File Matcher

Table of Contents

1. Notes

- When passing the paths use forward slash not back slash.

- I have created this application based on the assumptions that were given. There are a lot of different ways of implementation

- Also, I have picked this algorithm based on its simplicity and efficiency, since a certain accuracy rate wasn't specified.

- I didn't want to add more details and complexity to the project to make it simple.

I have used Java 17 with Dependencies

- Spring Web Starter (Spring Web 3.4.5)

2. Introduction of the Algorithm Used

In this project, the Weighted Jaccard Similarity algorithm is utilized to measure the similarity between two files.
Every file gets translated into a map of words and their number of occurrences.

Then, it calculates the sum of the minimum and the sum of the maximum frequencies for all words, and computes the score as follows:

Score = (Sum of Minimums) / (Sum of Maximums)

Example:
Text 1
“ hello this is me “

Text 2
“ hello hello this is me away”

Set 1
(hello, 1), (this, 1), (is, 1), (me, 1)

Set 2
(hello, 2), (this, 1), (is, 1), (me, 1), (away, 1)

All Combined Words with Frequencies:
- hello → (1, 2)
- this → (1, 1)
- is → (1, 1)
- me → (1, 1)
- away → (0, 1)

Sum of Minimums = 1 + 1 + 1 + 1 + 0 = 4

Sum of Maximums = 2 + 1 + 1 + 1 + 1 = 6

Final Score = 4 / 6 = 66%

3. Project Structure

This project uses Spring Web With one end point that takes the file path and the directory path and then goes through its service layer to compare the given file with each file and produce a list of file names and their scores

Project Files Overview:

- **FileController**: has one end point FileMatcher that takes two params file path and directory path and pass it to the service and return the list of results

- **File Service**: Main service layer that has four methods:
- **fileReader**: takes the file path and reads it into a String and calculates the map of frequency

- **directoryReader**: takes the path of the directory and returns array of files

- **calculateScore**: Calculates the similarity score of the given file with each file in the directory

- **fileMatcher**: Main method it uses all the above methods and return List of results

- **Jaccard Service:** Calculates the matching score between two files

- **MapService**: filters a String of words and remove non alphabetic words and make it into lowercases words and then turn it into a map of words and their frequencies

- **FileRecord:** Holds the file attributes throughout the entire processing step.

- **TaskApplication**: The main class that launches the Spring Web app.

- **CustomException:** a custom exception extends exception to pass a custom message.

- **GlobalExceptionHandler**: has one method that handles exception.

4.Sample Input

This input is provided by AI

Original Text

Life is beautiful 2025! Coding@night & debugging till 3AM. Grab $20 for groceries @SuperMart. Sunshine&Rainbows with 100% joy.

File 1:

The world is full of wonders 2025! Dream@big and achieve more every day. Pizza@Midnight with $15 leftovers. Smile&Shine with 300% enthusiasm. Wander through the galaxy of ideas! Life_is_a_journey, embrace@every_step123.

File 2:

The universe is full of wonders 2025! Achieve_big dreams and more every day. Pizza in Midnight with $20 leftovers. Keep&Shine with 250% motivation. Explore through galaxies of inspiration! Journey_is_life, embrace@new_steps123.

File 3:

The cosmos sparkles with stars 2040! Big@dreams await those who strive. Midnight snacks cost $30 at the shop. Glow&thrive with 200% effort. Fly toward galaxies of knowledge! Walk_the_path, explore@each_step888.

File 4:

Discover the mysteries of life 2050! Achieve incredible dreams every@day. Late-night cravings require $50 in hand. Shine&grow with boundless enthusiasm. Journey_through endless horizons! Adventure_is_a_key to success@1step.

File 5:

Embrace the unknown paths@in_life. Bright mornings bring courage&hope. $10 for a cup of coffee at Midday. Travel across time and galaxies! Live_inspiring moments everywhere! Take_the_lead and shine@forever.

File 6:

The world is full of wonders in 2025! Dream@big to achieve everything each day. Pizza_at_Midnight with just $15 savings. Keep&Smile with 300% positivity. Wander through galaxies of creativity! Life_is_a_journey, take@steps123forward.

Output:

File name: File1.txt Matching score is: 100.0%

File name: File2.txt Matching score is: 50.0%

File name: File3.txt Matching score is: 16.666666666666664%

File name: File4.txt Matching score is: 14.285714285714285%

File name: File5.txt Matching score is: 9.375%

File name: File6.txt Matching score is: 100.0%