Plagiarism Validation

Appendix: O/P Unification

This document summarizes some points to unify the output of the project to be aligned with the released output of the test cases. It covers three parts:

1. adjustments to the average similarity calculation,
2. Format of the statistics file
3. Format of the MST file.

# 1. Average Similarity

## Rounding Rule:

At every step involving average similarity, ensure that the value is rounded to one decimal point. This standardizes the similarity scores across the project and simplifies further analysis.

# 2. Format of the statistics file

## Count of Component Elements:

* Add a new column indicating the count of elements in each component. This additional information helps understand the size of each component.

## Sort the Components:

* Sort the components in the statistics file by their average similarity in descending order. This ensures that components with higher similarity appear first, providing a clear hierarchy.

## Store Only IDs:

* When writing the elements of each component, only store the numeric IDs, not the full path. This approach aligns with standard test cases, avoiding excessive information in the output.

## Sort the Component Elements:

* Sort elements within each component in ascending order based on the numeric part in the path. This ensures correct order, especially when elements contain various formats or structures.

Check the following example for the expected output for the given example.

### Example:

|  |  |  |
| --- | --- | --- |
| **File 1** | **File 2** | **Lines Matched** |
| [D:/Source/1/ (25%)](http://moss.stanford.edu/results/0/4012916089008/match0.html) | [D:/Source/2/ (36%)](http://moss.stanford.edu/results/0/4012916089008/match0.html) | 84.00 |
| [D:/Source/2/ (75%)](http://moss.stanford.edu/results/0/4012916089008/match1.html) | [D:/Source/3/ (76%)](http://moss.stanford.edu/results/0/4012916089008/match1.html) | 52.00 |
| [D:/Source/3/ (64%)](http://moss.stanford.edu/results/0/4012916089008/match2.html) | [D:/Source/1/ (13%)](http://moss.stanford.edu/results/0/4012916089008/match2.html) | 28.00 |
| [D:/Source/4/ (64%)](http://moss.stanford.edu/results/0/4012916089008/match3.html) | [D:/Source/5/ (13%)](http://moss.stanford.edu/results/0/4012916089008/match3.html) | 28.00 |
| [D:/Source/5/ (47%)](http://moss.stanford.edu/results/0/4012916089008/match4.html) | [D:/Source/6/ (49%)](http://moss.stanford.edu/results/0/4012916089008/match4.html) | 40.00 |
| [D:/Source/6/ (47%)](http://moss.stanford.edu/results/0/4012916089008/match5.html) | [D:/Source/4/ (48%)](http://moss.stanford.edu/results/0/4012916089008/match5.html) | 41.00 |
| [D:/Source/7/ (25%)](http://moss.stanford.edu/results/0/4012916089008/match6.html) | [D:/Source/8/ (23%)](http://moss.stanford.edu/results/0/4012916089008/match6.html) | 22.00 |
| [D:/Source/8/ (31%)](http://moss.stanford.edu/results/0/4012916089008/match7.html) | [D:/Source/9/ (41%)](http://moss.stanford.edu/results/0/4012916089008/match7.html) | 37.00 |
| [D:/Source/10/ (37%)](http://moss.stanford.edu/results/0/4012916089008/match8.html) | [D:/Source/11/ (8%)](http://moss.stanford.edu/results/0/4012916089008/match8.html) | 19.00 |
| [D:/Source/11/ (19%)](http://moss.stanford.edu/results/0/4012916089008/match9.html) | [D:/Source/8/ (15%)](http://moss.stanford.edu/results/0/4012916089008/match9.html) | 29.00 |

### Expected Output of the Statistics File:

|  |  |  |  |
| --- | --- | --- | --- |
| **Component Index** | **Vertices** | **Average Similarity** | **Component Count** |
| 1 | 1, 2, 3 | 48.2 | 3 |
| 2 | 4, 5, 6 | 44.7 | 3 |
| 3 | 7, 8, 9, 10, 11 | 24.9 | 5 |

# 3. Construction & Format of the MST File

## MST: Handle Edges with the Same Weight

* During the MST algorithm, if two edges within a component have the same weight, select the one with the greatest `Line Matches`. This rule helps differentiate edges with similar weights based on additional criteria.

## Sort the Component:

* Components in the MST file should be shown in descending order of average similarity. This provides a clear sequence, with the highest similarity components displayed first.

## Elements within each Component:

* Plot each component and its edges together, completing component by component. This approach helps maintain consistency and structure.

## Sort Edges within Each Components:

* Sort the edges of each component internally by the `LineMatches` value in descending order. This approach organizes components by the frequency of matches, allowing for a clear hierarchy within each component.

Check the following example for the expected output for the given example.

|  |  |  |
| --- | --- | --- |
| **File 1** | **File 2** | **Line Matches** |
| [D:/Source/2/ (75%)](http://moss.stanford.edu/results/0/4012916089008/match1.html) | [D:/Source/3/ (76%)](http://moss.stanford.edu/results/0/4012916089008/match1.html) | 52 |
| [D:/Source/3/ (64%)](http://moss.stanford.edu/results/0/4012916089008/match2.html) | [D:/Source/1/ (13%)](http://moss.stanford.edu/results/0/4012916089008/match2.html) | 28 |
| [D:/Source/5/ (47%)](http://moss.stanford.edu/results/0/4012916089008/match4.html) | [D:/Source/6/ (49%)](http://moss.stanford.edu/results/0/4012916089008/match4.html) | 40 |
| [D:/Source/4/ (64%)](http://moss.stanford.edu/results/0/4012916089008/match3.html) | [D:/Source/5/ (13%)](http://moss.stanford.edu/results/0/4012916089008/match3.html) | 28 |
| [D:/Source/8/ (31%)](http://moss.stanford.edu/results/0/4012916089008/match7.html) | [D:/Source/9/ (41%)](http://moss.stanford.edu/results/0/4012916089008/match7.html) | 37 |
| [D:/Source/11/ (19%)](http://moss.stanford.edu/results/0/4012916089008/match9.html) | [D:/Source/8/ (15%)](http://moss.stanford.edu/results/0/4012916089008/match9.html) | 29 |
| [D:/Source/7/ (25%)](http://moss.stanford.edu/results/0/4012916089008/match6.html) | [D:/Source/8/ (23%)](http://moss.stanford.edu/results/0/4012916089008/match6.html) | 22 |
| [D:/Source/10/ (37%)](http://moss.stanford.edu/results/0/4012916089008/match8.html) | [D:/Source/11/ (8%)](http://moss.stanford.edu/results/0/4012916089008/match8.html) | 19 |