# Language Model Analysis

Name: Faiza Alee

Section: 05

Github Account: <https://github.com/faiza-alee>

## Corpus Reference

The corpus used for this project is 'austen-emma.txt' from the NLTK Gutenberg collection. We limited the text to the first 1,000 sentences to make predictions using unigram, bigram, and trigram language models. This corpus, derived from Jane Austen's Emma, provides a structured language base for modeling English text patterns.

## Analysis of Predicted Words from Unigram, Bigram, and Trigram Models

1. Unigram Model:

- The unigram model generates predictions by selecting words based on their frequency in the corpus, without considering any context from preceding words.  
 - As a result, high-frequency words like 'the,' 'of,' and 'and' often appear in predictions, regardless of the specific input sequence.  
 - This leads to predictions that may not align with the sequence's intended meaning, making the unigram model the least context-aware of the three models.

2. Bigram Model:

- The bigram model improves upon the unigram approach by considering the last word of the input sequence. This introduces a basic level of context, enabling the model to produce predictions that are somewhat relevant to the sequence.  
 - For example, if the input sequence is 'machine,' the model may predict 'learning' as the next word if 'machine learning' frequently occurs in the corpus.  
 - However, the model’s accuracy depends on the availability of the specific bigram in the corpus. When such bigrams are absent, it falls back to unigram predictions, reducing contextual accuracy.

3. Trigram Model:

- The trigram model provides the highest level of context by taking the last two words of the sequence into account, which often leads to more contextually accurate predictions.  
 - For instance, with an input like 'artificial intelligence,' the model might predict 'applications' or 'techniques' if those trigrams appear in the corpus.  
 - Despite this improvement, the model may still revert to bigram or unigram predictions when specific trigrams are unavailable, potentially impacting the quality of the prediction.

## Summary

- Unigram Model: Provides predictions without any context, leading to generalized results that are often irrelevant to the sequence.  
- Bigram Model: Adds minimal context by considering the last word, yielding moderate relevance but limited adaptability to longer sequences.  
- Trigram Model: Offers the most accurate and contextually appropriate predictions by leveraging the previous two words, though it still falls back to simpler models if specific trigrams are missing.

In conclusion, while the trigram model generally offers the best predictive quality among the three, its effectiveness depends on the corpus size and diversity. For applications requiring high accuracy in next-word predictions, higher-order models such as trigrams are preferable, assuming a comprehensive corpus is available.