PA 3 - Shakespeare Text generation using N-grams

Due Mar 3 11:59 PM

Overview

In this assignment you will implement n-gram models and train them to generate text using Shakespeare's past works. All graded tasks are marked with TODO comments in PA3_starter.ipynb

Learning Goals

- Understand text preprocessing and tokenization for language modeling
- Implement a Unigram model (frequency-based word generation)
- Implement a Bigram model (next-word prediction with one word of context)
- Generalize to an N-gram model with Laplace smoothing
- Evaluate language models using perplexity
- Generate Shakespeare-like text from your trained models

Allowed Libraries

- Only the already imported packages are allowed.

Deliverables

- Submit PA3_starter.ipynb with all TODOs completed.
- Do not rename the notebook (the autograder expects this filename).

Point Distribution (100 points)

Part 1: Text Preprocessing and Tokenization (10 points)

Part 2: Unigram Language Model (40 points)

Part 3: Bigram Language Model (50 points)

Part 4 (Extra Credit): N-gram Language Model (20 points)

Notes

- Some functions are already provided.

Submission Checklist

- All TODOs are implemented.
- Notebook runs end-to-end without errors.
- You did not change function signatures.

Tips

- Submit each part as you go, as previous functions are used in later parts of the assignment.