IST 652 Scripting for Data Analysis

Pokémon Academic Project

(Group Project)

Course Learning Objectives:

* Write scripts to access and amass data from fields in structured data, access fields in

Semi-structured data, and define and find patterns of data in unstructured data;

* Prepare and transform data to produce data summaries, lists, and networks;
* Analyze and solve data access problems for the three types of data and to find and deploy appropriate software packages that can be integrated into the problem solution;
* Frame real-world data questions and show how they can be answered from data

Course skill sets:

* Data wrangling, scripting needed to solve problems of accessing and preparing data in a variety of formats and situations
* Data science pipelines, from acquiring and cleaning data to accessing data and transforming data for analysis or visualization

The goal for this project is to work on the semi-structure and unstructured data, and extract valuable insights from a web site or social media through data wrangling and data science pipelines.

In this project, I have learnt how to scrap a table from a web page through web scrapping packages, how to access data from a social media platform through the social media API, and how to manipulate with JSON structured data, Python dictionaries and lists.

The Pokémon project contains fours datasets from different data sources. The first dataset, main dataset, is from Kaggle, (<https://www.kaggle.com/abcsds/Pokémon>). Two datasets are Pokémon GO dataset and Pokémon moves dataset. Both datasets are from the same Pokémon database website, (<https://pokemondb.net/go/pokedex>) and (<https://pokemondb.net/move>). The last dataset is Tweets that contain Pokémon names, collected from Twitter. We used packages to scrap tables from the webs and utilized Twitter API to collect tweets. Then, we converted those collected data into structured data and semi-structured data for analysis. Data science pipelines were built to clean and transform data for visualization and for loops statements. We created functions, such as splitting a string into two for a column, inserting a bracket for special names, and breaking a string on camel case (a lowercase followed by an upper case). Data type and missing values were converted and replaced with new value. Some additional columns were added to answer the research questions. Further, plots and for loop outputs were implemented to answer the research questions.