Technical analysis

Overview

* 50 Countries
* 15 E-commerce Sites
* 100+ Sellers in Each E-com site

Tools used

* Google Data Studio
* Google Spread Sheet
* Tableau
* Python (Pandas, NumPy, Matplotlib, Seaborn)
* Data Science Techniques
* Machine learning Algorithm`s

Description

* Getting the ecommerce sites data from google for all 50 countries.
* Cleaning the data to get 15 Ecommerce sites that contains the seller’s details per country.
* Using Scraping techniques to get the product data for each Seller or shop from each ecommerce site.
* Analysing that data and representing in graphical format using Google spreadsheets and google data studio to help client for better future understanding.

About Tools

Google Spreadsheet

Google sheets is a spreadsheet application included as part of the free, web-based Google Docs Editor Suite offered by Google. The App is compatible with Microsoft Excel file formats. Sheets is thoughtfully connected to other Google App’s. Assistive features like Smart fill and formula suggestions help you analyse faster with fewer errors. And get insight quickly by asking questions about your data in simple language.

Google Spreadsheet makes sure that very member of team has the same data and makes changes in all the files if attempted in anyone.

And the changes will reflect the overall results.

Google Data Studio

Google data Studio is a web-based data visualization tool that helps user build customized dashboards and easy-to-understand reports. It helps in tracking key KPI’s for customers, visualizing trends, and comparing performance over time.

Google data Studio is used by every digital marketer to conduct data analysis.

Benefits

* Easier setup with simpler recording
* Creates customized and engaging reports.
* Pulls multiple data source into a single report.
* Makes sharing and collaboration simple.
* Reduces turnaround time on reporting.

Python

Pandas Pandas is an open-source library in Python that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library of Python.

#Role in our project.

Pandas will help us in maintaining and cleaning the scrapped data, stored in excel as .xlsx or .csv format or any other format. i.e. we can drop and remove the unwanted data and empty data for further analysis.

The Changes made during the process will also affect the final google spreadsheet, therefore we won’t have to update the file every time. The changes will be visible to everyone.

Machine learning Tools

* Some of the scraped data is in pictorial format or are images and to make a use of it, image processing technique has been used to extract the factors like Dominating colour in the image, Texts present in the image etc. The OCR engines are used to finish the tasks.
* Another machine learning Algorithm used in our program is the Speech recognition algorithm. The data we have scraped is in raw textual form and it’s impossible to analyse the text data hence, to make a report and statistical analysis we need to convert that data into some useful numerical value and to do so we are using algorithms like speech recognition and sentimental analysis to get the proper values that could be useful for us.
* Sentimental analysis is the process of examining the sentiments of an individual behind the phrase. Giving the different values to words and combining them together to get the average value, which leads the analysts to get to the final sentiment of a user.

Challenges/Failures

1. Before Analysing or making any report it was necessary to check and change the data type of all the data we have scraped.
2. To determine what variables to use to get the proper results. Google data studio allows to change the data type to work with proper data.
3. Size of every scraped data varies. Analysing the data of different length, type, etc. could result in failure of preparation of report.
4. The data scraped is stored in different excel sheets and google data studio can read one sheet at a time. Hence to make a combine result of all the data it is essential to combine all the data in single sheet which could be tough and challenging as there’s no straight way to concat different sheets directly in Google spreadsheet, some code and functions are involved.
5. For image processing, all the images vary in dimensions and some of the images are too small even for a machine to read in that case either an exception is being raised or and error been generated by a model.
6. Sentiment analysis or speech recognition are the techniques to analyse the text, but in most of the situation lot of punctuations are being used by people, different languages are used, connectors, helping verbs, special words are used that could be hard for machine or model to understand in that case model can generate error and the results might not be accurate. Separating all the words removing special case letters could be challenging.