

Python Homework 3: Report

Aman Shrivastava(as3ek)

Introduction

Superheroes have dominated pop culture for a long time now. Across different comic book universes, superheroes have a plethora of different strengths and weaknesses. They have different skill statistics across categories like intelligence, strength, power, combat, speed and durability. This project aims to provide an overview about heroes and their physical as well as power characteristics, helping us identify trends and patterns.

Data Source

The data is scraped from the website <https://www.superherodb.com> by extracting the list of heroes from the /characters page and detailed statistics extracted from individual details page of each character.

Content

The dataset extracted has a total of 743 characters and each character has the following 27 attributes :

- Name
- Url
- Intelligence
- Strength
- Speed
- Durability
- Power
- Combat
- Full name
- Alter Egos
- Aliases
- Place of birth
- First appearance
- Creator
- Alignment
- Gender
- Race
- Height
- Weight
- Eye color
- Hair color
- Occupation
- Base
- Team Affiliation
- Relatives
- Skin color
- Total Power

Libraries Used

- Pandas
For importing data into dataframes.
- Requests
For making http requests to web pages to be crawled through the script.
- BeautifulSoup
For parsing and navigating the source code of the web page to be scrapped.
- Matplotlib
For making plots and pie charts during data visualization.
- Jupyter notebook
For displaying the results of data analysis and visualizations made.

Approach

Step 1. Extracting names and urls to their respective details page

Using the requests library sent a request the the base url (<https://www.superherodb.com/characters/>) to extract the source code of the web page with list of all available characters on the website. After converting the source code to a soup object, extracted the names and url to their respective details page of the characters. Stored the collected data will all the names and personal urls to a dataframe.

Step 2. Extracting statistics and info from the individual details page

Iterated over the dataframe created in step 1. to visit the page for each character which contained information and detailed statistics. For each page used the soup object obtained to find and extract relevant information about the character across all the parameters mentioned in the content section. For every loop, appended the collected data to a dataframe.

Step 3. Creating a master dataframe and writing it to a csv file

Merged the dataframes generated from step 1 and step 2 to create a master dataframe of all the collected information about the characters across the parameters mentioned in the content section. Wrote this dataframe to a csv file for storage.

Step 4. Reading the csv file and data cleaning

Read in the data from the csv file generated in the above step to a dataframe to perform further analysis. Changed the data type of the columns to match the type of data contained in them. Took care of missing values by replacing them with relevant data.

Step 5. Generating new parameters from existing information

Used the available character power statistics to create a new column called Total Power to represent the cumulative strength of the character and used it to find the Top 10 most powerful superheroes in the dataset.

Extra Credit**Step 6. Display details about a character selected by the user**

Takes the name of any character as an input from the user and displayed the characteristics and all details about that character

Step 7. Battle two user selected characters and predict winner

Takes the name of two different characters as input and display the more proficient character across all fighting statistics and predict the overall winner of the fight.

Step 8. Generating visualizations and further analysis

Used jupyter notebook (.ipynb) to extract interesting insights and visualize the data further. Have submitted the notebook for consideration and have attached the generated pdf of the visualizations at the end of this report.

Conclusion

The script extracts an exhaustive dataset of all superhero characters, presents some interesting insights in form of visualizations, displays details of requested character and simulates a battle between two user selected characters.

Scope

Apart from these functionalities, the dataset can be used to do more advanced analyses of comic book characters as the dataset contains various other attributes—about the characters' personal traits and history—that can be explored, like race, eye color, height, weight, base, alter egos, team affiliations etc.