Tennis Player

Document

For the first identification of traits, we consulted a few Wikipedia pages. Then, using the Python module Beautiful Soup, we began web scraping a webpage. We chose not to use the Selenium library because the website is not dynamic. We written some code in python and then saved the output as a CSV file. In a CSV file, we obtained about 17000 entries with the properties of15 columns.

Next step is to clean the data i.e CSV file. we manually cleaned the raw data, and created a Sweetviz (**Sweetviz** is an open-source Python library that generates beautiful, high-density visualizations to kickstart EDA (Exploratory Data Analysis)) report.

Translation/Transliteration: The data in the CSV file must be transliterated into Telugu because the articles must be written in that language. We completed it by utilising the Google Translator tool. Data that was not transliterated or that was incorrectly transliterated were manually corrected. Bing Translator Tool was used.

We had written jinja template for article generation. The jinja template consists of various different sentences formed by using the attributes values.

Final we generated xml files using jinja template

Problems Faced:

It took a long time to web scrape the page and generate the 17000 entries in the CSV file. We manually reviewed each entry after using Google Translator to transliterate the CSV file. For the incorrectly transliterated data, we used Bing Translator. Due to the size of the data, this step took a long time.

In collected data most of the attributes are not available.

In creating article the infobox part is very problematic. First we created normal Infobox with every values of data available by using lables. But sir told to use tennis biography infobox and not general infobox.In the Tennis biography infobox template some data are not translated automatically.

XML part also felt difficult because there is no proper reference video .I had reviewed the fellow internship code to understand.