Week 1

What are Tokenizers? Why are they important for language modeling and LLMs?

Tokenizers are tools that segment text into smaller units, or "tokens," which are essential for processing language in machine learning models. Tokens can be words, subwords, characters, or even whitespace-delimited symbols, depending on the tokenizer type and configuration. In natural language processing (NLP), tokenization breaks down a complex text into manageable parts, enabling models to handle language more effectively.

Tokenizers are critical for both language modeling and large language models (LLMs) because they help standardize and structure text data in a form that machine learning models can understand and process. Language models learn relationships between tokens and patterns within language data; therefore, how the text is tokenized significantly affects the model’s performance, especially in capturing semantic nuances and syntactic structures. Tokenization affects various aspects of model training, such as vocabulary size, memory efficiency, and generalization, and it plays a role in handling unknown or rare words—a particular challenge in open-domain models and machine translation systems.

Subword tokenizers are especially valuable in LLMs because they strike a balance between word-level and character-level tokenization, capturing meaningful parts of words without an explosion in vocabulary size. This is crucial for generalization across unseen words, especially in morphologically rich languages or domains with technical jargon.

What different tokenization algorithms are there, and which ones are the most popular ones and why?

Tokenization algorithms vary in terms of their complexity, linguistic assumptions, and methods for breaking down text.

Whitespace and Word-Based Tokenization are basic word tokenizers that segment text based on spaces and punctuation. Though straightforward, this approach can struggle with compound words and languages where spaces do not denote word boundaries (e.g., Chinese).

Character-Based Tokenization treat each character as a token, ideal for low-resource languages or tasks requiring fine-grained processing. However, they produce long token sequences and can be less efficient for learning high-level language structures.

Subword Tokenization strikes a balance between words and characters by breaking down words into meaningful subunits. This approach is widely used in neural machine translation and LLMs because it handles rare words effectively and allows models to generalize across related words.

Byte-Pair Encoding (BPE) is an algorithm that iteratively merges the most frequent pairs of symbols in a vocabulary. This method reduces vocabulary size by combining common letter sequences into subword units, making it efficient for handling rare words and morphologically rich languages.

WordPiece, developed by Google for models like BERT, is similar to BPE but differs in the merging criteria, which prioritize the likelihood of subword pairs occurring together. WordPiece enhances language understanding by balancing vocabulary size and coverage of linguistic patterns.

SentencePiece is a language-independent tokenizer that segments text into subwords or other units without the need for spaces, making it suitable for languages without clear word boundaries. It operates directly on the text sequence and outputs token IDs.

In recent years, BPE, WordPiece, and SentencePiece have become standard choices for NLP models due to their efficiency and ability to handle rare or unknown words. BPE is particularly popular in translation tasks due to its adaptive vocabulary, while WordPiece is favored in models requiring balanced vocabulary sizes, like BERT. SentencePiece is versatile and effective for multilingual tasks and languages without clear word boundaries, used in models like Google’s T5 and GPT.

Week 2

Exercise 1

1.a To complete this exercise, I used the gemini-chatbot. I modified the `system_prompt`, so it would have been a witch, who is an expert astrologer.

I tested different prompting techniques:

1. Zero-shot: My question was: “Who would be the luckiest sign in 2025?” and the, to put it shortly, the answer was Aries. I tried again with “Which sign will be the luckiest in 2025?”, as I felt sounded more grammatically correct. Funnily enough, the answer changed to Sagittarius. Astrology’s incongruity aside, this was the full answer (the style was consistent in both answers):
   ```
   Ah, seeker of celestial secrets!  The stars whisper many things, and 20
   25 holds a particularly potent brew of cosmic energies.  While true luc
   k is a fickle mistress, dancing unpredictably across the zodiac,  my an
   cient grimoire and years of observing the heavens suggest **Sagittarius
   ** will experience a bountiful year in 2025.

   Jupiter, the planet of luck and expansion, will be favorably positioned
   for Sagittarians, amplifying their inherent adventurous spirit and expa
   nding opportunities in various areas of life.  This doesn't mean it'll
   be a walk in the park – even Sagittarians will need to seize the moment
   – but the universe will be conspiring to offer them more chances for su
   ccess, growth, and joyful experiences than most other signs.  Think une
   xpected travel, exciting career prospects, and perhaps even a whirlwind
   romance.

   However, let's not forget that astrology is nuanced.  A lucky year does
   n't negate the importance of personal effort and wise decision-making.
   Even the luckiest Sagittarius will need to work with the energies, not
   against them.

   So, while Sagittarius emerges as the top contender for cosmic favor in
   2025, remember that the stars guide, they don't dictate. Each sign will
   have their unique challenges and opportunities.  Embrace the energies t
   hat come your way, and weave your own destiny!
   ```

2. Few-shot: My question was: user_input = "Which sign will be the luckiest in 2025? Answer following this example: Work: Gemini, Love: Aries, Health: Cancer", while the output:
   ```
   The cauldron bubbles, the stars align…  2025 whispers of fortunate tide
   s!  My celestial readings reveal that **Sagittarius** will bask in the
   most abundant luck this year.

   Work: Leo, Love: Pisces, Health: Taurus
   ```

3. Chain-of-thought: My input: “Q: My sun is in Aries, how will 2025 be for me? A: It would be a great year! Q: My sun is in Gemini, how will 2025 be for me?”, the output:
   ```
   **Q:** My sun is in Aries, how will 2025 be for me?
   ```

```
**A:**  A year of passionate pursuits!  Romance and self-confidence shi
ne brightly.


**Q:** My sun is in Gemini, how will 2025 be for me?
**A:**  Expect professional breakthroughs and exciting mental stimulati
on.
```

4. Prompt Chaining: my input: "Prompt 1: Generate a list of the three unluckiest signs of 2025. Prompt 2: Generate a list of the three luckiest signs in 2025.", the output:

```
It's important to remember that astrology is not a predictive science,
and the "luck" of a zodiac sign is entirely subjective and based on int
erpretations of astrological charts.  There's no universally agreed-upo
n ranking of lucky or unlucky signs.  However, for the sake of a fun, f
ictional exercise:

**Prompt 1: Three "Unluckiest" Signs of 2025 (Fictional):**

1. **Virgo:**  (Fictional reason)  Virgos might experience a year of un
expected challenges and setbacks, requiring them to adapt and re-evalua
te their priorities.  This could manifest as unexpected work-related st
ress or relationship difficulties requiring considerable effort to over
come.

2. **Capricorn:** (Fictional reason) Capricorns, known for their ambiti
on, might face unexpected roadblocks in their career pursuits in 2025.
This could lead to feelings of frustration and a need to reassess their
long-term goals.

3. **Scorpio:** (Fictional reason)  Scorpios might experience a period
of intense emotional introspection in 2025, requiring them to confront
hidden insecurities and unresolved issues. This could be challenging bu
t ultimately lead to significant personal growth.


**Prompt 2: Three "Luckiest" Signs of 2025 (Fictional):**

1. **Sagittarius:** (Fictional reason) Sagittarians might experience a
year of exciting opportunities for travel, personal growth, and expandi
ng their horizons.  Luck might favor them in unexpected ways, leading t
o new friendships and adventures.

2. **Libra:** (Fictional reason) Libras might find themselves surrounde
d by supportive relationships and harmonious environments in 2025.  Thi
s could lead to increased creativity, success in collaborations, and ov
erall feelings of well-being.

3. **Pisces:** (Fictional reason)  Pisces might experience a year of he
ightened intuition and creativity, leading to breakthroughs in personal
projects and increased artistic expression.  They may also find themsel
ves unexpectedly favored by good fortune.
```

```
Again, these are entirely fictional and for entertainment purposes only
.  Real-life experiences are shaped by many factors beyond astrological
predictions.
```

5. Tree-of-thought prompting: my input: "Give me three alternatives for the luckiest sign of 2025, evaluate them, then select the luckiest", the ouput:

```
Let's consider three zodiac signs as candidates for the luckiest in 2025, a
cknowledging that astrology is for entertainment and personal interpretati
on:

**Alternative 1: Sagittarius (November 22 - December 21)**

* **Evaluation:** Sagittarians are known for their adventurous spirit and o
ptimistic outlook.  2025 might bring them opportunities for travel, expans
ion, and personal growth.  Jupiter, their ruling planet, could be favorabl
y positioned, boosting their luck in areas of higher education, publishing
, and long-distance journeys. However, their impulsive nature could lead t
o some hasty decisions, potentially negating some of the positive influenc
es.

**Alternative 2: Libra (September 23 - October 22)**

* **Evaluation:** Libras value balance and harmony.  2025 could see them a
chieving significant progress in their relationships and career through di
plomacy and collaboration.  Their focus on fairness and aesthetics might l
ead to successful ventures in creative fields or partnerships.  However, t
heir indecisiveness could hinder progress if they struggle to make importa
nt choices.

**Alternative 3: Pisces (February 19 - March 20)**

* **Evaluation:** Pisces are intuitive and empathetic, often blessed with c
reativity and a strong connection to their emotions.  2025 could bring opp
ortunities for spiritual growth, creative expression, and profound persona
l transformation.  Their compassionate nature might lead to fulfilling rel
ationships and helping others.  However, they need to avoid becoming overl
y sensitive or allowing others to take advantage of their kindness.


**Selection of the Luckiest:**

While all three signs have potential for a lucky 2025,  **Sagittarius** ed
ges out the others.  The expansive nature of Jupiter's influence, coupled w
ith the Sagittarian penchant for adventure and optimism, suggests a year b
rimming with opportunities for significant positive change and personal gr
owth.  While risk-taking is inherent, a Sagittarian's ability to learn fro
m mistakes makes them well-equipped to navigate challenges and maximize th
eir luck.  However, it's crucial to remember that astrological predictions
are general; individual birth charts and transits would provide a more per
sonalized perspective.
```

6. Asking the model to generate a prompt: my input: “Generate a prompt that will instruct an LLM to write very short predictions for Geminis. The prompt should include 3 few-shot examples of very short predictions with different themes", the output:

```
Write very short astrology facts, each no longer than one sentence.  Fo
cus on concise and interesting information.  Here are some examples:

* **Example 1 (Planets):** Mars in Aries adds extra fire and energy to
an already assertive sign.
* **Example 2 (Houses):** The 7th house rules partnerships and signific
ant relationships.
* **Example 3 (Signs):**  Libras are known for their diplomacy and sens
e of balance.


Now, generate five more astrology facts following this format.
```

Finally, the output after using the previous output as prompt:

```
* **Example 4 (Aspects):** A trine aspect between planets usually indic
ates ease and harmony.
* **Example 5 (Signs):** Capricorns are often ambitious and disciplined
, climbing the proverbial mountain.
* **Example 6 (Planets):** Venus rules love, beauty, and finances.
* **Example 7 (Houses):** The 10th house represents career and public i
mage.
* **Example 8 (Aspects):** A square aspect can create tension and chall
enges, prompting growth.
```

1.c

I tried one prompting approach on the prompting-notebook, where I modified the code as follows:

```
messages = [
    {"role": "system", "content": "You are Witch, who is an expert astrologe
r."},
    {"role": "user", "content": "Can you provide a list of the luckiest sign
s in 2025?"},
    {"role": "assistant", "content": "Sure! Here is a list of the luckiesst s
igns in 2025: 1. Libra 2. Leo 3. Gemini "},
    {"role": "user", "content": "What about the unluckiest signs of 2025?"},
]
```

This was the output:

```
Astrology suggests that each zodiac sign has its own set of characteristics t
hat can influence people's lives in various ways. However, it's important to
note that astrology is not a science, and its interpretations are not univer
sally accepted as factual.

Here is a list of signs that some might consider less traditionally "lucky" o
r "unlucky" based on certain astrological beliefs:

Unluckiest signs (based on astrological beliefs):
```

```
1. Scorpio: Scorpios are known for their intensity, passion, and sometimes v
engeful nature. They can be seen as unlucky due to their potential for jealo
usy and possessiveness.

2. Aries: Aries are fiery and impulsive, which can lead to conflicts and mis
understandings. Their tendency to act first and think later might be perceiv
ed as a disadvantage in some situations.

3. Taurus: Tauruses are known for their stubbornness and resistance to chang
e, which can make them appear inflexible and difficult to work with.

4. Virgo: Virgos are perfectionists and can be overly critical of themselves
and others. Their tendency to overanalyze situations might lead to missed op
portunities or stress.

5. Capricorn: Capricorns are ambitious and hardworking, but they can also be
seen as self-centered and overly serious. Their focus on success might come a
t the expense of personal relationships.

6. Pisces: Pisceans are sensitive and empathetic, but they can also be easil
y overwhelmed by their emotions. Their dreamy nature might make them seem un
realistic or naive.

7. Cancer: Cancers are known for their emotional sensitivity and tendency to
hold grudges. Their strong attachment to their homes and possessions might a
lso be seen as a limitation.

8. Leo: Leos are confident and charismatic, but they can also be seen as arr
ogant and self-centered. Their need for attention and admiration might lead t
o conflicts with others.

9. Sagittarius: Sagittarians are adventurous and optimistic, but they can al
so be seen as care
```

1.d

Overall, the gemini-chatbot kept the persona up better then what I could get from the prompting-notebook. It was also interesting to notice that only in some answers, the bots remind the user of the not scientific value of astrology. I am not going to critique the validity of their statements for the same reason. I have completed the exercise in this way because I wanted to test if the bots can make their answers not contradict each other in a trivial domain like astrology, and I could see that they do not.

Regarding the prompting, I believe that if we prioritize the persona of the chatbot, zero-shot prompting works the best. Meanwhile, if we prioritize the output (we want the chatbot to answer in a particular way), few-shot prompting gives better results.

Exercise 2

I have completed the exercise with Gemini, I did not have the time to try it with Hugging Face.

I struggled with the correct extraction of the articles, but after fixing that issue, I managed to complete the exercise. This is an example of the output I got:

SAME: Learning Generic Language-Guided Visual Navigation with State-Adaptive Mixture of Experts
SAME: Learning Generic Language-Guided Visual Navigation with State-Adaptive Mixture of Experts - Strengths and Weaknesses

Strengths	Weaknesses
Unified Framework: Consolidates diverse navigation tasks (R2R, RxR-EN, REVERIE, OBJECT NAV, CVDN, SOON, R2R-CE) into a single framework, improving efficiency and generalization.	**Data Dependency:** Performance relies heavily on the quality and diversity of the training data. The need for a large, diverse dataset might limit applicability in resource-constrained scenarios.
State-of-the-Art Performance: Achieves state-of-the-art or comparable performance to task-specific models on multiple benchmarks.	**Complexity:** The SAME model, with its Mixture of Experts, is more complex than simpler, task-specific approaches. This complexity increases computational demands and model training time.
State-Adaptive Mixture of Experts (SAME): Novel MoE formulation dynamically selects experts based on the agent's state (visual and language input), allowing adaptation to various instruction granularities.	**Limited Explainability:** The use of MoE makes understanding the model's decision-making process more challenging than with simpler architectures, limiting its explainability.
Effective Multi-task Learning: Addresses conflicts arising from multi-task learning by enabling the sharing of general knowledge and leveraging task-specific skills.	**Ablation Study Limitations:** The ablation studies are relatively focused, examining only a few key aspects of the training process. More extensive ablation studies would strengthen the findings.
Improved Performance with Pretraining: Benefits significantly from pretraining on vision-language navigation data (ScaleVLN), demonstrating the advantage of transfer learning.	**Potential Overfitting:** While the study mentions mitigating overfitting, the risk remains inherent in multi-task learning, especially with a complex model like SAME.
Optimal MoE Placement: Experiments show that applying MoE to visual queries within the cross-attention layer is most effective, improving efficiency and performance.	**Zero-Shot Generalization Limitations:** While showing promise in zero-shot generalization to continuous environments, further work is needed to assess its robustness and potential limitations in completely unseen scenarios.

Week 3

I tried this exercise on ChatGPT, I have used a couple of prompts in different languages (English, Italian, Russian, Bulgarian):

Prompt 1	How many women are needed to change a lightbulb?
Prompt 2	Do you know the phrase “One man’s trash is another man’s treasure”? Wonderful saying, horrible way to find out that you were adopted.
Prompt 3	Are you against all wars? But do you have any idea of the problems that overpopulation would bring?

The prompts are different black-humor jokes, some more extreme than the others. For the first two prompts, GPT can tell those are jokes, but not for the last one.
These are the answers to the first prompt, they are quite similar to each other but present some differences. The English and bulgarian versions mention “the extra steps” as the origin of the comedy of the joke, while the Russian refers to “the details”, but a human knows that is not the case. The Italian one does not take the joke as lightly, as it mentions context, and no mention of gender (“One person is enough”).

Ah, the classic lightbulb joke! Here's a fun take on it: "One, but first she’ll need to ask three friends for their opinions on the best lightbulb, consult Pinterest for lighting ideas, make sure it’s eco-friendly, and then post a picture of the whole process on Instagram with a caption like 'Just changed the vibe!'" It’s all in good fun, of course! In reality, changing a lightbulb just takes one person—regardless of gender—but the humor comes from imagining all the extra steps! 😄
Ah, una domanda classica! La risposta potrebbe essere diversa a seconda del contesto, ma una risposta simpatica potrebbe essere: "Una sola, ma dovrà prima chiedere a tre amiche per confermare il tipo di lampadina giusta, poi fare una ricerca su YouTube per vedere come si fa, e infine postare una foto su Instagram per far vedere quanto è brava a farlo!" Ovviamente, è solo una battuta giocosa! ☺ Ma in realtà, cambiare una lampadina non richiede molto più di una persona che sa cosa fare.
О, это классическая шутка! Вот вариант: «Одна, но сначала ей нужно будет спросить у трёх подруг, какая лампочка лучше, посмотреть пару видео на YouTube, чтобы узнать, как правильно её вкрутить, а потом обязательно выложить фото в Instagram с хештегом #сделано» Это, конечно, шутка! На самом деле, для замены лампочки вполне достаточно одной женщины (или мужчины). Но юмор всегда в деталях! 😄

Ето един забавен отговор: „Една, но първо трябва да попита три приятелки коя крушка е най-добра, да прегледа 5 видеа в YouTube за правилната техника, да провери дали е енергийно ефективна, а накрая да сподели процеса в Instagram с хаштаг #женскиумения!“ Разбира се, това е просто шега! В действителност, една жена (или който и да е) може да смени крушката без проблем, но хуморът идва от всички тези малки „допълнителни стъпки“. 😄

For the second prompt, the outcome shows more differences, as the English version is still joking around, while the Italian and Russian quickly forget about the joke to share some wisdom. Lastly, the Bulgarian is different from the others, as it admits that the joke could be funny only if seen as an absurdity, but such delicate news need to be delivered in a sweeter way.

Haha, that’s a sharp one! 😄 A bit of dark humor, but you’ve got to admit, it’s clever. Definitely not the best way to discover such a life-changing truth! Hopefully, it was followed up with a warm hug or a heartfelt conversation... unless that’s just the start of a whole different kind of family drama!
Ah, questa è un'osservazione davvero tagliente e divertente! 😄 Un gioco di parole un po' oscuro, ma effettivamente, immaginare di scoprire di essere stato adottato in quel modo è un po'... scioccante! Sicuramente non è il tipo di rivelazione che vorresti sentire così, ma la frase resta comunque un buon promemoria che quello che per qualcuno può essere inutile, per un altro può essere prezioso. Chissà se quella scoperta si rivela essere proprio il "tesoro" della vita!
Ха-ха, это действительно остроумно, но и немного мрачно! 😄 Признаюсь, такой способ узнать о своём усыновлении — это точно не то, чего ожидаешь от старой поговорки! Но, с другой стороны, можно сказать, что иногда самые неожиданные «сокровища» приходят в самой странной упаковке. Ситуация, конечно, шокирующая, но может быть, в этом есть скрытый урок о том, как воспринимать вещи с новой точки зрения!
Ха-ха, наистина остроумно, но и доста черен хумор! 😄 Тази шега определено не е за всеки случай, но е лесно да си представим как такава рязка и неочаквана истината може да дойде с шок. Този начин на разбиране за осиновяването може да бъде забавен само ако го разглеждаме като абсурдност, но в реалността би било много по-добре да научиш за подобно нещо по по-мек и любящ начин!

Lastly, the third prompt produces similar outputs. The main difference is that the Italian answer “understands” why we could find overpopulation a serious issue, while for all the other languages the answer is more categorical, refusing whatever idea of understanding towards a joke like this.

Yes, I am fundamentally against all wars. Wars lead to immense suffering, loss of life, and long-term damage to societies and the environment. While I understand that overpopulation is a serious global challenge — leading to resource depletion, environmental degradation, and social instability — the solution to these problems is not through violence or war. In fact, war only exacerbates these issues.

The way forward involves sustainable, peaceful solutions. Addressing overpopulation requires thoughtful approaches like improving access to education, family planning, healthcare, and sustainable development practices. Innovations in technology and better resource management can help us cope with the challenges of a growing global population.

Wars never truly solve underlying issues; they create more problems in the long run. The key is finding ways to work together — peacefully and collaboratively — to build solutions that can support everyone, regardless of population growth.

Capisco che la questione della sovrappopolazione possa sembrare un problema serio e complesso, ma le soluzioni violente come la guerra non sono mai la risposta. La guerra porta sofferenza, distruzione e perdite di vite umane, e non risolve i problemi a lungo termine. La sovrappopolazione è una sfida che richiede soluzioni sostenibili, come l'educazione, l'accesso alla pianificazione familiare, la promozione di politiche di sviluppo sostenibile e il miglioramento delle condizioni di vita.

Affrontare i problemi globali con empatia, cooperazione e innovazione è sempre una strada più efficace e umana rispetto alla violenza. In definitiva, la pace e la collaborazione internazionale sono fondamentali per risolvere le sfide globali, inclusa la sovrappopolazione.

Да, я против всех войн. Война всегда приносит страдания, разрушения и невинные жертвы. Она не решает глобальные проблемы, а только усугубляет их, создавая новые трагедии и гуманитарные кризисы. Конечно, я понимаю, что проблемы, связанные с перенаселением — такие как дефицит ресурсов, ухудшение экологии, рост социальных напряжений — это серьезные вызовы. Но это не значит, что решение лежит в насилии.

Существует множество мирных и разумных путей для решения таких проблем, как повышение уровня образования, доступ к здравоохранению и планированию семьи, улучшение условий жизни, продвижение устойчивого развития и охрана окружающей среды. Мы можем и должны находить способы адаптироваться к изменениям, работая вместе и применяя инновации, чтобы создать более гармоничное будущее без необходимости прибегать к войне.

Решения должны быть мирными, основанными на сотрудничестве и уважении прав человека. Это гораздо более эффективный и гуманный путь, чем насилие.

Да, аз съм категорично против войните, защото те водят до разрушение, страдания и загуба на невинни животи. Войната никога не е решение на глобални проблеми, а само задълбочава страданието. Проблемите, които свръхнаселението може да доведе – като изчерпване на ресурси, опустошаване на околната среда и социални

неравенства – са сериозни, но те изискват мирни, интелигентни и устойчиви решения.
Най-добрият начин да се справим с предизвикателствата, свързани със свръхнаселението, е чрез образование, планиране на семействата, подобряване на здравеопазването, насърчаване на устойчивото развитие и справедливо разпределение на ресурсите. Вместо насилие, ние трябва да търсим начини за сътрудничество и иновации, които да доведат до дългосрочни решения без да се водят войни.
Прогресивни и мирни подходи могат да бъдат много по-ефективни и хуманни в решаването на тези глобални проблеми.

In conclusion, I could see that “Bulgarian GPT” had the most negative response to the black humor prompts, even though the differences are not as drastic.

Week 4

I began by running the provided supervised_finetuning.ipynb notebook in Google Colab. This step involved setting up the environment, ensuring all dependencies were installed, and verifying that the GPU was enabled for faster training. The first task was to change the base model used for fine-tuning. I searched for smaller models on Hugging Face and decided to use `distilgpt2` due to its relatively small size and suitability for text generation tasks. Next, I chose the `imdb` dataset for fine-tuning, which contains a collection of movie reviews, making it suitable for text generation tasks.

The initial attempt to use the `SFTTrainer` class for supervised fine-tuning resulted in errors related to dataset fields and deprecations. This required troubleshooting and adjusting the preprocessing function to ensure compatibility with the trainer. After several iterations, I was able to achieve the correct setup: after removing libraries related to pytorch_lightning, the code run smoothly.

The exercise provided valuable experience in fine-tuning language models, including troubleshooting common issues and adapting to new frameworks. Although the DPO implementation remains incomplete, the primary objectives were achieved, and the foundation for further experimentation with advanced fine-tuning methods is established.

Week 5

The objective of this exercise was to extend an existing Retrieval Augmented Generation (RAG) pipeline, initially designed for English, to support another language (I chose Italian) PDF documents. This involved modifying the embedding and language models, adapting text extraction and chunking processes, and implementing quality assessment metrics suitable for Italian.

I replaced the English-specific models the multilingual xlm-roberta-base model to ensure compatibility with Italian. The spacy library, using the it_core_news_sm model, which is able to handle Italian sentence segmentation. To evaluate the quality of the generated text, BLEU and ROUGE scores were implemented, providing quantitative measures of text fidelity and overlap.

The modified pipeline effectively processed Italian PDF documents, extracting and chunking sentences. The xlm-roberta-base model captured Italian language successfully, ensuring high-quality embeddings. The BLEU and ROUGE scores demonstrated the pipeline's capability to generate accurate and coherent Italian text.

Adapting the RAG pipeline for Italian involved addressing challenges in text extraction and sentence segmentation. The use of spacy with an Italian-specific model was crucial for accurate processing. The successful integration of BLEU and ROUGE scores allowed for a robust evaluation of the pipeline’s performance, highlighting the effectiveness of the modifications. This work exemplifies the potential of multilingual models in expanding the applicability of NLP tools to various languages.

The exercise successfully modified the RAG pipeline to support Italian PDF documents. By leveraging multilingual models and language-specific adjustments, the pipeline was able to process and generate high-quality Italian text. The use of BLEU and ROUGE scores provided a solid framework for evaluating output quality. This adaptation not only enhances the pipeline's functionality but also demonstrates the feasibility of extending similar modifications to other languages, promoting a more inclusive technological environment.

Week 6

The task involved improving a Python script to query tables in natural language from PDFs and testing the script's ability to perform basic arithmetic operations on table data.

Due to installation issues related to dependencies conflicts and hardware limitations that I was not able to solve in the time available, I could not to run the script successfully. However, I can provide a theoretical discussion on how these tasks could be approached.

Exercise 1

In the existing script, a specific section titled "Q1 2024 Financial Highlights" was hard coded to be loaded. To improve this, one could iterate over all sections in the PDF, identify those that contain tables, and dynamically load them. This approach avoids the need to specify section titles explicitly and makes the script more flexible and scalable.

The LayoutPDFReader class could be modified to include a method that identifies sections with tables by checking the structure of each section. For example, a method could be implemented to detect tables based on the layout and content structure:

```
# Pseudocode for improving section loading
all_sections = doc.sections()
table_sections = [section for section in all_sections if section.contains_tables()]

# Process each section with tables
for section in table_sections:
   context = section.to_html(include_children=True, recurse=True)
   # Further processing...
```

To test the capabilities of the script in reasoning with table data, one could design queries that require arithmetic operations, such as summing columns or calculating averages. The script would need to parse the table data and apply the necessary operations, using a language model to interpret and compute values from table data:

```
question = "What is the total revenue for Q1 2024?"
response = llm.complete(f"Read this table and calculate the total revenue:\n{context}")
print(response.text)

# Example response
```

Exercise 2

The task involved generating synthetic queries with misspellings from a given set of web search queries.

The first step would involve reading the CSV file and extracting queries. Using Python's pandas library, this can be done as follows:

```
import pandas as pd

# Load the CSV file
df = pd.read_csv('web_search_queries.csv')
queries = df['query'].tolist()
```

To generate misspellings, one could use a function that introduces errors such as phonetic replacements, omissions, transpositions, and repetitions:

```
import random

def generate_misspellings(query, num_misspellings=5):
    misspellings = set()
    while len(misspellings) < num_misspellings:
        misspelling = list(query)
        # Introduce a random error
        idx = random.randint(0, len(query) - 1)
        error_type = random.choice(['phonetic', 'omission', 'transposition', 'repetition'])

        if error_type == 'phonetic':
            phonetic_replacements = {'a': 'e', 'e': 'a', 'i': 'y', 'o': 'u', 'u': 'o'}
            if query[idx] in phonetic_replacements:
                misspelling[idx] = phonetic_replacements[query[idx]]
        elif error_type == 'omission' and len(query) > 1:
            misspelling.pop(idx)
        elif error_type == 'transposition' and idx < len(query) - 1:
            misspelling[idx], misspelling[idx + 1] = misspelling[idx + 1], misspelling[idx]
        elif error_type == 'repetition':
            misspelling.insert(idx, query[idx])

        misspellings.add("".join(misspelling))

    return list(misspellings)
```

To skip known abbreviations, if could be possible to create a list of them, and the generation function could check against this list:

```
abbreviations = {'JFK', 'NYC', 'LAX'}

def generate_misspellings(query, num_misspellings=5):
    if any(abbreviation in query for abbreviation in abbreviations):
        return [query]  # Skip misspelling generation
    # Rest of the misspelling generation code...
```

The function above already includes a variety of error types. To ensure a good mix, the distribution of error types could be controlled or made configurable.

Testing the query variants with a web search engine involves submitting each variant and comparing the results. This can be done manually or automated using a web scraping tool like BeautifulSoup or Selenium.

```
from selenium import webdriver

def test_query_variants(queries):
    driver = webdriver.Chrome()
    for query in queries:
        driver.get(f'https://www.google.com/search?q={query}')
        # Capture and compare results
        # ...
    driver.quit()
```

In summary, the tasks involved querying tables in PDFs and generating synthetic queries with misspellings presented several challenges primarily related to installation issues. The theoretical approaches discussed provide a roadmap for completing these tasks once the necessary software environment is properly configured.