Interpolation Model for Language Modeling

1. Introduction

This report details the implementation of an interpolation model for language modeling. The goal is to improve language model performance by combining different n-gram probabilities using optimized interpolation weights. We also analyze execution time and efficiency improvements.

2. Problem Description

Given a dataset, we construct an n-gram language model and employ linear interpolation to combine probabilities from unigram, bigram, and trigram models. The interpolation technique optimizes the weight values (λ) to minimize perplexity on a validation set.

3. Implementation Details

3.1 Dataset Processing

- Tokenized text into unigrams, bigrams, and trigrams.
- Constructed frequency distributions to estimate probabilities:
    - **Unigram Model**
    - **Bigram Model**
    - **Trigram Model**

3.2 Interpolation & Perplexity Calculation

- Used linear interpolation to compute smoothed probabilities.

- Evaluated performance using perplexity.

4. Results

4.1 Perplexity Scores

Model	**Average Perplexity**
Unigram Model	85,023.8164
Bigram Model	1,643,368,727.4968
Trigram Model	6,495,576,203.9299
Interpolation Model	35,222.4560

The interpolation model significantly reduces the perplexity compared to individual n-gram models, demonstrating its effectiveness in improving language modeling performance.

5. Execution Time Analysis

We measured execution time for each major component:

Tasks	**Time Taken (seconds)**
Tokenization and Preprocessing	43.12
N-Gram Frequency Calculation	51.44
Perplexity Calculation	0.40
Interpolation Optimization	10.17
Total Execution Time	105.13

6. Efficiency & Optimization Strategies

- **Text Normalization:** All text was converted to lowercase for consistency.
- **Preprocessed Data Storage:** Stored preprocessed text to avoid redundant computations.
- **Efficient N-Gram Storage:** Stored different n-grams along with their probabilities for improved visualization (optional but beneficial for analysis).

6.1 Different Approaches Used

- **spaCy vs. NLTK:** Initially, spaCy was tested instead of NLTK, resulting in lower perplexity. However, preprocessing with spaCy was 50 times slower. Due to this significant time overhead, NLTK was preferred.
- **Stemming vs. Lemmatization:** Both stemming and lemmatization were evaluated. Lemmatization was chosen since stemming did not provide a notable improvement in processing speed.
- **Hyperparameter Tuning:** Instead of using a trial-and-error approach, **GridSearchCV** was employed to fine-tune the hyperparameter efficiently.

7. Conclusion

The interpolation model effectively combines different n-gram probabilities to improve language modeling. The results show that the interpolation model significantly reduces perplexity compared to individual unigram, bigram, and trigram models. Additionally, optimization strategies such as efficient text processing, hyperparameter tuning, and strategic algorithm selection contributed to improved efficiency and execution time.