CS 585 – Fall 2023 – Homework 1 (100 total points)

Due Wednesday September 13, 11:59pm

GOALS

- To apply the Information Theory concepts
- To begin to work with text data in python, and apply an open-source NLP package
- To gain exposure to an openly available NLP research dataset

DATA

For this homework you will work with data from one task in the GLUE benchmark, which was created to compare language models across various NLP tasks. This dataset was chosen because it provides a variety of English-language texts in a clean format; we are not going to complete any of the GLUE tasks in this HW. But, if you would like to learn more about GLUE, you can refer to the introductory paper.

For this homework you will need to download and unzip these datasets from the Tasks Download page:

- **SST** - Stanford Sentiment Treebank - Movie reviews extracted from Rotten Tomatoes and labeled by hand for sentiment
- **QNLI** - Question-Answering Natural Language Inference – Sentences from Wikipedia, matched with questions written by human annotators.

TOOLS

In this homework you are asked to apply the python Natural Language Toolkit, a commonly used open-source python package. You will use NLTK to separate a string into *tokens*, where a token is a sequence of one or more characters. Note that a token may contain letters, digits and/or punctuation (e.g. “1.25”)

You should use Python to complete this homework assignment. Other programming languages will not be accepted.

WHAT TO SUBMIT

Please upload or submit the following in Blackboard:

- For Problems 1-5, please upload to Blackboard:
  - One Jupyter notebook (.ipynb file) with cell output, showing your work for both datasets.
  - A PDF copy of the exact same notebook (same code and same output)
- For Problems 6 & 7: Enter your answers in Blackboard with your HW submission

PROBLEM 1 – Representing English Text (5 pts)

- Read in these two GLUE datasets (see section "DATA" above). Also convert alphabetical characters to lower case:

Dataset	Use file	Notes
SST	SST/train.tsv	Use column "sentence" (Ignore column "label")
QNLI	QNLI/dev.tsv	Use column "sentence" (ignore columns "question" and "label")

- Convert each dataset into a single list of tokens by applying the function "word_tokenize()" in the NLTK :: nltk.tokenize package. We will use these lists represent two **distributions** of English text.
- To show you have finished this step, print the first 10 tokens from each dataset.

PROBLEM 2 – Word probability (10pts)

- Write a **python function** that creates a probability distribution from a list of tokens. This function should return a dictionary that maps a token to a probability (I.e., maps a string to a floating-point value)
- Apply your function to the list created in Problem 1 to create **SST** and **QNLI** distributions.
- Show that both probability distributions sum to 1, allowing for some small numerical rounding error. Or, if they do not, add a comment in your notebook to explain why.

PROBLEM 3 – Entropy (20pts)

- Write a **python function** that computes the **entropy** of a random variable, input as a probability distribution.
- Use this function to compute the word-level entropy of **SST** and **QNLI**, using the distributions you created in Problem 2. Show results in your notebook.

PROBLEM 4 – KL Divergence (20pts)

- Write a **python function** to compute the KL divergence between two probability distributions.
- Apply this function to the distributions you created in Problem 2 to show that **KL divergence is not symmetric**. [This is also question 2.12 of M&S, p79].

PROBLEM 5 – Entropy Rate (20 pts)

- Write a **python function** that computes the per-word **entropy rate** of a message relative to a specific probability distribution.
- Find a recent movie review online (any website) and compute the entropy rates of this movie review using the distributions you created for both **SST** and **QNLI** datasets. Show results in your notebook.

PROBLEM 6 – Observed Entropy Rate (10pts – Answer in Blackboard)

Refer to your results from Problem 5. Which distribution gives you the lowest entropy rate for your movie review? Does this match what you expected? Why or why not?

PROBLEM 7 – Zero probabilities (10 pts – Answer in Blackboard)

Problem 5 required that you handle “zero probabilities” cases, where a token occurred in one dataset but not the other. How did you handle these tokens? (Hint: Dropping the word from both probability distributions is not an ideal solution).

CODE ORGANIZATION AND READABILITY (5 pts)

The receive full credit, please ensure that:

- Your notebook includes cell output, and does NOT contain error messages
- It is easy to match a problem with its code solution in your notebook via markdown or comments (e.g., “Problem 2”)
- You have re-run your notebook before submitting, and cells are numbered sequentially starting at [1].

GETTING HELP

- If you are new to Python, you may use the “HW1_GettingStarted.ipynb” notebook that is posted with these directions. Please note that a starter notebook will **NOT** be posted for all class homework assignments.
- Please post questions about homework directions in the Blackboard discussion group.
- For questions about python setup and runtime errors, please reach out to TAs during posted office hours (TA office hours will start the week of Tuesday, September 5; see Blackboard for times)