Artificial Intelligence for Software Engineering

Assignment#1

Recommending code tokens via N-gram models

Instructor: Antonio Mastropaolo, PhD

Data Collection: -

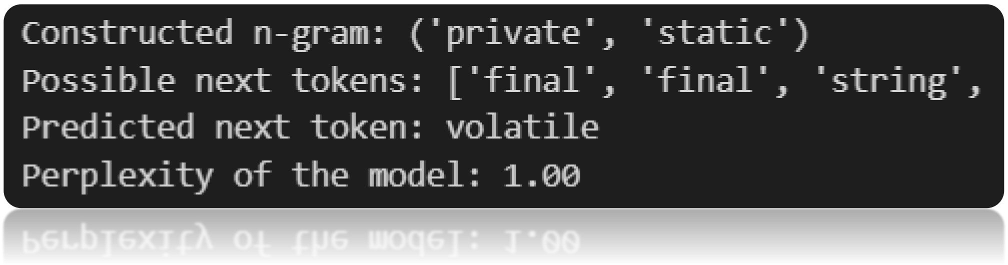
For this project, we were asked to collect at least 25k Java methods. I used SEART to collect the Java methods. I've collected overall 2k repositories from that website and I used only 20 repositories to collect the data. The training dataset size was more than 400 MB and for the testing dataset the size was 18 MB.

Data Pre-processing: -

Cleaned data is very crucial for the model training. As the data was in the java file, I used some regular expression to extract only the java methods. After pre-processing I saved the data to a text file.

Model Training Methodology: -

First of all, I created tokens. The total number of tokens was 380731 and the number of unique tokens was 69141. Then, I build a n-gram model to predict the next word based on inputted word.



```
Constructed n-gram: ('private', 'static')
Possible next tokens: ['final', 'final', 'string',
Predicted next token: volatile
Perplexity of the model: 1.00
```

Code: -

The code is available in the following address: https://github.com/robiul-islam-rubel/ngram