Deep Learning 2

Section - 1

Overview

The code is divided into two main sections: importing Kaggle data and processing it to summarize text using natural language processing techniques.

1. Data Import and Setup

1. **Imports and Initialization**:
   - Libraries such as os, sys, shutil, urllib, and others are imported for file handling, downloading, and extracting data.
   - Constants like CHUNK_SIZE and DATA_SOURCE_MAPPING are set up for downloading data.
2. **Download and Unpack Data**:
   - The code sets up paths and symlinks for Kaggle data.
   - It downloads files from URLs provided in DATA_SOURCE_MAPPING and extracts them based on their type (ZIP or TAR).

2. Text Processing and Summarization

1. **Library Imports and Setup**:
   - Libraries such as numpy, pandas, nltk, gensim, scipy, and networkx are imported for data manipulation, tokenization, word embedding, and graph-based operations.
2. **Load and Clean Data**:
   - A CSV file with medium articles is read into a DataFrame.
   - The DataFrame is cleaned by removing duplicates and unnecessary newline characters.
3. **Text Summarization**:
   - **Preprocessing**: Sentences are tokenized, cleaned of punctuation, and stopwords are removed.
   - **Word Embeddings**: A Word2Vec model is trained on the tokenized sentences to get word vectors.

   - **Sentence Embeddings**: Each sentence is converted to an embedding by averaging the word vectors of its words.
   - **Similarity Matrix**: A matrix is created to store the cosine similarity between each pair of sentence embeddings.
   - **Graph Construction**: A graph is built from the similarity matrix, and PageRank is used to rank sentences based on their importance.
   - **Summarization**: The top-ranked sentences are selected to create a summary.

4. **Generate Summary Function**:
   - A function generateSummary is defined to summarize text based on the same process used earlier.
5. **Save Summarized Data**:
   - The script writes the summaries into a CSV file with columns for title, summary, and content.
   - A callback function applies generateSummary to each row of the DataFrame and writes the result to the CSV.

Key Points:

- **Data Import**: Ensures data is downloaded and available in the right format.
- **Text Processing**: Involves cleaning, tokenizing, and embedding sentences.
- **Summarization**: Uses a combination of word embeddings and PageRank to extract important sentences.
- **CSV Output**: Writes the processed summaries back to a CSV file.