Problem set 3.

Exercise 1. Basic Word Tokenization

Getting Started with Google Colab

Google Colab is a free cloud-based platform that allows you to write and execute Python code in your browser. It's a great tool for learning, experimenting, and collaborating on data science and machine learning projects.

Here's how to get started:

1. **Access Google Colab:**
   - Open a web browser and go to colab.research.google.com.
   - If you're not already signed in to your Google account, you'll be prompted to do so.
2. **Create a new notebook:**
   - Click on "New notebook" to create a fresh Colab notebook.
3. **Familiarize yourself with the interface:**
   - The Colab interface is similar to the Jupyter Notebook. You'll see cells where you can write and execute code.
   - You can add new cells by clicking the "+ Code" or "+ Text" buttons.
4. **Start coding:**
   - In a code cell, you can write Python code and execute it by pressing Shift+Enter or clicking the play button.
   - Colab provides many pre-installed libraries for data science and machine learning, such as NumPy, Pandas, and TensorFlow.
5. **Save your work:**
   - Your notebook is automatically saved to your Google Drive. You can also download it in various formats (e.g., .ipynb, .py).
6. **Collaborate with others:**
   - You can share your notebook with others for collaboration, just like you would with a Google Doc.

Additional Tips:

- Use the "Table of contents" feature to navigate your notebook easily.

- Explore the "Code snippets" panel for helpful code examples.
- Check out the Colab documentation for more advanced features and tutorials.

Your goal is to practice tokenizing using the NLTK or transformer library. Use code agents to help you with the code.

Example with Transformer library:

https://github.com/Unisvet/haf_ai/blob/main/tokenizer_101.ipynb

Tokenization with NLTK library:

Tokenization is the first step in text processing. It's the process of breaking text into parts or tokens. You will use the Natural Language Toolkit (NLTK), a powerful Python library for working with text.

Steps:

1. Install NLTK: **!pip install nltk**
2. Import the **word_tokenize** function from **nltk.tokenize**.
3. Create a variable ‘text’ that contains a paragraph of at least 4 sentences about any topic of your choice.
4. Use the **word_tokenize** function to tokenize the text into words.
5. Print the list of tokens.
6. Count the number of tokens in the text.
7. Identify the frequency of each token using **nltk.FreqDist** or other function.

Write the documentation to your code and upload it to GitHub.

Exercise 2. Word Cloud (Google Colab)

Objective: To learn how to generate and customize word clouds from text using Python.

Tools:

- Google Colab (colab.research.google.com)
- A text file (.txt) - You can upload this to your Google Drive.

Instructions:

1. **Create a new Colab Notebook:**

   Go to colab.research.google.com and create a new notebook.

2. **Install the wordcloud library:**

   In the first code cell, run the following command to install the necessary library:

```
!pip install wordcloud
```

3. **Import libraries:**

   In the next cell, import the required libraries:

```
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
```

4. **Upload your text file:**

   Upload your .txt file to your Google Drive (you can create a new folder for this exercise).

5. **Mount Google Drive:**

   Run the following code to mount your Google Drive in the Colab environment:

```
from google.colab import drive
drive.mount('/content/drive')
```

   Follow the authorization steps to allow Colab to access your Drive.

6. **Load the text file:**

   Replace "path/to/your/text_file.txt" with the actual path to your file in Google Drive:

```
with open("/content/drive/My Drive/path/to/your/text_file.txt", "r") as file:
    text = file.read()
```

7. **Create a WordCloud object:**

```
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color="white",
    stopwords=set(STOPWORDS),
    min_font_size=10,
).generate(text)
```

8. **Display the word cloud:**

```
plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```

9. **Observe and analyze:**
   - The word cloud should be displayed in the output of the cell.
   - Which words are the most prominent? Why do you think that is?

Challenge:

- **Customization:** Experiment with the WordCloud parameters in your Colab notebook to customize your word cloud. Try different:
  - background_color
  - colormap
  - max_words
  - stopwords
- **Mask:** Research how to use the mask parameter to create a word cloud in a specific shape. Upload a mask image to your Drive and use it in your code.

Use the code agents to improve your code and to help with documentation. Upload your code to GitHub.