Properties extraction of Dismissal Event in a Cricket Match using its Commentary

Vanshika Sharma, Samyak Bhagat, and Tanisha Upadhyay

This is project report for the subject 'CS1302: Cognitive Dashboards' at JK Lakshmipat University, Jaipur

ABSTRACT

The project aims to extract the details of Dismissals Event in the Cricket Match. Cricket is one of the games with a lot of scope in analysis, prediction and decision-making. Our project will serve one of the motives by analysing the dismissals. This will be helpful in strategising the game. The data collection is done through web scraping and data analytics through various python tools.

Introduction

Cricket is one of the most loved sports across the globe. The reach of the game attracts every individual. The individuals involved can be classified as, Players (called Cricketers), Viewers, Bidders, and other professionals involved. The project involves text analytics done on the commentary of the cricket match.

The commentary used in the text analytics is web scraped through UI path studio, followed by data processing. In the future, there can be models where in commentary will be collected as a speech-to-text model and simultaneously analysed. This helps in analysing players. The efficient analysis will yield better decision making such as selection of players, strategy, etc., contributing to betterment of the cricket team.

The objective of the project is to analyse the Dismissal Event of Cricket Matches. The report describes the methodology, results and future scope of the project.

Background

Cricket is amongst the most popular field games in world. The project uses Text Analytics to refine the commentary of the cricket matches to extract details of dismissal events. The commentary was taken from the cricbuzz.

Data Properties

The data was scraped as a structured data of the cricket match. It included details like playing teams, batsman (striker and non-striker) and bowler, number of runs per ball, extra runs, innings details, team score and the commentary. The whole data can be classified as alphanumeric. As some fields like balls, runs, etc were numeric whereas fields like name of team, players, commentary were in of string data type. The cricket matches were identified by their match id, as the name of the data files suggest.

Game of Cricket

Cricket is one of the many games that include a bat and a ball. The game is played on a field at the centre of which is a 20 metre pitch with a wicket at each end. The toss to decide which team bats first is determined by a coin. The game is refereed upon by two umpires, third umpire and a match referee. The game is extensively played across three different formats i.e.T20,ODI(One Day International) and Test matches. T20 is the shortest format with each side playing 20 overs. ODI's include 50 overs per team while Test Cricket is played for unlimited overs in the span of 5 days.

Text Analytics

Text Analytics is the process of drawing meaning out of written communication. It is also called Text mining. It identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed, or presented directly using clustered HTML tables, mind maps, charts, etc. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

NLTK

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning.

Methodolgy

Step 1: Data Collection

We have collected data from www.cricbuzz.com through web scrapping, the scrapping was performed using UI path studio where data was unstructured, then we have cleaned and procced the data to convert it into tabular form so that we can do calculations on it. UI path studio

Step 2: Processing the data we have collected in the data frame

We got the csv file now we have imported the file in python using pandas and converted it into a data frame.

Step 3: String operation

We are iteration over each commentary for checking how many players are out and using hard code we have found when a player is out what type of method of dismissal was there.

Results and Discussion

We were able to make a tabular data containing all the information.

Batsman	**Bowler**	**Runs**	**Type of Dismissal**	**Extra Man**	**Overs**
Azhar Ali	Starc	5	Lbw	-	2.5
Haris Sohail	Starc	8	Caught	Paine	4.4
Shafiq	Pat Cummins	0	Caught	Smith	6.4
Masood	Pat Cummins	42	Caught	Paine	28.3
Iftikhar Ahmed	Hazlewood	0	Caught	Paine	29.3

Table 1. Dismissal Details

The cricket match we performed the analysis on was between Pakistan and Australia.
We were able to extract the details of the dismissal events

Future Scope

The following is the future scope of the project:

1. Proper Web Scraping: The scraping we performed could not include all the details given on the web page. Using a proper model like BeautifulSoup (Python Library) can provide us with better efficiency.

2. Machine Learning Models: We can apply different ML models like clustering, (NLP) natural language processing, etc.

3. Speech-to-Text model: We can make a speech-to-text model through which the commentary will be automatically collected as the commentator speaks.

Appendix

Metadata

Number of fields (columns)	2
Field 1	Commentary
Data Type	Categorical (String)
Field 2	Number of Over
Data type	Numerical

Table 2. Meta Data