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 confirming completing the quiz and attaching your Quiz3-CSCE771-answers.pdf.
- Total points = 70 + 30 = 100
- Obtained =

Student Name:

Brendan Reidy

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
    - For text summarization neither perform great, this is mainly due to the fact that the input is too large for LLM-based approaches, so we summarize the sentences in batches, and then summarize the summaries.
    - For sentiment analysis, the LLM-based approach is more consistent
    - For entity extraction the non LLM-based approach performs the best

- What issues, if any, do you see with the LLM methods
    - They are much slower than non LLM based methods
    - They can often be trained for domain specific tasks and cannot generalize outside of those tasks
    - They have fixed input sizes

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

a) Is the university safe? How did you arrive at the conclusion?
    a. The university is moderately safe although there have been several fires on campus and several crimes on campus

```
The following fires occurred at a University residential
facility in the previous three years 2021 - None 2020 - None
2019 - (1) Delta Zeta House (514 Lincoln Street) - fire
resulting from a plastic cutting board catching fire while on
a stove being used to prepare meals
label: NEGATIVE, with score: 0.9976
(2) Honors Residence (1215 Blossom Street) - a mechanical fire
occured on a trash compactor located in a garbage shoot
label: NEGATIVE, with score: 0.9991

This information was found using sentiment analysis and
checking for negative scores over 0.99
```

b) Are the rights of the accuser and victim same ? If not, the policies are skewed towards whom? How did you arrive at the conclusion?

The rights of the victim and an accuser are not the same. The policies are skewed towards in favor of the accuser. This is from the TF-IDF based text summary

TF-IDF based text summary:

```
An accuser/reporter that reports a crime to a CSA does not have to
prove that they were the victim or witness of a crime.
```

c) Is it better to report a crime openly or anonymously? How did you arrive at the conclusion?
    a. It is better to report a crime in openly
    b. DLES does not generally investigate anonymous allegations of criminal activity outside of CRIMESTOPPERS and the RAVE Guardian app. Anonymous tips without further corroboration are not included in the University's Annual Security and Fire Safety Report

Reference:

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