

Documentation for Code in main.py

Output: **The lesser the distance. The more relevant the article is to the query given.**

```
Search results: data: [{"id": 6, "distance": 0.0, "entity": {"title_id": 6}}, {"id": 5, "distance": 1.1799043416976929, "entity": {"title_id": 5}}, {"id": 3, "distance": 1.3024373854584395, "entity": {"title_id": 3}}, {"id": 2, "distance": 1.5484514236458195, "entity": {"title_id": 2}}, {"id": 4, "distance": 1.6702942848205566, "entity": {"title_id": 4}}]
Title IDs: [6, 5, 3, 2, 4]
Title: Navigating the changing landscape of BTK-targeted therapies for B cell lymphomas and chronic lymphocytic leukaemia
Abstract: The B cell receptor (BCR) signalling pathway has an integral role in the pathogenesis of many B cell malignancies, including chronic lymphocytic leukaemia, mantle cell lymphoma, diffuse large B cell lymphoma and Waldenström macroglobulinaemia. Bruton tyrosine kinase (BTK) is a key node mediating signal transduction downstream of the BCR. The advent of BTK inhibitors has revolutionized the treatment landscape of B cell malignancies, with these agents often replacing highly intensive and toxic chemotherapy regimens as the standard of care. In this Review, we discuss the pivotal trials that have led to the approval of various covalent BTK inhibitors, the current treatment indications for these agents and mechanisms of resistance. In addition, we discuss novel BTK-targeted therapies, including covalent, as well as non-covalent, BTK inhibitors, BTK degraders and combination doublet and triplet regimens, to provide insights on the best current treatment paradigms in the frontline setting and at disease relapse.
```

Overview

This script automates the process of scraping articles from the specified website, storing them in a MySQL database, generating text embeddings, storing the embeddings in a Milvus vector database for efficient retrieval, and performing similarity searches. Each part of the process is modularized into functions for easy understanding and maintenance.

Step-by-Step Explanation

1. Configuration Loading

- The script reads configuration settings from a config.ini file using the configparser module.
- The config.ini file is expected to store MySQL and Milvus connection details, which are then loaded and used throughout the script.

2. Database Connection: MySQL

- MySQL Database Connection Setup: Configures MySQL database connection using settings from config.ini.
- This is used in various functions to save, retrieve, and manage article data.

3. Milvus Connection

- Connecting to Milvus: Establishes a connection to a Milvus server using parameters from the configuration file.
- Milvus is a vector database used here for efficient similarity search of embedded article titles.

4. Function save_to_database

Purpose: Saves article details into the MySQL database, specifically into the journals table.

Process:

Connects to MySQL.

Checks the current maximum id in the journals table to assign a sequential id to each new record.

Inserts the article's title, authors, published_date, and abstract as a new record.

Error Handling: Catches MySQL errors and ensures that connections are properly closed.

5. Function get_article_details

Purpose: Extracts article details from a specific article URL and calls save_to_database to store them.

Process:

Fetches the page using requests.

Parses the HTML to extract:

Title

Authors

Published Date

Abstract

Calls `save_to_database` to store these details.

6. Function `get_latest_research_urls`

Purpose: Scrapes a main URL to find links to the latest research articles.

Process:

Fetches the main page.

Finds the "Latest Research and Reviews" section.

For each article link found, calls `get_article_details` to retrieve and store its details.

Structure: Uses a fixed base URL (<https://www.nature.com>) and concatenates it with relative links for complete URLs.

7. Function `recreate_milvus_collection`

Purpose: Ensures that the Milvus collection for storing article embeddings is set up.

Process:

Checks if the `journal_titles` collection exists and deletes it if it does.

Defines the schema for a new collection with two fields:

`title_id`: ID of the title, primary key.

`title_embedding`: 384-dimensional vector field to store embeddings.

Creates a new collection in Milvus based on this schema.

8. Function `embed_and_store_in_milvus`

Purpose: Generates embeddings for each article title and stores them in Milvus.

Process:

Retrieves all titles from the MySQL journals table.

For each title:

Generates an embedding using SentenceTransformer.

Stores both the title ID and embedding in Milvus.

Output: Displays the ID and title being stored, as seen in the output logs you shared.

9. Function create_index

Purpose: Builds an index on the title embeddings in Milvus to speed up similarity searches.

Process:

Defines indexing parameters.

Creates an index on the title_embedding field using IVF_FLAT with L2 distance metric.

Loads the collection into memory for fast searching.

10. Function search_articles

Purpose: Searches for articles similar to a given query based on title embeddings.

Process:

Generates an embedding for the query using SentenceTransformer.

Uses this embedding to search for the most similar articles in the Milvus collection.

Retrieves the top K matches and then fetches their details from MySQL by matching IDs.

Output:

Displays the query embedding and found title IDs.

Prints details of the top matching articles, including titles and abstracts.

11. Execution in main Block

Data Retrieval: Calls `get_latest_research_urls` to fetch and store the latest articles.

Milvus Setup and Embedding:

Calls `recreate_milvus_collection` to set up the Milvus collection.

Calls `embed_and_store_in_milvus` to store title embeddings.

Calls `create_index` to index embeddings.

Search Example: Runs a sample query using `search_articles`.

Explanation of the Provided Output

The output provides feedback during the embedding and search processes:

Embedding Storage Messages:

"Storing title: [Title] with ID: [ID]": Shows that the script is generating and storing embeddings for each title in the database, confirming which title (and corresponding ID) is processed.

Embedding Storage Confirmation:

"Embeddings stored in Milvus": Indicates all title embeddings have been successfully stored in the Milvus vector database.

Query Embedding and Search Results:

"Query embedding: [First few values]...": Shows the embedding generated from the search query to help confirm the embedding model's output.

"Search results: data: [...], Title IDs: [IDs]": Lists the IDs of the top matching titles from the search. The distance values indicate similarity, with lower values being more similar.

Retrieved Articles:

Each "Title: [Title] \nAbstract: [Abstract]" section shows a matching article found in the search, with its title and abstract. This provides the user with context and detail on similar articles based on the original search query.

```
Storing title: Prostaglandin E2-EP2/EP4 signaling induces immunosuppression in human cancer by impairing bioenergetics and ribosome biogenesis in immune cells with ID: 6
Embeddings stored in Milvus.
Query embedding: [0.019746916368603706, 0.06938496977090836, -0.044301752001047134, -0.06438110768795013, 0.06508477032184601]...
Search results: data: [{"id": 6, "distance": 0.0, "entity": {"title_id": 6}}, {"id": 5, "distance": 1.1799043416976929, "entity": {"title_id": 5}}, {"id": 3, "distance": 1.3024373054504395, "entity": {"title_id": 3}}, {"id": 2, "distance": 1.5484514236450195, "entity": {"title_id": 2}}, {"id": 4, "distance": 1.6702942848205566, "entity": {"title_id": 4}}]
Title IDs: [6, 5, 3, 2, 4]
Title: Navigating the changing landscape of BTK-targeted therapies for B cell lymphomas and chronic lymphocytic leukaemia
Abstract: The B cell receptor (BCR) signalling pathway has an integral role in the pathogenesis of many B cell malignancies, including chronic lymphocytic leukaemia, mantle cell lymphoma, diffuse large B cell lymphoma and Waldenström macroglobulinaemia. Bruton tyrosine kinase (BTK) is a key node mediating signal transduction downstream of the BCR. The advent of BTK inhibitors has revolutionized the treatment landscape of B cell malignancies, with these agents often replacing highly intensive and toxic chemoimmunotherapy regimens as the standard of care. In this Review, we discuss the pivotal trials that have led to the approval of various covalent BTK inhibitors, the current treatment indications for these agents and mechanisms of resistance. In addition, we discuss novel BTK-targeted therapies, including covalent, as well as non-covalent, BTK inhibitors, BTK degraders and combination doublet and triplet regimens, to provide insights on the best current treatment paradigms in the frontline setting and at disease relapse.

Title: Differences in axillary response and treatment implications in HER2 positive node positive breast cancer during neoadjuvant HER2 targeted dual therapy
Abstract: Explore whether the axillary outcomes differ among HER2 positive subgroups receiving standard dual-targeted therapy, aiming to identify subgroups exhibiting enhanced sensitivity to NAT among HER2-positive/node-positive breast cancer patients. HER2 positive female patients with biopsy-proven node-positive disease from April 2020 to May 2023 were included. All patients underwent standard Neoadjuvant HER2-targeted dual therapy and axillary lymph node dissection (ALND) at Breast Surgery Center of Sichuan Cancer Hospital. Univariate and multivariate analyses were used to identify factors associated with axillary pathological complete response (ApCR). Statistical analysis and graphing were performed using SPSS 24.0 and GraphPad Prism 9.0 software. This study enrolled 215 HER2 positive patients with a total ApCR rate of 76.7%, which included 49 HER2 2+/FISH + and 166 HER2 3 + cases with approximate ApCR rates of 63.3% and 80.7% (P = 0.011). Univariate and multivariate analysis indicated that HER2 3 + disease (OR = 2.43, 95% CI 1.21-4.88, P = 0.012), Ki-67 ≥ 20% disease (OR = 3.00, 95% CI 1.26-7.13, P = 0.013) and NAC regimen of TCB (OR = 2.71, 95% CI 1.39-5.38, P = 0.004) were more likely to achieve ApCR. Further subgroup analysis revealed that HER2 3 + patients receiving TCB regimen showed the highest ApCR rate of 88% compared to other subgroups. HER2 3 + breast cancer had a higher ApCR rate than HER2 2+/FISH + breast cancer during Neoadjuvant HER2-targeted dual therapy. HER2 positive patients could benefit from NAC regimen of TCB in axillary response.

Title: Risk factors for local recurrence following marginal mandibulectomy in gingival cancer
Abstract: Surgery is the first line of treatment in gingival cancers of the mandible, and bone resection is necessary in the majority of cases. In the less extensive surgical option, marginal mandibulectomy (MM), the mandibular base is preserved. In contrast, in a segmental mandibulectomy (SM) the mandible is divided and the continuity is not preserved. If MM can be performed with comparable oncological results to SM, it is the preferred method. The aim of the present study was to identify preoperative predictors for local recurrence (LR), to support the selection of candidates for MM. Outcome measures were local recurrence free survival (LRFS) and disease specific survival (DSS). 67 patients treated with MM between 2008 and 2021 were included. Cox regression analyses of LR with hazard ratios and adjustments for postoperative radiotherapy, pathological T-stage (pT) and soft tissue margins were performed. 5-years LRFS was 63% (95% CI 46.9-75.5) and DSS 80.6% (95% CI 64.7-89.9). In conclusion we found that edentulous patients, more advanced pT-stage and positive soft tissue margins
```

The lesser distance. The more relevant the article is to the query given.