Project 9: Neural Networks

Paul Beggs
GitHub Link

November 19, 2025

1 Markov Chain Analysis

Through testing different sentences, the Markov Chain was able to consistently identify the languages that it was trained on. For the other languages, it classified Somali and Yoruba as English, and Romanian and Portuguese as French. These choices make sense, as the Romanian, Portuguese, and French languages are all Romance languages. That is, they share bigrams like 'qu', 'ti', 'ca', etc. The elephant in the room is that Spanish is, of course, also a Romance language, so the fact a language like Portuguese does not show any similarity is confusing. This disparity is probably due to the common character 'ç' that both Portuguese and French share, but Spanish does not. The same explanation could also be applied between Romanian and French: both share the letter 'î', of which, Spanish does not. For Somali and Yoruba, I'm not entirely sure why they are correlated with English, but I assume it has to do with the lack of letters with accents for Somali, and a lack of common accents with Yoruba.