Abstract

Text analysis is the automated process that uses AI to obtain information from the text. This project intends to determine the name of players in a cricket match and complete the formation of the team of 11 players using text analysis. The programming language used in this project is Python as it is a high level language. This project utilises python libraries such as pandas, numpy, cricpy. The project reflects that from any Data-Frame containing a particular commentary, text can be extracted in such a way that players and the performance of the players can be found.

TEXT ANALYTIC ON CRICKET COMMENTARY

SUBMITTED BY

Yash Goyal(2018BTechCSE121)
Harsh Udai(2018BTechCSE106)
Aishwaryaditya Jha(2018BTechCSE123)

as Project
of

Institute of Engineering and Technology
JK Lakshmipat University, Jaipur
November 2019

INTRODUCTION

Text Analysis of any sports commentary is a popular topic. In today's time and compressed world, people prefer to view cricket updates instead of viewing an entire match. Cricket is the second most sought and played game in the world. Our interest in cricket was another motivating factor to take this as a project.

We faced many challenges in this project. For example, matching of full names from a certain web database was very challenging. We were challenged a lot while analysing text from a csv file. Extraction of names of players from the .csv file was really challenging.

1 Background

In this project, we have used Text Analytic using Python to extract useful information from the **cricket commentary data** and display it in the tabular form.

1.1 Data-Set Profile

This data-set represents the data of commentary of a cricket match. The data-set contains 18 columns containing various data types such as int, float, string. There are total 1840 entries which is helpful for finding the required details.

1.2 Cricket - A Throbbing Sensation

Cricket is a bat and ball game played between two teams, 11 players each, on a field which has a rectangular 22-yard-long pitch in the center. The game is played by 120 million players world-wide making it the second most popular sport in the world. The purpose of the game is to score more runs than your opposing team.
A Cricket match is divided into periods called innings. It is decided before the game begins, if both teams will have one or two innings. During the innings one team bats while the other fields. All 11 players on the fielding team are on the pitch at the same time however only two batsmen are the field at any one time. Team captains toss a coin to decide who should bat first.
Cricket fields are oval. The end which is marked off is called the boundary, with the rectangle "pitch" in the center.
At each end of the pitch are the wickets, 22 yards apart. A bowling crease is

in line with the wicket and the batting or popping crease is 4ft in front of the wicket

1.3 Text Analytics

Text Analytics is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.
Text Analytics 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 Analytics employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

1.3.1 NLP

Natural Language Processing is manipulation or understanding text or speech by any software or machine. An analogy is that humans interact, understand each other views, and respond with the appropriate answer. In NLP, this interaction, understanding, the response is made by a computer instead of a human.

1.3.2 NLTK

NLTK stands for **Natural Language Toolkit**. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. It contains packages like Tokenization, Stemming, Lemmatization, Punctuation, Character count and word count.
Five main Component of Natural Language processing are:

- Morphological and Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Discourse Integration
- Pragmatic Analysis

1.4 Libraries Used

1.4.1 NumPy

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code

- Useful linear algebra, Fourier transform, and random number capabilities

Array in Numpy is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In Numpy, number of dimensions of the array is called rank of the array.A tuple of integers giving the size of the array along each dimension is known as shape of the array. An array class in Numpy is called as ndarray. Elements in Numpy arrays are accessed by using square brackets and can be initialized by using nested Python Lists.

1.4.2 Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal. Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

The well known Pythagorean theorem $x^2 + y^2 = z^2$ was proved to be invalid for other exponents. Meaning the next equation has no integer solutions:

$$x^n + y^n = z^n$$

Meta-Data :

Title	Cricket match between India and Australia
Subtitle	Cricket Match
Description	The data-set is the about the match between India and Australia, data set also contains the commentary, innings, bowler, Striker, runs, non-striker
Data Columns	Total 18 Columns
Range Index	1840 entries
Data Types	int type, float type, object type