Natural Language Processing Assignment 2

Report on Text Summarization:

1. The written analysis for Steps 1, 10, and 11 is included in this report.
2. The code implementation for Steps 2 to 9 is provided in the script file, with step-wise comments embedded. The following two files contain identical content but in different formats:
   a. The output of **nlp_assignment_2.py** is available in **nlp_output.txt**.
   b. The **nlp_assignment_2.ipynb** file includes both the code and output.

Prerequisites:

1. Required packages for running the script:
   pip3 install gensim==3.6.0
   pip3 install nltk
   pip3 install Wikipedia-API
2. Update gensim's **dictionary.py** file as mentioned in "***Problems Faced***" section.

Summarization methods explored in this assignment:

1. Manual Summary
2. Wikipedia Summary
3. NLTK Summary
4. GenSim Summary
5. Open source LLM Summary

Written work Steps:

Step 1: Analysing the Text and Summarizing manually:

I've selected a Wikipedia page on a historical site *Elephanta_Caves* for this assignment. While summarizing manually, I focused on extracting key facts and generalizing similar texts to capture the most relevant insights. I also eliminated redundant information to enhance readability and coherence.

Manual Summary:

The Elephanta Caves are a network of sculpted caves located on Elephanta Island in Mumbai Harbour, approximately 10 kilometers east of Mumbai, India. These caves are renowned for their rock-cut sculptures depicting Hindu deities, primarily dedicated to Lord Shiva. The island, originally known as Gharapuri, was renamed 'Elephanta' by Portuguese explorers in the 16th century due to a large stone elephant statue they found there.

Key Features of the Elephanta Caves:

1. Main Cave (Cave 1): This is the most significant cave, featuring a large hall with numerous sculptures, including the famous 7-meter-high Trimurti, a three-headed depiction of Shiva symbolizing his roles as creator, preserver, and destroyer.
2. Other Caves: There are several smaller caves on the island, some of which are incomplete or have deteriorated over time.

Historical Significance:

The exact origins of the caves are uncertain, but they are believed to have been constructed between the 5th and 8th centuries AD. The intricate carvings reflect the artistry and religious traditions of the time, providing insight into the cultural history of the region.

Preservation Efforts:

The caves have faced challenges over the centuries, including damage from natural elements and human activity. Recognizing their cultural importance, UNESCO designated the Elephanta Caves as a World Heritage Site in 1987, leading to increased conservation efforts to preserve this historical landmark.

Step 10: Check output with the output you wanted. See how far it matches:
After generating summaries using different methods, I compared the results with the manually crafted summary to assess their effectiveness.

Findings:

1. **NLTK Summarizer:** Extracted key sentences effectively but lacked deep contextual understanding. It sometimes included less relevant sentences based on frequency analysis.
2. **GenSim Summarizer:** Provided more concise summaries with improved coherence but occasionally omitted important details.
3. **Open Source LLMs:** Generated highly readable summaries with better contextual flow but were sometimes too verbose. Tried multiple values for tuneable parameters like chuck_size, max_length, min_length to acheive less verbose yet acceptable results.

Step 11: Now, feed similar text and check if summarizer works as you expect.
Observations:

To assess consistency, I applied the summarization methods to similar texts and evaluated their performance.

1. **NLTK:** Performed well with structured texts but struggled with complex sentence structures.
2. **GenSim:** Effectively condensed information while maintaining coherence.
3. **Open Source LLMs:** Adapted well to different texts, generating fluent summaries but requires fine-tuning for conciseness.

Conclusion:

Each summarization method has its own strengths and limitations:

1. **NLTK:** Efficient for quickly extracting key points but lacks deep contextual understanding.
2. **GenSim:** Strikes a balance between conciseness and coherence but may overlook some details.
3. **LLMs:** Best for human-like summaries but requires careful tuning.

For large-scale summarization tasks, a hybrid approach that integrates LLMs with extractive summarization techniques could yield the best results.

Problems faced:

There were multiple issues with the gensim package. After debugging the issue, I discovered that in gensim version 3.6.0, modifying the import statement in dictionary.py resolved the problem. After below modification, everything worked as expected.

Fix:

Existing code:

```
from collections import Mapping, defaultdict
```

Changed to:

```
from collections import defaultdict
from collections.abc import Mapping
```

File Path: /Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/sitepackages/gensim/corpora/dictionary.py