

AIM 5011-1 Natural Language Processing (3 credits)

Assignment 4

Adam Faulkner

11/09/2025

E-mail: adam.faulkner@yu.edu

Office:

Class Hours: Tuesday 5:30-7:30pm

Classroom: 205 Lexington Avenue Rm. 700

1 Instructions

Complete sections 2 and 3 *by hand* and show all of your work in a separate scratch sheet or in the question itself. Use a calculator for complex calculations such as taking the natural log or softmax. Scan or take a picture of the completed sections and upload it to Canvas.

Section 4 describes a coding assignment. Upload this as a Jupyter notebook directly to Canvas.

Due: November 25

2 Sampling

For the next three problems, you'll be using the following dataset:

| Token | Probability |
|-------|-------------|
| a | 0.32 |
| b | 0.22 |
| c | 0.45 |

2.1 Greedy sampling

Which of the three tokens would be generated next using greedy sampling?

2.2 Temperature sampling

Temperature sampling involves three steps:

1. Divide each logit by the chosen “temperature” value. For the purposes of this exercise, use the logit function to first convert the probabilities to logits: $\text{logit}(p) = \log(p/(1-p))$ where p is the probability and \log is the natural log. Then round to two decimal places to keep things manageable.
2. Apply the softmax function to the temperature-scaled logits to convert them into a probability distribution.
3. Randomly select the next token based on the re-weighted distribution.

If we set the temperature to 0.2, what is the new token distribution?

2.3 Top-k sampling

Top-k sampling involves taking the top-k rank-ordered probabilities, and then applying *softmax* to those k probabilities to create a new probability distribution. A token is then selected randomly from this new distribution.

What is the new distribution of our dataset if we set $k=2$?

3 Byte-pair encoding

HuggingFace provides [a walkthrough](#) of the BPE algorithm using

- the corpus ["hug", "pug", "pun", "bun", "hugs"]
- and the base vocabulary ["b", "g", "h", "n", "p", "s", "u"]

We'll also be using this corpus and vocabulary, but our frequency distribution will be

- ("hug", 5), ("pug", 15), ("pun", 20), ("bun", 2), ("hugs", 4)

Calculate the first three merge rules the learned by the BPE for this corpus, vocabulary, and new frequency distribution.

4 Programming Assignment

In this programming assignment, you'll be asked to complete the notebook

`Assignment_4_programming_assignment_RAG.ipynb` The goal is to create a RAG-based recipe generation system using Alibaba's KingNish/Qwen2.5-0.5b-Test-ft LLM, a small but high-accuracy LLM that can be run on a CPU.

For all lines marked `### WRITE YOUR CODE HERE ###` insert your own code and make sure that the cell runs. Additionally, answer all questions at the bottom of the notebook. Upload the completed Jupyter notebook to the Assignments section in Canvas.