Krannert School of Management

Individual Assignment 2

Total Points: 10

Part 1. Image Representation (total 5 points)

There are 10 images in a collection saved under the attached compressed file on Brightspace. All those images are in png format. In this assignment, you are required to process those images, so they will be converted to arrays as the final output. Please follow the steps listed below:

1. Read in all of the 10 images in python, resize each of them a 100 by 100 pixels format.
2. Following step 1, convert them to greyscale arrays (no color information needs to be kept, so the arrays will be 2-D).
3. Following step 2, flatten the 2-D array to a 1-D array (vectors), draw a histogram to present the intensity value distribution for each image.
4. Following step 3, conduct a histogram equalization to normalize each image, draw a histogram to present the intensity value distribution of each image after the normalization.
5. Compare the histogram in step 4 with step 3, discuss what the difference is.

   Tips: you may consider using a "for loop", so you could process the whole collection at once.

Please submit two files:

1. A word file includes python code with your comments #, and one screenshot on your PyCharm showing that your code has run through successfully for **each of the first four steps** (4 screenshots in total). Also, includes the histogram in step 3 and step 4, and discuss what the difference is (step 5) at the end of the word file.
2. A CSV file saves flattened arrays (step 3), each image should be a row, and there should be 100*100 columns.

Part 2. Topic Model (total 5 points)

There are 1000 reviews for restaurants and films in a collection under the attached csv file on Brightspace. All those reviews are saved as text files. In this assignment, you are required to investigate the topics of those reviews. Please follow the steps listed below:

1. Transform those reviews into a term-document matrix, lemmatize all the words, remove the stop-words and punctuations, set the minimal document frequency for each term to be 5 and include unigram and bi-gram.
2. Following step 1, use the LDA model to extract the topics of each document assuming there are 6 topics.
3. Report the topic distribution and the top-2 topics of the first 10 restaurant reviews (id = [1:10]) and the first 10 movie reviews (id = [501:510]).
4. Find the top-5 terms (terms with the top-5 highest weights) for each of the 6 topics. Based on those terms, describe what those topics are about.
5. Based on finding in 3 and 4, describe what review 1 [ID=1] and review 501 [ID=501] are about (i.e., the themes of their focal topics)?

Please submit 1 file:

A word file includes python code with your comment #, and one screenshot on your PyCharm showing that your code has run through successfully for **each of the first four steps** (4 screenshots in total). Also, **report** your answers to question 3, 4, and 5 at the end of the word file.