

Preview - Levenshtein

The screenshot shows the Spoon IDE interface. The main window displays a document titled 'lab04.pdf' with a table of language pairs and their average Levenshtein distances. The 'Examine preview data' window is open, showing the same data. The table has 8 rows and 3 columns: language, language2, and avg_measure.

language	language2	avg_measure
portuguese	spanish	0.8571428571
spanish	portuguese	0.8571428571
italian	spanish	1.75
german	swedish	2.5
swedish	german	2.5
dutch	swedish	2.9285714286
french	spanish	3.0357142857
english	italian	3.1071428571

81. Change the transformation to use the Damerau-Levenshtein distance. Do the results change? What conclusions can you draw from here?

The screenshot shows the Spoon IDE interface with the transformation changed to Damerau-Levenshtein distance. The 'Examine preview data' window displays the updated results. The table has 8 rows and 3 columns: language, language2, and avg_measure.

language	language2	avg_measure
portuguese	spanish	0.8571428571
spanish	portuguese	0.8571428571
italian	spanish	1.75
german	swedish	2.5
swedish	german	2.5
dutch	swedish	2.9285714286
french	spanish	3.0357142857
english	italian	3.1071428571

Yes, some pairs have changed

For the pair Portuguese-Spanish, the result did not change, indicating that transpositions were not significant for this pair, or the languages were similar enough that additional operations did not affect the result. For pairs like french-Spanish and Dutch-Swedish, the Damerau-Levenshtein distance increased, indicating that when considering transpositions as a possible operation, these languages were found to be less similar than initially calculated using just Levenshtein.

82. Change the transformation to use the Needleman-Wunsch measure. Do the results change? Note: The Needleman-Wunsch measure is better if its values are higher. So you will have to change the way the avg_measure is being sorted at two different places in the transformation.

The screenshot shows the Spoon IDE interface. On the left, a list of exercises is visible, with exercise 82 selected. The main workspace displays the configuration for a transformation step. A table shows the results of the transformation, with columns for language, language2, and avg_measure. The data is sorted by avg_measure in descending order.

#	language	language2	avg_measure
1	portuguese	spanish	0.8571428571
2	spanish	portuguese	0.8571428571
3	italian	spanish	1.75
4	german	swedish	2.5
5	swedish	german	2.5
6	dutch	swedish	2.9285714286
7	french	spanish	3.0357142857
8	english	italian	3.1071428571

An "Examine preview data" window is open, showing the same data table. The window also displays the transformation step configuration, including the transformation name and the transformation type.

The Needleman-Wunsch algorithm is based on alignment and allows for both positive and negative scoring based on matches, mismatches, and gaps. The presence of negative scores indicates that some alignments between the language pairs might have been heavily penalized, possibly due to many mismatches or gaps, leading to a lower similarity score compared to the previous methods. This change in the scoring reflects the different approach taken by Needleman-Wunsch in evaluating a string similarity.

83. Change the transformation to use the Jaro measure. Is there any change in the results, besides the measure values?

The screenshot shows the Spoon IDE interface with the 'Sort rows 3' step selected. The 'Examine preview data' window displays the following data:

#	language	language2	avg_measure
1	portuguese	spanish	0.8571428571
2	spanish	portuguese	0.8571428571
3	italian	spanish	1.75
4	german	swedish	2.5
5	swedish	german	2.5
6	dutch	swedish	2.9285714286
7	french	spanish	3.0357142857
8	english	italian	3.1071428571

The 'Examine preview data' window also shows the 'Rows of step: Sort rows 3 (8 rows)' and the 'avg_measure' column values, which are all positive, ranging from 0.8105154436 to 0.9212410605.

All avg_measure values are now positive, which contrasts with the negative values observed in the Needleman-Wunsch analysis. This is consistent with how the Jaro similarity metric works, where the similarity score ranges from 0 (no similarity) to 1 (exact match).

The scores are generally higher across the board compared to the previous methods. For example, Portuguese-Spanish now has a similarity score of 0.9212410605, which is higher than the scores seen in both the Levenshtein and Damerau-Levenshtein analyses.

The similarity scores are relatively close to each other, ranging from around 0.81 to 0.92. This suggests that, according to the Jaro metric, the language pairs being compared have a fairly high degree of similarity.

