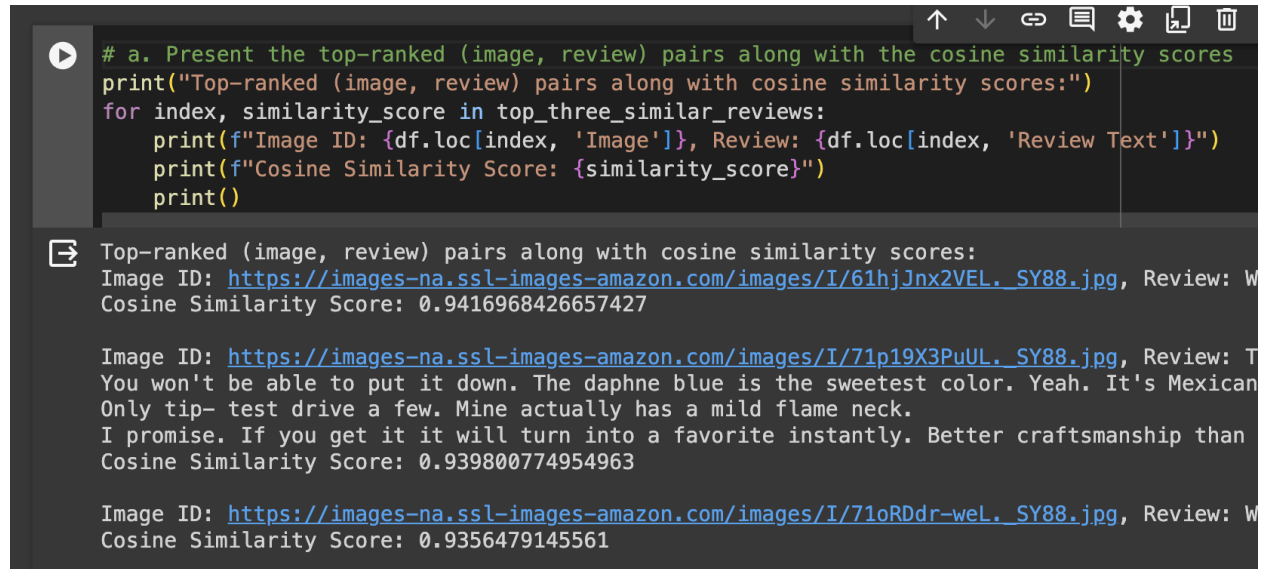


# Assignment 2

## Part 5 (A)



```
# a. Present the top-ranked (image, review) pairs along with the cosine similarity scores
print("Top-ranked (image, review) pairs along with cosine similarity scores:")
for index, similarity_score in top_three_similar_reviews:
    print(f"Image ID: {df.loc[index, 'Image']}, Review: {df.loc[index, 'Review Text']}")
    print(f"Cosine Similarity Score: {similarity_score}")
    print()
```

Top-ranked (image, review) pairs along with cosine similarity scores:

Image ID: [https://images-na.ssl-images-amazon.com/images/I/61hjJnx2VEL\\_SY88.jpg](https://images-na.ssl-images-amazon.com/images/I/61hjJnx2VEL_SY88.jpg), Review: W  
Cosine Similarity Score: 0.9416968426657427

Image ID: [https://images-na.ssl-images-amazon.com/images/I/71p19X3PuUL\\_SY88.jpg](https://images-na.ssl-images-amazon.com/images/I/71p19X3PuUL_SY88.jpg), Review: T  
You won't be able to put it down. The daphne blue is the sweetest color. Yeah. It's Mexican  
Only tip- test drive a few. Mine actually has a mild flame neck.  
I promise. If you get it it will turn into a favorite instantly. Better craftsmanship than  
Cosine Similarity Score: 0.939800774954963

Image ID: [https://images-na.ssl-images-amazon.com/images/I/71oRDdr-wEL\\_SY88.jpg](https://images-na.ssl-images-amazon.com/images/I/71oRDdr-wEL_SY88.jpg), Review: W  
Cosine Similarity Score: 0.9356479145561

## Part 5 (B)

Quantitative Comparison:

- Image retrieval using feature extraction yields superior results in quantitative analysis.
- The VGG16 model captures high-level features from images, encoding comprehensive visual information.
- Due to the high similarity among images, the extracted features likely encapsulate crucial visual characteristics.
- Cosine similarity scores computed from these features effectively measure image similarity, leading to accurate retrieval outcomes.

Qualitative Comparison:

- Reviews outperform in qualitative analysis as they offer more comprehensive insights.
- Reviews contain detailed textual descriptions encompassing various aspects of the images.
- These descriptions include information such as composition, context, emotions, and specific details not solely captured by visual features.

- Computing cosine similarity scores between TF-IDF representations of reviews captures semantic similarity effectively, providing a nuanced understanding of images.

Overall Assessment:

- While image retrieval based on feature similarity excels quantitatively due to image similarity, reviews offer richer context and detail qualitatively.
- Reviews enhance the overall similarity assessment by providing additional information beyond visual features.

## Part 5 (C)

Challenges:

- Ensuring the quality of text preprocessing and feature extraction
- Dealing with sparsity in TF-IDF vectors
- Handling large datasets efficiently
- High Dimensionality

Potential improvements:

- Experiment with different text preprocessing techniques
- Explore advanced feature extraction methods for images
- Use more sophisticated similarity measures
- Employ dimensionality reduction techniques to handle high-dimensional data
- Implement parallel processing for faster retrieval
- Fine-tune parameters and algorithms based on the specific dataset and requirements

In the initial phase of our analysis,

- we commence by importing the dataset,
- configuring it to include pertinent columns such as product IDs, image links, and associated review texts.
- Following this, we focus on refining the image data by transforming the links from strings into more manageable list formats, streamlining subsequent processing tasks.
- Subsequently, we proceed to transition from image links to tangible images, and downloading and integrating the images into our dataset.

With the images in hand, our focus shifts towards enhancing their quality and consistency.

- We undertake essential tasks such as resizing images to standardized dimensions and fine-tuning factors like contrast and brightness to ensure optimal visual representation across the dataset.
- Simultaneously, we use the VGG16 model to extract distinctive features from the images. These extracted features undergo normalization

Simultaneously,

- We do text analysis, employing similar methodologies to preprocess the textual reviews.(As per Assignment 1)
- We compute essential metrics such as term frequency (TF), inverse document frequency (IDF), and composite TF-IDF scores for each preprocessed review text, laying the groundwork for comprehensive textual analysis.

With both image and textual data prepared, we proceed to identify the most similar images and reviews.

- For images, we compute cosine similarity scores between the extracted image features and those of a selected input image, facilitating the identification of the top three most similar images.
- Likewise, for reviews, we calculate cosine similarity scores between the TF-IDF scores of the reviews and those of a chosen input review, enabling the identification of the top three most similar reviews.

In the final stage of our analysis, we consolidate these findings to derive a composite similarity score. By averaging the cosine similarity scores obtained from both image and review comparisons, we obtain a full evaluation of similarity.