Duplicate Detection Documentation

This document explains the functions, parameters, and usage of Duplicate Detection

# Functions Overview

|  |  |  |  |
| --- | --- | --- | --- |
| Function | Parameters | Return | Description |
| record\_post(post\_id, text, topic=None) | post\_id (str): unique ID text (str): content topic (str, optional): group label | None | Create or update a post and store its embedding in SQLite. Recomputes the embedding if text is updated. |
| delete\_post(post\_id) | post\_id (str): ID to delete | bool | Deletes a post and its embedding. Returns True if something was deleted. |
| similar\_posts(query\_text, top\_k=10, min\_score=0.45, topic=None) | query\_text (str): search query top\_k (int): number of results (default 10) min\_score (float): similarity threshold (default 0.45) topic (str, optional): restrict to a topic | List[(post\_id, score)] | Embeds the query and returns top-k similar posts with cosine similarity scores. |
| batch\_fill(topic, posts) | topic (str): group label posts (iterable): list of (post\_id, text) | None | Adds all posts in the batch, computes embeddings only for missing ones. Useful for bootstrapping a topic. |

# Step-by-Step Usage

1. Clone or copy the project folder structure:

.  
 ├── dupdet  
 │ ├── \_\_init\_\_.py  
 │ ├── batch.py  
 │ ├── config.py  
 │ ├── delete.py  
 │ ├── embedder.py  
 │ ├── record.py  
 │ ├── search.py  
 │ └── storage.py  
 └── run\_tests.py

**2. Create and activate a virtual environment on the server:**  
 python -m venv .venv  
 source .venv/bin/activate

**3. Install dependencies:**

pip install -U llama-index-core

pip install -U llama-index-embeddings-huggingface

pip install -U transformers

pip install -U numpy

**then**

python -m pip install --upgrade pip wheel

python -m pip install --extra-index-url https://download.pytorch.org/whl/cpu "torch==2.1.2"

python -m pip install "llama-index-embeddings-huggingface==0.2.0"

python -m pip install "transformers==4.43.3" "accelerate<1.0" sentencepiece "huggingface-hub>=0.19.0”

**4-Set up Hugging Face access**  
Since access to google/embeddinggemma-300m is already approved for the project,

Use this command: “huggingface-cli login”

Then you have to enter the TOKEN I’ll provide you with the **(Hugging Face** **account email password and token in teams)**

**5. Run the test script:**

**-First test**: “python3 -m dupdet.test”  
 -**second test**:”python3 test\_each\_function.py”