NLP PROJECT 1 REPORT

Project is coded with Python language on Google Colab environment. We uploaded our training data to Google Drive. With json and os packages we have read datas from there and stored json objects in a list.

Then we parsed strings in those json objects and stored them in a string list. We removed stop words and expressions like ".", ",", "/" etc.

Later, we split data into training and testing sets. %80 for training and %20 for testing. For NLP tasks we used nltk library. It is very useful. We formed 2-grams with padded_everygram_pipeline() function which makes data ready to train. We constructed our models using Laplace class. Our models are word based and character based. They apply laplace smoothing approach then we trained our data.

For testing, we created a test folder on Google Drive. Every json file in that folder is used to test the model. For testing we calculate perplexity of the model.