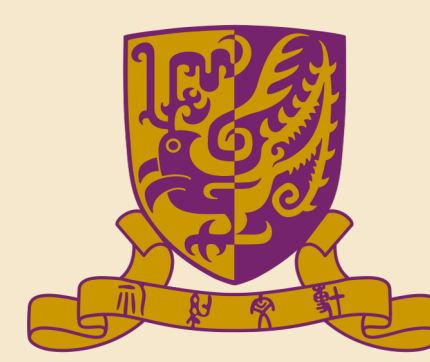




Data Analytics Practice Opportunity 2024/25

Lü Zu Quan Shu Topic Retrieval Engine

呂祖全書主題搜尋器

CHIU Chi Chung (AIST/4), CHAN Yuk Chuen (AIST/4), IP Chun Shing (AIST/4)

OBJECTIVE

Studying Ancient Chinese texts has long been a challenging task due to their concise nature and the absence of grammatical markers such as tense, articles, and prepositions. However, this field remains an interesting and inspiring area of research, as it holds vast philosophical wisdom and rich literary and artistic value waiting to be uncovered. Therefore, by leveraging the digital resources of the 呂祖全書 (Lü Zu Quan Shu), we aim to develop a topic retrieval system designed to assist anyone interested in exploring this hidden gem.

METHODOLOGY

今古仙佛
哀愍衆生

Text Segmentation

Dataset Preparation and Preprocessing

- **Adding Punctuations:** Augments sentence and paragraph boundaries for clarity.
- **Re-documenting:** Converts the text into coherent sentences and paragraphs.
- **Data Cleaning:** Removes unwanted punctuations (e.g., “@”, “】”) and extraneous symbols for consistency.

Segmentation Method

- **Volume-Based Division:** Text is divided by volumes (e.g., marked by “卷二”) to preserve the document's inherent structure.
- **Algorithmic Segmentation:** Computational algorithms identify logical sentence or paragraph boundaries after the volume-based split.
- **AI-Based Refinement:** AI algorithms enforce paragraph word limits (e.g., under 256 words) to ensure consistency and optimal chunk sizes for machine learning.

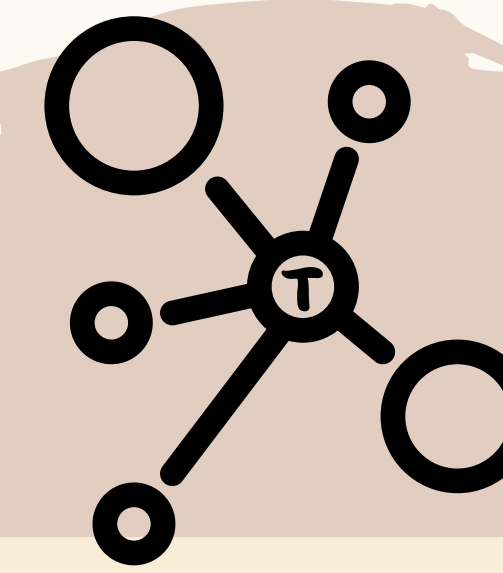
Visualization

Text Embedding

- Each paragraph from classical Daoist texts is embedded into a high-dimensional vector space.
- The embedding process uses the pre-trained model “BAAI/bge-large-zh-v1.5” to generate 1024-dimensional vectors.
- This representation captures subtle semantic nuances, positioning similar paragraphs closer in the vector space.

Visualization in 3D space

- High-dimensional embeddings are projected into 3D space using **TensorBoard** for interactive exploration.
 - **Features of the visualization panel:**
 - Interactive Navigation: Users can zoom, rotate, and pan to explore spatial relationships.
 - Search and Highlight: Users can search and highlight specific paragraphs for easy identification.
 - Similarity Exploration: Users can select a paragraph to locate the k most similar paragraphs based on Euclidean or cosine distance.
- The system improves cross-text searches and enhances the accuracy of topic retrieval in classical texts.



Topic Retrieval

Embedding and Topic Retrieval

- Text chunks are converted into numerical embeddings using a **deep learning model**.
- **Similarity scores** between the user query and text chunks are computed.
- The top three most relevant text chunks are retrieved based on the query.
- Retrieved chunks are traced back to their corresponding text segments for final results.

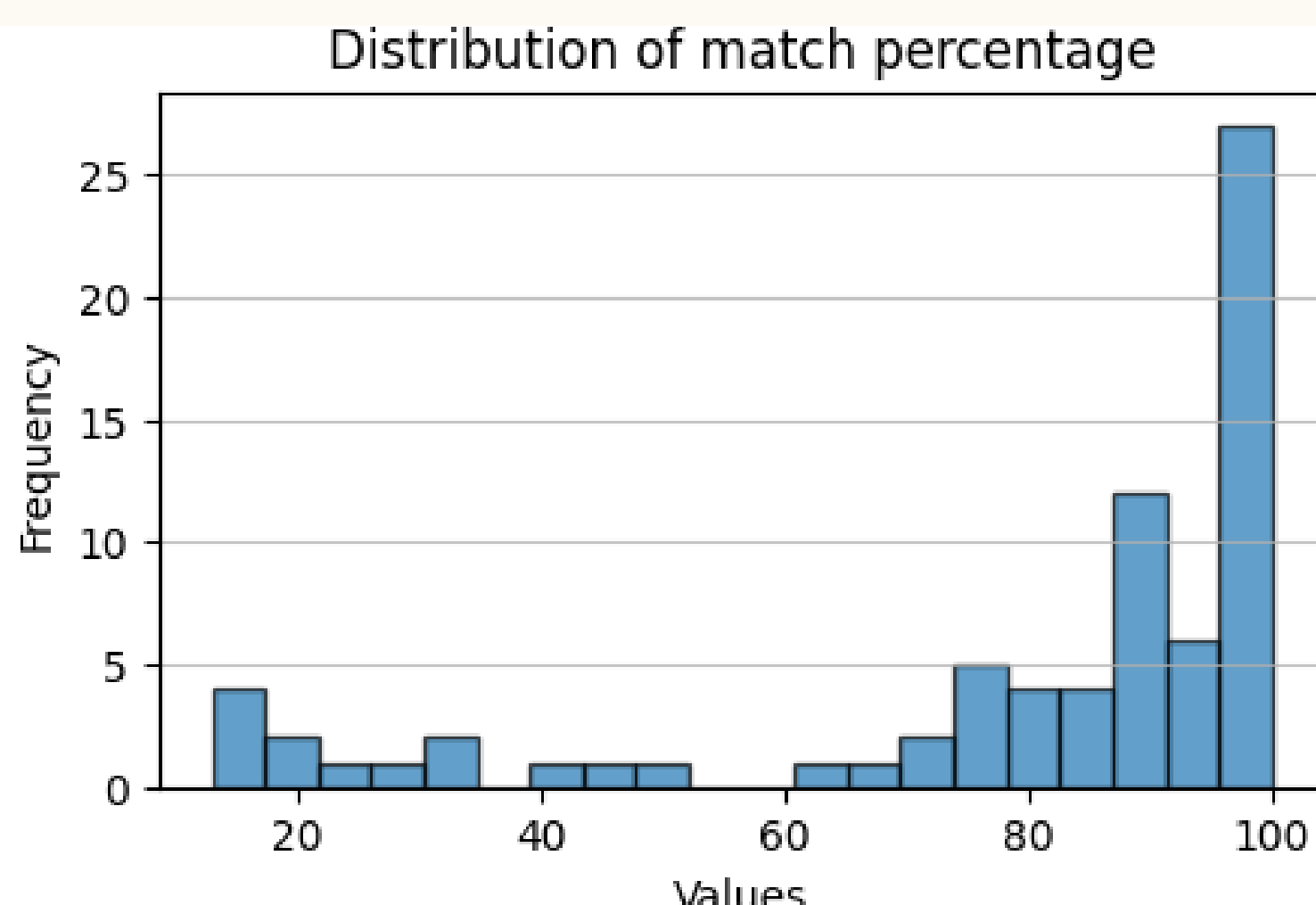
Evaluation Procedure

- A dataset of 25 annotated quotes and 53 descriptions from 《呂祖全書》 was used for evaluation.
- Steps in Evaluation:
 - **Retrieval:** The system retrieves the top three most relevant text chunks for each topic query.
 - **Segment Traceback:** Retrieved chunks are mapped to their original text segments.
 - **Similarity Evaluation:** Matching accuracy is measured using the Dynamic Time Warping (DTW) algorithm, ignoring punctuation differences.

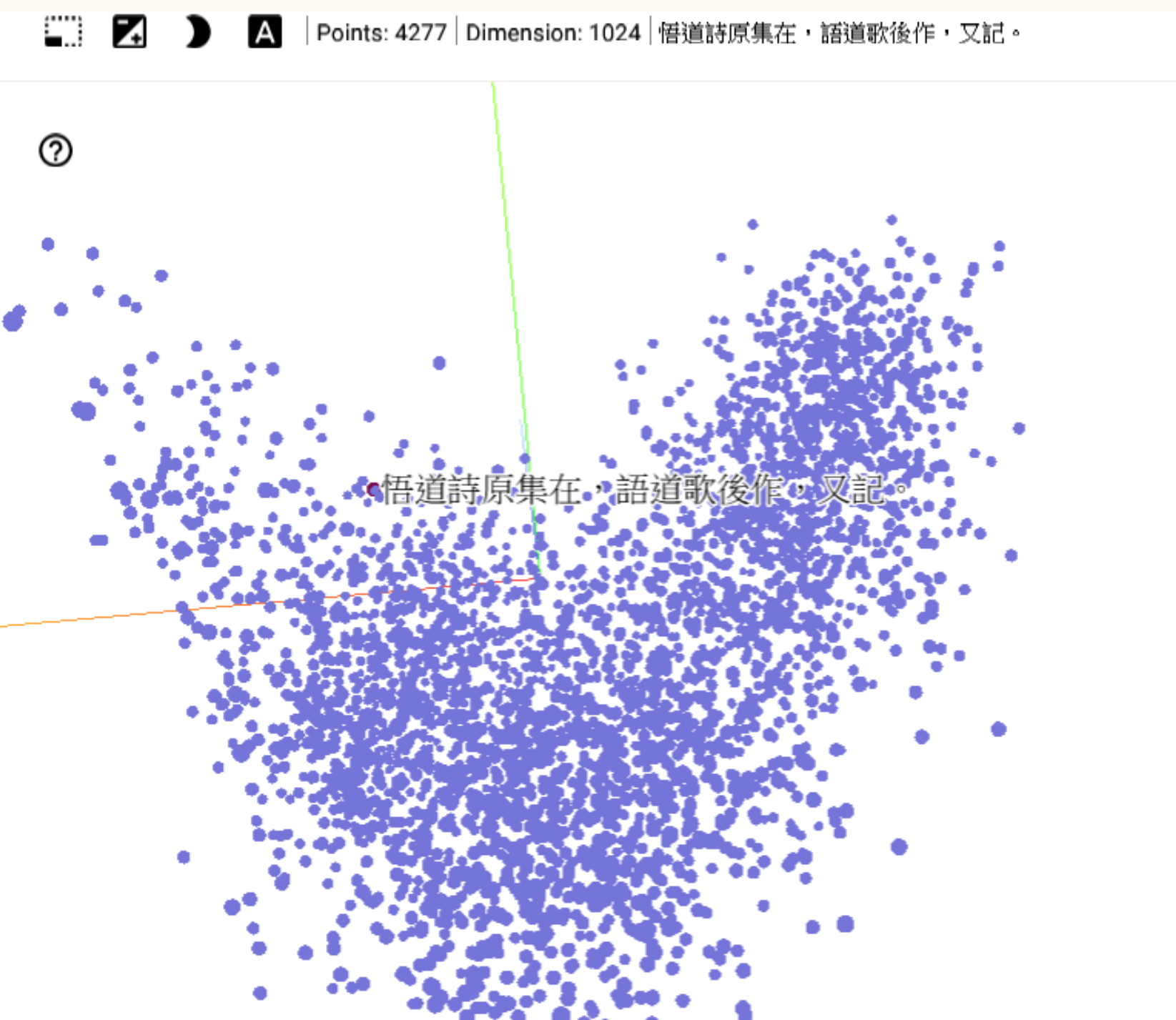
PROJECT OUTCOME

Text Segmentation

- Total Evaluation Segments: 75
- Exact Inclusions: 25 segments
- **Accuracy** (80% or above inclusion): Approximately **70.67%**

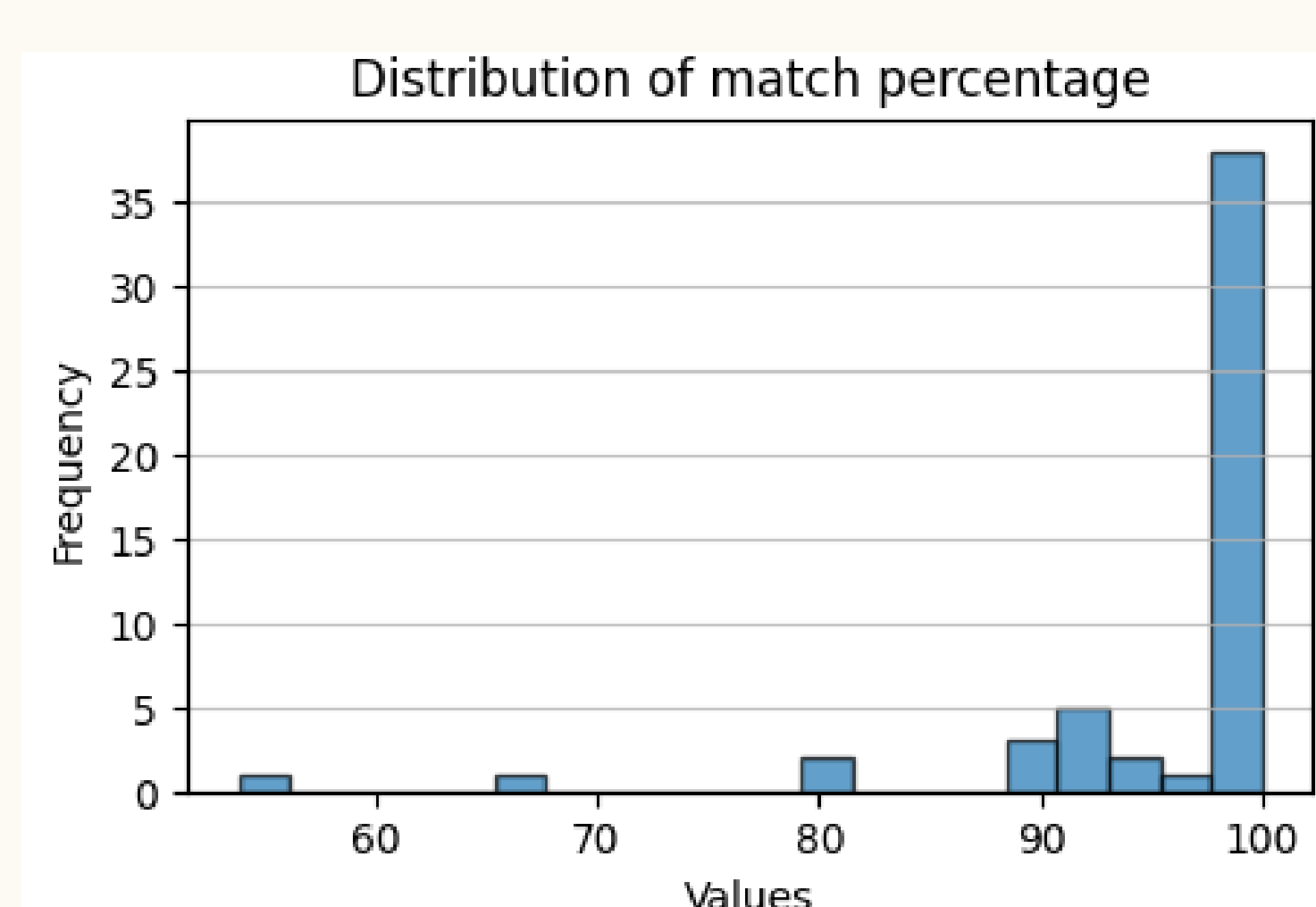


Visualization



Topic Retrieval

- **37** segments showed an exact inclusion of the ground truth.
- **More than 96%** of the segments contained **80% or more** of the ground truth text.



Interactive User-Interface

- **Ebook-Style Reading**
- **Retrieval by Topic or Description**
- **Search Capabilities:**
 - Text-Based Search
 - Volume and Ordinal-Based Search
 - Fuzzy Search Support
- **User Feedback Mechanism**
 - Users' feedback is used to further refine the model, allowing it to evolve continuously to meet actual user needs and interactions.

Interactive User Interface Screenshot:

卷: 卷七 段: 4

主題: 修真 原文:

下一句 跳至 原文搜尋 主題搜尋 回饋

選擇要回饋的段落: 但當年問過應尚有秘密口訣，不能筆之於書者，所謂... 分數 (10分為最高): 0

以下是與「修真」相關的原文：

1: 但當年問過應尚有秘密口訣，不能筆之於書者。所謂口口相傳，不記文，今皆不可得見矣。原本多訛，今將外音嘴錯者改正。其餘闕疑，分十八篇為上下二卷。論真仙第一。呂祖曰，人之生也，安而不病，壯而不老，生而不死，何道可致如此，鍾祖曰，人之生，自父母交會，而二氣相合，即精血為胎胞。於太初之後，而有太質，陰承陽生，氣，隨胎化，三百日，形圓，靈光入體，與母分離。自太素之後，已有升降，而長黃芽五千日，氣足，其數自滿八十一丈，方當十五，乃曰童男。是時陰中陽半，可比東日之光，過此以往，走失先陽，耗散真氣，氣弱則病老死絕矣。平生愚昧，自損靈光，一世兒頑，時除壽數。所以來生，身有等殊，壽有長短，既生復滅，既滅復生，轉轉不悟，而世世墮落，則失身於異類，透靈於別殼。至真之根性，不復於人，旁道輪迴，永無解脫。或遇真仙至人，與消其罪報，除皮脫殼，再得人身，方在癡。癡音因愚昧之中，橫行百劫，升在福地，猶不免饑寒殘患。進音以進音里，升遷漸得完全形貌尚居奴隸卑賤之中。苟或復作前孽，如立板走丸，再入旁道輪迴。--- 來自《卷二九》



More about this project: <https://dsprojects.lib.cuhk.edu.hk/en/projects/daoist/introduction/>
If you would like to reuse the data, please contact data@cuhk.edu.hk