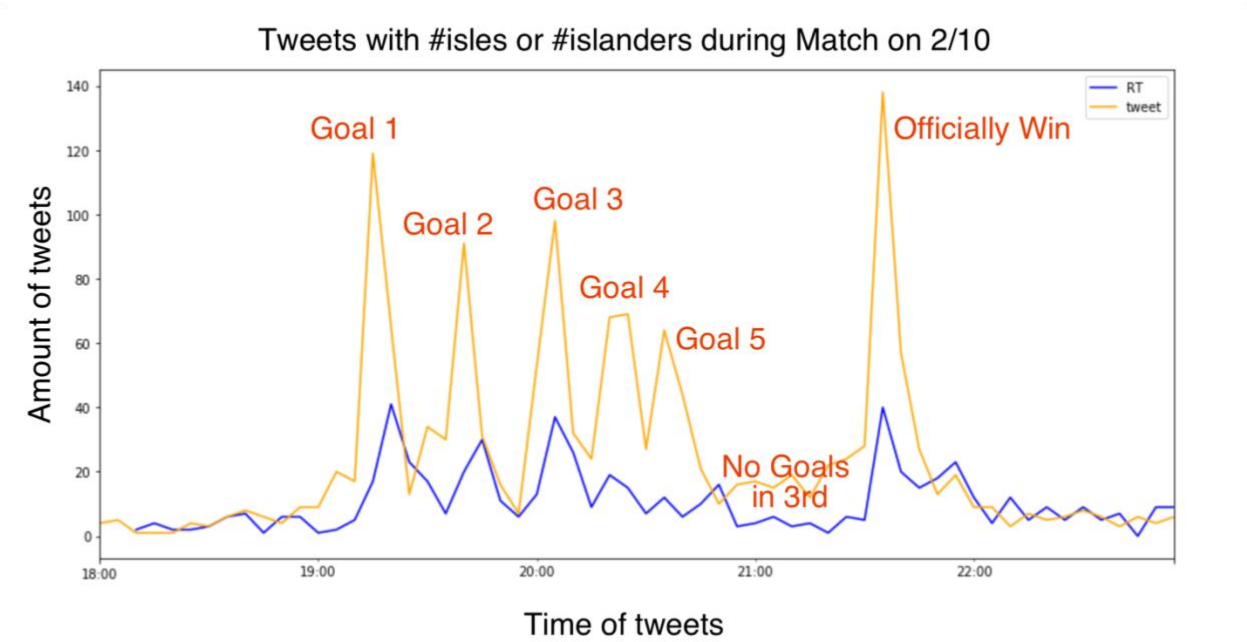
Mateo Correa
Islanders Sports Tech
Individual Technical Deliverable
Due Date: 2/20/20

1 Examining time relationship between tweets and match development



The above plot has some very interesting points as they relate to the match. The game wa s slated to start at 7:00PM and ended before 10:00PM. The Islanders scored a total of 5 goals that game; three were in the first period, two were in the second period, and none were in the third pe riod. The first five peaks in the plot each correspond to increased tweet activity as the fans celebr ated the goals. After these five peaks, there was a "dead" period in regards to twitter activity beca use no goals or otherwise notable events ocurred. At the end of the game, there was a final peak where fan activity increased greatly to celebrate the victory.

2 How plot was achieved

Notebook Set Up

This cell authenticates the notebook with your credentials.

```
[ ] #@title Authenticate
    from google.colab import auth
    auth.authenticate_user()
    print('Authenticated')
```

Authenticate

```
Authenticated
```

This cell imports some important libraries and links the notebook to the appropriate Google Cloud Project.

```
[ ] #@title Library Imports & Client Setup
    from google.cloud import bigquery
    from google.cloud import storage
    import logging
    from datetime import datetime
    import pandas as pd
    import json
    import numpy as np
    import math

    project = 'mit-islanders'
    bq_client = bigquery.Client(project=project)
    gcs_client = storage.Client(project=project)
```

Library Imports & Client Setup

Matchday 2/10

```
[ ] #@title Set up matplotlib package for data visualization

    import matplotlib.pyplot as plt
    %matplotlib inline
    import seaborn as sns
```

Set up matplotlib package for data visualization

```
[ ] #@title Run SQL query to get tweets during the match time
    sql = """
        SELECT created_at, id, text, EXTRACT(HOUR FROM created_at) as hour
        FROM tweets.tweets_week_5_12_Feb
        WHERE (text LIKE "%#islanders%" OR text LIKE "%#Islanders%" or text LIK
              AND (cast(date as STRING) LIKE "%02-10%")
              AND (EXTRACT(HOUR FROM created_at) BETWEEN 18 AND 22)
      """

    # Below we run the above generated SQL through the python BQ client
    df_1 = bq_client.query(sql).to_dataframe()
    df_1['hashtag'] = 'islanders or isles'
    df_1['is_retweet'] = df_1.apply(lambda x: True if x['text'].startswith('RT'

    df_1
```

Run SQL query to get tweets during the match time

```
[ ] #@title Plot the tweets and retweets during the game time, in five minute b

    fig = plt.figure(figsize=(15,7))
    ax = fig.add_subplot(111)
    df_rt = df_1.loc[df_1['is_retweet']==True].rename(columns={'id':'RT'})
    df_no_rt = df_1.loc[df_1['is_retweet']==False].rename(columns={'id':'tweet'
    df_rt[['created_at','RT']].set_index('created_at').resample('5T').count().r
    df_no_rt[['created_at','tweet']].set_index('created_at').resample('5T').cou
```

Plot the tweets and retweets during the game time, in five minute buckets

3 Other Insights

Number of tweets on Matchday 2/10 that used #isles or #islanders and number of distinct users:

date	hashtag	tweets_with_hashtag	distinct_users
2020-02-10	islanders	75	47
	isles	2760	1036

Top 100 tweeters:

	user_name	followers	acts_following	tweets	amt_languages
0	CordUpTime	3036	1796	63	5
1	stefen_rosner	181	1004	48	2
2	DaveBismo	3421	5070	47	4
3	AGrossNewsday	31828	613	42	3
4	ob1moroney	322	476	42	5
...	...	...	...	...	...
95	inthefade	704179	923	6	2
96	massjmcd67	134	216	6	1
97	GwenIsles	601	294	6	1
98	aaronfeigin	266	1525	6	1
99	RynoOnAir	7329	7838	6	2

Suspicious data?

There was definitely some noise in our data. Scraping Twitter for data is great, but some of the data that is picked up is done so unintentionally. For example, in our table, we found “island” related data that had nothing to do with the New York Islanders. There was data about “Rhode Islanders for Bernie,” and the hashtag “#islanders” was used regarding the television show Love Island. Going forward it could be useful if we could filter out some of this data, potentially by removing certain tweets if they also have other words in them.

4 Brainstorming “Fan Engagement” Project for the Semester

There are many interesting ways to approach our problem this semester

A Few Ideas:

- Analyze the engagement of fans by location
    - Is there a difference in activity between fans local to the city/state and those that live elsewhere?
    - Is there any specific locations where a lot of negative activity is ocurring? (like a rival's town)
- Analyze how fan engagement differs by type of game
    - Blowout vs very tight game --> do fans stop paying attention if the team starts losing badly? winning badlY?
    - Are rivalry games followed more closely? What about if the team is out of contention, who/to what degree do people still pay attention and interact?
- Analyze fan engagement as it correlates to roster moves
    - Is there outpouring of positive/negative moves when specific players join/leave the team
    - How do fans on twitter react to coaching changes? Or other organizational decisions
- Analyze what kind of fans there are?
    - Can we cluster fans into buckets like: super fan, regular fan, light fan?
    - Do different clusters of fans engage with the team differently? (e.g. are super fans more critical?)
- Can we analyze fan favorites based on tweets?
    - Are there certain players whose actions trigger significantly more (or at least more proportionally to their fame) tweets?
    - Like if a certain player scores a goal does he cause the biggest increase in activity?
    - If certain players do something bad, do they get hated on more on twitter? (opposite of fan favorite)