CS410 Team Crickets Project progress report

Project Title: 2021 Cricket World Cup Tweet Analysis

Below are the high-level tasks planned initially, we broke down the sub-tasks further as below:

1. Pull data from Twitter API
   1. Identify keywords to extract useful tweets
   2. Compile a dataset with refined tweets using dates identified from Task 2 schedule
2. Scrape data from the world cup schedule
   1. Parse html content and create a csv file with match schedule information
   2. Identify appropriate parsing keyword to identify which teams will win
3. Cleanse the data into a normalized format
4. Run sentiment analysis against tweets
   1. Assign sentiment scores using text blob
   2. Validate sentiment scores for insights
   3. Data visualization insights for data analysis
5. Use historical data for evaluation of the results
   1. Compare results to real outcomes
6. Create video demonstration with appropriate documentation

Progress Report:

1. Which tasks have been completed? Above tasks 1a, 2a, 3, 4a
2. Which tasks are pending? Above tasks 1b, 2b, 4b, 4c, 5a, 6
3. Are you facing any challenges? Yes
   1. Tweepy API retrieves only 7 days of past tweets, and this limitation will hamper our tasks (4b, 4c, 5a) to run sentiment analysis for all matches. We are mitigating this issue by exploring options to use Kaggle dataset for world cup tweets
   2. We are still exploring the parsing keyword from the tweets to identify which team will win & associate the sentiment score accordingly

CS 410 Team Crickets Project Proposal

Project Title: 2021 Cricket World Cup Tweet Analysis

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Adam Ruther (aruther2@illinois.edu) - Captain,

Prasanna Kumar ([pkumar15@illinois.edu)](mailto:pkumar15@illinois.edu)

Surabhi Choudhary ([surabhi5@illinois.edu)](mailto:surabhi5@illinois.edu)

1. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

Free Topic: Social Media Analysis, Sentiment Analysis

Detailed Description: social media has evolved as a platform where fans express their feelings towards their favorite sports teams. Our goal is to analyze tweets from Twitter users discussing the ongoing T20 world cup in Dubai. In order to analyze the tweets, we will pull tweets from using a predetermined interval (4,6,8 hours...) before the match until the start of the match and use those tweets to determine if there is a correlation between who won the match and the sentiment of the tweets that were sent before the game. We will get the start of the match from scrapping the game schedule from a source online.

Tasks/Planned approach:

* + Pull data from Twitter API
  + Scrape data from the world cup schedule
  + Cleanse the data into a normalized format
  + Run sentiment analysis against tweets
  + Use historical data for evaluation of the results

Why is it important or interesting?

Daily, we see a lot of social media posts before a certain match. These tweets include game discussion, predictions, and positive encouragement for their team. We think it would be very interesting to see if there is any correlation between the sentiment of these tweets and who will win the match. We believe this topic will allow us to use much of what we have learned so far by gathering a large amount of data from multiple sources and then mining that data for insights.

What tools, systems or datasets are involved?

* Python (Pandas, Numpy, SciPy, Selenium, MeTaPy, visualization packages)
* [TweePy](https://docs.tweepy.org/en/latest/)
* Scrape Game Data from Cricket Website (<https://www.t20worldcup.com/fixtures/men)>

What is the expected outcome?

To build a script that enables to make analytical insights on the correlation of tweets to the match results.

1. Which programming language do you plan to use?

Python

1. Please justify that the workload of your topic is at least 20\*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

We expect this project to take anywhere from 55-65 hours. The breakdown is approximately as follows:

* + Pull data from Twitter API (10%)
  + Scrape data from the world cup schedule (10%)
  + Cleanse the data into a normalized format (30%)
  + Run sentiment analysis against tweets (25%)
  + Use historical data for evaluation of the results (25%)