HIT220: Group Assignment 3.3 - Algorithm and Complexity

Oct 2024

Project Title: Keyword Search and Frequency Analysis Tool

Assignemnt Overview:

In this project, your group will develop a tool that analyzes a set of documents (text files) to search for specific keywords, count their occurrences, and identify common phrases. The tool will employ various text processing techniques to achieve its objectives efficiently.

Objective：

To assess your ability to apply text processing techniques to solve a problem, demonstrating your understanding of algorithms and data structures related to maps, dictionaries, hash tables, sets, tries, and pattern matching.

Assignment Tasks:

1. Keyword Frequency Count using Hash Tables (7 Marks)

1) **Task:**
   Implement a hash table to store keywords (given as input) and count their occurrences in a set of provided documents(doc1.txt, doc2.txt). The program should read the documents, extract words, and populate the hash table with the count of each keyword.
2) **Deliverables:**
   Submit the code for the hash table implementation and a brief explanation (150 words) describing how the frequency count works and its time complexity.
3) **Assessment Criteria (Rubrics):**
   a) Correctness and functionality of the hash table implementation (4 Marks)
   b) Clarity and accuracy of the explanation (3 Marks)

2. Common Phrase Detection using Tries (8 Marks)

1）**Task:**

Use a trie to store and detect common phrases (of length 3 words) in the provided documents. The tool should output all the phrases that appear in more than one document.

2）**Deliverables:**

Submit the code for the trie implementation along with a brief explanation (150 words) on how phrases are stored and detected.

3）Assessment Criteria (Rubrics):

a) Correctness and efficiency of the trie implementation (5 Marks)

b) Quality of the explanation and understanding of trie operations (3 Marks)

Submission Requirements: submit each document and code as **individual files**; do not compress them into a zip file.

1) One source code file for each task (in Python), submitted separately.
2) A brief report for both tasks, clearly highlighting the title of each task.