Mid Term project steps and challenges

Anirban Pal

Bellevue University

Author Note

The topic detail has been selected on personal interest and the data has been collected from online resources mentioned in the reference section.

Mid Term project steps and challenges

The data for the mid term project is collected from Kaggle. The data file location is: <https://www.kaggle.com/lava18/google-play-store-apps> and the file names are

* googleplaystore.csv (main data file)
* googleplaystore\_user\_reviews.csv

The data contains details of android applications on Google Play store.

**Step 1: Import Data**

The data files are read into ‘pandas’ dataframe as ‘appinfo’ and ‘reviewinfo’ respectively, using ‘pandas read\_csv’ method.

**Step 2: Replace Header**

The review file headers are acceptable. But in the main file, some of the headers have space in them. The spaces are replaced by ‘\_’.

**Step 3: Outliers**

To identify outliers, data profiling is done using the ‘pandas\_profiling’ library. The library provides information about the dataset, variable types, distinct and missing value count for each field, distribution and descriptive statistics of each variable.

After the profiling, I updated the ‘Last\_Updated’ field from string to date time. I also replaced unrelated values in the ‘Category’ field.

**Step 4: Finding and Handling Duplicate Rows**

I found the rows that are exactly duplicates of each other. There were 888 such rows in the dataset. After examining the duplicates, I removed the duplicates from the dataset to proceed with unique values only.

**Step 5: Fuzzy Matching**

Fuzzy matching matches the pattern and assigns the matching % as score. Exactly similar string will have a score of 100. I have used the “fuzzywuzzy” library for comparing the Category column.

Références

Difference between Data Visualization and Business Intelligence,

Retrieved from, <https://www.educba.com/data-visualization-vs-business-intelligence/>

What are the Different Types of Business Intelligence Tools?

Retrieved from, [https://selecthub.com/business-intelligence/key-types-business-intelligence-tools/https://selecthub.com/business-intelligence/key-types-business-intelligence-tools/](https://selecthub.com/business-intelligence/key-types-business-intelligence-tools/https:/selecthub.com/business-intelligence/key-types-business-intelligence-tools/)