IR Assignment 2

(Jay Saraf 2020438)

1. I have resized the images to (300,300). The contrast was increased and the brightness was also changed. I have used two methods for normalizing. One is sklearn's standard scaler and the other is custom min max scaling.
2. I have done preprocessing of the text and store the tfidf vectors in a new column.
3. Files are saved to fetch the extracted features and the tfidf scores. For text based retrieval, the image giving the most cosine similarity value was considered out of the given list of urls.
4. I have taken average of the cosine similarity of image and text based retrieval for the composite similarity score.
5. a. Done in the code.
   b. Image retrieval was giving better results. There could be various reasons like feature extraction was better. Reviews were less in number as compared to the number of urls.
   c. There were various challenges faced. For example data cleaning was required to remove the 8 redundant urls. The 'Review Text' column had text written in multiple lines. The ids were mapped to multiple urls.
   Potential improvements could be applying super resolution on images to get better image quality. Using some kind of feedback mechanism to improve the results. Finding some advanced techniques for getting the correlation between the image and the review (text) provided.