IR Assignment 2

Nitesh Kumar Chaurasia
MT23053

Packages Used

Pandas - For Data Manipulation
Numpy - For Numerical Computing
Tensorflow - For Image Classification
Sklearn - Model Evaluation and Preprocessing
NLTK - Text Processing libraries for tokenization, stemming
PIL - Manipulating Image format

Question 1

For adjustment of contrast, brightness packages Image and Image Enhance from module PIL has been used which are then resized to 224 to 224 in order to maintain uniformity. This is done for each image in the dataset.
Numpy linalg.norm module is used for feature normalization the images and store them back.

Pipeline

- First the image is preprocessed
- The preprocessed image is selected and relevant features are extracted
- The features extracted are normalized

Question 2

NLTK library is used for preprocessing of text. The textual data is converted to lowercase. Then we tokenize the data and further perform stemming and lemmatization after removing stopwords. Then we calculate the the TF-IDF score and return the score and document frequency. These are then stored to a csv file along with word frequency, term frequency, and tf-idf score.

Question 3

Using the images in the database, cosine similarity between the input images and other images is calculated. The similarity scores generated are then stored in a list named similarity_images. The list is then sorted in descending order ensuring the most similar image appearing at the top. Top 3 similar images are saved in another csv with each row to a retrieved image with its relevant details.

Question 4

Here we are retrieving similar images and text reviews based on the input image URL and review text. The URL and review text are taken as user input. On the retrieved images we are calculating the similarity scores between the existing and the input images along with existing and input text both.

Preprocessed data is loaded from the already created CSV. and then scores for both image and text similarity is calculated. The DataFrame is sorted based on text similarity scores, and the top three similar reviews are retrieved and then composite similarity score is calculated as the average of image and text similarity scores for each pair and top three pairs having highest composite score are retrieved.

Question 5

Performance comparison (B)

The retrieval method based on review text similarity, specifically using TF-IDF scores, consistently yields higher similarity scores compared to the method relying on image features, assessed through cosine similarity. When employing TF-IDF scores for review text similarity, the cosine similarity scores often approach 1.0, indicating a significant likeness between the input review text and the retrieved reviews. This underscores the effectiveness of the TF-IDF approach in capturing semantic similarity by considering the content and meaning of the reviews. Conversely, the cosine similarity scores for image features exhibit more variability and tend to be lower than those for review text similarities. While certain pairs may exhibit high image similarity scores, others might demonstrate relatively lower scores.

(c)

difficulties arise in precisely capturing semantic significance, effectively representing

features, and ensuring scalability. Enhancements could entail the utilization of sophisticated models, refining semantic comprehension, integrating hybrid methods, adjusting models meticulously, and deploying efficient algorithms.