Jairo Nicolás Gómez

Juan Pablo Martinez

20N Dataset

In the reading of the files, we tokenized sentences whenever there was a point. In the preprocess function, we applied the rules mentioned in the instructions and return a list of sentences from all the corpus.

Then divide this sentence into training and test sets. Then construct the N-Gram models applying the Laplace smoothing applying the corresponding formula for each one of the models, returning txt files with the unigram, bigram or trigram and its corresponding probability.

To calculate perplexity we took the unigram, bigram and trigram respectively from the test corpus, and then see if there were in the training corpus in order to get their probabilities and apply the perplexity formula. We noted that the best perplexity was the unigram model, because the others are way to high.

BAC Dataset

The reading of the files and preprocessing was very similar. As this dataset had a larger number of sentences, the N-gram models used where the same as in 20N, but they took longer to execute specially the bigrams and trigrams.

Drive with the result files

Tarea 2 NLP