DS 5983: Large Language Models (Summer 2025) Roi Yehoshua

Student name: (Due) May 22, 2025

PA1: N-Gram Models

In this task you are going to build a simple language model that predicts the next word in a sentence based on n-gram statistics.

1. Data collection and preprocessing:
   (a) Collect a dataset of text. You can use the text corpora available in the `nltk` library, such as `reuters` or `gutenberg`.
   (b) Clean and preprocess the data (e.g., remove special characters, convert to lowercase).
   (c) Tokenize the text into words.
2. Model implementation:
   (a) Create n-grams from the tokenized text and calculate their frequencies in the dataset.
   (b) Write a function to calculate the probability of a word following a given $(n-1)$-gram.
   (c) Write a function to predict the next word given a sequence of words based on these probabilities.
   (d) Write a function to generate a sentence of a specified length given a prefix of $(n-1)$ (or less) words.
   (e) Implement smoothing techniques (like Laplace smoothing) to handle the issues of zero probabilities for unseen n-grams.
3. Testing and evaluation:
   (a) Test the model by inputting various prefixes (with at most $n-1$ words) and evaluating its ability to generate text.
   (b) Compute the perplexity of the model on a test set that was not used during training.
   (c) Compare the performance of models with different values of n (e.g., bigrams vs. trigrams vs. 4-grams). Discuss which model achieves lower perplexity and provide insights into why certain n values might be more effective in various contexts.
4. Write a short report discussing your results and any challenges encountered during the implementation.
5. *Bonus* (up to 5 points): Create a simple user interface (e.g., using Streamlit) where users can enter some prefix and get a completion of words up to a specified length.

Submission instructions:

- Submit your source files (.ipynb or .py) and the PDF file with the report to Canvas separately (do not compress them into one zip).
- You are welcome to discuss the assignment problems with other students in class, but you must write up the solution yourself, *and* indicate who you discussed with (if any).
- While tools like ChatGPT can be used for brainstorming ideas or clarifying concepts, they must not be used to generate substantial portions of code, analysis, or written sections of any assignment. If you use AI assistance, you must explicitly document its usage in your submission. Failure to disclose AI-generated content will be considered an academic integrity violation. All submitted work should reflect your own understanding and effort.
- Assignments may be handed in up to one day late (24-hour period), penalized by 10%. Submissions later than this will not be accepted. Contact the teaching staff if there are *extenuating* circumstances.