Term Project Summary

Sucharitha Puppala

Data Science, Bellevue University.

DSC540-T301 Data Preparation.

Professor Catherine Williams.

August 10, 2022

Summary

Abstract

This paper provides the summary of the Term Project data preparation. The data collected is about the Cricket data. The data preparation is done on three types of data i.e. Flat File data. Web data and from API. We will know the steps that are involved in the merging of the three forms of data into a data base and doing some visualizations.

Summary

Data Preparation

Data preparation is the process of cleaning and transforming the raw data collected prior to processing the data and analyzing. It is an important step prior to processing the data and mostly involves in reformatting the data, making corrections to the data and then combining of data sets to enrich the data.

Data Preparation involves the following steps:

1. Access the data.
2. Fetching the data.
3. Cleaning the data.
4. Formatting the data.
5. Combine the data.
6. Analyzing the data.

Cricket is a bat-and-ball game played between two teams of eleven players each on a field at the center of which is a 22-yard (20-metre) pitch with a wicket at each end, each comprising two bails balanced on three stumps. Cricket stands in the second place of the top most popular sports in the world. There are Twenty20, ODI, Test formats in the game Cricket.

During this course I have learned many new things that are related to the subject of Data Sciences. I have learned to collect data from three different sources like flat file, website and API , which I felt was interesting, as this is the first time for me to do analysis by collecting the data from three different sources and working for a project. I have learned how to identify the data from different sources and structuring the data, cleaning the data, transforming the data, visualizations, web scraping, to get API data and work on Sqlite.

Summary

I felt working on the website data and the API data, interesting and learned how to insert all the three data forms into the data base and working on it

For this project I have collected the data from different sources on the topic of “Bowlers with 300 or more Test wickets” in Cricket.

I have collected the data in a flat file form, API data form and Web data form.

The data collected from the flat file is from espncricinfo.com, and is saved in Excel file. The data is saved in a CSV file and is read for the data analysis. The CSV file is then read and cleaned by identifying the missing values if any, finding the duplicates if any and removed them, checked for any null values and removed if any present, identifying the outliers and bad data, fixed the inconsistent data present in the data set. The data is finally kept in a readable format for inserting into the database.

The data collected from the Website is read from the URL from which the data is to be read. Then the data is read and cleaned by identifying the missing values if any, finding the supplicates if any and removed them, checked for any null values and removed if any present, identifying the outliers and bad data, fixed the inconsistent data present in the data set and the final cleaned data is converted into a CSV file for inserting into the database.

The data collected from the API is read using the key provided for reading the file and data cleaning and identifying the outliers and fixing the inconsistent values is done and the final data is converted into a CSV file for the inserting into the database.

In the final milestone all the cleaned data files are inserted into the data base by creating three individual tables. Different visualization are plotted from all the three data tables individually and by extracting the data from two tables are also plotted. All the three data tables are ready for printing into a human readable format. All the three datasets are joined into single data set into a human readable

format and printed. With the help of the weekly assignments I was able to insert all the three data files into the data base and do the needful visualizations.

All the data that is used in this project is being collected from the trusted sites that have been updated using the actual records of the matches that are being played by the bowlers. During the data preparation for the project the data manipulation is done keeping in view of not changing the actual data.

References

Google, Cricket,

https://en.wikipedia.org/wiki/Cricket