

Code the Game – Stage 1 Report

Team ID: CO1051

Aspects taken into account for selecting top events –

We have followed the data driven approach for this natural language processing task. We have used quantitative methods to discover relations. After converting the data to structured form, we have gone through methods like word frequency counting, ranking by means of the Term Frequency – Inverse Document Frequency metric, n-grams, and clustering. We effectively selected the events corresponding to balls which makes the game biased towards change like balls with sixes, fours, wickets, no-balls etc. After that we evaluated the sentiment score of commentary corresponding to each ball and selected the balls which gave us the highest score. We have chosen this strategy so that we may be able to analyze any special events that have occurred with a particular delivery.

We used IBM Watson Natural Language Understanding API for different tasks like categorization of content into five-level classification hierarchy. (However there is no need of such categorization but to finally provide the output to user (who will use the gui) with full perfection.) , identifying high level concepts that are not directly referenced in the text, identified the entities mentioned in the content, this includes people, places, events and other types of entities and also done the emotional analysis of each entity. We searched the relevant keywords from the document so that particular analysis can be done on those keywords. Identification of semantic roles is done by parsing the text into subject-object and action form, and identified entities and keywords that are subjects and objects for an action.

We are looking forward to work on –

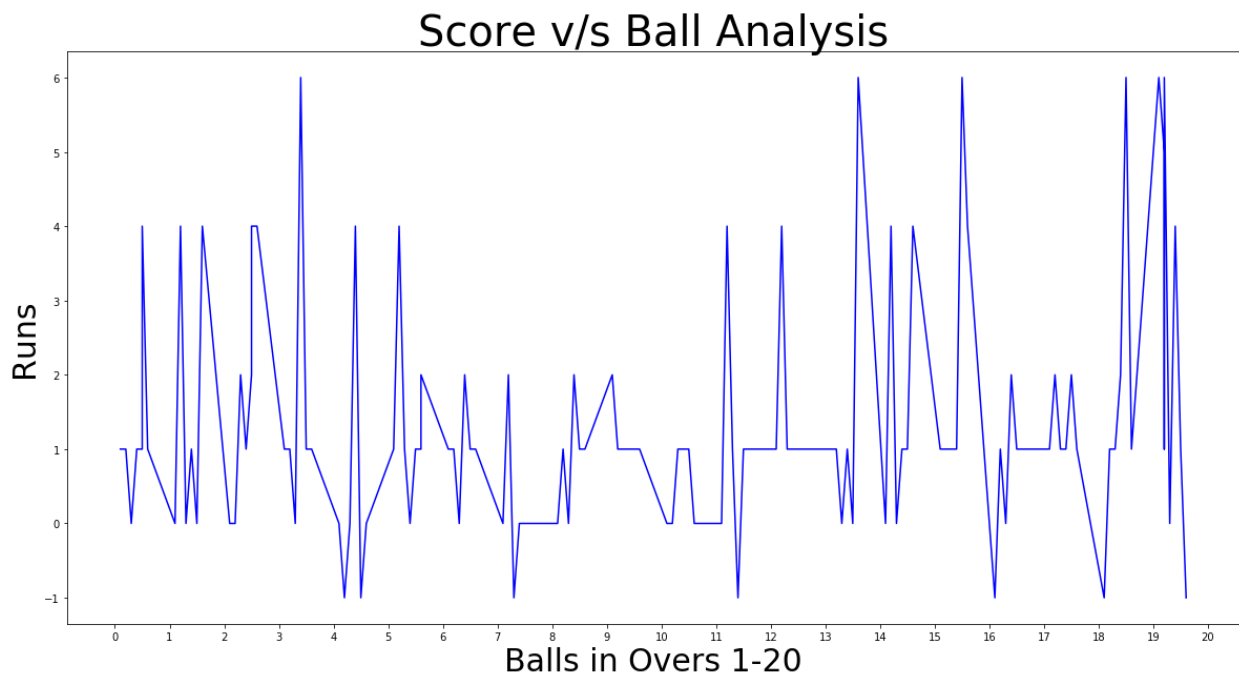
We have selected the top 50 events till now. Now we have the challenging task to make the best team. For this we will go through the corresponding commentary of top events and we will apply named entity recognition on that to extract the name of players from those events. After that we will focus on the keywords around that entity. (We will use regex parsing for that and will prepare keywords on our own in a separate file which will include “bat”, “bowl”, “sixes”, “fours”, “field”, “keeping”, “stumping”, etc. because ultimately we need the finalized team which have best players from every segment. After that we will evaluate the parameters like average strike rate, bowling economy, catches taken, stumping made, run out, and for bowlers we will consider the parameters like no-ball, Yorker, short pitch, good-length, long pitch. This will give use features corresponding to each player. After that we will calculate the score by allocating weights to each parameter and players with best score will be selected. We will use Watson api for drawing insights from the data and applying Natural Language processing concepts.

For doing this task we will create the gui using flask in python which have the option to select multiple pdf's file, after that we will convert those pdf into text. Then different parsing methods will be applied that will produce three different csv files and covert unstructured data into structured form. One will contain the commentary data, another will contain the runs/wicket on each ball along with commentary corresponding to that ball, and one will contain the end of over

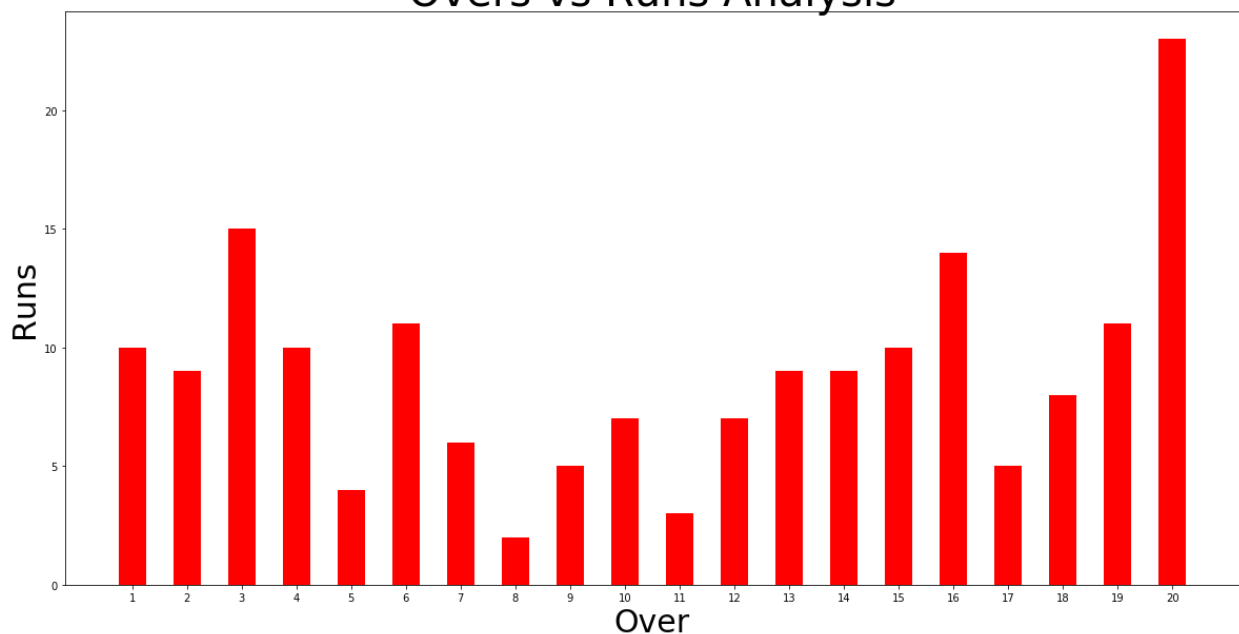
summary. After that we will integrate the second csv file with IBM Watson natural language tool kit api, Watson discovery api to again select the top 50 events and follow the strategy as discussed above, the second and third csv file will be used to plot the visualization and directly analyzer can get aware with the insights of the data.

Alternate Approaches –

This is the alternative approach that we can use for the major event extraction. Unlike data driven approach which we have used to extract the 50 major events from the data, it is often on patterns that express rules representing expert knowledge. It is inherently based on linguistic and lexicographic knowledge, as well as existing human knowledge regarding the contents of the text that is to be processed. In this approach we could have divided the commentary by means of three keywords (score, ball, batsman) that together describe the major events. For example let's take the instance if a batsman score 6 runs in the last ball of the last over it is considered as a major event rather than the ball of any middle over. Also the batsman that is already playing good throughout the series, if that batsman gets out then it will be the major event. So to find these relations we need the expert knowledge and create the rules manually. This approach is known as knowledge driven approach which we could have use for the event extraction.



Overs vs Runs Analysis



Word Cloud

