Big Data Project Update

Analyzing Real-Time Tweets to Generate Game Report

By
Clouded Minds

Project Goal

Game Report generation based on tweets posted by users watching game live.

- Important moments detection
  - will be detected by analysis of spikes in the volume of tweets.
  - sentence analysis is performed on those moments to describe the moment (What exactly happened at that moment).
  - Sentence analysis: Clustering on tweets of important moments.
- Sentiment Analysis
  - Collected tweets will be divided into two groups for each team based on hashtags.
  - Sentiment value for each team is calculated.
  - Correlation of sentiment value to actual match result (by what margin one team lost or won).

Our goal has not changed

Thus far...

- Data Collection
  - English Premier League Soccer games.
  - Collected tweets of about 10 games.
  - Filtered using official game hashtags, team names and their short names.
  - Working to separate tweets into tweets for each team (sentiment analysis)
  - Collect more data sets
    - Some previous data corrupted
    - Language parameter
- Data Preprocessing
  - Stored in mongodb.
  - Conversion of twitter json format into csv format.
  - tweet timestamp translation into game minutes.
  - Building data format that sentiment analyzer will accept

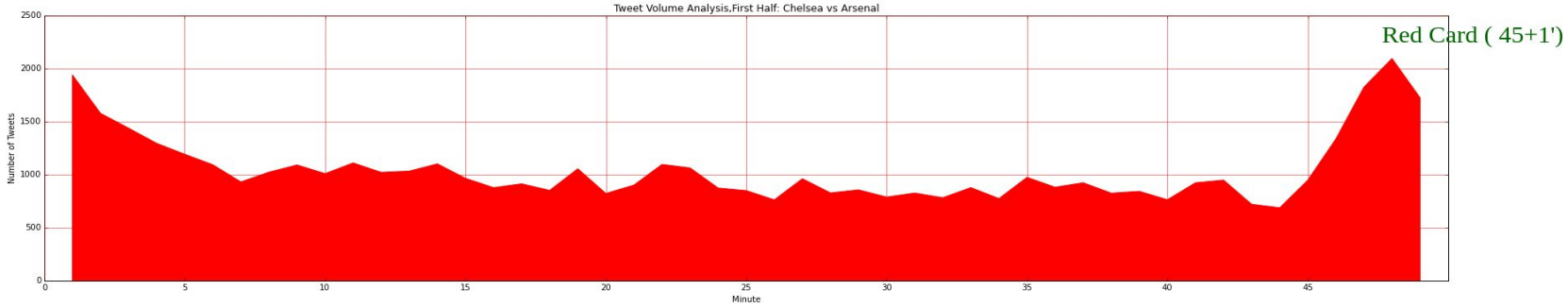
Thus far...

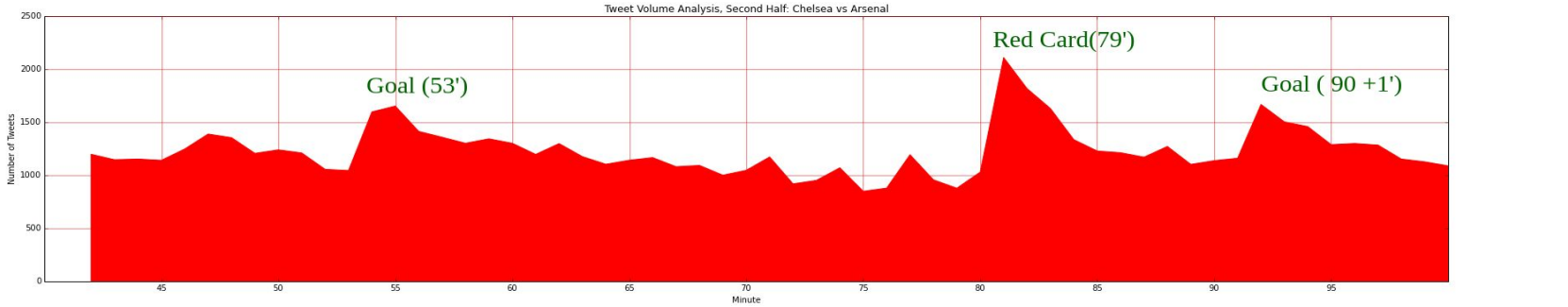
Analysis

- Based on tweet volume
- Tweet Count:
  - Gives high accuracy in detection of top moments of the game.
  - Goals, Red Cards and missed chances are detected with high accuracy.
- Results
  - Chelsea vs Southampton
  - Chelsea vs Arsenal

Thus far...

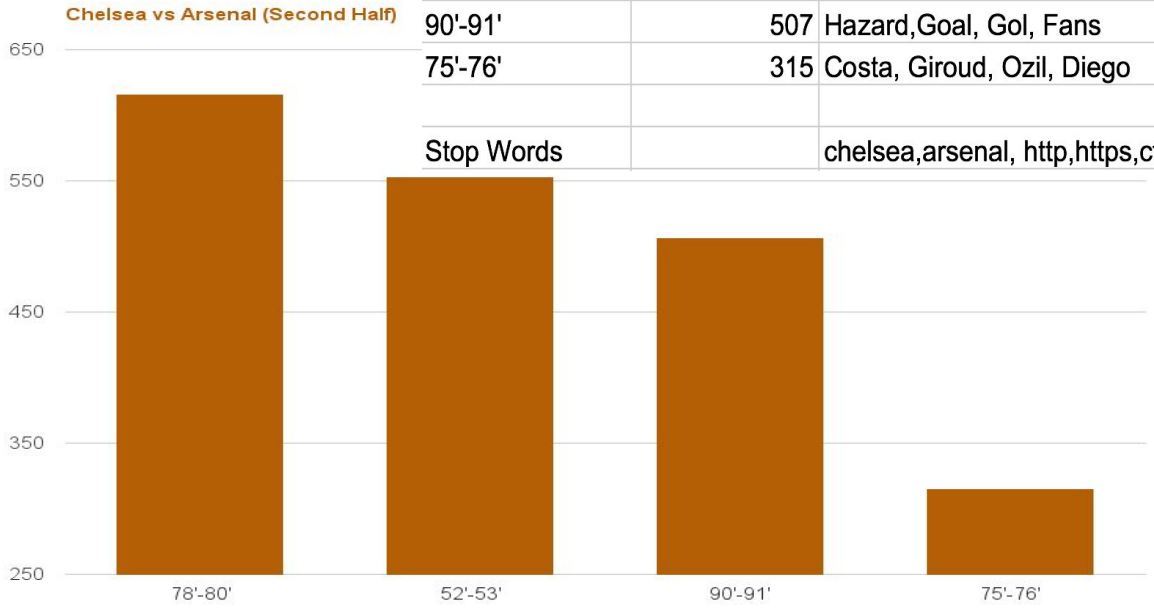
Chelsea vs Arsenal





Thus far...

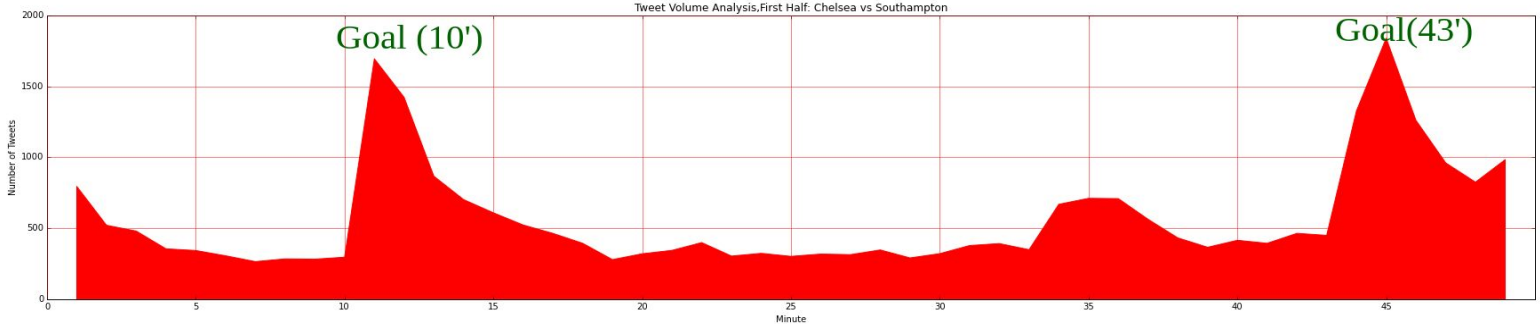
Spike	Slope Value	Top Words	Real Incident
78'-80'	616	Men, Red, Cazorla, lol	Red Card to Cazorla
52'-53'	553	Zouma, Gol,Goal, Costa	Goal by Zouma
90'-91'	507	Hazard,Goal, Gol, Fans	Goal by Hazard
75'-76'	315	Costa, Giroud, Ozil, Diego	Substitution: Ozil off, Giroud On
Stop Words		chelsea,arsenal, http,https,cfc,afc,chears	

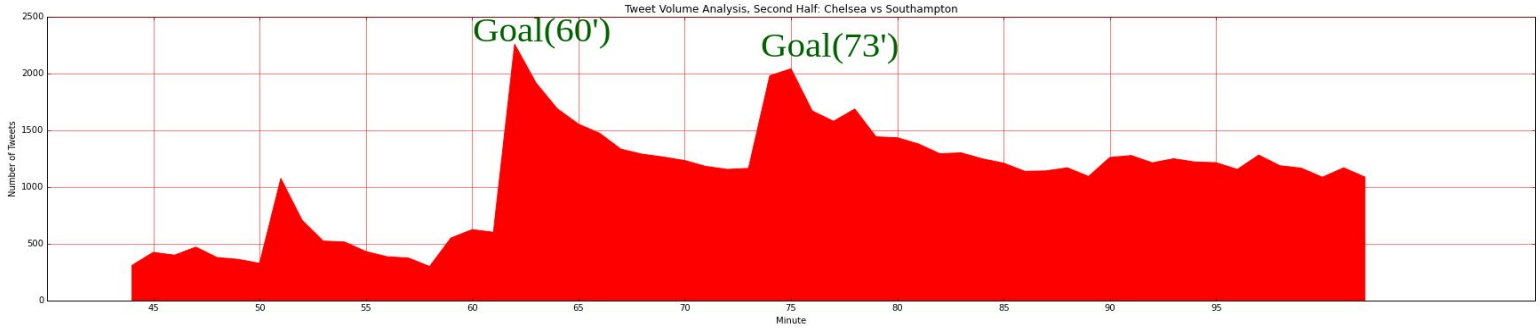


Slope (Tweet/Interval) for each spikes

Thus far...

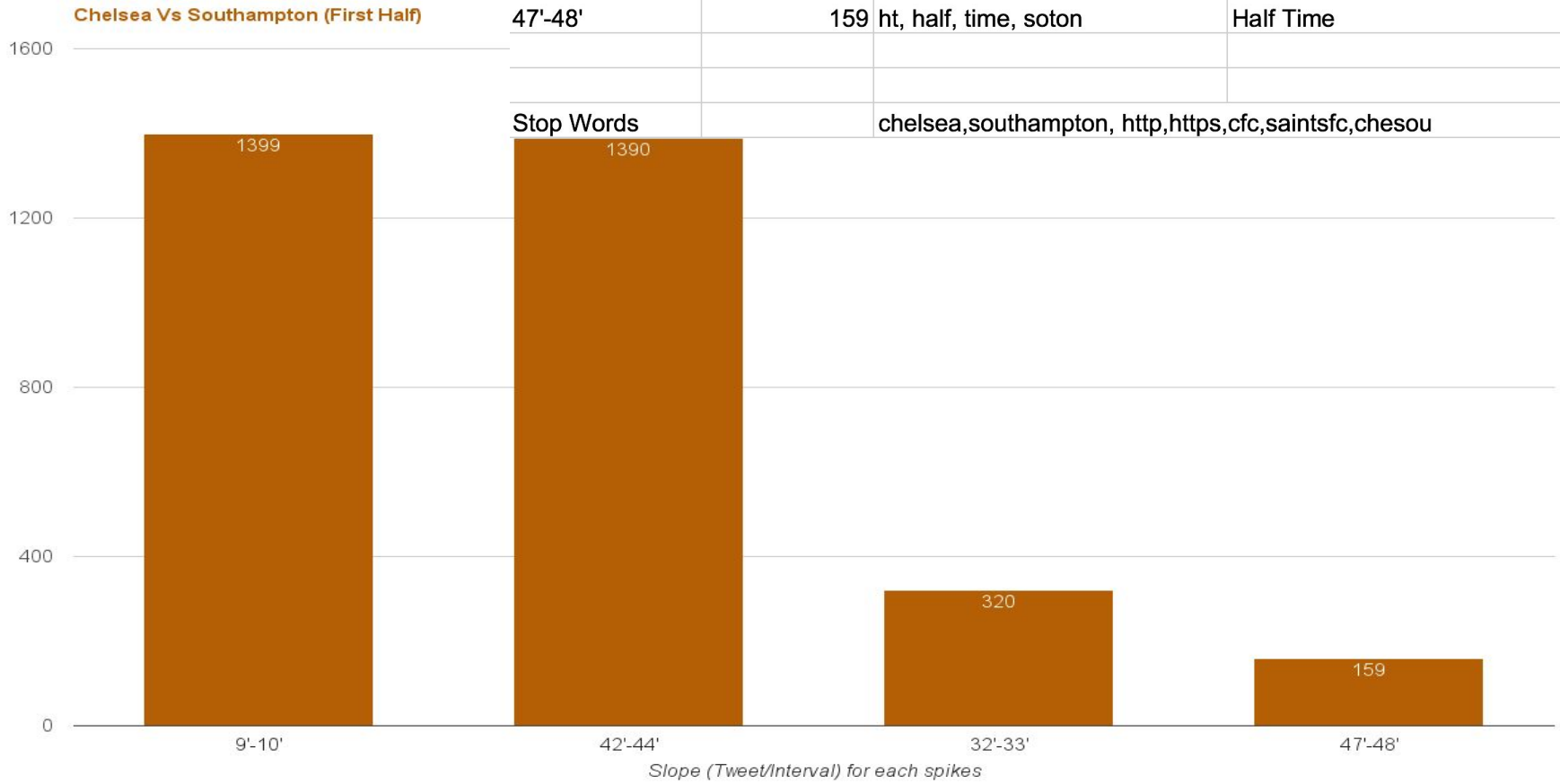
Chelsea vs Southampton





Thus far...

Spike	Slope Value	Top Words	Real Incident
9'-10'	1399	Willian, Goal, Free, Kick	Willian scores Goal on Free Kick
42'-44'	1390	Goal, Devis,Gol, Deserved	Devis Scores Goal deservedly
32'-33'	320	Penalty, Penalties,Ref,ungguli	Penalty Appeal, but denied
47'-48'	159	ht, half, time, soton	Half Time
Stop Words		chelsea,southampton, http,https,cfc,saintsfc,chesou	



Thus far...

Analysis

- Sentiment Analysis
  - Attempted to use online sentiment analysis API.
  - Due to the limit on number of requests, still searching for alternatives.
  - Using sentiment140
    - only supports English and Spanish
    - Language parameter

To do...

Detailed volume analysis

- Tweet Count Completed
  - May not sufficient to detect less important moments such as yellow cards, substitutions.
  - But, evaluation of slope of each spikes in the volume of tweets can give better result for less important moments.
  - Spikes can be detected using sliding window on tweets over the time.
- Sentence Analysis by clustering on tweets of detected moments
- Sentiment Analysis

To do...

Noise Reduction Still researching improvement

- Removing off-topic tweets in the detected moments is huge challenge.
- Similarity detection for words like ‘Goal’ or ‘Goooooaaal’ for sentence analysis.
- Removing spellings errors before sentiment analysis.

To do...

- Date to stop collecting data
- How many games to be in final analysis (5-6)
- May use sentiment analysis for tweets within detected moments
    - Detect which team scored
    - Predict overall result of game
    - Predict range of victory based on fan reactions