HW3
Team: Semanticons
Team members: Gayatri Balaji, Joshua Jose, Naga Hemachand Chinta, Vaishnavi
Link for Colab:
https://colab.research.google.com/drive/1UumyzpzoXbcMqgxfBchndEv9rTGGEHRX?usp=sharing

Task-1:
We have used the four decoding algorithms:
1. Greedy Search
2. Top-K Sampling
3. Top-P Sampling
4. Beam Search

Model Parameters:

Greedy Search:

```
[7]  1 greedy_outputs = model.generate(input_ids, do_sample=False, max_length=30, pad_token_id=tokenizer.eos_token_id)
```

Beam Search:

```
[8]  1 beam_outputs = model.generate(input_ids, num_beams=4, max_length=30, no_repeat_ngram_size=2, pad_token_id=tokenizer.eos_token_id )
```

Top-k Sampling:

```
[9]  1 top_k_outputs = model.generate(input_ids, do_sample=True, max_length=30, top_k=50,  pad_token_id=tokenizer.eos_token_id)
```

Top-p Sampling (Nucleus Sampling):

```
[10]  1 top_p_outputs = model.generate(input_ids, do_sample=True, max_length=30, top_p=0.9, pad_token_id=tokenizer.eos_token_id)
```

Model outputs:
Prompt: "Today I believe we can finally"

Greedy Search: ['Today I believe we can finally get to the point where we can make a difference in the lives of the people of the United States of America.\n']

Beam Search: ['Today I believe we can finally get to the bottom of this issue.\n\n"We need to find a way to make sure that we don\'t']

Top-k Sampling: ['Today I believe we can finally reach an agreement on the conditions that are in place for Greece to fulfil its obligations in this international agreement." This was followed']

Top-p Sampling (Nucleus): ["Today I believe we can finally find our future together. It's time to do something that the future may not hold: I'm gonna take out my"]

Avg. Perplexity score: 69.63790130615234

For other prompts as the following the average perplexity score of all the four models is the corresponding value.

```
prompts = [
    "Prompt 1: Today, I believe we can finally",
    "Prompt 2: Another interesting topic is",
    "Prompt 3: In the future, technology will",
    "Prompt 4: One sunny day, I decided to",
    "Prompt 5: The mystery began when I found"
]
```

Prompt	**Avg perplexity score**
Prompt-1	69.63790130615234
Prompt-2	423.43316650390625
Prompt-3	55.82099914550781
Prompt-4	75.8462142944336
Prompt-5	124.12049865722656

Link for task-1 excel:
https://docs.google.com/spreadsheets/d/1wE54OpFnp5uKRvEtmkdDO3XbaPD23QOet-OuI-xR_5k/edit?usp=sharing

Task-2:

In the context of decoding for downstream generation tasks, we employed the "Samsum" dataset. Our model of choice was "bart-large-cnn-samsum." To generate the outputs, we harnessed four distinct decoding methods, as detailed in task 1. As a result, we've compiled the generated outputs for 50 prompts from the dataset into an Excel file “decoding_results” for your convenience.
Link for task-2 excel:
https://docs.google.com/spreadsheets/d/1wE54OpFnp5uKRvEtmkdDO3XbaPD23QOet-OuI-xR_5k/edit?usp=sharing

Task-3:

Next, we compute BLEU and ROUGE Scores for each of the 50 prompts and their respective outputs. The resulting average BLEU and ROUGE Scores for each model are as follows:

Content Overlap metrics:

	Greedy Search	Top-K Sampling	Top-P Sampling	Beam Search
Average BLEU Scores	0.0958	0.0984	0.0921	0.0995
Average Rouge Scores	0.3738	0.3975	0.4019	0.3721

Model-based metrics:

The BERT Scores on the 50 outputs for each model are as follows:

	avg_beam_f1	avg_beam_p	avg_beam_r	avg_greedy_f1	avg_greedy_p	avg_greedy_r	avg_topk_f1	avg_topk_p	avg_topk_r	avg_topp_f1	avg_topp_p	avg_topp_r
Average BERT Scores	0.846164167	0.911702931	0.921484411	0.846164167	0.911702931	0.921484411	0.846164167	0.91170293	0.9214844	0.84616417	0.91170293	0.9214844

Model Comparisons:
Based on the average BLEU and ROUGE scores computed for various decoding methods, including Greedy Search, Top-K Sampling, Top-P Sampling, and Beam Search, we can draw the following comparisons among different Language Model Models (LLMs):

BLEU Scores: The highest average BLEU score is achieved when using the Beam Search decoding method, followed by Top-K Sampling and Greedy Search. On the other hand, Top-P Sampling yields the lowest average BLEU score. This implies that Beam Search tends to generate summaries with greater n-gram overlap with human reference summaries, while Top-P Sampling is less proficient at producing high n-gram overlap summaries.

ROUGE Scores: The highest average ROUGE score is observed when applying Top-P Sampling, followed by Top-K Sampling and Beam Search. Greedy Search yields the lowest average ROUGE score. This suggests that Top-P Sampling excels at creating summaries that closely align with human reference summaries in terms of content overlap and recall, whereas Greedy Search is less effective at generating high content overlap and recall summaries.

In summary, the choice of decoding method significantly impacts the performance of different Language Model Models (LLMs) concerning BLEU and ROUGE scores. Beam Search demonstrates strength in the BLEU metric, while Top-P Sampling is more effective in the ROUGE metric. To optimize performance on both metrics, it may be beneficial to consider a combination of these decoding methods and carefully fine-tune their parameters.

Human evaluation:

Greedy search scores:

fluency check	3.4
coherence check	3.36
formality check	3.27
typicality check	3.2325

Top k scores:

fluency check	3.095
coherence check	3.125
formality check	3.22
typicality check	3.2525

Top p scores:

fluency check	3.2925
coherence check	3.445
formality check	3.1925
typicality check	3.33

Beam search scores:

fluency check	3.1825
coherence check	3.215
formality check	3.345
typicality check	3.2575

Average sum for every individual is the calculation for the inter-annotator agreement.
The following is the image for reference which is present in the excel sheet “human evaluation”

AVERAGE SUM		3.15	3.15	3.4	3.25	3.1	3.35	3.5	3.1	3.15	2.9	3

Justification for Model Metrics:
ROUGE: It assesses the similarity of n-grams between generated and reference summaries, exhibiting a strong alignment with human assessments of summary quality. This metric is widely embraced in the research community as a standard for evaluating summarization tasks.

BLEU: It gauges the overlap of n-grams between generated and reference summaries and is prominently employed in machine translation evaluations. Despite some limitations, it can offer valuable insights into the performance of different models.

BERT score: This metric computes the cosine similarity between summary embeddings using the BERT model. It demonstrates a robust correlation with human judgments of text quality and has gained substantial recognition in recent research as a measure for assessing the quality of generated text.

Reasoning for choice of aspects:
We provide human annotations in a separate file included within the final folder. The human annotations are evaluated based on the following criteria:

Factuality: This assessment measures the accuracy of the summary in representing the facts contained in the source text. A summary that includes inaccurate or misleading information would receive a lower score in this category.

Formality: This evaluation gauges the level of formality in the language employed within the summary. A summary that employs more formal language would score higher in this aspect, while one using more informal language would receive a lower rating.

Typicality: This metric assesses the extent to which the summary encapsulates the typical or representative content of the source text. A summary that predominantly includes unusual or atypical information from the source text would receive a lower rating.

Coherence: This criterion evaluates the degree of coherence and logical consistency within the summary. A summary that includes disjointed or contradictory information would receive a lower score in terms of coherence.

- Factuality: The highest average score is obtained with the greedy search decoding algorithm, while the lowest score is obtained with the Top-K Sampling algorithm.
- Coherence: The highest average score is obtained with the Beam search decoding algorithm, while the lowest score is obtained with the Top-K algorithm.
- Formality: The highest average score is obtained with the Beam search Sampling decoding algorithm, while the lowest score is obtained with the Top-P Sampling algorithm.

● Typicality: The highest average score is obtained with the Top-P Sampling decoding algorithm, while the lowest score is obtained with the Greedy Search algorithm.

Overall Beam search algorithm performs better except at factuality and typicality where its performance is next best.

Difference between Human and Automatic Evaluation:
Objective assessments, like BLEU and ROUGE scores, are automated measures that hinge on comparing generated summaries to reference summaries using predefined criteria. In contrast, subjective evaluations rely on human assessors to gauge the quality of summaries across diverse criteria. Automated evaluations are typically quicker and more cost-effective than human evaluations but may not encompass all facets of summary quality. Although there is some correlation between automated and human evaluations, human assessments are generally regarded as the benchmark for appraising the quality of generated summaries.

Bonus points:

- Advanced decoding algorithms implemented: Temperature scaling
  Temperature Scaling:
  Temperature scaling can be used to control the randomness of generated text. Higher values (e.g., 2.0) make the output more random, while lower values (e.g., 0.7) make the output more deterministic.

- For content overlap: BLUE and ROGUE are implemented.
- For model metrics: Bert score implemented and METEOR is implemented too.