CSE 102 Spring 2025 – Computer Programming Assignment 9

Due on May 7, 2025 at 23:59

In this assignment, you will implement a system that reads a text file containing matched question-answer pairs, parses the content, tokenizes the sentences, encodes the sentences into numerical vectors, and finally writes the resulting embeddings into a structured output file. Additionally, the output file will contain a metadata section, which summarizes information about the data such as the maximum embedding length across all sentences, the dimension of word embeddings, and the total number of question-answer pairs.

The input will be a text file named dataset.txt, formatted as follows:

```
Question: How many months are there in a year?
Answer: 12
---
Question: What is Earth's satellite?
Answer: Moon
---
…
```

Students are required to implement the following steps:

1. Read the input file line by line and parse it into question and answer pairs. The separator between entries is a line containing three hyphens (- - -).
2. For each question and answer, perform tokenization by splitting the text into words. You can assume that all the words are separated by a single whitespace.
3. Encode each token into a simple numerical representation. You must implement a character-based one-hot encoder, where each character in the sentence is represented as a one-hot vector over the set of all characters found in the dataset. To ensure that all word embeddings have the same size, shorter sentences must be left-padded with zero vectors so that all embeddings match the length of the longest word in the dataset.
4. Generate a sentence embedding for each question and answer by concatenating the one-hot vectors of the words in the sentence while exporting to the file.
5. Generate an output text file named embeddings.txt, which should contain two sections:

I. A metadata section at the beginning, providing information such as:
   - The maximum length of a sentence embedding across all questions and answers
   - The dimension of the word embeddings
   - The number of question-answer pairs

II. A data section, listing the paired embeddings of each question and its corresponding answer which are separated by three hypens as in the input file:

```
…
Metadata part
…
#####
Question: 001101010111010100….
Answer: 11000101010000111001101 0….
---
Question: 000101010010001010…
Answer: 11100111…
---
…
```

The output file must be organized clearly so that it can be easily read by another program later. You must structure the code modularly, using functions for parsing, tokenizing, encoding, and writing the file. Additional user-defined function/s can be used. Note that, the embedding data must be stored in a multi-dimensional array.

IMPORTANT NOTES:

- Submit your homework as a zip file named as your student id (StudentID.zip) and this file should include:
  - YourStudentID.c file
  - A reports containing the screenshots of running code and generated outputs.
- Programs with compilation errors will get 0.
- The output format must be as given, do not change it.
- Compile your work with given command “gcc --ansi your_program.c -o your_program”.
- For any questions and problems use Teams page of the course.