Exploratory Data Analysis

1 Overview

Soundcloud is a website that allows users to upload and share songs. We will collect information about the artists, bands, podcasts and music composers on Soundcloud through this project.

Specifically, you will collect the data from https://soundcloud.com/ by parsing HTML. After your data set is ready, you continue analysing it.

2 Purpose

- Hands on collecting data from a website by parsing HTML.
- Hands on preprocessing, analyzing and getting insights about data sets.

3 Detail description

The project has 2 main part. There are Data Collection and Exploratory Data Analysis.

3.1 Data Collection

In the first part, you will get the informations of 3 objects which are User, Playlist, Track.

You can parse HTML to crawl data directly in the Soundcloud website.

- The file such as User,Tracklist,... should be related to each other.
- Each file should be at least 1000 records.
- You can use any python library.

3.2 Exploratory Data Analysis

Before exploring the data, write down a short list of what you expect to know in the data: distribution of key variables, relationships/correlation between pairs of variables, etc.

In the data analysis part, you should tabulate or summarize whatever is needed to know if it matches your expectations.

- Making the Data Analysis Checklist: It can be used as a guide during the process of a data analysis, as a rubric for grading data analysis projects, or as a way to evaluate the quality of a reported data analysis.
- Answer the questions about data set:
  - Did you clearly understand the attributes in your data sets?
  - Have you considered whether the question can be answered with the available data?
- Data preprocessing:
  - Duplicate data
  - Missing data
  - Data Outliers
  - ...
- Data Visualization:

  - Visualize univariate relationships (histogram, distplot, boxplot,...)
  - Visualize bivariate correlations (scatterplot, jointplot, kdeplot, correlation matrix,...)
- Multivariate Analysis (Option): If you use some machine learning basic algorithms to answer a question, you will be add 1+ to total grade of this project.

4 Require

4.1 Code

You should code directly on notebook file .ipynb. In your code, you have to use markdown cell to describe your report.

Your format should be had the following sections:

- Project title
- Assignment tables
- Data Collection
- Exploratory Data Analysis
- References

With data collection part, you must describe in the notebook about the collected data (number of collected samples, meaning of features,...).

With exploratory data analysis, you must have more than 5 question to explore your data sets. After making a question, you must answer it

4.2 Submit

4.2.1 Data

After collecting data, you have to save it to the folder with name is "SoundcloudData". The folder must have files track.csv, playlist.csv, user.csv

With parsing HTML methods, you need to crawl as much data fields (columns) as possible.

Playlist can have multiple tracks, just leave one column of data: trackIds is a string list of ids. For example: playlists[1]["tracks"] = "345,376,389". When you need it, just separate the numbers in the string to have a list of tracks of a playlist.

4.2.2 Project

The submission folder needs:

- The data folder
- The notebook file of the code (<ID1>_<ID2>_<ID3>_<ID4>_<ID5>.ipynb)
- The pdf of the above notebook files (<ID1>_<ID2>_<ID3>_<ID4>_<ID5>.pdf)

Compress the folder into a file, name it with the following syntax and submit it via moodle:
<ID1>_<ID2>_<ID3>_<ID4>_<ID5>.zip

4.3 Important Note

- **The same projects are 0 points for the whole subject.**
- Score distributions:
  - Data Collection: 4 points
  - Data Analysis: 6 points

- Assignment tables:
  - You should divide the work properly, each member needs to take on the same amount of work.
  - Assignment tables consist of 4 columns, such as: ID's student, Name, Task, Contribute rate
  - **Member's grade = Team's grade * Contribute rate**

5 Contact

- Teacher: Nguyen Thi Thu Hang
- Email: ntthuhang0131@gmail.com