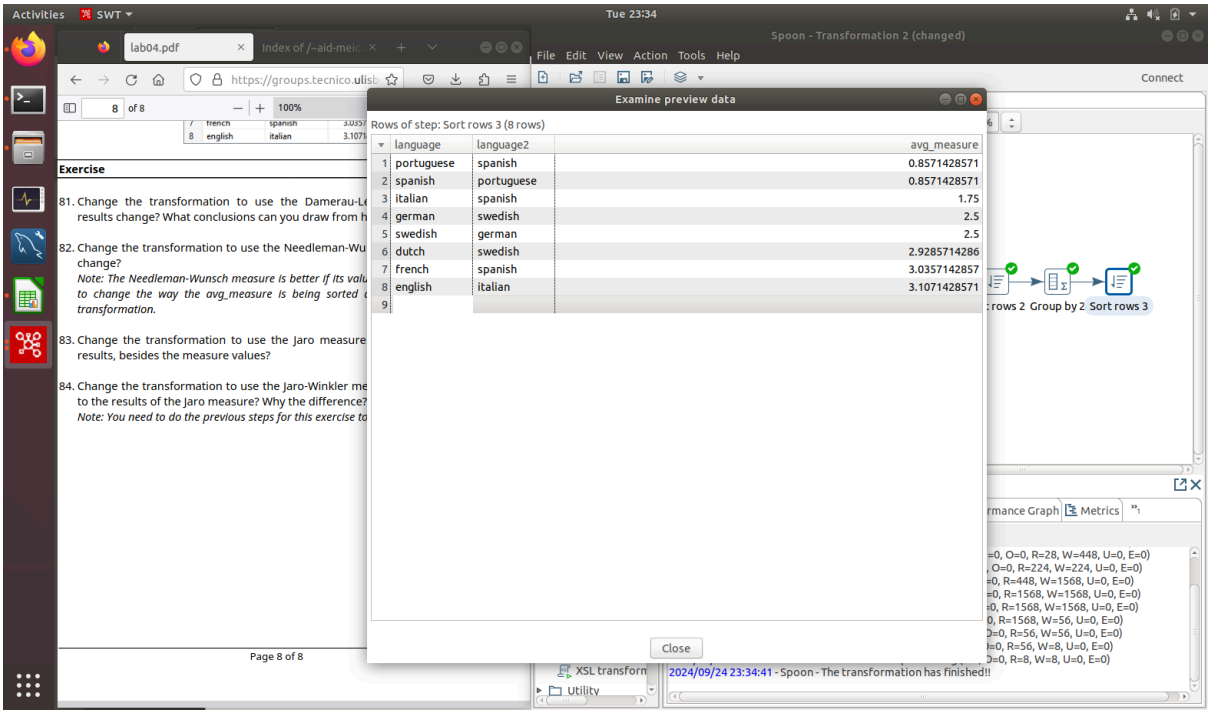
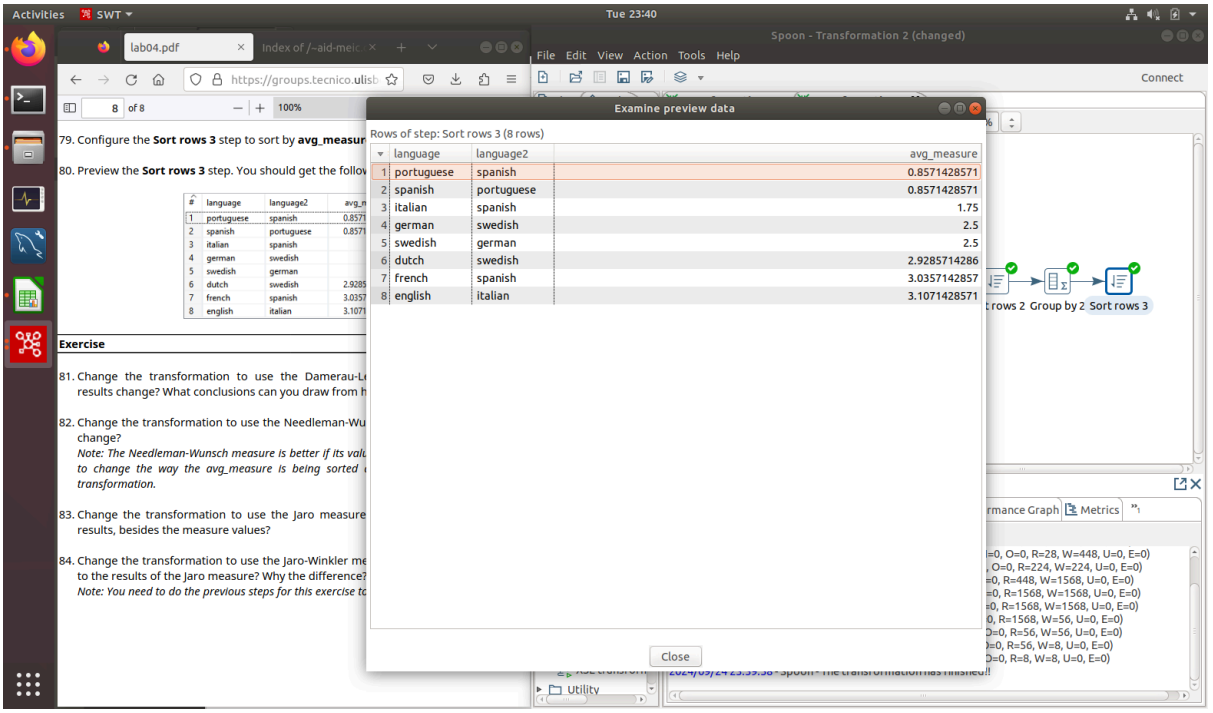
Preview - Levenshtein



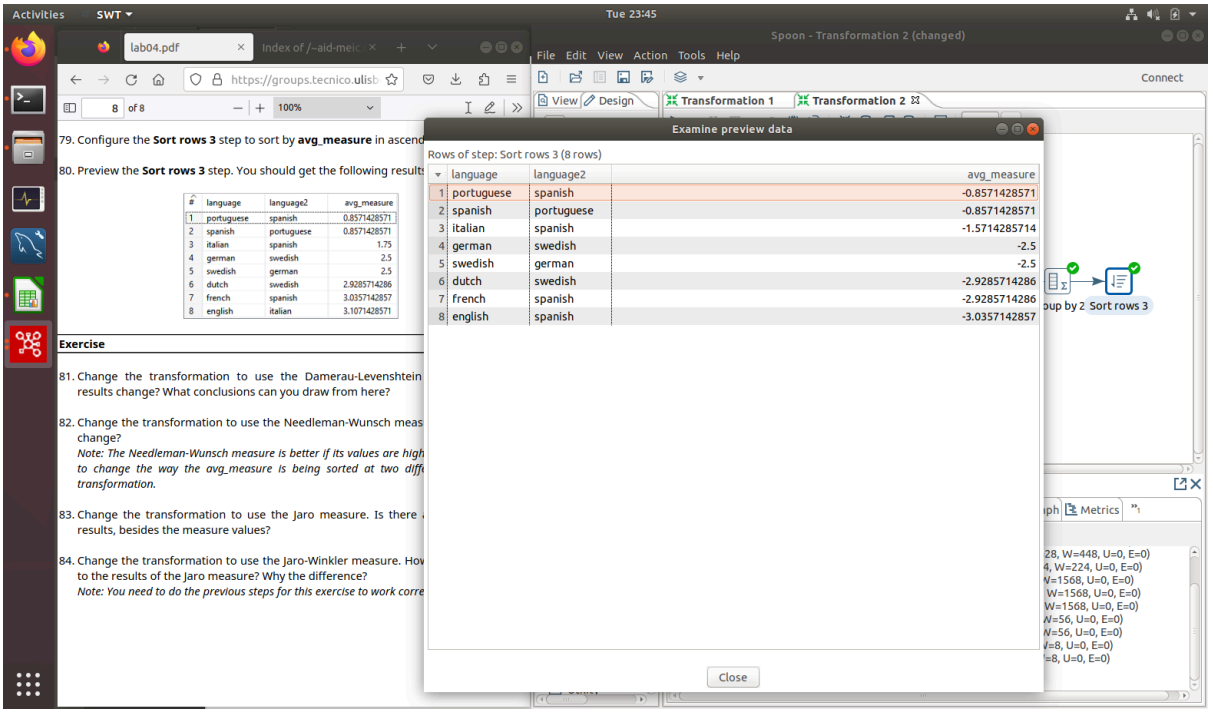
81. Change the transformation to use the Damerau-Levenshtein distance. Do the results change? What conclusions can you draw from here?



Yes, some pairs have changed

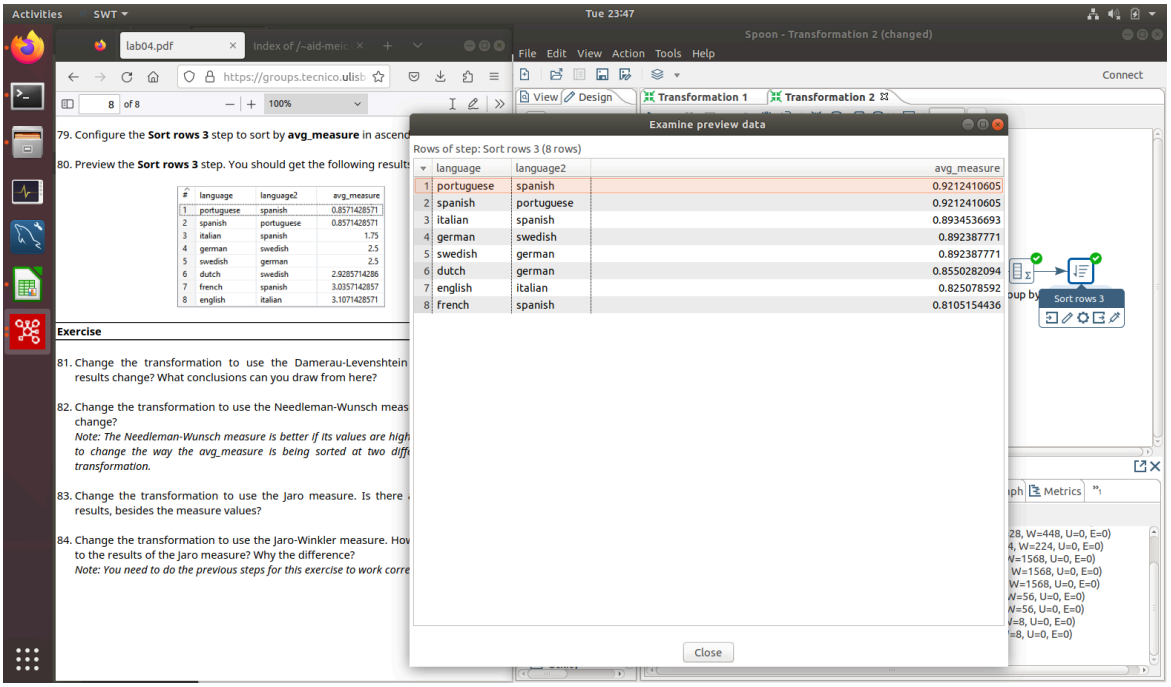
For the pair Portuguese-Spanish, the result did not change, indicating that transpositions were not significant for this pair, or the languages were similar enough that additional operations did not affect the result. For pairs like french-Spanish and Dutch-Swedish, the Demerau-Levenshtein distance increased, 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 Needleman-Wuncsch 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?

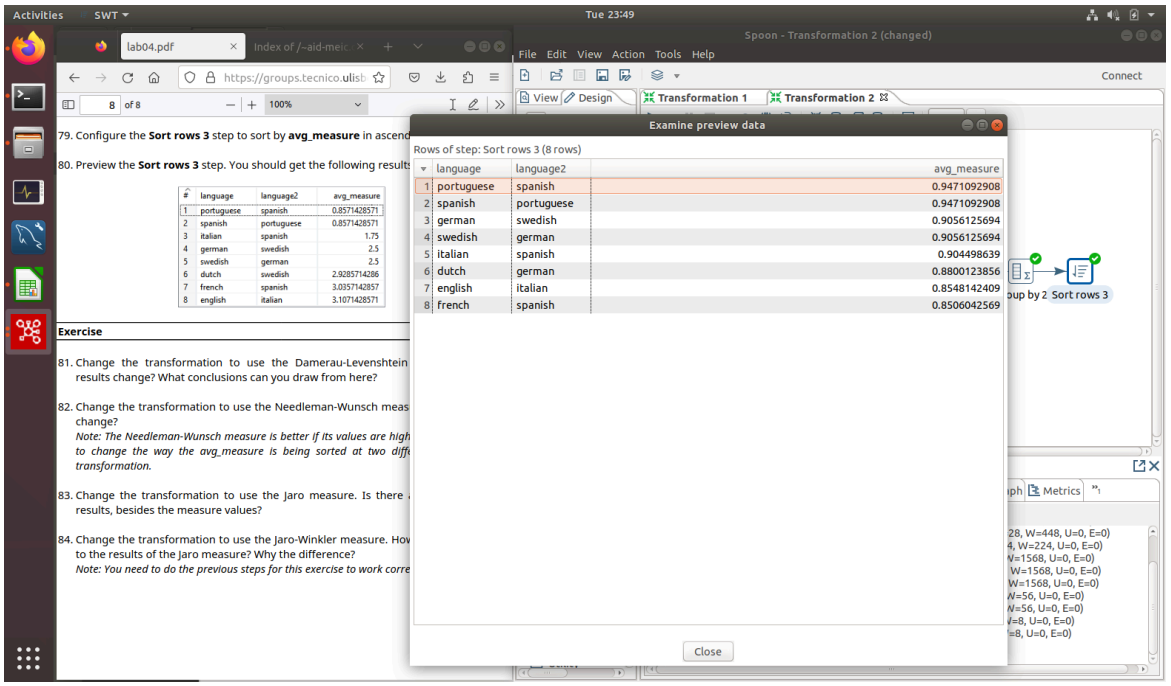


All avg_measure values are now positive, which contrasts with the negative values observed in the Needleman-Wunsch analysis. This is consistent with how the Jaro similarity metric works, where the similarity score ranges from 0 (no similarity) to 1 (exact match).

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.

84. Change the transformation to use the Jaro-Winkler measure. How do they compare to the results of the Jaro measure? Why the difference? Note: You need to do the previous steps for this exercise to work correctly



The scores in the Jaro-Winkler measure are slightly higher than those in the Jaro measure. For example, Portuguese-Spanish increased from 0.9212410605 (Jaro) to 0.9471092908 (Jaro-Winkler). This increase is consistent across all language pairs, with all scores being somewhat higher in the Jaro-Winkler measure. The relative ordering of language pairs by similarity is mostly preserved between the Jaro and Jaro-Winkler measures. The language pairs that were most similar according to Jaro are still the most similar according to Jaro-Winkler, but with a higher score.