Customer-reviews-webS

About

Hello, my name is Jacques Benhur ESSISSONGO, I am a BSc student in Data Science and certified in the Google Data analytics program.

The sole purpose of this project is to make available to the community of analysts and data scientists, a dataset of customer reviews on the largest companies from the Trustpilot site in order to perform analysis, build models and make recommendations.

This data is very useful as it contains a lot of information that can be used in different ways

About

Customer reviews are comments given to a company based on a customer's experience with the organization. By obtaining and analyzing customer reviews, companies can measure customer satisfaction, identify recurring customer issues, determine areas for improvement in existing strategies, and even discover new trends that can be exploited.

Dataset description

Customer-reviews-webS is a directory that contains 8 csv files.Each file is named as follows"company_page_number_extracted".

Each file contains 4 columns of data which are.

country: country where the publication was made

date: date of publication

rating out of 5: number of stars

comment: customer's comment

Method

In order to build this database, I had to do some webscrapping on the trustpilot website. The code allows to collect data in real time on the site.

The code is available on my github and kaggle account. It is quite detailed to allow everyone to understand the process.

For readability purposes, I had to create a class called Trustpilot which contains 5 methods and which does all the extraction as soon as it is called.

Code Structure

The Trustpilot class, once initialized, gives access to the following functions.

- Trustpilot.extract(page_source) extracts the raw data from the source page
- Trustpilot.harmony() which uses another function is_same_length() to measure the sizes of each data list. If the lists are not the same size, the function adds null values to the list that is missing to create the balance.
- Trustpilot.save() creates a data table from the lists and saves the data in a csv file.
- Trustpilot.clean_df() cleans the data table and converts the columns with the right datatype.

Code Structure

The Trustpilot class also gives access to the following attributes.

- df: which is the raw data table
- cleaned_df: which is the data table after cleaning.

Code Structure

NB: The Trustpilot class has a very essential argument called n_run which is used to track the number of times the extract function is executed in the case that there are iterations on several pages. Its default value is 0 and increments by 1 for each iteration. But if it is specified at initialization with a value other than 0, the extract function does not create a new list but rather tries to add to the existing list the data collected during the n_run iteration.