Task 1.1. Import datasets into Jupyter Environment:

See jupyter notebook file.

Task 2.1. Constructing high dimensional centroids:

What will be the size of the n-gram input vector in conventional (local) representation?

The size of the input vector in conventional (local) representation is equal to the amount of possible N-GRAMS we can create with our alphabet, in this case it would be 27 to the power of 3, which is equal to: 19683

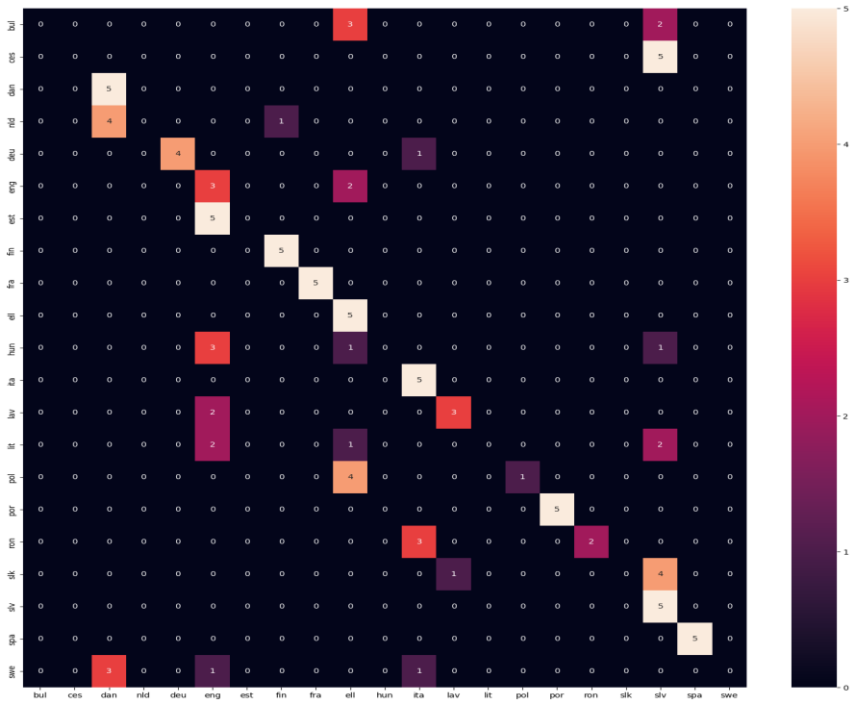
Identify difficulties of working with conventional representations of n-grams in the machine learning context.

A main difficulty posed by working with conventional representations of N-GRAMS comes from the massive dimensionalities we work with, making it very computationally expensive and also due to a very high amount of extra parameters it makes it harder to generalize to unseen data and rather overfits to the training data, specially when there are few samples to train with.

Task 2.2. Classification using centroids:

Hd = 100:

The accuracy of the language detection algorithm using fixed size hd encoding of N_GRAMS with dimension 100, is 60.952380952380956%, while the F1-Score is 0.5323301037586752



Hd = 1000:

The accuracy of the language detection algorithm using fixed size hd encoding of N_GRAMS with dimension 1000, is 70.47619047619048%, while the F1-Score is 0.6402045666751549