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: Sai Teja Paladi

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

**TextRank Algorithm**



**BERT**



A ROUGE score close to one indicates strong similarity between candidate and references. As we can observe that TextRank gives better Rouge score than BERT.

It is evident that TextRank is performing better.

* What issues, if any, do you see with the LLM methods

While doing the summaraization using bert we got the summary of the document with words 30000 truncated down to 1649 words. Which is much more concise and was giving effective summary of the document. However during some sentences it missed the context beacuse we are using a pre-trained model. It was not effectively trained for our use case.

Using the summary given from PageRank I conclude that the given summary was very large to read which consisted of nearly 7626 words. The summary was effective because it doesn't depend on training data and is a method to find summary using PageRank algorithm. i. e same root words are ranked togeather to form a summary.

In conclusion BERT will perform better when we have trained data on a similar dataset and PageRank can be effective for smaller documents. Because our data was not that large Pagerank performs better than the BERT.

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

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

**Answer**:

For the above question I have parsed the pdf document and converted to text and extracted all the passages by splitting at '.' .

Dowloaded the distilbert model and caluculated the query embeddings by giving the query that we want to ask and also calculated the scores of each passages based on cosine similarity between the query embedding and passage embeddings.

1.From the top ranked extracted sentences I can say that the university is safe.



2.It seems that the rights are different. The policies are skewed towards the accuser.



3.It's better to report crime openly



**Reference:**

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