Kettering UNIVERSITY

College of Sciences & Liberal Arts
Department of Computer Science
CS 465/665 Information Retrieval and Data Mining
(Winter 2024)
Project # 1 (10 points)
Group Project (3 students)
Course Instructor: Dr. Jamal Alhiyafi
Due: Thursday, February 15, 2024

Instructions:

- Grade will be based on the running code, code comments, naming convention, etc.
- This is a group assignment. Each member will fill out an evaluation form evaluating other members of the group. This will be used to determine your grade for this assignment.
- You will submit a statement listing each member's contribution to this assignment.
- Make sure to add your names, CS465/665, W24, and Project # 1 at the beginning of each file.
- Title of each file should start with CS465/665-W24-IRproject-Group#-……
- Make sure to add your references.
- Submit only one WinZip or WinRAR file containing all your code files, a simple report describing your system, and how to compile, run, and use the program.

- Design and implement a simple IR system.
- The system should:
  - Create the inverted index (the dictionary and postings lists) for your collection of documents
  - Parse and execute simple queries
  - Perform simple tokenization and normalization of the text such as removing digits, punctuation marks, etc.
  - Statistics:
    - ✓ Report the number of distinct words observed in each document, and the total number of words encountered.
    - ✓ Report the number of distinct words observed in the whole collection of documents, and the total number of words encountered.
    - ✓ Report the total number of times each word is seen (term frequency) and the document IDs where the word occurs (Output the posting list for a term).
    - ✓ Report the top 100^{th}, 500^{th}, and 1000^{th} most-frequent word and their frequencies of occurrence.
    - ✓ Create postings and assign a term frequency to every document in the postings list.
    - ✓ Provide a simple GUI to test the system.

Please ask questions if anything is not clear.

- **If interested, you may contact me to do an extra task for an extra credit.**