**This project is a Python-based text analysis pipeline designed to process cryptocurrency-related articles, extract meaningful insights, and provide answers to user queries. The pipeline begins with acquiring data from web articles, converting them into PDFs, extracting text, preprocessing it, vectorizing it for analysis, performing sentiment analysis, and retrieving ranked answers using the BM25 algorithm. The codebase is modular, with contributions from multiple authors, and leverages a variety of NLP and machine learning libraries.**

**OBJECTIVES:**

**- Scrape and gather cryptocurrency-related articles from the web.**

**- Extract and refine text from these articles for analysis.**

**- Analyze sentiment and identify cryptocurrency-specific entities.**

**- Deliver ranked, relevant answers to user queries based on the processed text.**

**DATA ACQUISITION**:

This step scrapes article URLs from the cryptocurrency news website, thenewscrypto.com for five specified coins: Dogecoin, Bitcoin, Ethereum, Solana, and Hamster. The script fetches up to 10 article links per coin from the first page of search results and saves them into individual CSV files (e.g., dogecoin\_news\_urls.csv). It incorporates error handling, logging, and a user-agent header to mimic a browser request, ensuring robust and reliable web scraping.

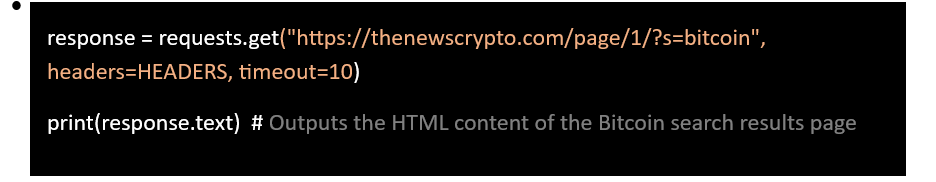
**Key Features**

* Scrapes only the first page of search results per coin.
* Limits extraction to 10 articles per page to maintain consistency.
* Uses a user-agent header to avoid being blocked by the website.
* Implements logging to track progress and errors.
* Saves results in CSV format with coin names and URLs.

**Libraries used:**

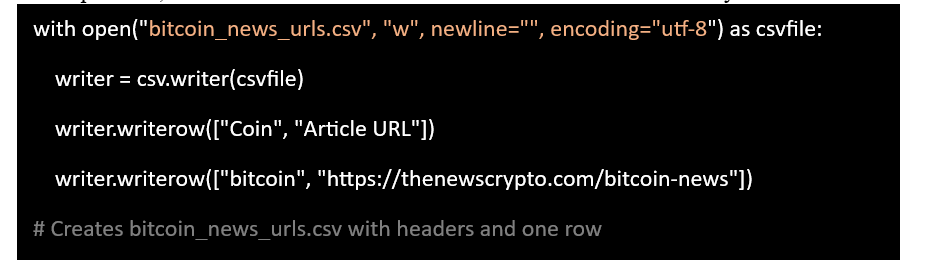
1. **Requests**: https://pypi.org/project/requests/

* It is used for sending the HTTP request to the website and then retrieving the HTML content of the search results.
* Used in the ***get\_article\_link()*** function to fetch the webpage content for a given coin and page number via ***requests.get(url, headers= HEADERS, timeout=10)***
* **Efficiency**: requests simplify HTTP interactions, abstracting away low-level socket programming and connection handling. This makes it highly efficient for fetching web pages compared to manual alternatives like **urllib.**
* **Reliability**: The library handles redirects, connection errors, and timeouts gracefully. The timeout=10 parameter ensures the script doesn’t hang indefinitely, while the try-except block catches ***RequestException*** errors (e.g., network failures).
* **Flexibility**: The ability to pass custom headers (e.g., ***HEADERS***) is crucial for avoiding anti-scraping measures like IP blocking, as many websites reject requests without a valid user-agent.
* Without requests, the script would require complex manual HTTP request construction, reducing readability and increasing the likelihood of errors. Its use ensures reliable page retrieval, critical for subsequent parsing.

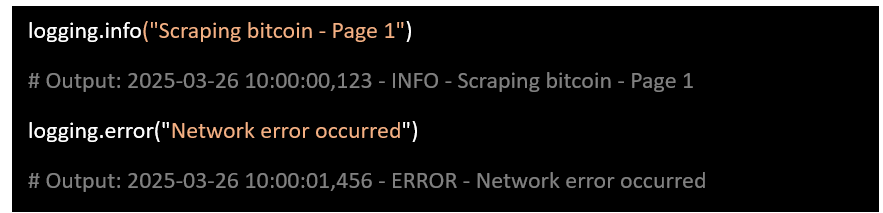


1. **CSV**: https://docs.python.org/3/library/csv.html

* For providing functionality to read and write CSV (Comma-Separated Values) files.
* Used in the ***scrape\_articles\_for\_coin()*** function to save scraped article URLs into a CSV file (e.g., dogecoin\_news\_urls.csv) with columns "Coin" and "Article URL" via ***csv.writer***.
  + **CSV Writer**: Creates a writer object to write rows to the CSV file.
  + **newline=""**: Ensures consistent line endings across platforms (e.g., Windows, Linux).
  + **Encoding**: Uses utf-8 to support special characters in URLs or coin names.
* **Simplicity**: csv provides a straightforward interface for writing tabular data, eliminating the need to manually format comma-separated strings. This reduces errors like missing delimiters or improper escaping.
* **Portability**: The library ensures the output CSV files are compatible with standard tools (e.g., Excel, pandas), making them reusable for further processing.
* **Robustness**: Handles edge cases like special characters or quotes in URLs automatically, thanks to utf-8 encoding and built-in escaping mechanisms.
* **Context in Script**: Without csv, the script would need to manually write comma-separated lines (e.g., f.write(f"{coin},{link}\n")), increasing complexity and error risk. Its use ensures clean, structured output files, critical for downstream tasks like PDF conversion or analysis

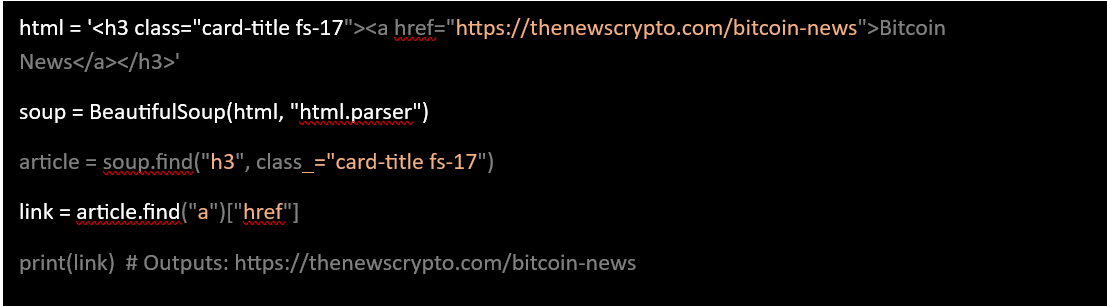
****

1. logging: <https://docs.python.org/3/library/logging.html>

* Enables logging of events (info, warnings, errors) during execution for debugging and monitoring.
* Log messages like "Scraping bitcoin - Page 1" or errors if the request fails.
* Configured at the start with ***logging.basicConfig()*** and used throughout to log scraping progress (e.g., ***logging.info***), warnings (e.g., ***logging.warning***), and errors (e.g., ***logging.error***).
  + **Level Setting**: Sets the logging level to INFO to capture all events.
  + **Format**: Customizes log messages with timestamp, level, and message ***(%(asctime)s - %(levelname)s - %(message)s).***
  + **Log Types**: Uses info for progress, warning for non-critical issues, and error for exceptions.
* **Debugging**: Provides a detailed trace of the scraping process (e.g., "Scraping bitcoin - Page 1"), making it easy to identify where failures occur (e.g., HTTP errors).
* **Monitoring**: Logs the number of articles found per page, helping verify the script’s success without manual inspection of output files.
* **Robustness**: Captures exceptions (e.g., network timeouts) with ***logging.error***, ensuring issues are documented rather than silently failing.
* Without ***logging***, debugging would rely on print statements, which lack timestamps and severity levels, making it harder to track issues in a long-running process. Its use enhances maintainability and reliability, especially for scraping multiple coins.
* **Example**

1. Bs4 [beautiful Soup]: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

* Used for scraping websites. And parsing the HTML content
* **Ease of Use**: ***BeautifulSoup*** simplifies HTML parsing with an intuitive API, allowing quick extraction of article links without writing complex regular expressions or DOM traversal logic.
* **Flexibility**: Handles malformed HTML gracefully, which is common on real-world websites, ensuring the script doesn’t break if the page structure varies slightly.
* **Precision**: The use of ***class\_="card-title fs-17***" targets specific article titles, reducing the risk of scraping irrelevant links (e.g., ads or navigation).
* Without ***bs4***, parsing HTML would require manual string manipulation or a less robust library like ***re***, increasing complexity and fragility. Its effectiveness lies in reliably extracting structured data from unstructured HTML, making it indispensable for web scraping.



1. **Text Extraction:**
2. trafilatura: https://trafilatura.readthedocs.io/en/latest/

A web scraping and content extraction library designed to fetch and extract clean, readable text from HTML pages, removing boilerplate (e.g., ads, navigation). Unlike general-purpose scraping tools like BeautifulSoup, trafilatura specializes in extracting the main article content efficiently, which is ideal for news articles.

* Used in ***process\_coin()*** to download article pages with ***trafilatura.fetch\_url(url)*** and extract their main content with ***trafilatura.extract(downloaded)***.
* **Key Features Utilized**:
  + **URL Fetching**: Downloads HTML content from a given URL.
  + **Content Extraction**: Strips away non-essential elements to isolate the main article text.

**Effectiveness**

* **Specialization**: Unlike general-purpose scraping tools like BeautifulSoup, trafilatura is optimized for extracting main article content, making it highly effective for news articles where boilerplate removal is critical. It uses heuristics and machine learning to identify primary content, reducing manual parsing effort.
* **Robustness**: Handles diverse webpage structures and gracefully returns None if downloading or extraction fails, allowing the script to continue without crashing. This is evident in the checks for downloaded and content.
* **Efficiency**: Combines fetching and extraction in a streamlined process, minimizing the need for separate HTTP requests (e.g., via requests) and parsing steps. This reduces latency compared to a requests + BeautifulSoup approach.
* **Context in Script**: Without trafilatura, the script would need to fetch pages with requests and manually parse HTML with BeautifulSoup, requiring custom logic to filter boilerplate. Its use ensures clean, relevant text extraction, crucial for generating meaningful PDFs for downstream analysis.

1. pdfkit: https://pypi.org/project/pdfkit/

Converts HTML files to PDF format using the wkhtmltopdf command-line tool (requires wkhtmltopdf to be installed separately).

D**ATA PREPROCESSING**

This step cleans, tokenizes, stems, and lemmatizes the text while preserving cryptocurrency-specific terms, preparing it for downstream tasks like vectorization or sentiment analysis. It processes the text sentence-by-sentence, provides progress updates, and saves the cleaned output to cleaned\_data.txt.

Key Features:

* Splits text into sentences using a simple newline split for efficiency.
* Applies comprehensive cleaning: lowercase conversion, special character removal, tokenization, stemming, and lemmatization.
* Uses a progress bar to monitor processing of large datasets.
* Preserves a predefined set of cryptocurrency terms to maintain domain-specific context.

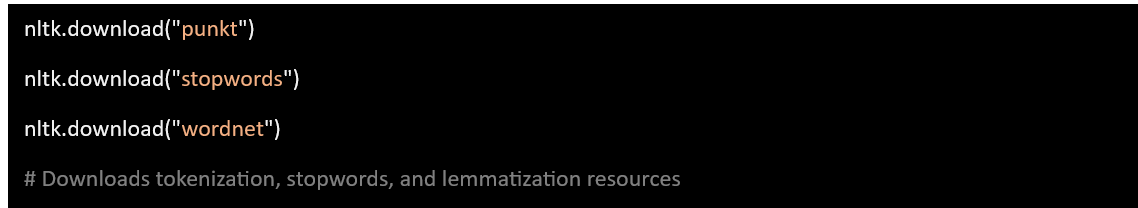
Libraries Used:

1. nltk (Natural Language Toolkit): https://www.nltk.org/

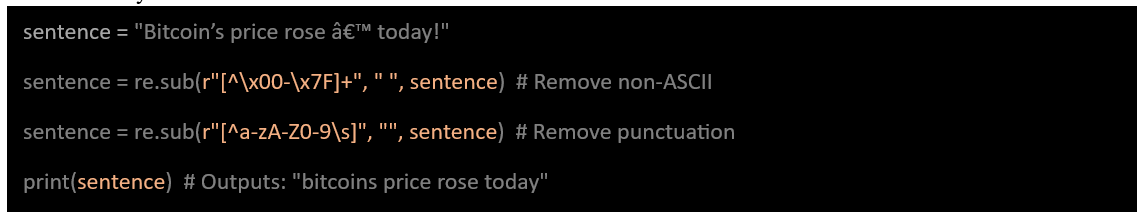
This library is used for tasks like tokenization, stopword removal, and lemmatization. It provides pre-built tools and resources (e.g., tokenizers, stopwords, lemmatizers) that simplify text preprocessing, avoiding the need to write these from scratch.

* Provides a broad suite of natural language processing tools and resources, serving as the backbone for tokenization, stopwords, and stemming/lemmatization in this script.
* Used indirectly through its submodules (***nltk.tokenize, nltk.corpus, nltk.stem***) and to download required resources (***punkt, stopwords, wordnet***) via ***nltk.download(***).
  + **Resource Downloads**: Ensures necessary datasets are available for tokenization and lemmatization.
  + **Foundation**: Acts as the parent library for specialized NLP tasks.
* **Versatility**: ***nltk*** is a comprehensive NLP library, offering pre-built tools that save time compared to custom implementations. Its modular design allows specific submodules to handle distinct tasks.
* **Reliability**: The downloaded resources (***punkt, stopwords, wordnet***) are well-tested and widely used, ensuring consistent preprocessing results.
* Without nltk, the script would require manual tokenization, stopword lists, and lemmatization logic, significantly increasing complexity. Its role is foundational, enabling the use of specialized submodules for efficient text processing.

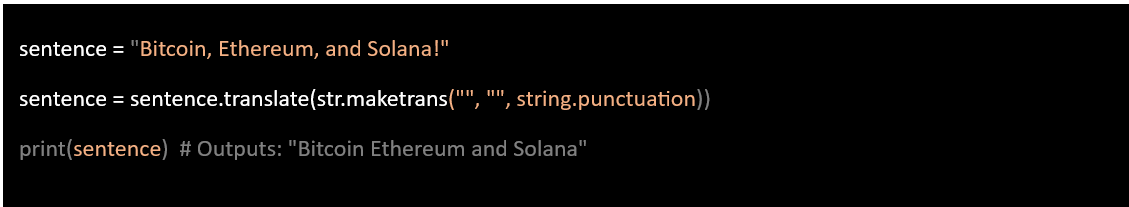
**Example**



1. **re:** https://docs.python.org/3/library/re.html

* Provides regular expression matching operations to remove unwanted characters from text.
* Used in ***preprocess\_sentence()*** to remove non-ASCII characters and punctuation via ***re.sub().***
  + **Pattern Matching**: Removes specific character sets (e.g., non-ASCII, non-alphanumeric) using regex patterns.
  + **Substitution**: Replaces matched patterns with spaces or empty strings.
* **Precision**: re allows fine-grained control over text cleaning, effectively targeting non-ASCII characters ***([^\x00-\x7F])*** and punctuation ***([^a-zA-Z0-9\s]).***
* **Speed**: Regular expressions are optimized for string operations, making them faster than iterative character checks for large texts.
* Without re, cleaning special characters would require slower, less flexible methods (e.g., multiple ***replace()*** calls). Its use ensures robust text normalization, critical for consistent tokenization and analysis.

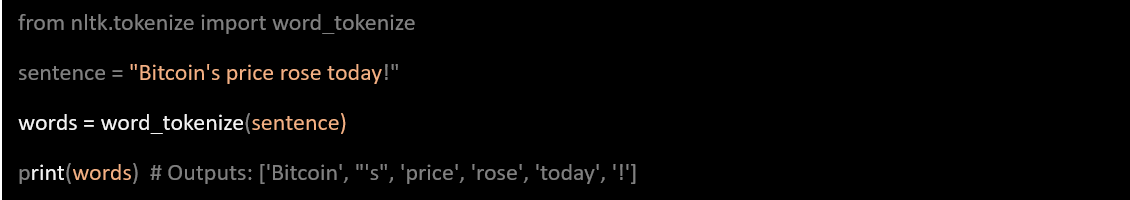
1. tring: <https://docs.python.org/3/library/string.html>

* Supplies string manipulation utilities, specifically for punctuation removal.
* **Usage in Code**: Used in preprocess.***sentence***() to create a translation table for removing punctuation via ***str.maketrans("", "", string.punctuation).***
  + **Punctuation Constants**: Provides ***string.punctuation***, a predefined set of punctuation characters ***(e.g., !"#$%&'()\*+,-./:;<=>?@[\]^\_{|}~`).***
  + **Translation Table**: Generates a mapping to strip punctuation efficiently.
* **Simplicity**: ***string.punctuation*** offers a ready-made list of characters to remove, avoiding the need to define them manually.
* **Efficiency**: The ***translate()*** method with a translation table is faster than ***regex*** or iterative replacements for punctuation removal.
* While re could handle punctuation removal, string provides a cleaner, more efficient alternative for this specific task. Its use complements ***re*** by focusing on punctuation, enhancing the overall cleaning process.

**4. nltk.tokenize: https://www.nltk.org/api/nltk.tokenize.html**

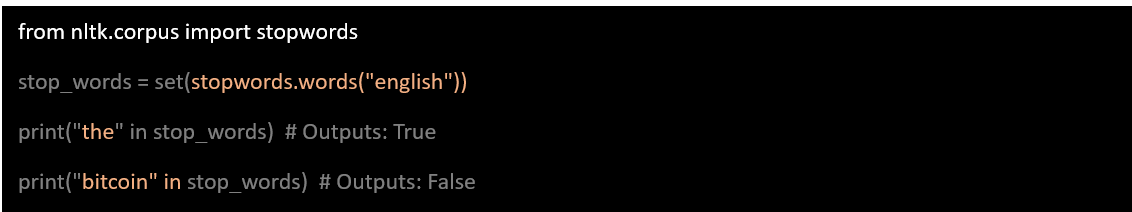
**Role**

* **Purpose**: Provides tokenization tools to split text into words or sentences.
* Used in preprocess\_***sentence()*** to tokenize a sentence into words via ***word\_tokenize(sentence)***.
  + **Word Tokenization**: Splits a sentence into individual words based on the punkt tokenizer.
* **Accuracy**: ***word\_tokenize*** leverages the punkt model, trained on diverse English texts, to accurately handle punctuation and contractions (e.g., "Bitcoin's" → ["Bitcoin", "'s"]).
* **Convenience**: Eliminates the need for custom tokenization logic, which would be error-prone and less robust (e.g., simple ***split()*** misses edge cases).
* Without ***nltk.tokenize***, the script might use ***split(),*** which fails on complex punctuation or multi-word crypto terms. Its effectiveness ensures precise word-level processing, critical for stemming and lemmatization.

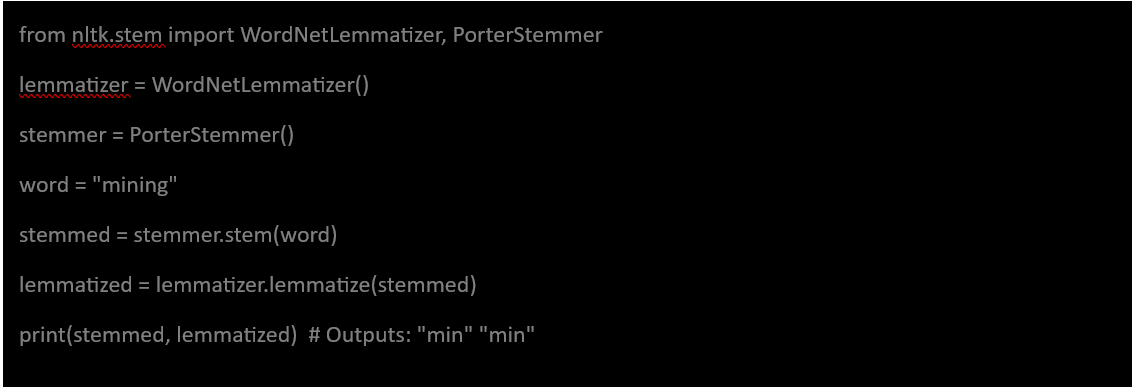


[**nltk.corpus**](https://www.nltk.org/api/nltk.corpus.html)

* Provides access to linguistic resources, such as a list of stopwords.
* Used to load English stopwords into a set via ***stopwords.words("english"***) in the setup.
  + **Stopwords**: Supplies a precompiled list of common English words (e.g., "the", "is", "and") to filter out.
* **Efficiency**: Loading stopwords into a set enables fast lookup during processing, crucial for large datasets.
* **Standardization**: Uses a widely accepted stopword list, ensuring consistency with NLP best practices.
* Without ***nltk.corpus,*** the script would require a custom stopword list, risking omissions or inconsistencies. Its use reduces noise in the cleaned text, improving the quality of downstream analysis.

**5. nltk.stem : https://www.nltk.org/api/nltk.stem.html**

* **Purpose**: Provides stemming and lemmatization tools to reduce words to their base forms.
* **Usage in Code**: Used in preprocess\_***sentence()*** to initialize and apply WordNetLemmatizer and ***PorterStemmer*** via ***stemmer.stem(word)*** and ***lemmatizer.lemmatize(stemmed\_word).***
* **Key Features Utilized**:
  + **PorterStemmer**: Applies rule-based stemming (e.g., "running" → "run").
  + **WordNetLemmatizer**: Uses WordNet to lemmatize words to their dictionary form (e.g., "better" → "good").
* **Normalization**: Combining stemming and lemmatization ensures words are reduced to consistent base forms, improving text similarity for analysis (e.g., "mining" and "mined" → "mine").
* **Precision**: ***WordNetLemmatizer*** leverages a lexical database for accurate base forms, while ***PorterStemmer*** handles simpler cases efficiently.
* Without ***nltk.stem***, words like "prices" and "pricing" would remain distinct, reducing the effectiveness of vectorization or retrieval. Its use enhances text consistency while preserving crypto terms.



S**entiment Analysis using Blob**

This step performs sentiment analysis on the sentences prepared, adding sentiment scores and categories (positive, negative, neutral) to the DataFrame. It then visualizes the distribution of sentiments across all sentences, providing insights into the tone of the crypto news articles.

The libraries used:

1. **textBlob: https://textblob.readthedocs.io/en/dev/**

A simple NLP library that provides tools for tasks like sentiment analysis, part-of-speech tagging, and text processing.

It offers an easy-to-use interface for sentiment analysis, calculating polarity (positive/negative) and subjectivity (objective/subjective) without requiring complex model training.

Analyzes the sentiment of raw sentences using **TextBlob(x) sentiment, polarity,** and subjectivity**.**

Used in ***analyze\_crypto\_sentiment()*** to create a ***TextBlob*** object and compute polarity via ***blob.sentiment.polarity.***

 **Simplicity**: Requires no model training, making it ideal for quick sentiment analysis compared to complex libraries like transformers.

 **Accuracy**: Uses a pre-trained pattern-based analyzer, effective for general English but may miss nuanced crypto-specific sentiment (e.g., "mining" as positive or negative).

 **Context in Script**: Without ***textblob,*** sentiment analysis would require a custom model or external API, increasing complexity. Its ease of use makes it effective for rapid prototyping and visualization.

*Example: TextBlob("Dogecoin prices rose today!").sentiment returns:*

* + *polarity: ~0.25 (positive, as "rose" suggests growth).*
  + *subjectivity: ~0.2 (fairly objective).*

*Category: positive (since 0.25 > 0.1).*

1. **matplotlib.pyplot**: https://matplotlib.org/3.5.3/api/\_as\_gen/matplotlib.pyplot.html
   * **Purpose**: A plotting library for creating visualizations like bar charts, line graphs, etc. Used to plot sentiment distribution with ***plt.bar()*** in the main execution.

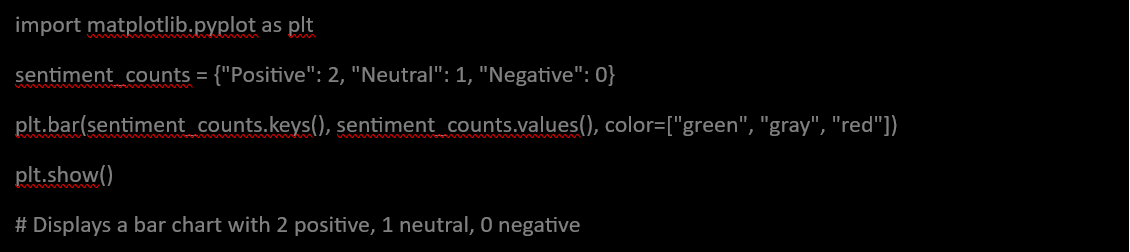
It’s widely used for quick, customizable data visualization, ideal for showing sentiment distribution.

Generates a bar chart to display the count of sentences in each sentiment category.

Visualizations: Counts all categories (e.g., 50 positive, 30 neutral, 20 negative) and plots them in a bar chart.

* + **Visualization Example**:
    - Bar chart with:
      * Green bar (positive)
      * Gray bar (neutral)
      * Red bar (negative)
  + Without ***matplotlib***, sentiment distribution would require manual counting or external software, reducing interactivity. Its use enhances insight generation from sentiment data.

**Example:**



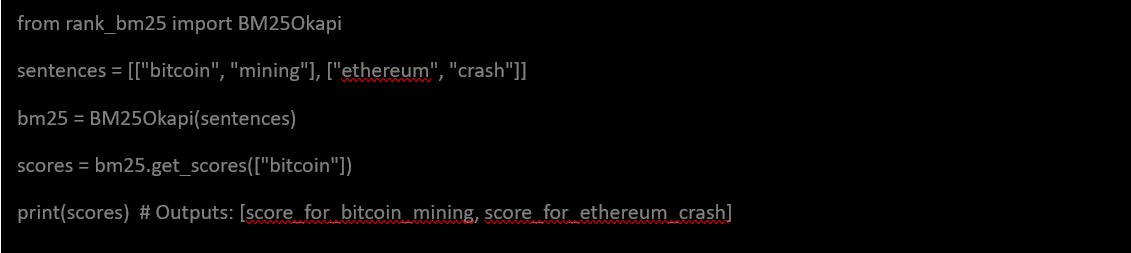
**Implementation of BM25**

This step creates a searchable index of preprocessed sentences using the BM25 algorithm, processes user queries, and retrieves the most relevant sentences as answers. It integrates sentiment data from Step 5 and provides a ranked list of responses with metadata.

Libraries used:

1. **rank\_bm25**: https://pypi.org/project/rank-bm25/
   * Implements the BM25 (Best Matching 25) algorithm, a ranking function used in information retrieval to score document relevance based on query terms.
   * BM25 is efficient and effective for text search, balancing term frequency and document length, making it ideal for retrieving relevant sentences without needing a complex machine learning model.
     + B**M25Okapi:** Creates an index from tokenized sentences and scores them against queries. Used in ***CryptoAnswerRetrieval*** to initialize a ***BM25Okapi*** index with tokenized sentences and score them against a query via ***self.bm25 = BM25Okapi()*** and ***bm25.get\_scores().***

The process followed:

1. **Creating the BM25 Index (Function: create\_bm25\_index)**:
   * Convert preprocessed sentences into a tokenized corpus.
   * Initialize a ***BM25Okapi index*** with the corpus.
2. **Query Preprocessing (Function: preprocess\_query)**:
   * Apply the same preprocessing to the user’s query to ensure consistency with the indexed data.
3. **Searching with BM25 (Function: search\_with\_bm25)**:
   * Preprocess the query and tokenize it into words.
   * Compute relevance scores for all sentences using the BM25 index.
   * Add scores to a copy of the input DataFrame, sort by score, and filter to the top N results.
   * Warn if the highest score is below a threshold (0.05), indicating poor matches.
   * Return the top N results.
4. **Execution**:
   * Create the BM25 index from ***sentiment\_df.***
   * Test with an example query ("What is the latest trend in Bitcoin price?").
   * Display the top 3 results with source, score, sentiment, and raw sentence.