INFO-533 Homework 4 (100 points)

Deadline: TBD, 2024 11:59 pm in Brightspace

1. Consider the two metrics Precision & Recall. For an IR system, which of the following is possible:
    a. Precision = 1
    b. Recall = 1
    c. Both = 1

    Justify your answer.

2. Calculate Cosine Similarity between each document and query pair provided below:
    Documents:
    1. fast car win more race
    2. sport car and fast car win
    3. car win major car race
    4. formula race win major car bet

    Queries:
    1. formula
    2. sport fast drive

3. How storing term indexes in sorted manner helps in searching the terms? Give an example where sorted index can be searched quickly compared to unsorted.
4. Is stemming good in every situation? Explain.
    Give a stemmed example query where the document 4 from question 2 will be retrieved because of stemming but the query has no relation with the document 4.
5. Consider all the documents and 2nd query from question 4. Use formulas from the slides.
    a. Calculate the TF-IDF scores of query with document 1 and 4 using natural (raw) term frequency count, no document frequency, and cosine similarity
    b. Calculate the TF-IDF scores for all documents using logarithmic term frequency and inverse document frequency
    c. Give the ranking for the query for all the documents. Show your calculations.
    d. Explain with an example, the problems with using natural (raw) frequency