Homework 3: Applying tf-idf scheme to non-text data

Logistics

- This is a group assignment. Submit one zip file per project group.
- Due date: 12:30pm Tuesday, November 9, 2021
- Method: Upload a zip file to eTL. The zip file name should be [group number].zip such as [10.zip] without brackets.
- Zip file content (file names are case sensitive):
  - report.pdf (document that contains your story and visualizations)
  - *.py or *.ipynb (any Python code or Jupyter Notebook file you created for this homework)

Background

The tf-idf scheme is not only applicable to text mining. It is a flexible scheme to be applied to a non-text data setting as long as the data can be mapped to the paradigm of "document-word" pairs.

You will be using a dataset compiled from the class survey where you submitted top 5 preferred Kaggle datasets with different weights. You will be asked to convert it into a document-term matrix and apply several text mining techniques.

The mapping is:

- Each student is considered a document.
- Each Kaggle dataset is considered a term.
- The weight is the number of occurrences of the term (= Kaggle dataset) in a given document (= student).

Thus, each person is a document of the same length of 100. Through this homework, you will better understand the concept of tf and idf.

To do list

- Download the csv file: sp21_student_kaggle.csv
- The following items will be checked for the grade.
  - Compute tf, idf, and tf-idf. Print in the report the top 10 student-kaggle data pairs in each of tf, idf, and tf-idf. Print the value of tf, idf, tf_idf along with each student-kaggle data pair. Each accounts for 1 point. (Total 3 points)
  - Create two histograms of tf and tf-idf for one student in your group as a document. Each histogram accounts for 0.5 points. (Total 1 point)
  - Explain why the two histograms look different. (1 point)
  - Compute cosine similarities between you and all other students using the tf-idf statistics normalized by L2 norm. (0.5 points)
  - According to the cosine similarities you computed above, who (other than yourself) is most similar to you in terms of their weighted choice of Kaggle datasets? Name all if multiple people tie. (0.5 points)