Report WALS

In this assignment I chose to do 3 regular tasks and I used 2 files from Wals: languages.csv (contains languages and their genus) and values.csv (contains features and their values).

I have implemented the Hamming similarity – measuring it by dividing the number of common equal features by the number of features in both languages. I played with it, and it shows plausible results, for example the closest language for Ukranian is Slovene, for Hungarian – Finnish, for Italian – Portuguese, for French – Spanish, etc. Anyway, sometimes I was a bit surprised by results, for example, for English it is Russian, which looks a bit weird at the first glance.

For the second task, it was the same: I have calculated the similarity pairwise for all languages in a genus and for each language I've computed the sum of similarity scores. So, the language with the highest sum of scores is the centroid of the genus. For the weirdest language in a genus, it works the other way round - the language with the lowest sum of scores is the weirdest one.

From code point of view, there are 3 files:

- data_preparation.py - preprocessing of the data from languages.csv and values.csv.
- task1.py - implementation of the first task, should be run with command line parameter --language
- task2&3.py - implementation of the second and third tasks, should be run with command line parameter –genus