README

How to run this program:

To run this Java program, you should first make sure that all 4 files (3 datasets and 1 test document) is in the “assignment1-data” folder.
All codes are in the “src” folder, you can run the main function in the Main class.
The perplexity and prediction of each model would be printed in the console and the program will also generate 4 files in the “result” folder. “English.txt” is the count of English dataset. “Spanish.txt” is the count of Spanish dataset. “German.txt” is the count of German dataset. “ThProb.txt” is the probabilities of each “th*” word calculated by 4 trigram models.
The version of Java is JDK1.11

Resources:

[1] Bill MacCartney, NLP Lunch Tutorial: Smoothing,
https://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf
[2] Kneser-Ney smoothing explained,
http://www.foldl.me/2014/kneser-ney-smoothing/

Person I discussed with:

I discussed this project with two classmates, we just talked about how to understand the task and finished it independently.

Unsolved problems:

- The probability sum of the backoff model is not 1. I am still learning how to use good tuning to set the parameters.
- The underflow problem, I can use log to represent probability but in this way I cannot calculate the perplexity.