## CSCE 771: Computer Processing of Natural Languages

## Prof. Biplav Srivastava, Fall 2022

## Quiz 3 / Instructions

* This is a programming quiz. Code has to be submitted in a directory of your GitHub called “Quiz3” with sub-dir for code, data and doc. Code will have your source code, data will have any input or output generated, and doc will have a .pdf of this file (called Quiz3-CSCE771-answers.pdf) along with any answers
* Complete quiz by 9:00 am on Monday, Nov 21, 2022 by sending an email to [biplav.s@sc.edu](mailto:biplav.s@sc.edu) confirming completing the quiz and attaching your Quiz3-CSCE771-answers.pdf.
* Total points = 70 + 30 = 100
* Obtained =

Student Name: Bharath Chandra Muppasani

The objective of the Quiz is to learn usage of large language models on NLP tasks and their superiority, if any, over traditional methods. Then, to also solve a practical problem.

**Dataset**: South Carolina universities have “Annual Security and Fire Safety Reports”. For 2022, for University of South Carolina, it is publicly available as well as conveniently cached at: <https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/common-data/2022-uosc-securityandfirelreport-1001bcleryreport.pdf>

**Goals:** Your task is to use NLP techniques to provide specific information to prospective new students and their parents who do not have the background or time to read the document.

**NLP Tasks:** Entity extraction, sentiment mining, events, topic analysis and text summarization

**Activity:**

- Choose any 3 NLP task and corresponding goodness metrics. (You may use additional task for extra credits but mark it so in your report/ code)

- Use any LLM available from Huggingface like BERT, DistilBERT. Use [1] for reference.

- Use any one traditional NLP method (i.e., non-LLM) for the NLP tasks (like extractive summarization based on TF-IDF as discussed in class).

- Now answer the questions and their parts.

Q1: Comparison of methods [20 x 3 + 10 = 70 points]

* Which method (traditional or LLM-based) does better on the three NLP tasks
  + Summarization:
    - Given a small paragraph, the traditional TF-IDF method failed to generate a concise summary, where as LLM-based method was successful in extracting a summary for the given paragraph. Thus, LLMs seemed to work better for the task of summarization.
  + Topic Modeling:
    - Traditional based methods perform better in this regard. LLM-based approach didn’t work with large text and often gave topics that are irrelevant as output.
  + Entity Recognition:
    - In terms of number of Entities recognized, NLTK gave more entities. But the quality of entities obtained from SpaCy seem better.
* What issues, if any, do you see with the LLM methods
  + Does not work well with unseen data

Q2: Based on your analysis, answer the following questions:  
[10 + 10 + 10 = 30 points]

1. Is the university safe? How did you arrive at the conclusion?
2. Are the rights of the accuser and victim same ? If not, the policies are skewed towards whom? How did you arrive at the conclusion?
3. Is it better to report a crime openly or anonymously? How did you arrive at the conclusion?

**Reference:**

[1] <https://github.com/huggingface/notebooks/blob/main/transformers_doc/quicktour.ipynb>