OS Fall Lab Mid Exam

0014

String Similarity Analysis Tool

Input/Outputs should be done using system calls

1. **Introduction:** Display a welcome message and brief instructions on how to use the tool for calculating string similarity and performing operations based on user input.
2. Input Processing: The first process reads input strings from a file containing n pairs of strings.
3. **Pipes Creation:** Create pipes to establish communication channels between each pair of consecutive processes.
4. **Forking Processes:** Create four processes, each process will perform a specific task in the data processing pipeline.
5. **Data Preprocessing Process (Second Process):** This process receives the strings from the first process after exec call. It also receives a reference string from the user. It then calculates the similarity score between each pair of input strings and the reference string using a chosen similarity metric (e.g., cosine similarity), and passes the scores to the next process through a pipe.
6. **Sorting Process (Third Process):** This process receives the similarity scores from the second process via a pipe. It sorts the scores and passes the sorted scores and corresponding input strings to the next process through another pipe.
7. **Analysis Report Process (Fourth Process):** This process receives the sorted scores and corresponding input strings from the third process via a pipe. It generates an analysis report based on the received data and displays the top k elements with the highest similarity scores to the reference string, where k is taken from the user.
8. **Error Handling:** Implement robust error handling to handle potential failures during process creation, pipe establishment, or data processing. Ensure that your program gracefully handles any errors and provides informative error messages.
9. **Cleanup:** Properly close and release any open file descriptors or resources associated with pipes and child processes. Ensure proper cleanup to avoid memory leaks or resource exhaustion.

Expected Output:

Terminal 1

```
Input strings loaded from file:
String 1: "cat dog mouse elephant"
String 2: "dog mouse elephant tiger"
```

Terminal 2

```
Calculating similarity scores with reference string "dog cat elephant":
Similarity scores calculated:
String 1: Similarity Score = 0.816
String 2: Similarity Score = 0.952
```

Terminal 3

```
Sorting similarity scores:
Sorted scores:
String 2: Similarity Score = 0.952
String 1: Similarity Score = 0.816
```

Terminal 4

```
Generating analysis report:
Top 2 strings with highest similarity scores to reference string "dog cat elephant":
1. String 2: Similarity Score = 0.952
2. String 1: Similarity Score = 0.816
```

Helping Material

Cosine similarity is a measure of similarity between two non-zero vectors in an inner product space that measures the cosine of the angle between them. When dealing with strings, cosine similarity can be calculated based on the frequency of terms (words) in the strings, treating each string as a vector where each dimension represents the frequency of a specific term.

Here's the formula for cosine similarity between two strings A and B:

$$\text{cosine\_similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where:

- A_i and B_i are the frequencies of term i in strings A and B, respectively.
- n is the total number of unique terms across both strings.

Example:

Let's say we have two strings:

- String 1: "apple banana apple"
- String 2: "banana orange banana"

We'll first tokenize these strings into terms (words) and then calculate the frequency of each term. After that, we'll use the formula for cosine similarity to find the similarity between these two strings.

Tokenization and Frequency Calculation:

For the given strings:

- String 1: "apple banana apple"
- String 2: "banana orange banana"

The unique terms are: "apple", "banana", "orange".

The frequency of each term in the strings:

- String 1: {"apple": 2, "banana": 1}
- String 2: {"banana": 2, "orange": 1}

Using the formula:

$$\text{cosine\_similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

For our example:

- A represents the frequency vector of terms in String 1.
- B represents the frequency vector of terms in String 2.

Let's calculate:

$$\text{cosine\_similarity}(A, B) = \frac{(2 \times 1) + (1 \times 2)}{\sqrt{(2^2 + 1^2)} \times \sqrt{(2^2 + 1^2)}}$$

2x2 + 1x1

$$\text{cosine\_similarity}(A, B) = \frac{2 + 2}{\sqrt{5} \times \sqrt{5}}$$

$$\text{cosine\_similarity}(A, B) = \frac{4}{5}$$

$$\text{cosine\_similarity}(A, B) = 0.8$$