Instructions to Replicate Final Dataset

The final skincare product review dataset being analyzed consists of a filtered version of 6 separate datasets. One dataset contains information on products, while the remaining five have information regarding reviews submitted about products. Given that each review dataset contains at least 100,000 rows of data, they are too large to upload to the GitHub repository.

Here are the instructions to obtain the original datasets and augment them in a way that reproduces the final dataset that is used in the sentiment analysis:

Go to the link below:
https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews/data?select=reviews_750-1250.csv .

You should come across a page uploaded by the user Nady Inky. Download all 6 datasets (products_info.csv, reviews_0-250.csv, reviews_1250-end.csv, reviews_250-500.csv, reviews_500-750.csv, and reviews_750-1250.csv)

Using the dataset_construction.ipynb file in the SCRIPTS folder, replace the file paths of the read_csv code lines with the corresponding file path to the downloaded datasets in your computer.

So the resulting code to import the product information would look something like:
`products_df = pd.read_csv("/content/product_info.csv")` with the inner part of the read_csv() function replaced with your filepath.

Do this for all of the review datasets as well, which look like this,
`rev1 = pd.read_csv("/content/reviews_0-250.csv", low_memory=False)`

Run the file and the final dataset should be constructed. You can uncomment and run the last chunk in the document to output this file into your Google Drive if you are using Google Colab to run it. If not, run only this line: `Skincare_df.to_csv('Skincare_df.csv', index=False)`

This gives you the final dataset that can be read into the Sentiment_review_analysis.ipynb file as the skincare_reviews_all variable and used for the analysis.