HW2

Database: CogNet
Languages: English, Polish, Czech, French, German, Norwegian, Russian, Ukrainian, Latin, Norwegian, Italian - Indo-European languages
Scoring function: average of bigram probabilities and Levenstein distance.
Original CogNet database were parsed line by line to select entries containing only selected languages. Collected data lines were transformed into a table of concept rows (12,000+) and language columns. Vocabulary for all languages was transliterated into Latin characters and cleared from numbers. Then bigram probabilities of 885 unique character pairs were counted for each language. Then, the mean of the absolute difference between language pairs was computed and normalized into a range from 0 to 100. Moreover, the Levenstein distance was computed for language pairs as a normalized mean of all nonempty entries (1,700 - 11,300) for separate words. The total score is an average of bigram and Levenstein scores that show the level of difference between the two languages in a [0, 100] range.
Results: the most similar are Russian and Ukrainian, the most different are Latin and Ukrainian. The German language has the largest amount of common features to all other languages, as well as Czech and English, while the most unique languages are Latin, Ukrainian, and Norwegian. There are a few similarity clusters: German, French, Italian, and English; plus Slavic: Ukrainian, Russian, Czech, and Polish.