Milestone_Report

February 7, 2018

1 Milestone Report - NHL Sentiment Analysis

The NHL sentiment analysis will rank teams based on the positivity of the tweets by each team, tweets by bloggers/writers of the teams, tweets using specific team hashtags, and replies to team accounts. The tweets were collected from January 20th, 2018 until the All-Star break, which was January 25th, 2018. The purpose of this document is describe how the data was gathered and cleaned. The description of the initial exploratory analysis is also included.

2 Data Collection

The tweets for each NHL team were collected using Tweepy's streaming API. The API allows for key phrases or hashtags to be followed (like #nyr for the New York Rangers) and for accounts to be followed (like @NYRangers). Twitter handles of the teams and writers/bloggers were followed and team hashtags were tracked. In addition to this, the language can be specified and this was set to English. The stream was run for on a mostly continuous basis for the 6 days that it ran. There were a few breaks in the collection due to errors in the streaming code. The tweets were returned in JSON format.

3 Data Cleaning

Twitter provides all of their tweets in the JSON file format. The first step was to try to normalize the data. However, not every name/value pair exists for each tweet that is collected (example, not every tweet has a quoted tweet, therefore the quoted tweet name/value pairs do not exist for those tweets). Instead of normalizing, each name/value pair was put into a dataframe, which consisted of 340+ columns and 425k tweets, with a lot of rows containing NaN values. The data included columns for hashtags and mentions for tweets, retweets, quoted tweets, and retweeted quoted tweets. It also had columns for replies and two text columns (short and extended) for each of the four kinds of tweets. Once all of the data was loaded into a dataframe, hashtag and mentions columns had to be cleaned as they contained embedded JSON. Each column was cleaned up to contain a list of each hashtag and mention from the tweet, which was done for determining the team(s) associated to the tweet. After this was done, most columns were removed from the dataset except the date, all text columns, all reply columns, all mentions columns, and all hashtag columns (about 30 columns in total). The next step of the cleaning process was to determine the team(s) associated to each row of the dataframe. Using the hashtags, mentions, and reply columns, I was able to generate a list teams associated to each tweet based on the hashtags tracked and accounts followed for each NHL team. Once this was done, the reply, mentions, and hashtag columns were removed from the dataset. After generating a column for the teams, the next thing to do was to reduce the text columns to one column. Each tweet contained a short version of a tweet (140 characters) and an extended version (280 characters). These were combined by looking at whether both text columns contained text and if so, the extended text was used. If not, then the shortened version was kept because the tweet was less than 140 characters. This was done for all of the types of tweets and the number of text columns was reduced to 4, from 8. These 4 columns of text still contained duplicate text, like tweets and retweets and quoted tweets and retweeted quoted tweets. To reduce the duplication, the tweets and retweets were reduced by using the retweet text if both had text and using the tweet text if only the tweet had text. The same process was done for the quoted tweet and the retweeted quoted tweet. The remaining columns were removed which left 2 text columns. The quoted tweets were then added to the tweets column, which added another 40k tweets.

3.1 Cleaning Text Data

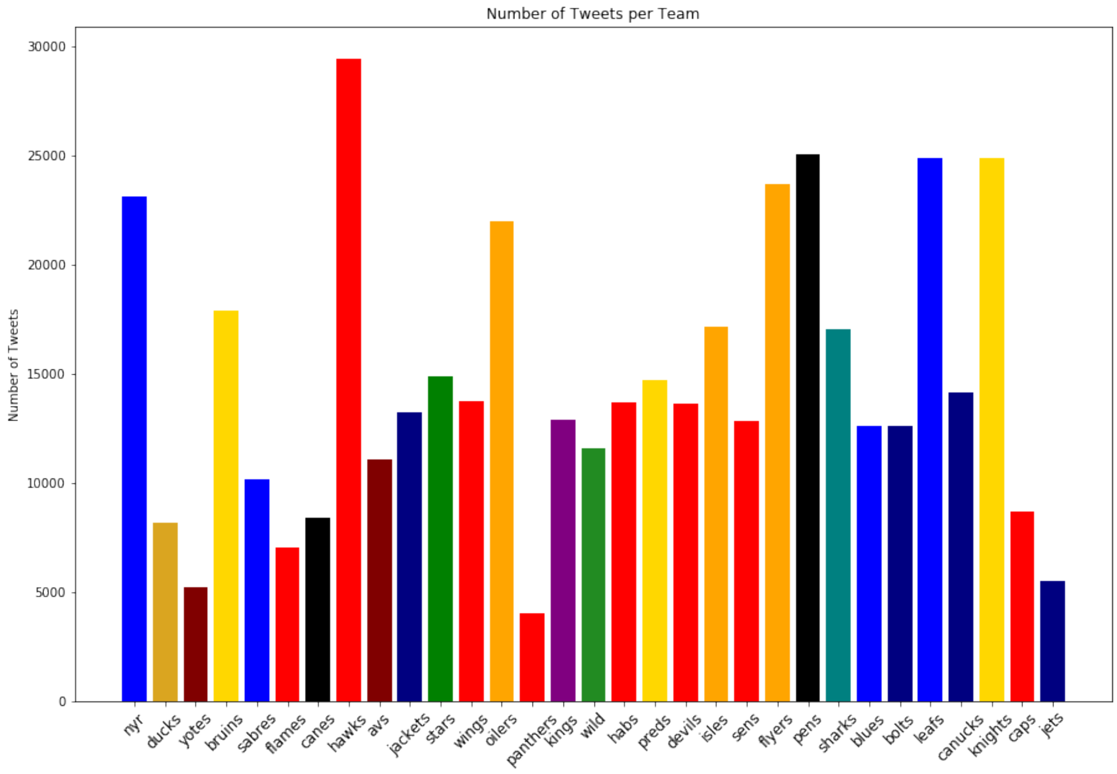
The next phase of the cleanup involved removing tweets that were irrelevant, correcting spelling errors, and spelling out acronyms. To clean out irrelevant tweets, I scanned the document looking for hashtags that did not reference hockey teams and tweets about politics. These tweets, which amounted to about 25k tweets, were removed from the dataset. Using a list of common acronyms in social media (lol, jk, etc.) and hockey (pp, ppg, etc.), each row of the dataset was cleaned up. In addition to cleaning up the acronyms, URLs were removed from the tweets so that the URLs would not cause issues with the sentiment analysis. The next step was to build a program to find potential misspellings. In order to do this, the text was put in a separate dataframe where hashtags, mentions (@), emojis, and URLs were removed. The text was then put through a spell check function in python, which returned 29k possible spelling errors. Using this 29k potential errors, I corrected about 60 errors, with more spelling fixes to come.

4 Exploratory Analysis

The initial exploratory analysis breaks down the total number of tweets per team, the number of tweets by hour for the NHL and each team (not all were included in this report as the file became too large), the most frequent words and bigrams for the NHL, and word clouds for each team (not all were included in this report as the file became too large).

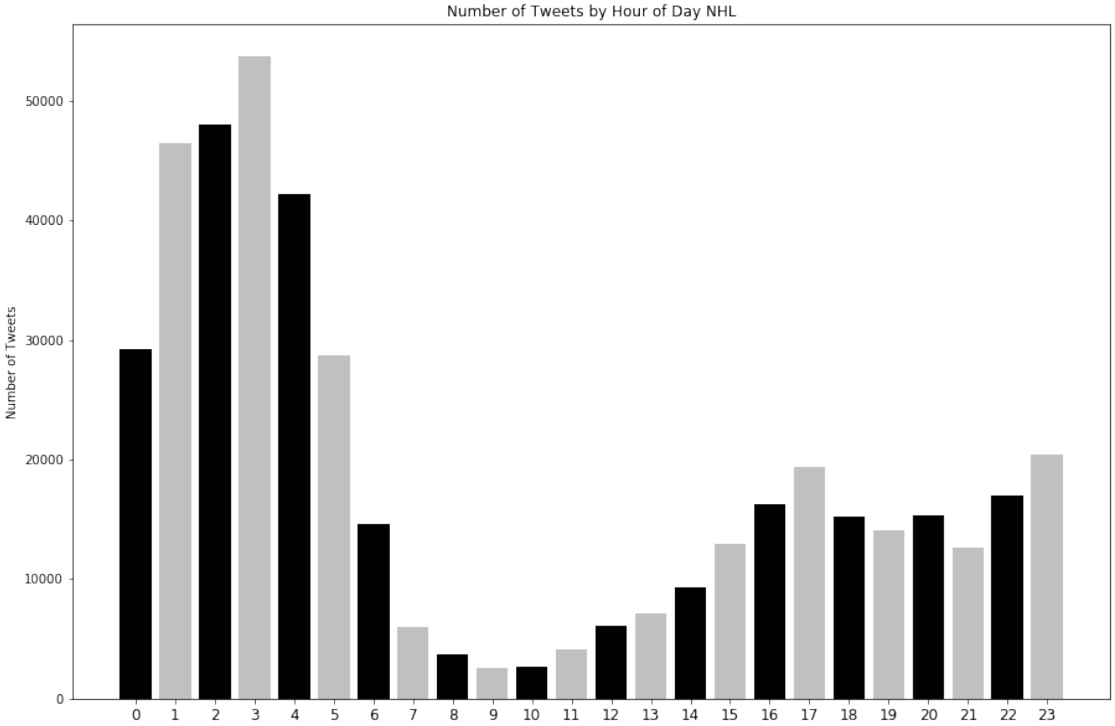
5 Tweets Per Team

The following bar plot shows the number of tweets collected per team after all of the data cleaning. The Chicago Blackhawks had the most tweets out of all of the teams and the Panthers had the fewest. The biggest surprise is the Golden Knights as this season is their first season in the NHL.



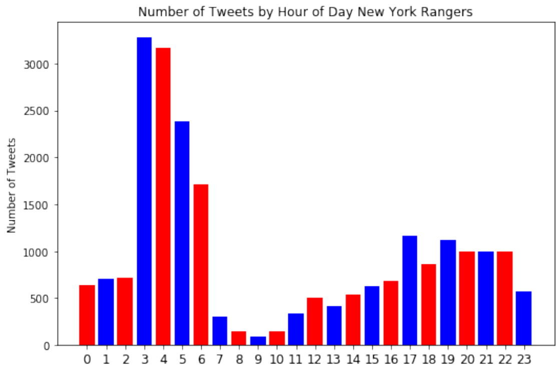
6 Tweets by Hour of Day for the NHL

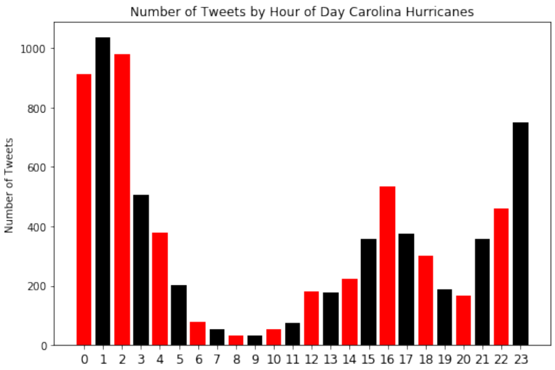
The bar plot below shows the most popular tweet times (24 hour clock) for the entire NHL. The most popular times were around 5PM through 4AM, which coincides with pregames, the actual games, and post-games.

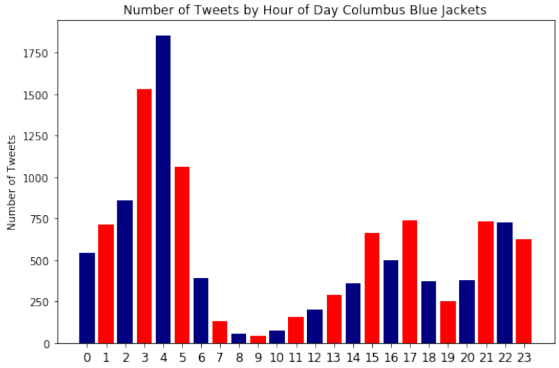


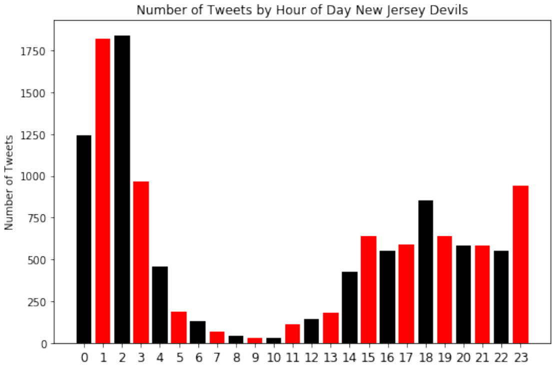
6.1 Tweets by Hour of Day - Metropolitan and Central Divisions

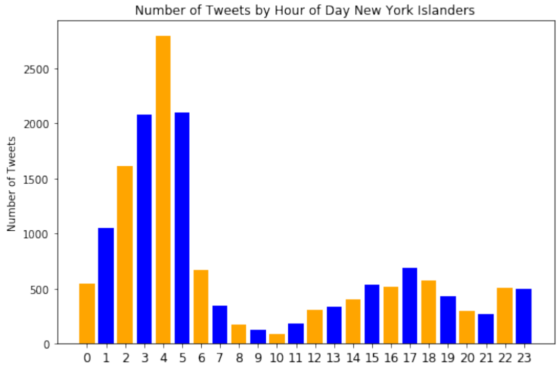
The following charts show the tweets by hour of the day for the 8 teams in the Metropolitan Division and the 7 teams in the Central Division. Most tweets are during games and after games end. The New York Rangers, although on the east coast, played on the west coast during this stretch and therefore most tweets happened after midnight.

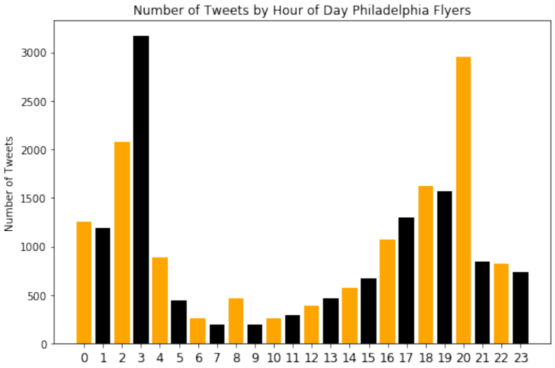

Number of Tweets by Hour of Day New York Rangers
Number of Tweets
0
500
1000
1500
2000
2500
3000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

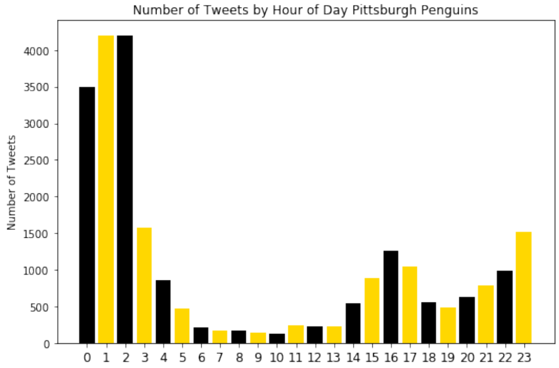

Number of Tweets by Hour of Day Carolina Hurricanes
Number of Tweets
0
200
400
600
800
1000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

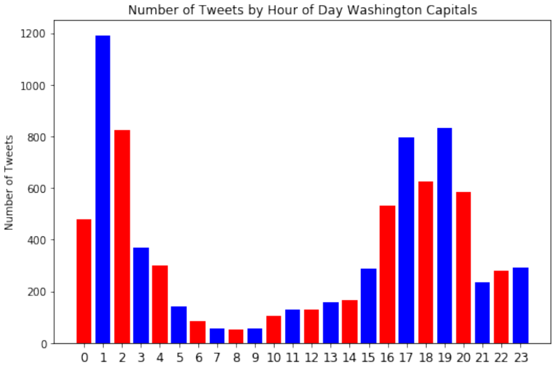

Number of Tweets by Hour of Day Columbus Blue Jackets
Number of Tweets
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Number of Tweets by Hour of Day New Jersey Devils
Number of Tweets
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

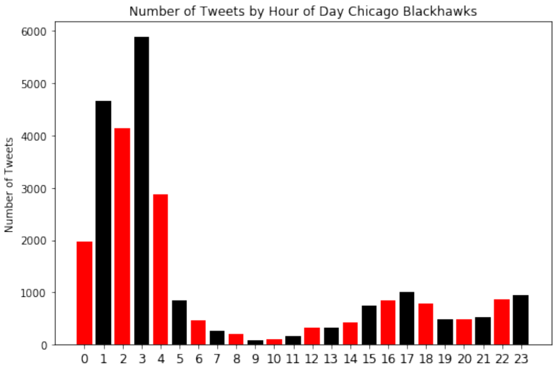

Number of Tweets by Hour of Day New York Islanders
Number of Tweets
0
500
1000
1500
2000
2500
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

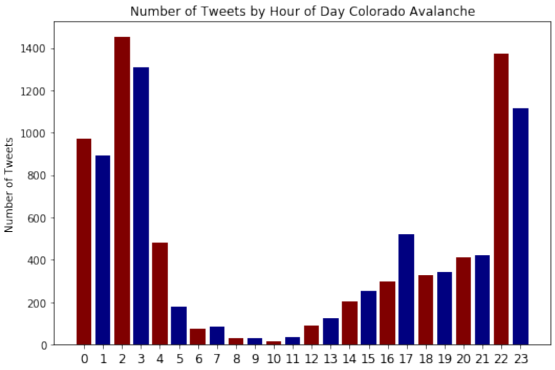

Number of Tweets by Hour of Day Philadelphia Flyers
Number of Tweets
0
500
1000
1500
2000
2500
3000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

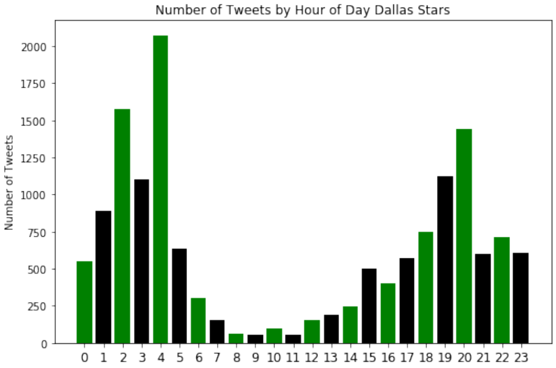

Number of Tweets by Hour of Day Pittsburgh Penguins
Number of Tweets
0
500
1000
1500
2000
2500
3000
3500
4000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

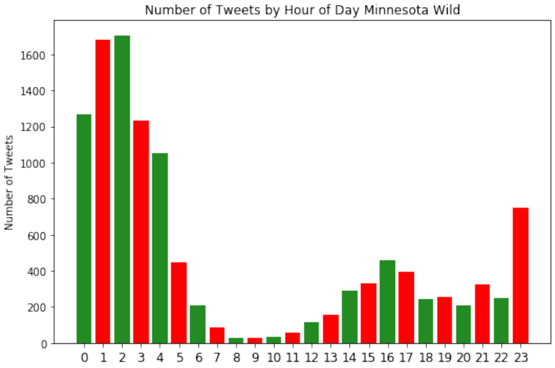

Number of Tweets by Hour of Day Washington Capitals
Number of Tweets
0
200
400
600
800
1000
1200
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

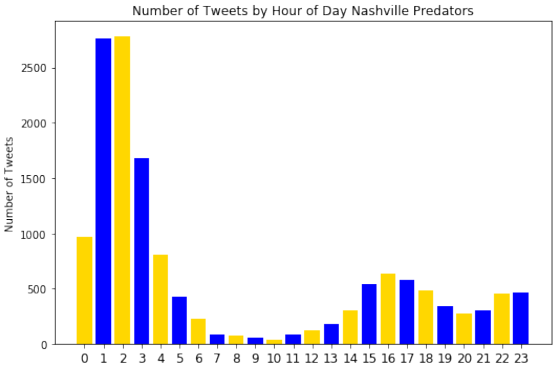
6.2 Central Division - Tweets by Hour of Day

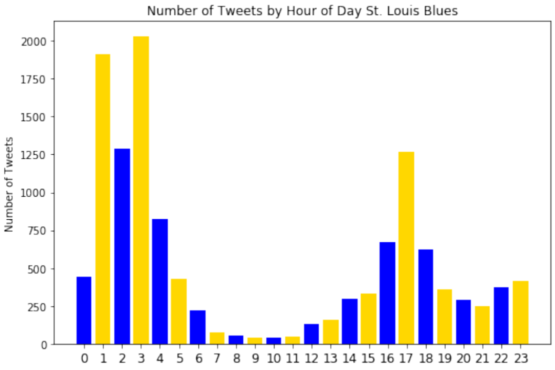

Number of Tweets by Hour of Day Chicago Blackhawks

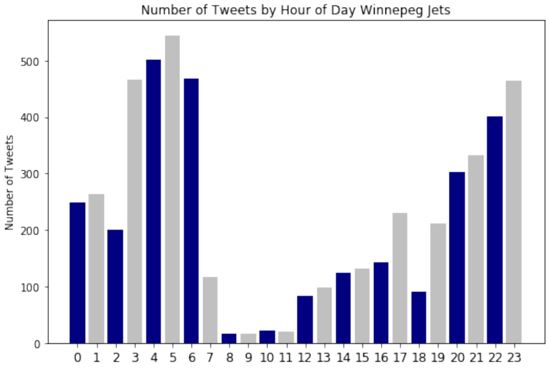

Number of Tweets by Hour of Day Colorado Avalanche


Number of Tweets by Hour of Day Dallas Stars


Number of Tweets by Hour of Day Minnesota Wild


Number of Tweets by Hour of Day Nashville Predators


Number of Tweets by Hour of Day St. Louis Blues

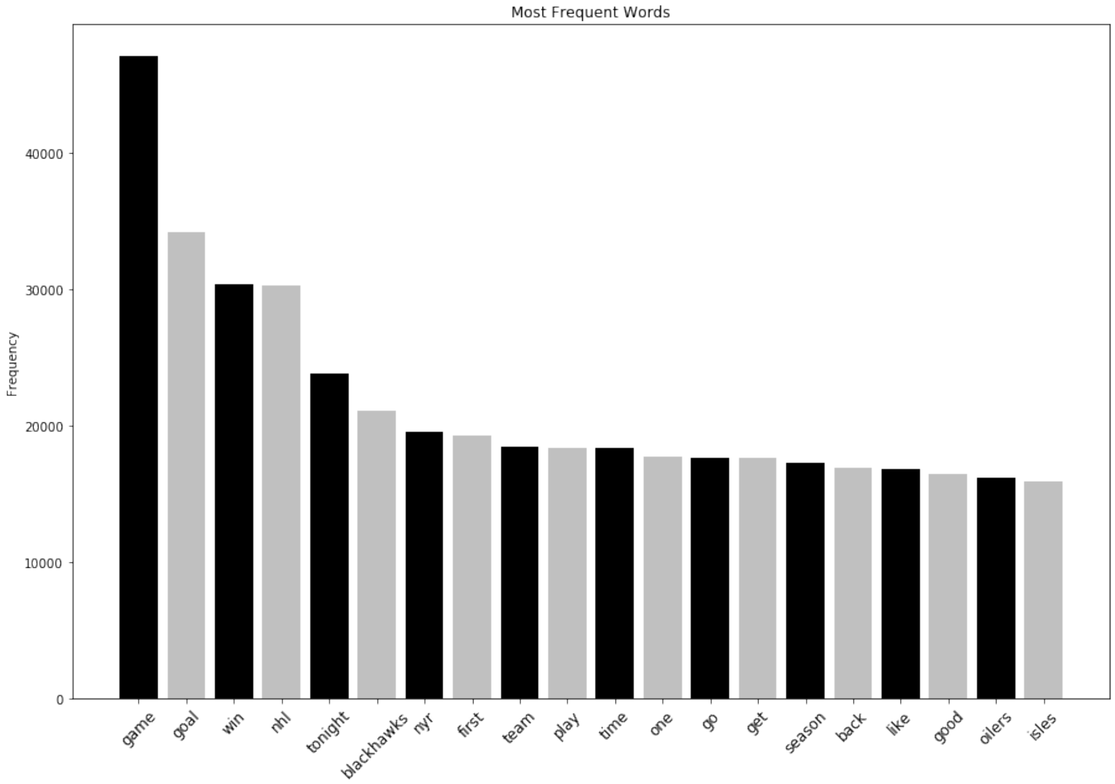

Number of Tweets by Hour of Day Winnepeg Jets

7 Frequency Counts for the NHL

In addition to the number of tweets, the most frequent words and bigrams were show for the NHL.

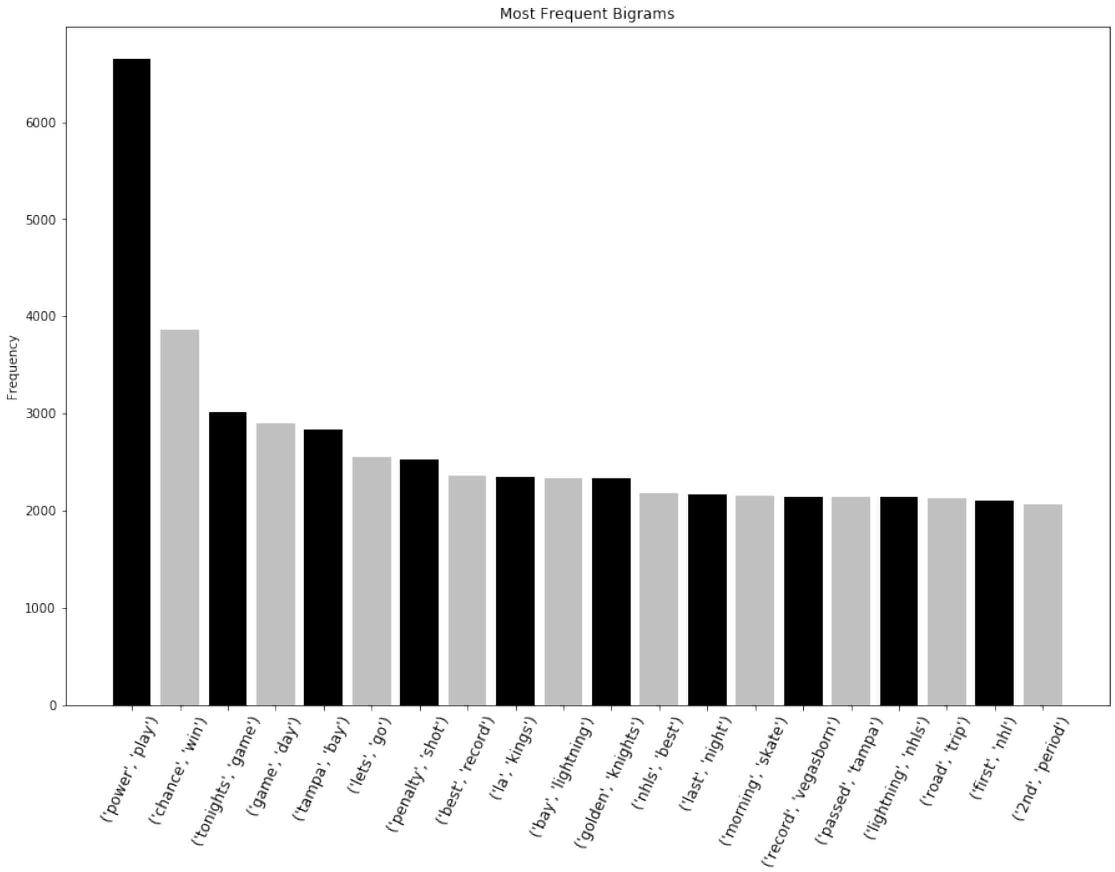
7.1 Single Words

The bar plot below shows the most popular words in the entire list of tweets for the NHL. The most popular words are game, goal, win, and nhl, all of which are not surprising.



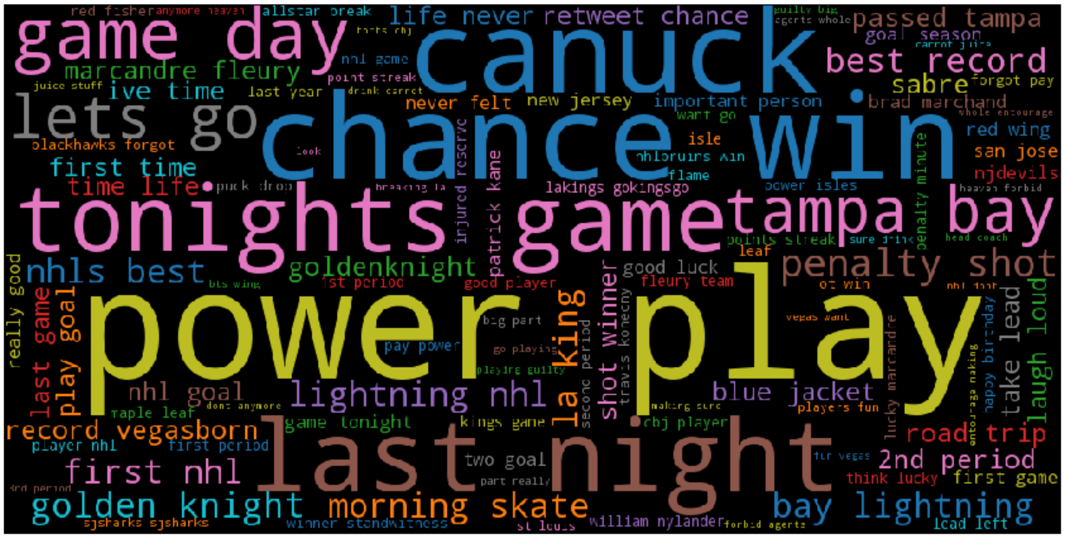
7.2 Bigrams

Bigrams are combinations of two consecutive words in a tweet. The bar plot shows the most popular bigrams in all of the tweets for the NHL. Power play is the most common bigram in the data, which is awarded when one team gets a penalty.



8 Word Clouds

The word cloud below shows the most prominent words for the entire set of tweets for the NHL. The size of the word means that the word occurs more frequently than others.

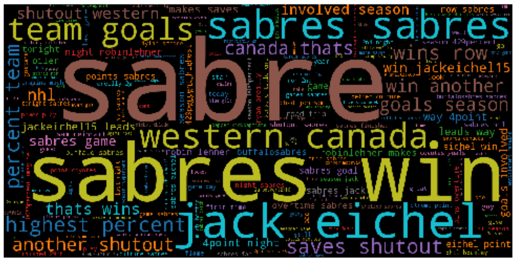


8.1 Atlantic Division

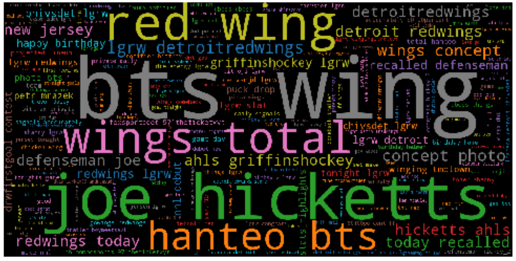
Bruins Word Cloud

nhlbruins win
points streak
brad marchand
power play
boston bruin
charlie mcavoy
two goals
streak gamesgoals pastrnak96
montreal two
win montreal
win devils
youre noldone
patriots youre
devils points
patrice bergeron
right patriots

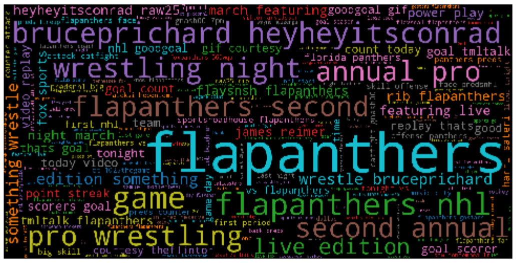
Sabres Word Cloud



Red Wings Word Cloud



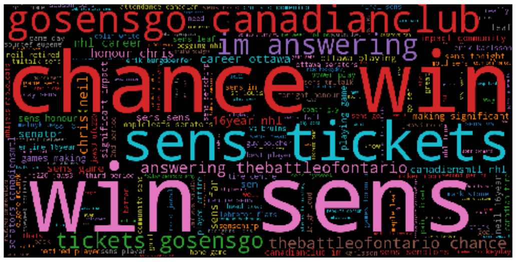
Panthers Word Cloud



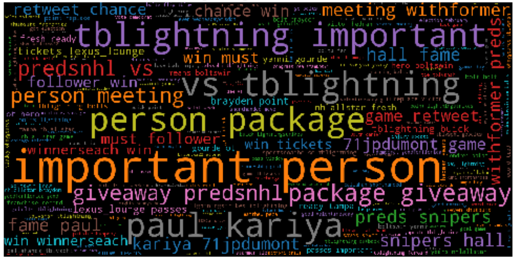
Canadiens Word Cloud

gohabsgo gohabsgo
bell centre
hab
red fisher
canadiensmtl mcdocanada
gohabsgo habs

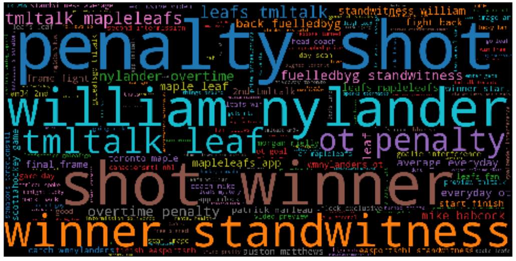
Senators Word Cloud



Lightning Word Cloud

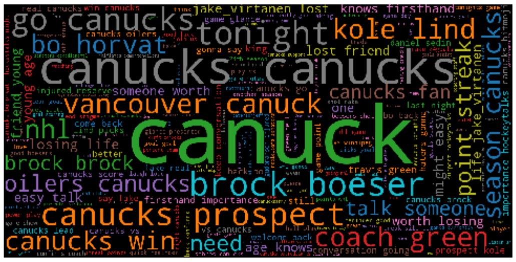


Maple Leafs Word Cloud



8.2 Pacific Division

Canucks Word Cloud



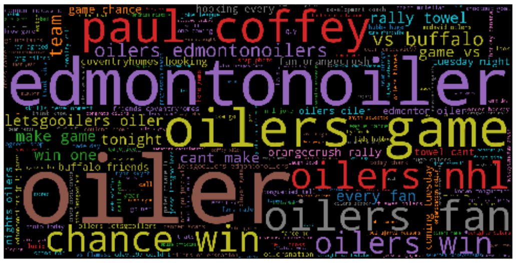
Sharks Word Cloud

joe thornton
hayden bradley
sjsharks sjsharks
power play
goal sjsharks
san jose
sanjosesharks love sharks

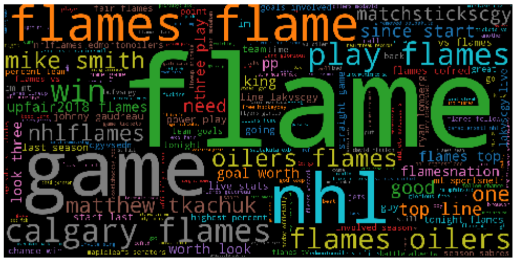
Kings Word Cloud

gokingsgo lakings breaking la
kings game
tanner pearson
la king
jonathan quick
lakings gokingsgo

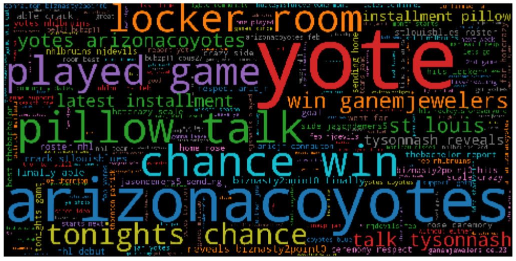
Oilers Word Cloud



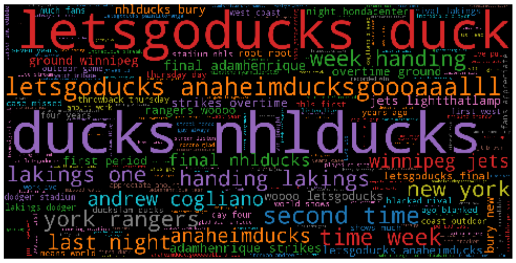
Flames Word Cloud



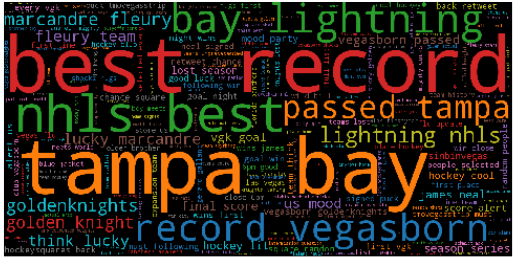
Coyotes Word Cloud



Ducks Word Cloud



Golden Knights Word Cloud



9 Next Steps

The next steps for the project are included in the following list: - Perform additional cleaning on the text - Look at removing mentions and hashtags - Correct more spelling errors - Replace more acronyms like team abbreviations (example: nyr is New York Rangers) - Use PySpark for Sentiment Analysis for the full NHL and each team - Use PySpark to tokenize the text - Use the ML package in PySpark to perform the sentiment analysis - Use the full NHL tweets to do sentiment analysis using a word2vec algorithm