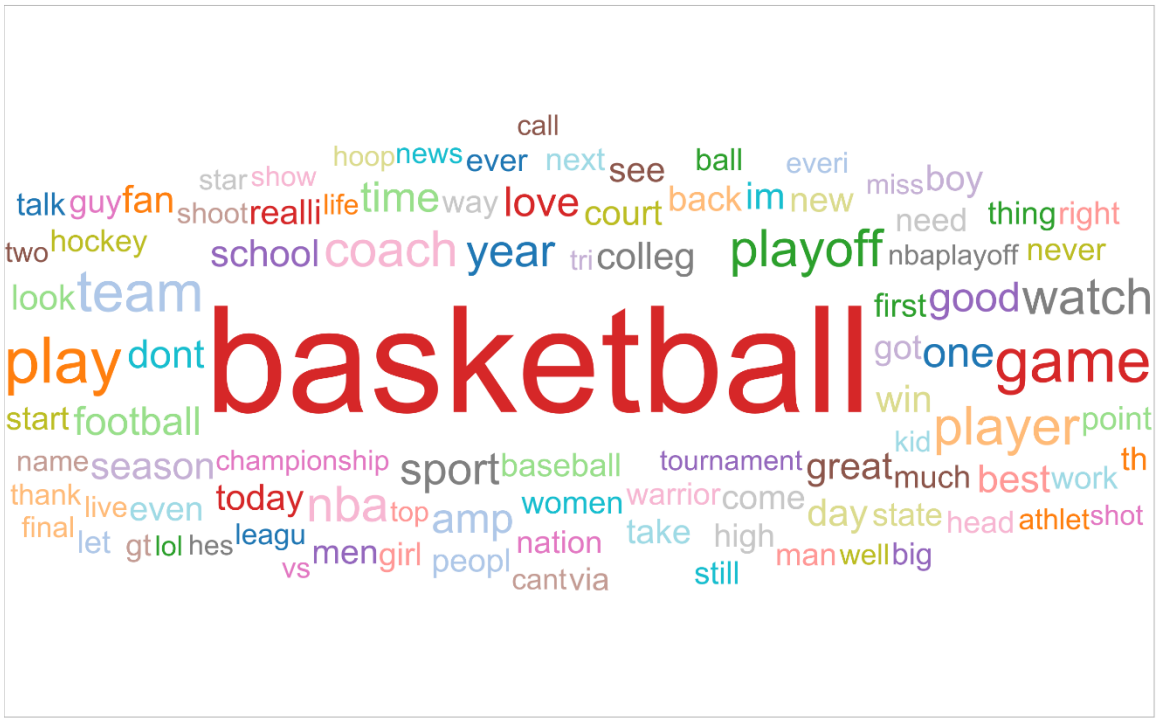
Word Count of NY Times (Tennis)



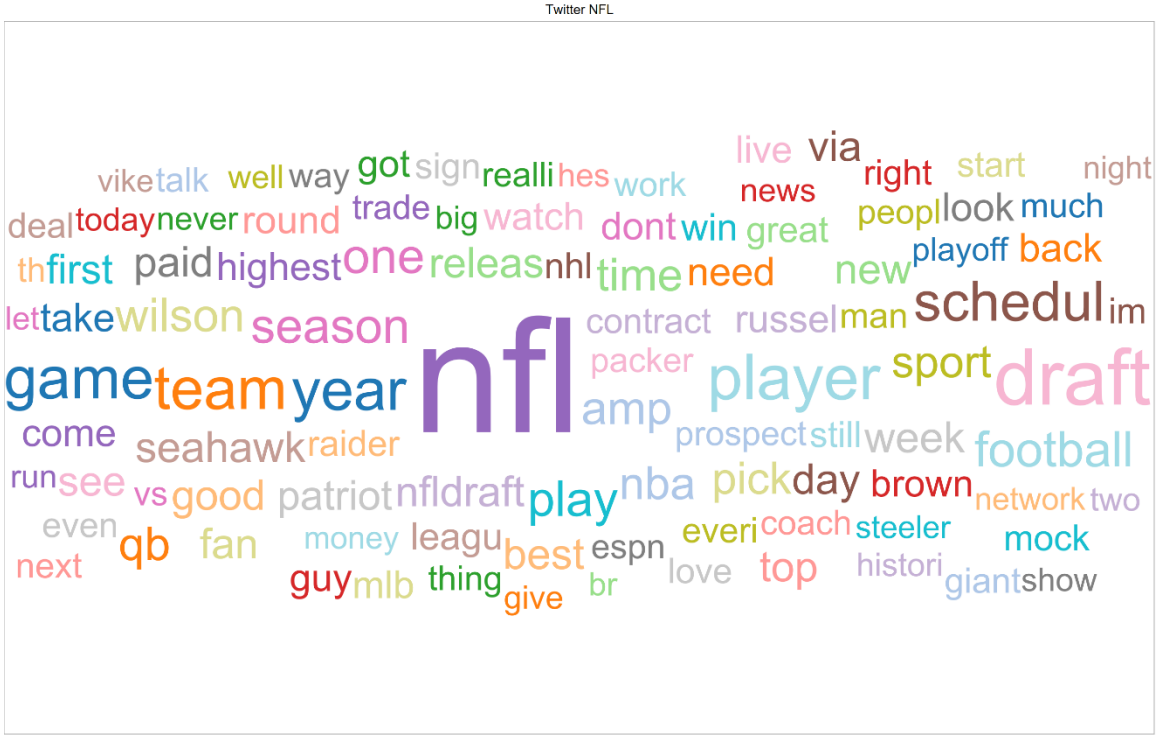
Word Count of Twitter (Baseball)



Word Count of Twitter (Basketball)



Word Count of Twitter (NFL)

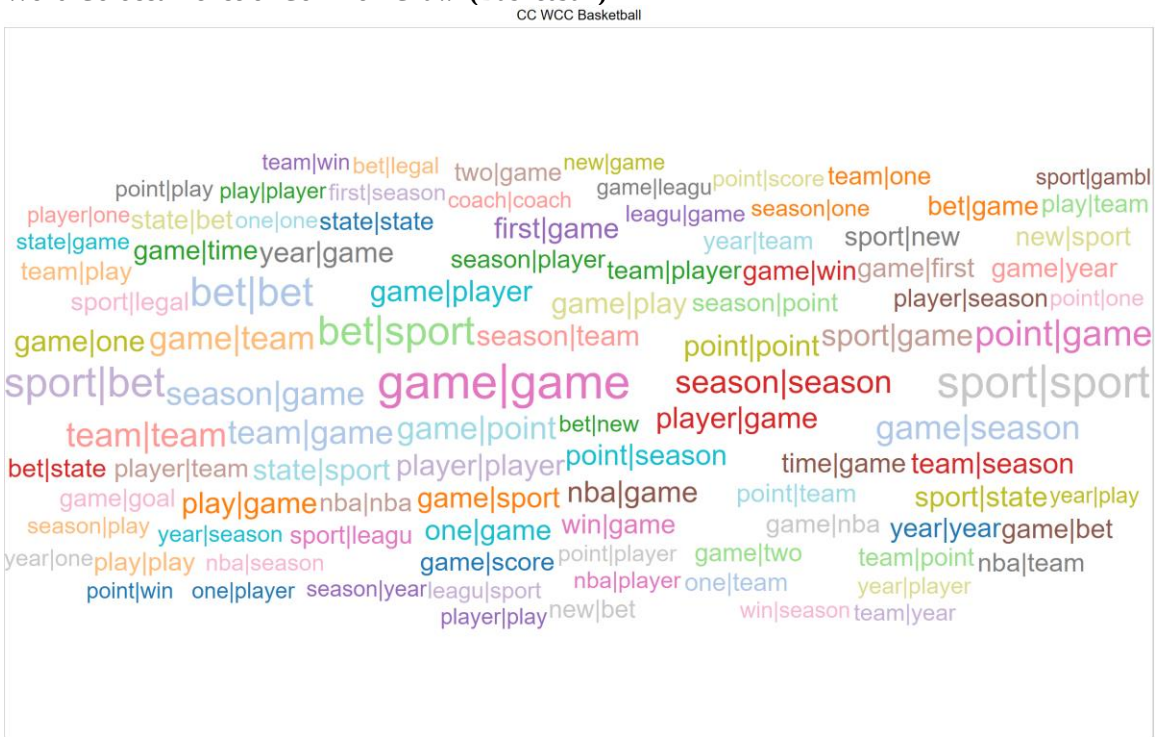


Word Count of Twitter (Soccer)

Word Count of Twitter (Tennis)

Word Co-occurrence of Common Crawl (Baseball)

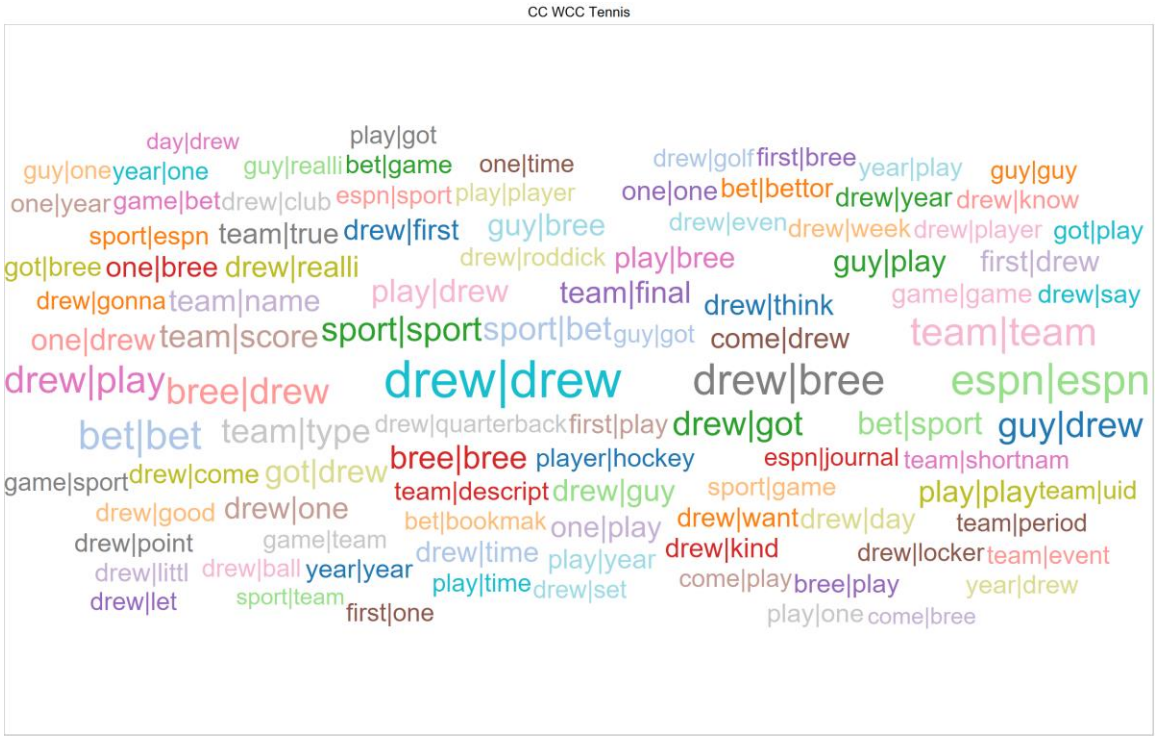
Word Co-occurrence of Common Crawl (Basketball)



Word Co-occurrence of Common Crawl (NFL)

Word Co-occurrence of Common Crawl (Soccer)

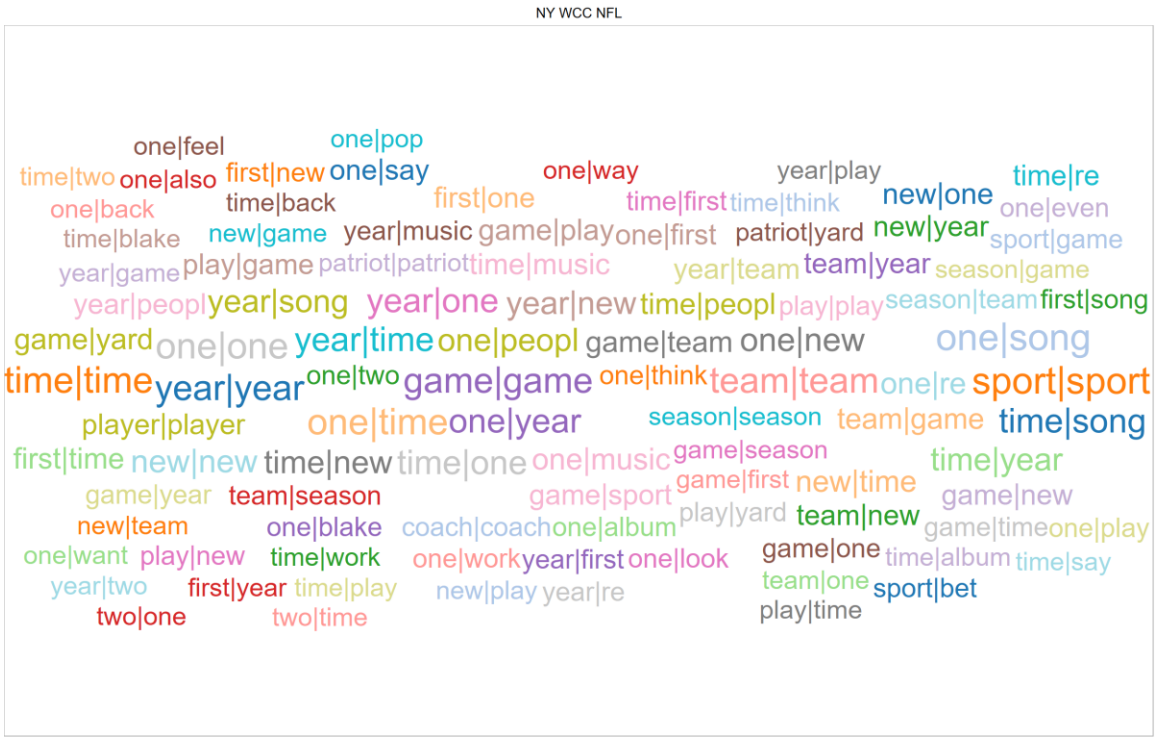
Word Co-occurrence of Common Crawl (Tennis)



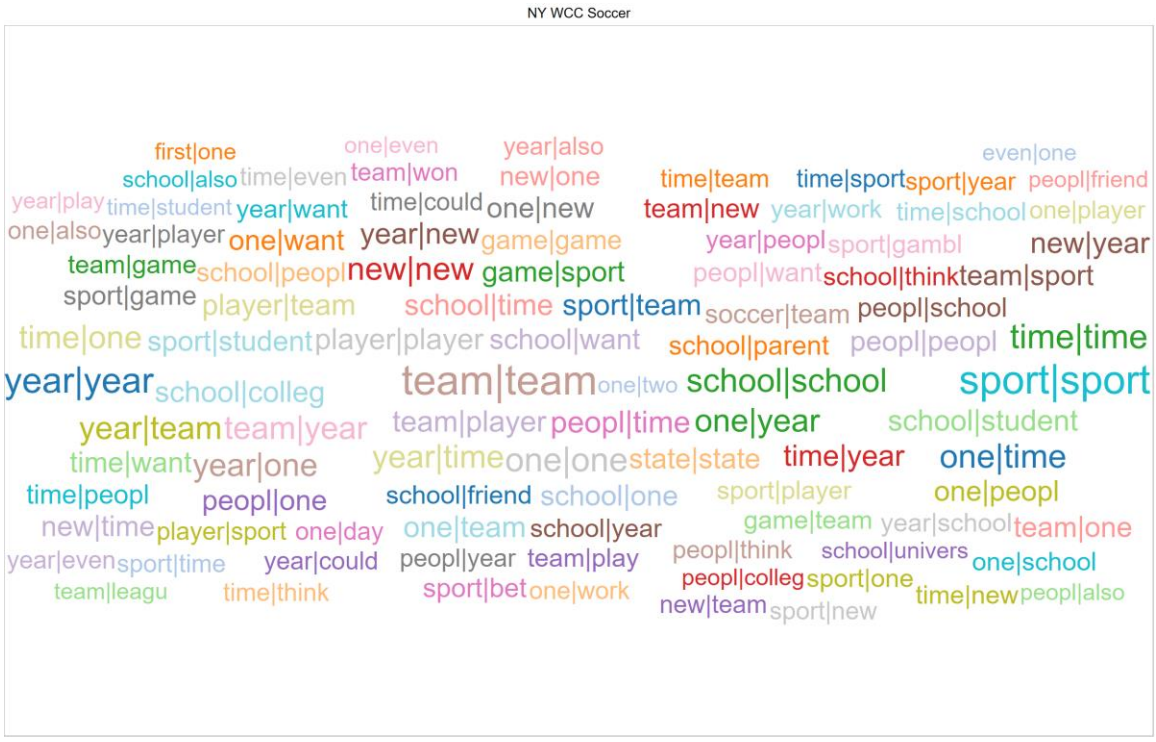
Word Co-occurrence of NY Times (Baseball)

Word Co-occurrence of NY Times (Basketball)

Word Co-occurrence of NY Times (NFL)

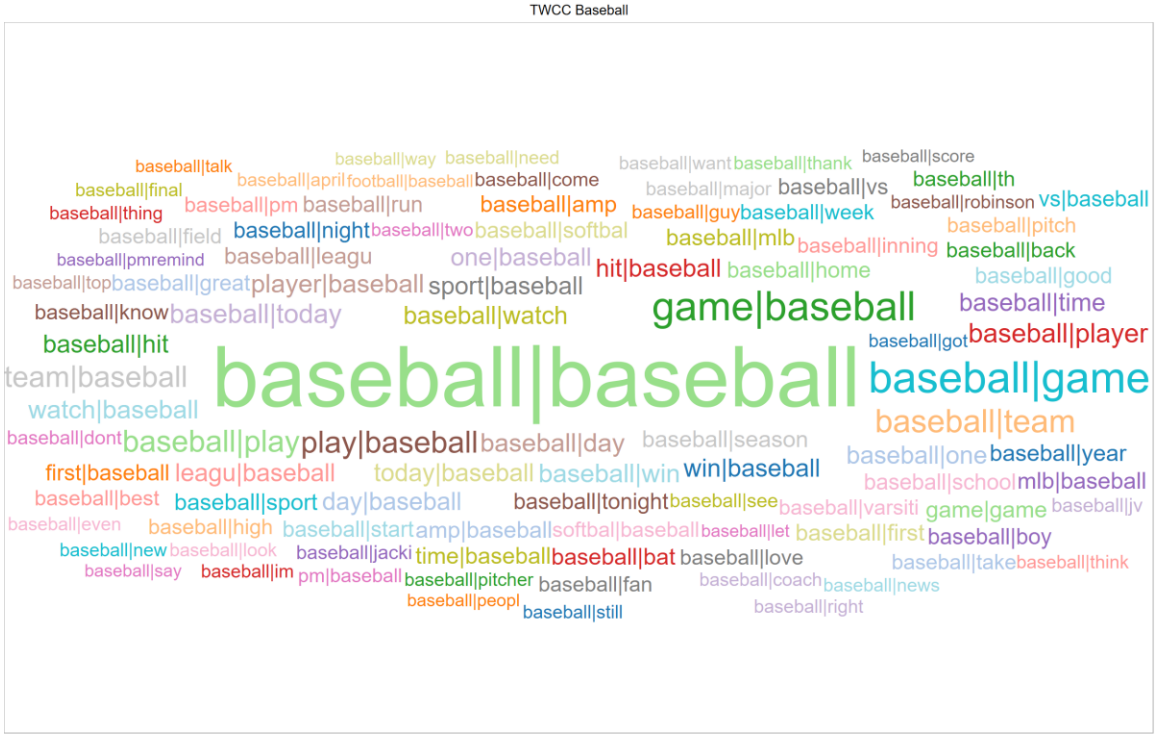


Word Co-occurrence of NY Times (Soccer)



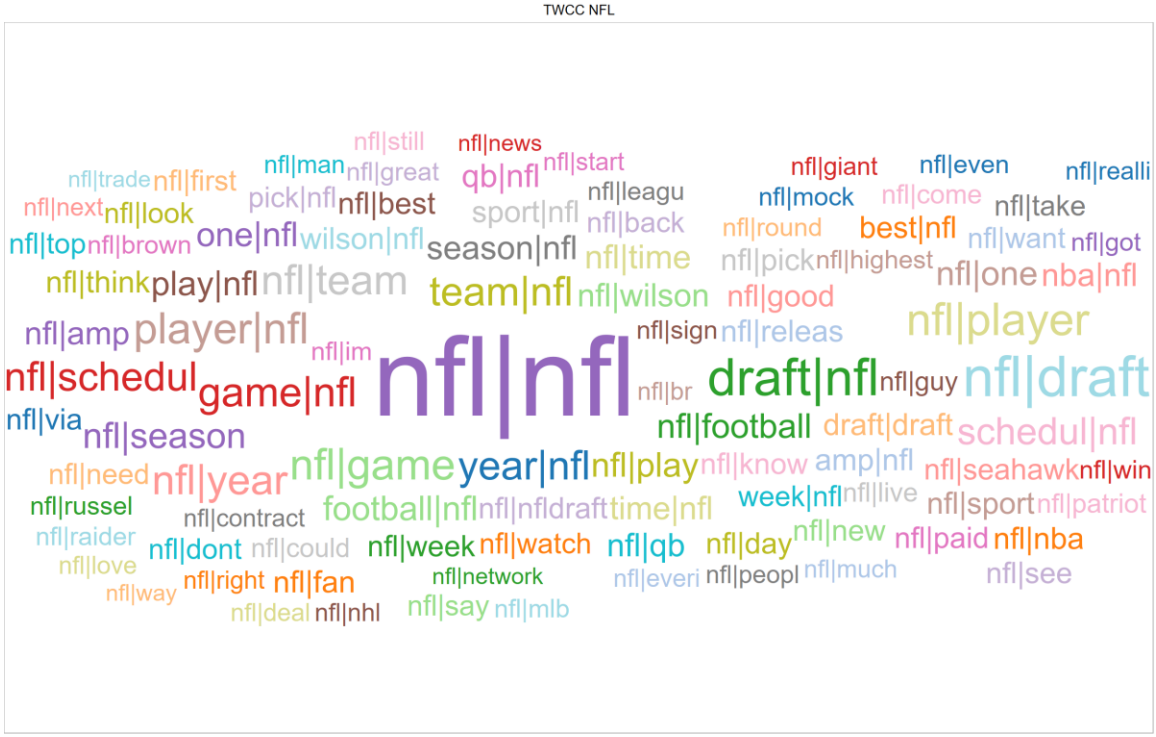
Word Co-occurrence of NY Times (Tennis)

Word Co-occurrence of Twitter (Baseball)

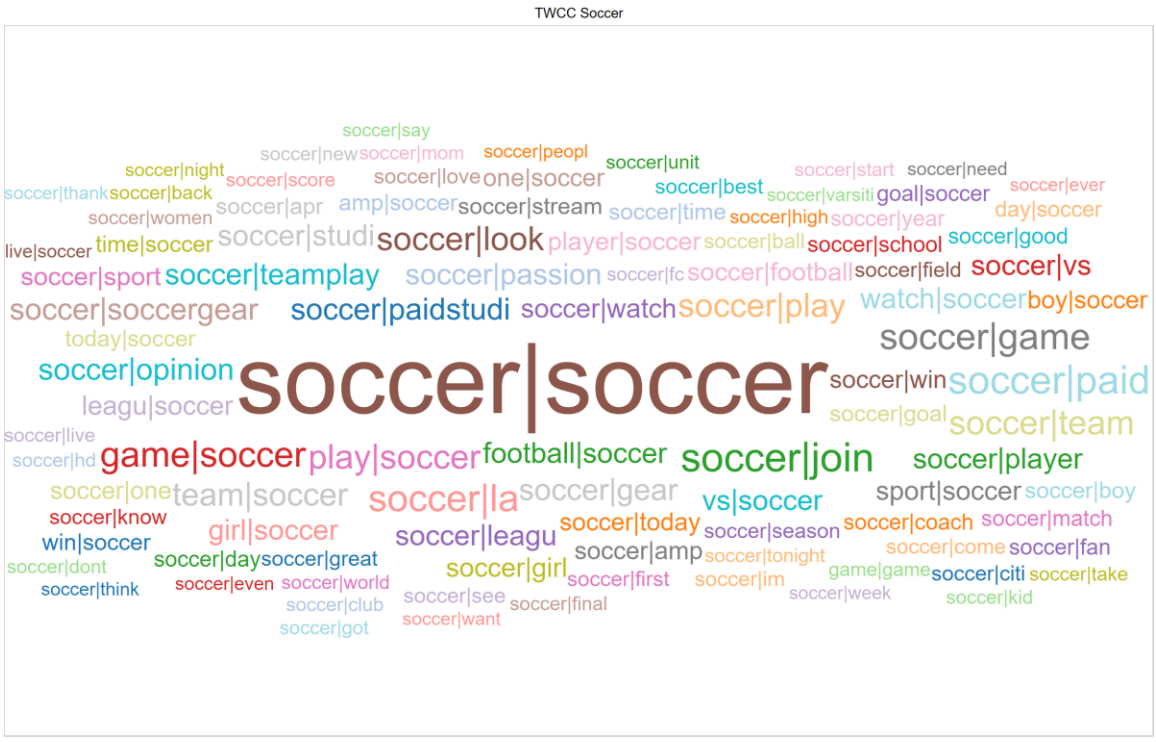


Word Co-occurrence of Twitter (Basketball)

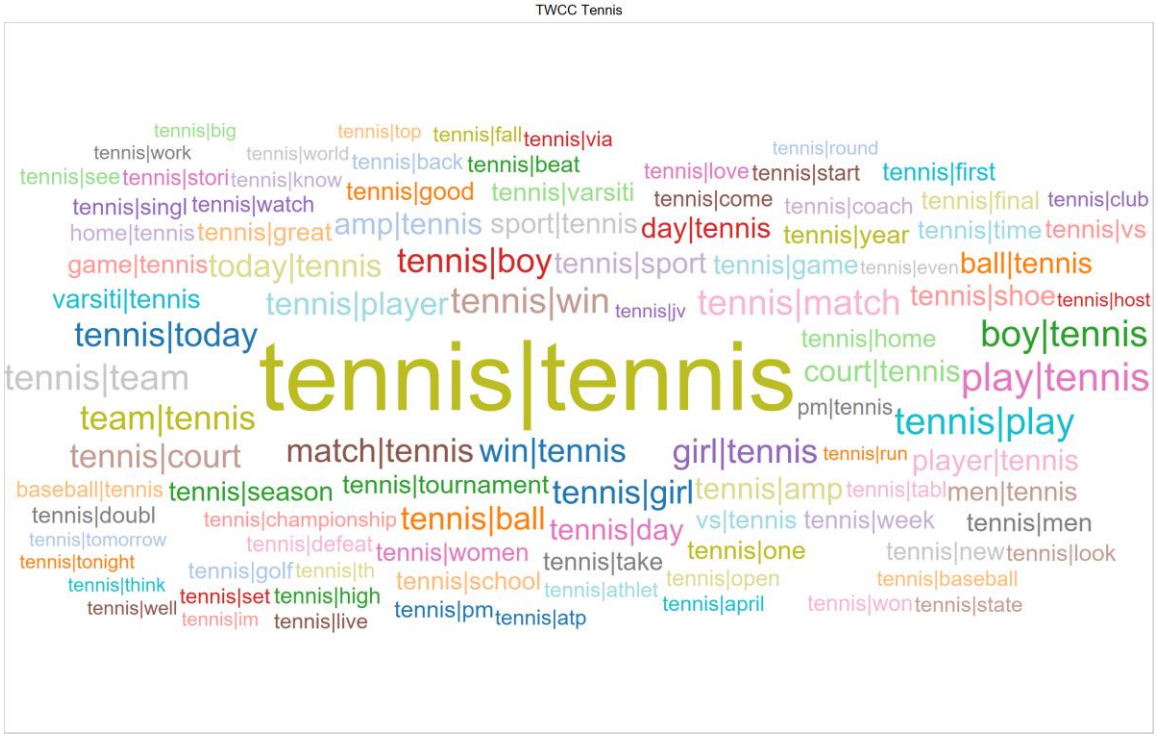
Word Co-occurrence of Twitter (NFL)



Word Co-occurrence of Twitter (Soccer)



Word Co-occurrence of Twitter (Tennis)

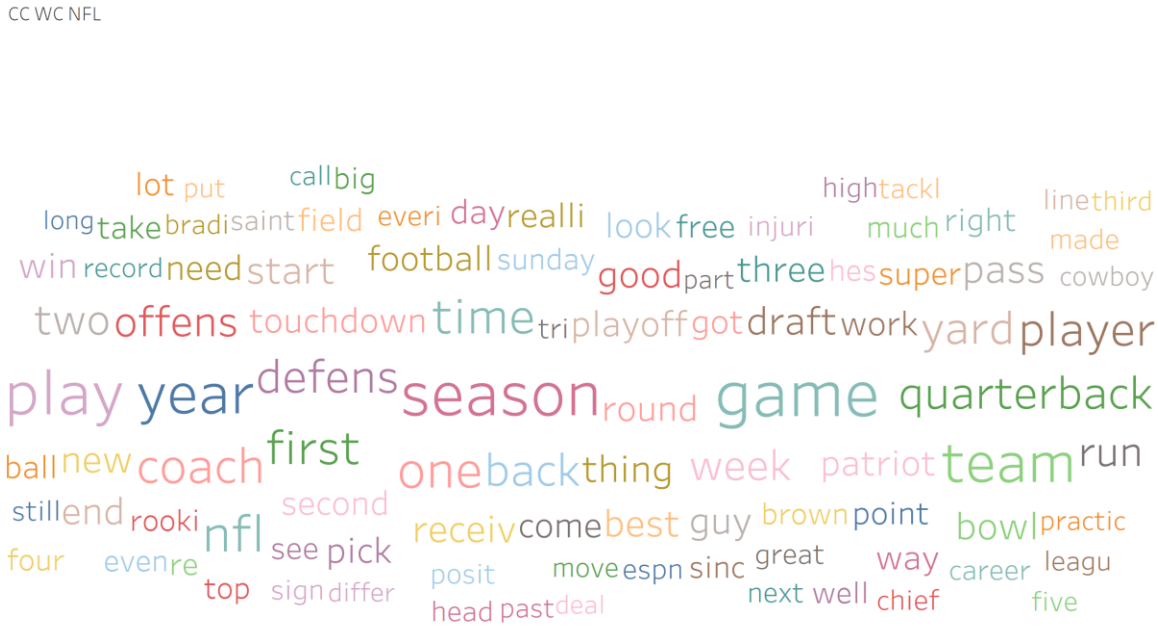


Small Data

Word Count of Common Crawl (Baseball)

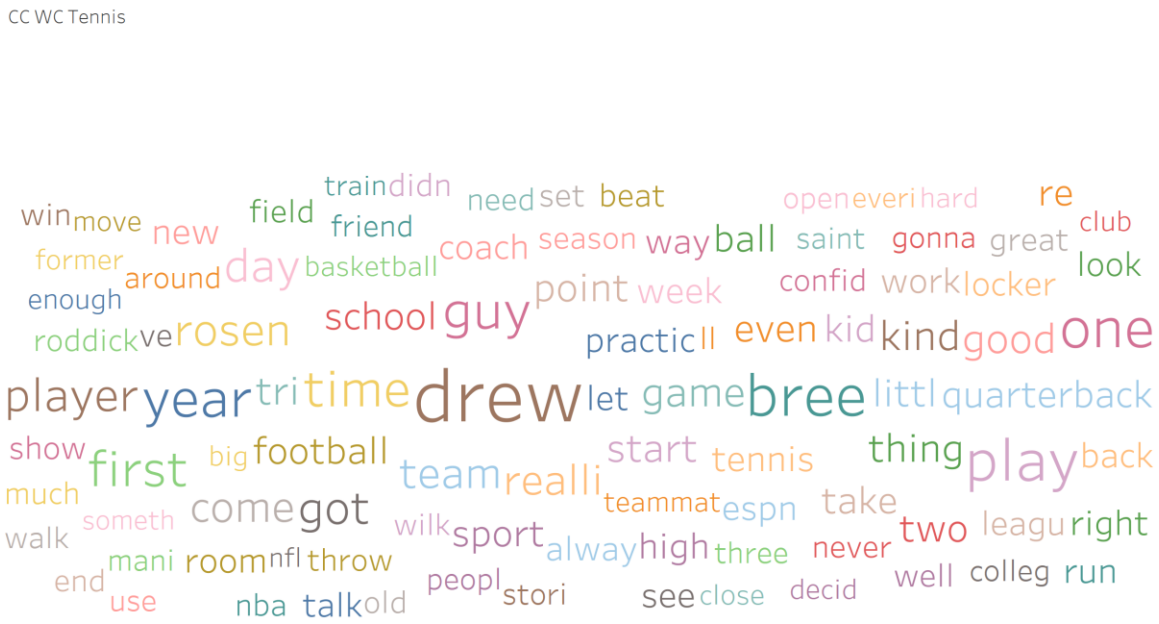
Word Count of Common Crawl (Basketball)

Word Count of Common Crawl (NFL)



Word Count of Common Crawl (Soccer)

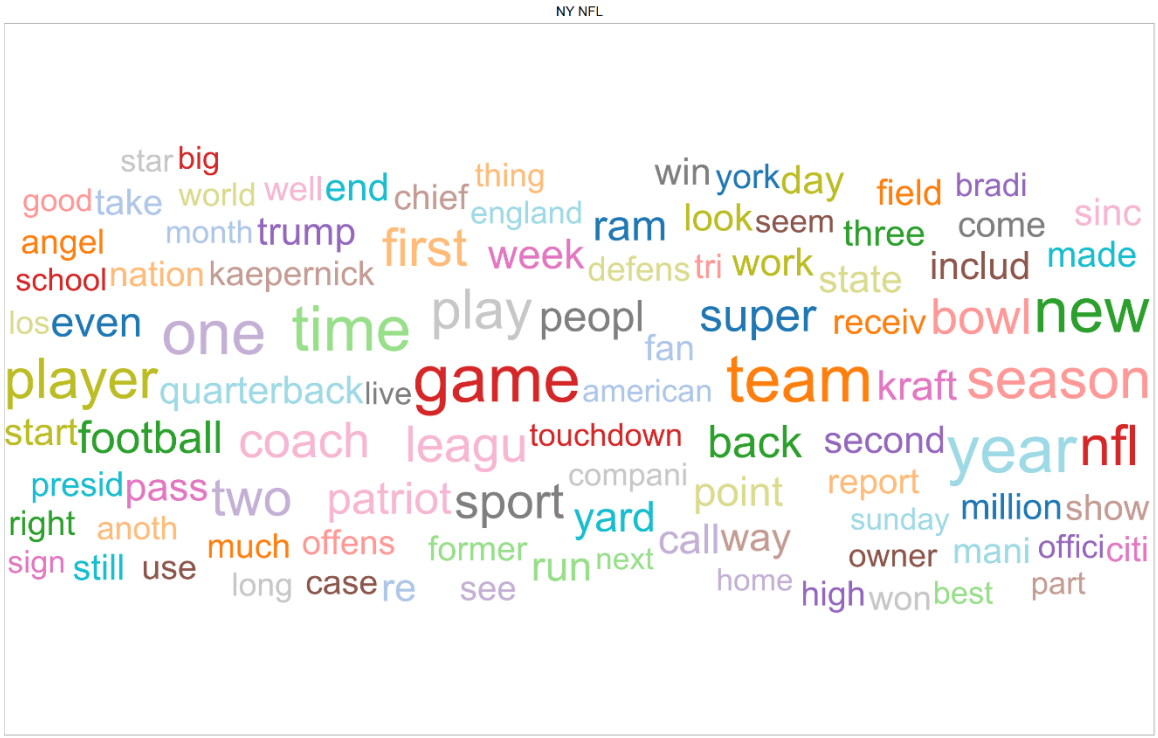
Word Count of Common Crawl (Tennis)



Word Count of NY Times (Baseball)

Word Count of NY Times (Basketball)

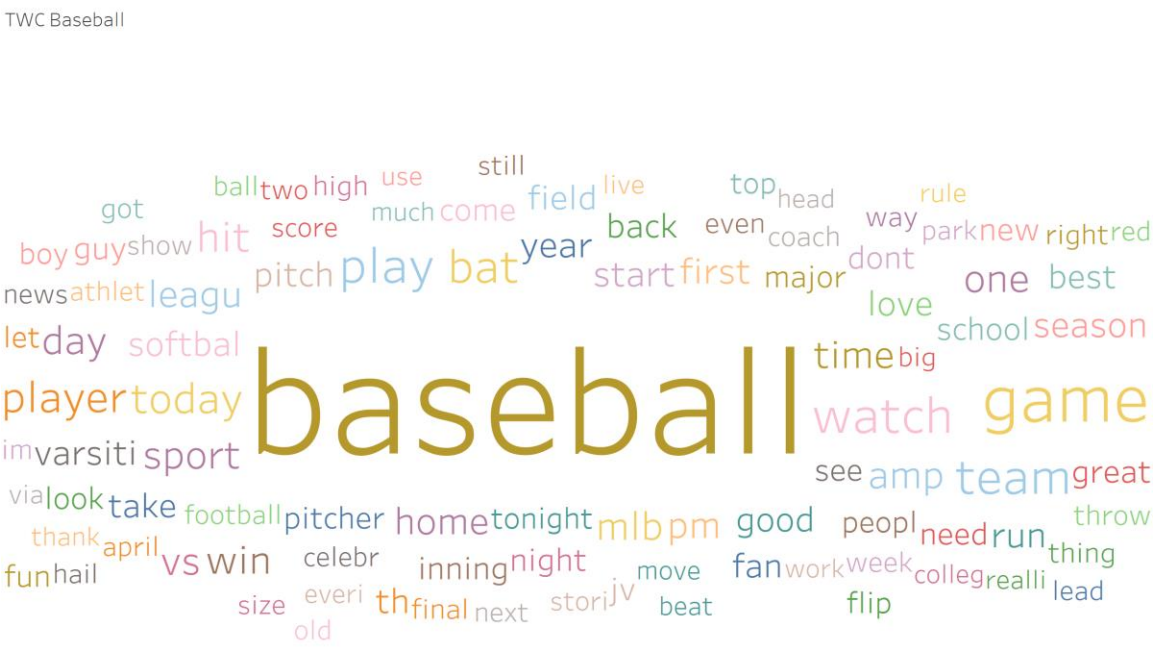
Word Count of NY Times (NFL)



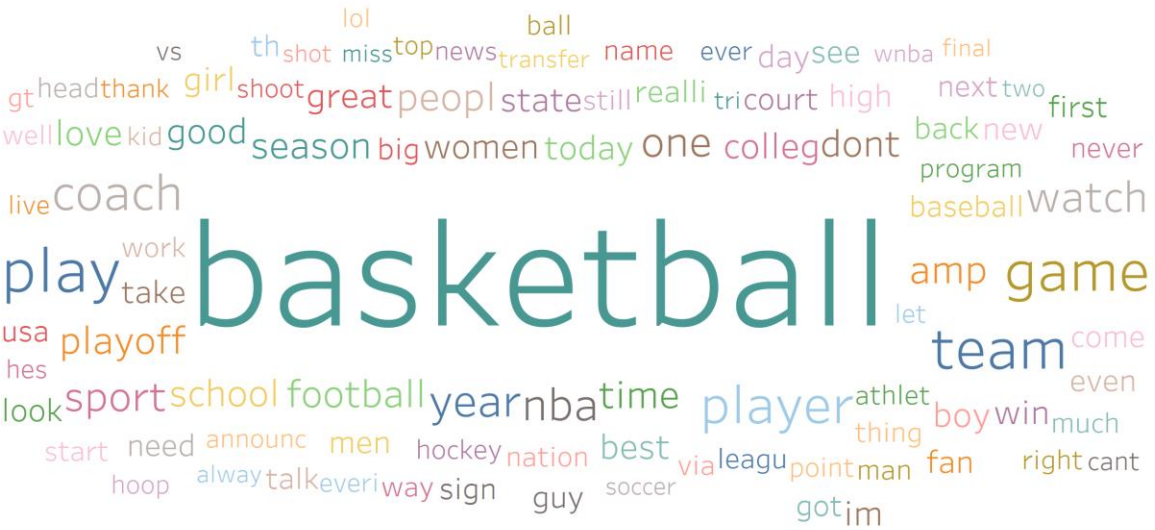
Word Count of NY Times (Soccer)

Word Count of NY Times (Tennis)

Word Count of Twitter (Baseball)



Word Count of Twitter (Basketball)



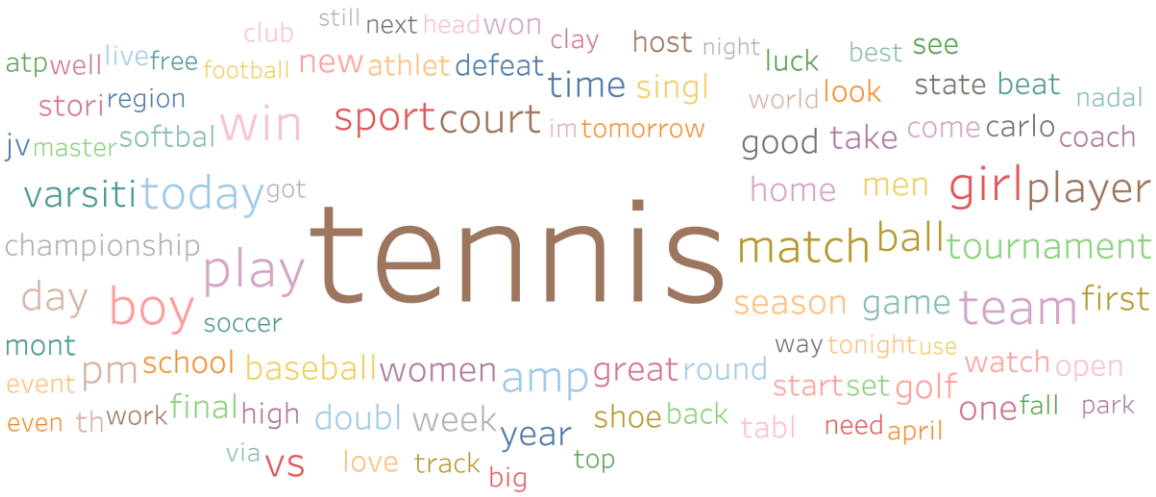
Word Count of Twitter (NFL)

TWC NFL

Word Count of Twitter (Soccer)

TWC 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)

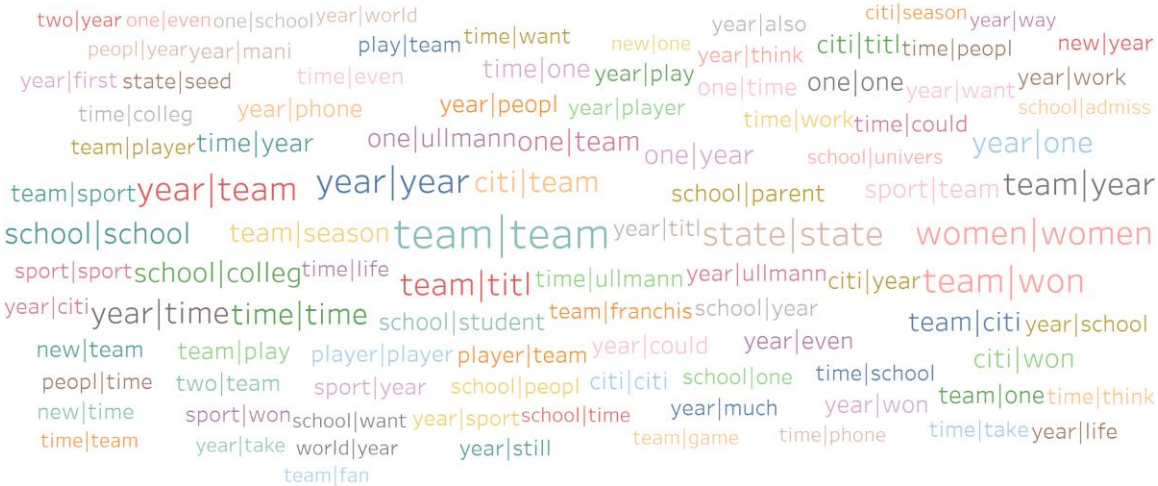
Word Co-occurrence of Common Crawl (Tennis)

Word Co-occurrence of NY Times (Baseball)

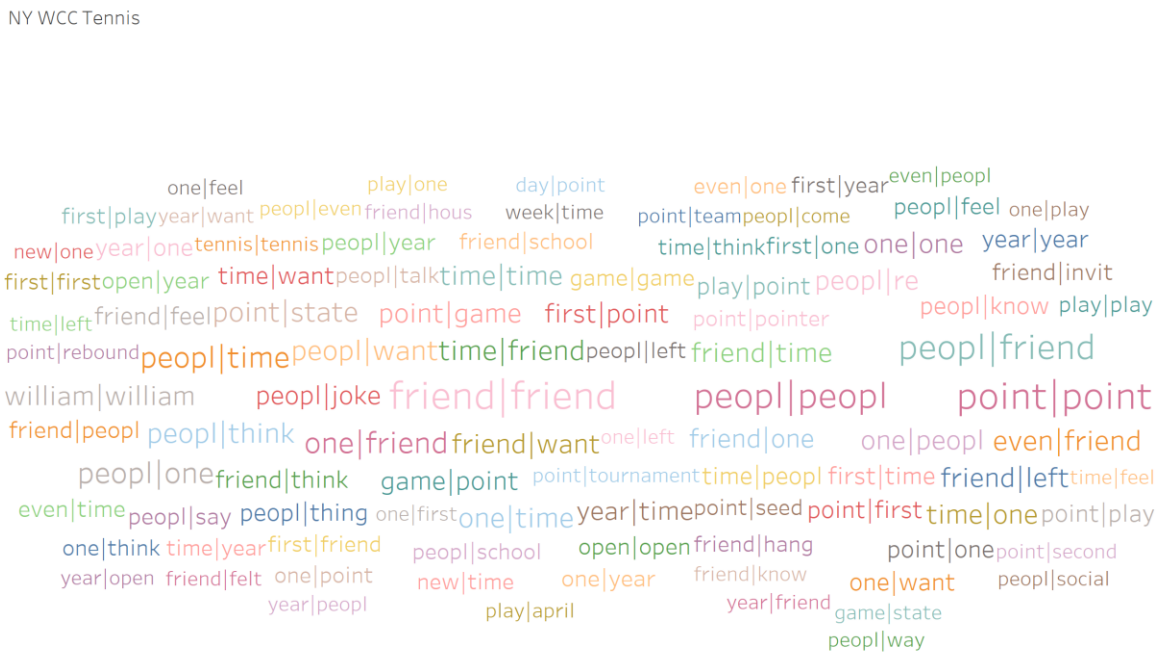
Word Co-occurrence of NY Times (Basketball)

Word Co-occurrence of NY Times (NFL)

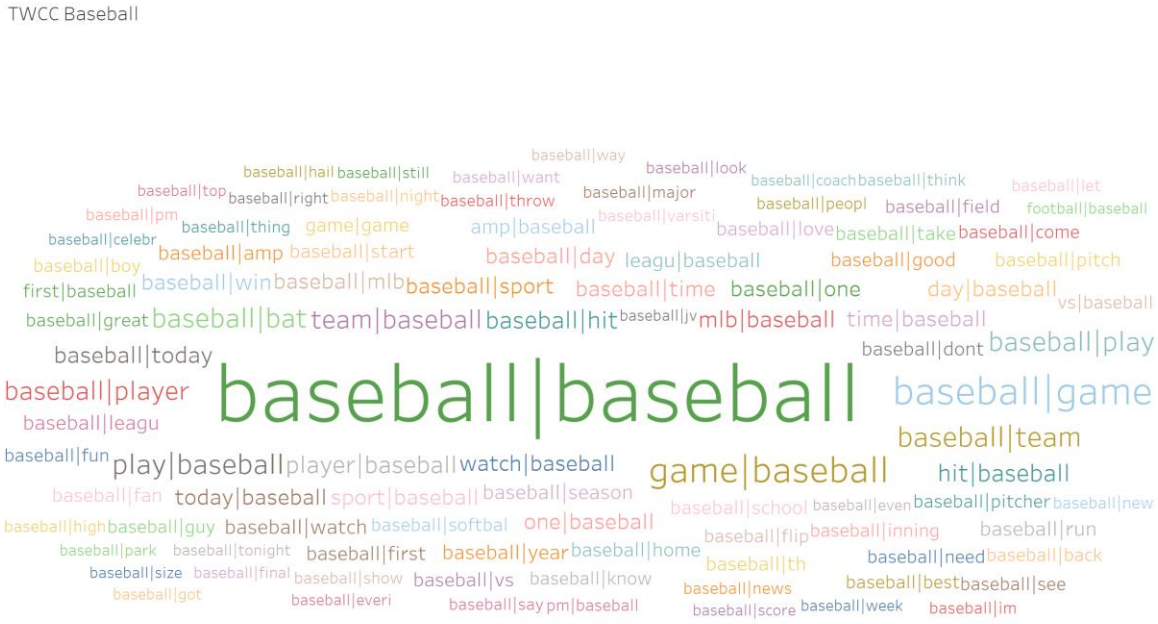
Word Co-occurrence of NY Times (Soccer)



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)