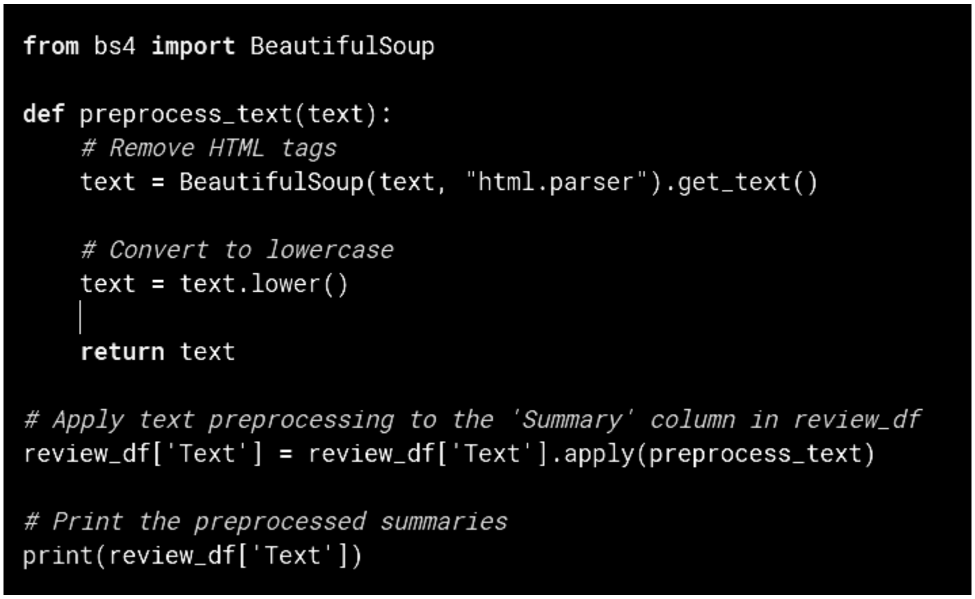
CSE508 Information Retrieval Winter 2024 Assignment-3 Report
AKASH KUMAR
MT23012
Date of Submission:-23-04-2024

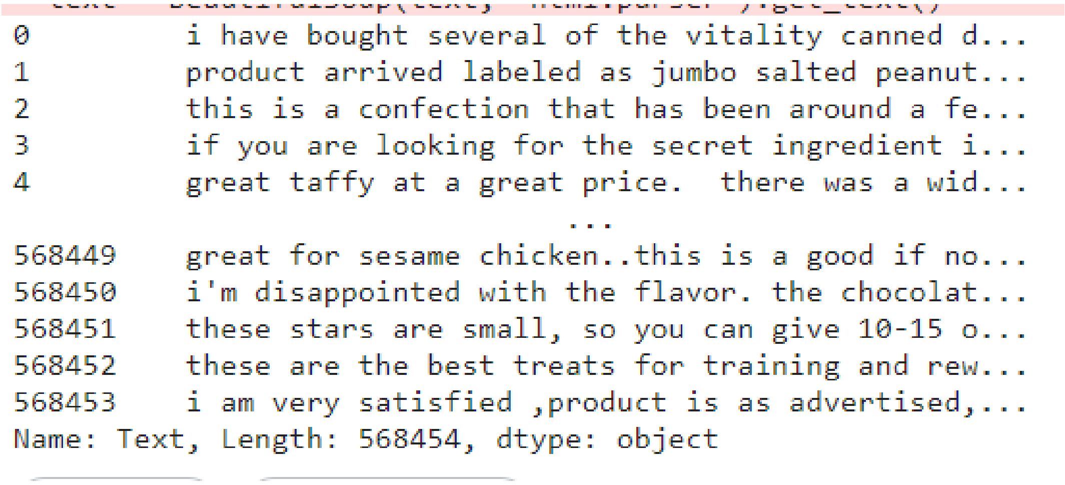
1. Data Preprocessing:

In the we preprocess the data on the two column that name is Title and summary and preprocess it. the



```
from bs4 import BeautifulSoup

def preprocess_text(text):
    # Remove HTML tags
    text = BeautifulSoup(text, "html.parser").get_text()

    # Convert to lowercase
    text = text.lower()

    return text

# Apply text preprocessing to the 'Summary' column in review_df
review_df['Text'] = review_df['Text'].apply(preprocess_text)

# Print the preprocessed summaries
print(review_df['Text'])
```

o/p



```
0         i have bought several of the vitality canned d...
1         product arrived labeled as jumbo salted peanut...
2         this is a confection that has been around a fe...
3         if you are looking for the secret ingredient i...
4         great taffy at a great price.  there was a wid...
                                ...
568449    great for sesame chicken..this is a good if no...
568450    i'm disappointed with the flavor. the chocolat...
568451    these stars are small, so you can give 10-15 o...
568452    these are the best treats for training and rew...
568453    i am very satisfied ,product is as advertised,...
Name: Text, Length: 568454, dtype: object
```

2. Methodology:

GPT-2 Setup:

Start by initializing the GPT-2 Tokenizer and Model from Hugging Face, leveraging their GPT2Tokenizer and GPT2LMHeadModel.

Data Preparation:

Split the dataset into 75% for training and 25% for testing. Develop a custom PyTorch dataset class to streamline data preprocessing, including tokenization and padding.

Model Fine-Tuning:

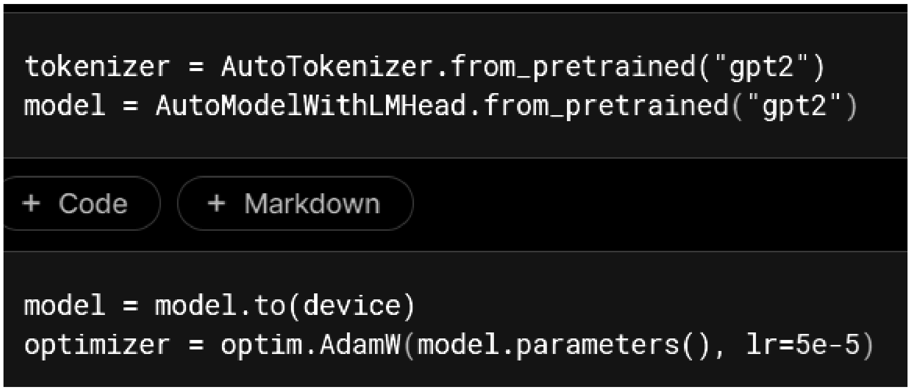
Fine-tune the pretrained GPT-2 model on the review dataset to generate summaries. Experiment with hyperparameters such as learning rate, batch size, and number of epochs for optimal performance.

Key Parameters:

For this setup, utilize the following parameters: epoch = 8, learning rate = 5e-5, batch size = 32, sample size = 20000, max length = 180. The model is trained on a dataset consisting of 20,000 data points, with training time averaging around 2 hours.

Query Processing:

Implement a function to process queries and present the output in the specified format.

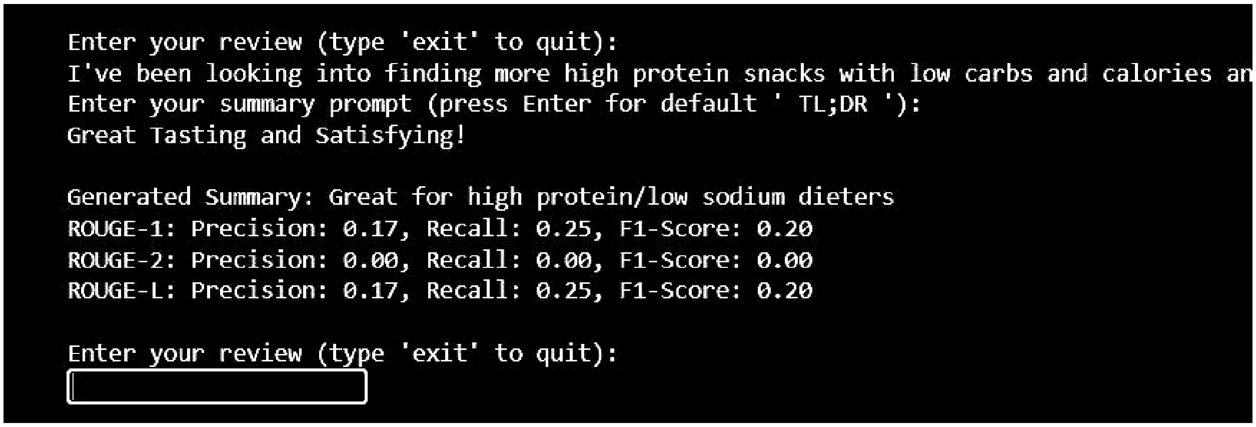


```
tokenizer = AutoTokenizer.from_pretrained("gpt2")
model = AutoModelWithLMHead.from_pretrained("gpt2")
```

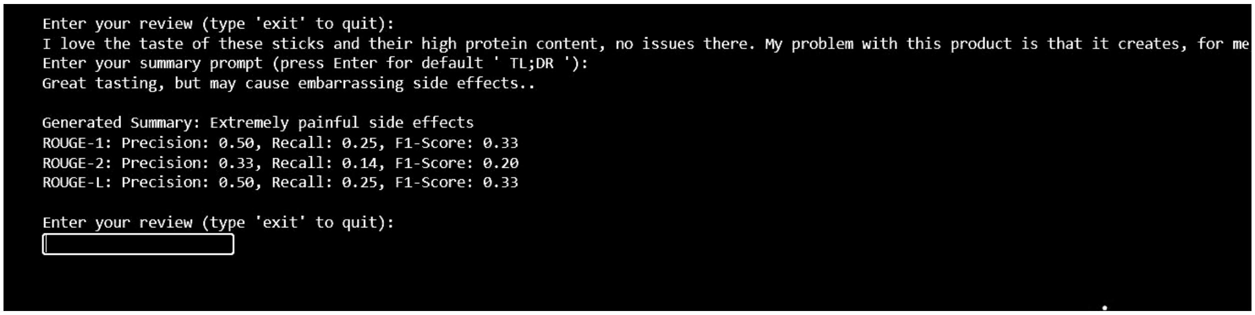
```
model = model.to(device)
optimizer = optim.AdamW(model.parameters(), lr=5e-5)
```

3. Evaluation:

- **ROUGE Score Assessment:**
  Post-training, evaluate the model's effectiveness by computing ROUGE scores on the test set. Measure the ROUGE scores for each predicted summary against the corresponding actual summary.

- 

```
Enter your review (type 'exit' to quit):
I've been looking into finding more high protein snacks with low carbs and calories an
Enter your summary prompt (press Enter for default ' TL;DR '):
Great Tasting and Satisfying!

Generated Summary: Great for high protein/low sodium dieters
ROUGE-1: Precision: 0.17, Recall: 0.25, F1-Score: 0.20
ROUGE-2: Precision: 0.00, Recall: 0.00, F1-Score: 0.00
ROUGE-L: Precision: 0.17, Recall: 0.25, F1-Score: 0.20

Enter your review (type 'exit' to quit):
```

- 

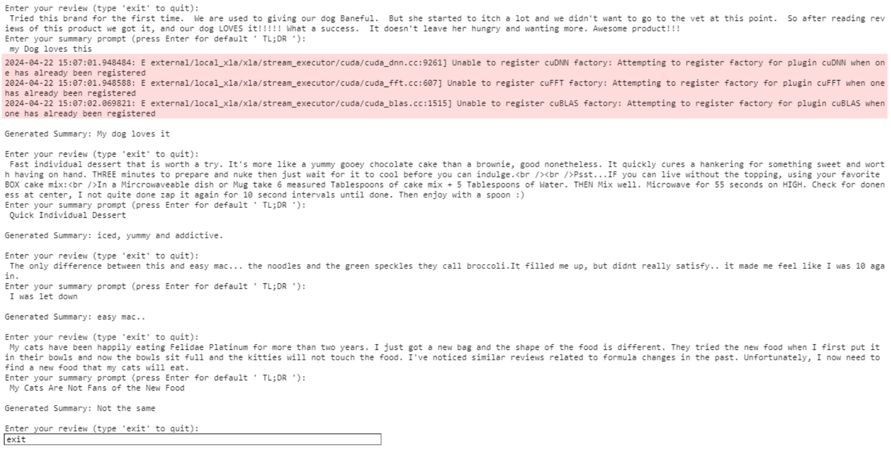
```
Enter your review (type 'exit' to quit):
I love the taste of these sticks and their high protein content, no issues there. My problem with this product is that it creates, for me
Enter your summary prompt (press Enter for default ' TL;DR '):
Great tasting, but may cause embarrassing side effects..

Generated Summary: Extremely painful side effects
ROUGE-1: Precision: 0.50, Recall: 0.25, F1-Score: 0.33
ROUGE-2: Precision: 0.33, Recall: 0.14, F1-Score: 0.20
ROUGE-L: Precision: 0.50, Recall: 0.25, F1-Score: 0.33

Enter your review (type 'exit' to quit):
```

- **Analyzing Results:**
  Quantitative Assessment: Look at ROUGE scores to see how well the model summarized the text.
  Qualitative Evaluation: Check a few generated summaries to see if they make sense and sound good.
  Hyperparameter Influence: How different settings affect how well the model works.
  Comparison with Other Methods: See how the model's ROUGE scores stack up against more straightforward methods.

Conclusion:

Sum up what you found, what the model did well, and where it could improve.

Suggest areas where the model could be improved, or more research could be done.



```
Enter your review (type 'exit' to quit):
 Tried this brand for the first time.  We are used to giving our dog Baneful.  But she started to itch a lot and we didn't want to go to the vet at this point.  So after reading rev
iews of this product we got it, and our dog LOVES it!!!!! What a success.  It doesn't leave her hungry and wanting more. Awesome product!!!
Enter your summary prompt (press Enter for default ' TL;DR '):
 my Dog loves this
2024-04-22 15:07:01.948484: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:9261] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when on
e has already been registered
2024-04-22 15:07:01.948588: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:607] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one
has already been registered
2024-04-22 15:07:02.069821: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1515] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when
one has already been registered

Generated Summary: My dog loves it

Enter your review (type 'exit' to quit):
 Fast individual dessert that is worth a try. It's more like a yummy gooey chocolate cake than a brownie, good nonetheless. It quickly cures a hankering for something sweet and wort
h having on hand. THREE minutes to prepare and nuke then just wait for it to cool before you can indulge.<br /><br />Psst...IF you can live without the topping, using your favorite
BOX cake mix:<br />In a Mircrowaveable dish or Mug take 6 measured Tablespoons of cake mix + 5 Tablespoons of Water. THEN Mix well. Microwave for 55 seconds on HIGH. Check for donen
ess at center, I not quite done zap it again for 10 second intervals until done. Then enjoy with a spoon :)
Enter your summary prompt (press Enter for default ' TL;DR '):
 Quick Individual Dessert

Generated Summary: iced, yummy and addictive.

Enter your review (type 'exit' to quit):
 The only difference between this and easy mac... the noodles and the green speckles they call broccoli.It filled me up, but didnt really satisfy.. it made me feel like I was 10 aga
in.
Enter your summary prompt (press Enter for default ' TL;DR '):
 I was let down

Generated Summary: easy mac..

Enter your review (type 'exit' to quit):
 My cats have been happily eating Felidae Platinum for more than two years. I just got a new bag and the shape of the food is different. They tried the new food when I first put it
in their bowls and now the bowls sit full and the kitties will not touch the food. I've noticed similar reviews related to formula changes in the past. Unfortunately, I now need to
find a new food that my cats will eat.
Enter your summary prompt (press Enter for default ' TL;DR '):
 My Cats Are Not Fans of the New Food

Generated Summary: Not the same

Enter your review (type 'exit' to quit):
exit
```