CSCI 182 Homework 4
Due Date: Feb 18th, 2022

All assignments MUST have your name, student ID, course name/number at the beginning of your documents.
Your homework MUST be submitted via Camino with the file format and name convention as follows:
HW#_Name.zip (both the python file and the report)
If you have any questions, please don't hesitate to contact me :)

TF-IDF

1. **(This question is a theory question)** Assume you have three documents:

    a. **Document 1:** It is going to rain today.
    b. **Document 2:** Today I am not going outside.
    c. **Document 3:** I am going to watch the season premiere.

Find the TF-IDF scores for each of the words in these three documents. Show the steps involved in the calculation (similar to example in class). Consider the TF to be the raw count of words in the document and normalize if required. For the IDF take base 2.
For each document state the words that are important for that document. What can you conclude about the documents based on the TF-IDF scores.

2. **(This question is a programming question**) Given the 15 documents in the Docs.zip, perform Keyword extraction using Python for each of the documents using TF-IDF.
    a. Report the top 5 keywords for each document.
    b. Are the keywords representative of the document? Report your findings.