Lab 2X report

Ludvig Noring ludno249
Michael Sörsäter micso554
Victor Tranell victr593

Implementation

The character map is stored in a dictionary.

A character based 2-gram model with Witten-Bell smoothing is implemented with the character map as the vocabulary.

The model trains on the training file and creates dictionaries for unigrams and bigrams.

When predicting the test input, all possible characters based on the input token are tested and the one with the highest probability is chosen.

Result

The pronunciation “yi” contains 484 different characters.

For the first line in the input, the ten first probabilities (%) are:

1	2	3	4	5	6	7	8	9	10
0.1	0.2	5.7	0.1	21.3	4.0	1.2	63.0	0.2	64.5

The total accuracy for the model is: 81.69%