Web Scraping and Machine Learning

Model- 2

Prepared by:

Sara AlAbdulsalam

Fatimah AlShammari

Introduction:

Online shopping today has become an important part of every business. Products presentations and alignments is a significant part because it impacts how the audience perceive their brands. One of the most popular shopping sites is iHerb that has more than 30K products [1] and we chose to help iHerb by implementing machine learning model specifically linear regression to predict the products ratings and reflect it on the products distribution on the site, making the higher rating products appears first and on top to the audience. The model will be trained and tested on data scraped from iHerb website.

Data Description:

A model performance deepens heavily on the data it was trained on. To acquire the data, we will be using web scraping on iHerb website. The features of the products dataset are the following:

- Name: the name of the product
- Categories: the name of the category the product belongs to (e.g., Beauty, Supplements, Grocery)
- Size: the size measurement of the products (measures with oz, grams…)
- Price: product cost in Saudi Riyals
- Number of reviews: the number of customers reviews on the product

The target will be:

- Ratings: the average rate of the product.

Tools Description:

To achieve our goal, we will analyze and explore the data in Python by using Jupyter, and we will use different packages such as: BeautifulSoup, Selenium, Pandas, Matplotlib, Seaborn, and numpy.

Conclusion:

We aim to improve the user experience and site sales by using a linear regression algorithm to predict product ratings, making the products with the highest rating appear first. By training and testing the model on iHerb website dataset. In this document, we reviewed the problem that we want to solve, a description of the data we will work on, and finally the tools that we will use.

References:

[1] https://www.iherb.com/info/about