

Description of the code

The script performs several tasks related to processing Jeopardy questions, renaming files, normalizing text, loading data, and interacting for reranking search results.

- **File Renaming (rename_files function):**
 - The rename_files method searches a defined directory (directory_path) for files with a specific extension (.txt).
 - It renames each file according to a predefined pattern, adding a numerical suffix to the original name.
 - The renaming process is used to avoid overwriting existing files.
- **Text Processing (advanced_normalize_text function):**
 - The advanced_normalize_text function accepts a text input and performs the following steps:
 - Changes the text to lowercase.
 - Removes all non-alphanumeric characters.
 - Tokenizes the text into individual words.
 - Uses NLTK to remove stopwords (common words such as "the," "and," and so on).
 - Each word is lemmatized into its fundamental form.
- **Loading Jeopardy Questions (load_jeopardy_questions function):**
 - Reads a file with Jeopardy questions (given by the filename variable).
 - Parses the file and stores three consecutive lines in a dictionary with the keys 'category,' 'clue,' and 'answer.'
 - Returns a list of dictionaries that represent Jeopardy questions.
- **Loading Wikipedia Pages (load_wikipedia_pages function):**
 - Reads many files containing Wikipedia page data from a given directory (directory_path).
 - Uses a regular expression (page_pattern) to extract titles and content from files.
 - Returns a list of dictionaries with the keys 'title' and 'content', which represent Wikipedia pages.
- **Indexing Wikipedia Pages in Elasticsearch (index_wikipedia_pages function):**
 - Indexes a list of Wikipedia pages into Elasticsearch.
 - The advanced_normalize_text function is used to preprocess and normalize the material prior to indexing.
- **Search and Rerank with ChatGPT (search_and_rerank_with_chatgpt function):**
 - Searches Elasticsearch for relevant pages using a Jeopardy clue.
 - Creates a prompt for ChatGPT, instructing it to rank the retrieved Wikipedia pages.
 - Interacts with OpenAI's GPT-3.5 Turbo model to produce reranked results.
 - Manages API rate limits by pausing between every two queries.
- **Evaluating Precision at 1 (evaluate_match_es_for_p_at_1 function):**
 - Takes a list of Jeopardy questions and calculates precision at 1 by comparing the top-ranked answer from Elasticsearch to the correct answer.
 - Prints details about each question, including the top-ranked answer, the correct answer, and if the match is correct.

- Calculates and publishes the overall precision as 1.
- **Main Function (main function):**
 - Creates an Elasticsearch connection with a specified host.
 - Sets the index name for Elasticsearch.
 - Loads Jeopardy questions via the `load_jeopardy_questions` function.
 - To evaluate precision at one, call the function `evaluate_match_es_for_p_at_1`.
- **Execution:**
 - If executed as the main module, the script executes the main function, which completes the workflow.

The script automates the process of importing Jeopardy questions, indexing Wikipedia pages into Elasticsearch, and using Elasticsearch and ChatGPT to rerank and evaluate search results.

Output

The code's principal output is an evaluation of precision at 1 for reranked search results using Elasticsearch and ChatGPT. The script displays information for each Jeopardy question, such as the Jeopardy clue, the top-ranked answer from the search, the correct answer, and if the match was correct. Furthermore, the script prints the overall precision as 1.

What can be expected in the output:

- Details for each Jeopardy question:
 - Jeopardy clue: Use the top-ranked answer from a search (e.g., Elasticsearch or GPT-3.5 Turbo model).
 - Correct answer and match status (correct or incorrect).
- Overall Precision at One:
 - The script calculates precision at 1 by determining the fraction of right answers among all questions.
 - The result is displayed as "Precision at 1 with Reranking: [precision_value]."
- Pause Information:
 - The script may print warnings indicating that it is halting for a set amount of time to handle API rate constraints, particularly when communicating with the GPT-3.5 Turbo model.

Because the script is intended for measuring precision at 1 in a specific use case, the detailed output sheds light on the effectiveness of the reranking procedure. When you execute the script, carefully examine the printed information to determine how successfully the system matches Jeopardy clues with right answers.