Paul Mitchell, Stephanie Omondiale

Gen AI Homework 1 N-Gram Model

Dataset Creation, Model Training, and Evaluation

The dataset was constructed using Seart GHS to identify Java repositories, filtered based on key characteristics such as commit count (≥100) and lines of code (30,000–50,000) while excluding forks. PyDriller was then used to extract Java methods from these repositories, with a cap of 100 methods per repository to maintain dataset balance. To ensure data quality, extracted methods were validated, removing duplicates, trivial methods, and non-ASCII entries. Initial filtering was overly aggressive, reducing the dataset to 2,725 methods, prompting refinements to preserve meaningful functions. The final dataset contained 3,224 cleaned methods, which were then split into 60% training, 20% evaluation, and 20% testing. Any test tokens absent from the training vocabulary were replaced with <UNK> to handle out-of-vocabulary cases.

The N-Gram Model was trained using 3-Gram, 5-Gram, and 7-Gram models. Initial training with Laplace smoothing led to unstable perplexity scores (inf), prompting a transition to Kneser-Ney smoothing, which improved generalization. The process involved troubleshooting multiple issues, including dataset size discrepancies, incorrect tokenization, and handling generator objects in perplexity calculations. We used the assistance of Chatgpt in certain troubleshooting and method implementations. Due to persistent issues with perplexity-based evaluation, model selection shifted to confidence-based scoring. The 7-Gram model outperformed others, achieving a median prediction confidence of 0.90015 and a mean of 0.5753. The model was then tested on the professor's dataset to ensure consistency across different data sources.

Model Evaluation & Results Analysis

Evaluation of the 7-Gram Model included a probability distribution histogram and a probability spread boxplot. The histogram showed that while many predictions had high confidence (close to 1.0), a significant number of tokens were assigned a probability of 0.0, likely due to out-of-vocabulary tokens or low contextual relevance. The probability spread between 0.1 and 0.9 was relatively sparse, indicating that intermediate-confidence predictions were less frequent.

The boxplot reinforced these findings, with a median probability of 0.90015, suggesting high confidence in at least half of the predictions. The mean probability of 0.5753 reflected overall moderate confidence, while the presence of 0.0 probabilities indicated some failures in token recognition. The interquartile range was skewed toward higher probabilities, showing that the model effectively captured patterns within the data. However, frequent 0.0 probability predictions suggested challenges with rare sequences and out-of-vocabulary tokens. Potential improvements could include expanding the training dataset, implementing backoff smoothing for unseen n-grams, and analyzing mispredicted tokens to address patterns in function names,

variable names, or specific coding styles. Despite these limitations, the 7-Gram Model demonstrated strong predictive performance and was the best-performing model in this study.