Project Plan: Analyzing Degrowth Concepts in WTO Documents Using FAISS

Objective

Identify the extent to which Degrowth principles are discussed (implicitly or explicitly) in a set of WTO documents on global food trade, despite possible differences in terminology.

Approach

Convert the documents into hierarchical vector embeddings stored in a local FAISS database. This allows for semantic searching at multiple granularity levels (document, section, paragraph, and sentence) to capture Degrowth-related concepts even without direct keyword matches.

Steps

1. **Prepare the Documents**
   - **Extract Text from PDFs**: Use Python libraries like PyMuPDF or pdfplumber to convert each document into text
   - **Implement Hierarchical Text Chunking**:
     - Split documents into logical sections based on headers
     - Further split sections into paragraphs
     - Split paragraphs into sentences
     - Maintain metadata about document structure and locations
2. **Generate Hierarchical Embeddings**
   - **Set Up an Embedding Model**: Use Sentence-BERT (or similar) for semantic embeddings
   - **Create Multi-level Embeddings**:
     - Generate embeddings for full documents
     - Generate embeddings for sections
     - Generate embeddings for paragraphs
     - Generate embeddings for sentences
   - **Store Location Metadata**: Track source document, page numbers, section numbers for each embedding
3. **Set Up Multi-Level FAISS Index**
   - **Install FAISS**: `pip install faiss-cpu`
   - **Create Separate Indices**:
     - Document-level index for broad matching

    - Section-level index for thematic matching
    - Paragraph-level index for detailed context
    - Sentence-level index for specific mentions
- **Implement Index Management**: Create system to query and filter across index levels

4. **Create Queries and Hierarchical Search**
    - **Generate Query Embeddings**: Convert Degrowth principles into query vectors
    - **Implement Multi-Stage Search**:
        a. Find most relevant documents
        b. Within those documents, identify relevant sections
        c. Within sections, locate specific paragraphs and sentences
    - **Result Organization**: Structure results to show context hierarchy
5. **Document and Analyze Findings**
    - Create structured output showing matches at each level
    - Provide context and location information for each match

Example Use Case

Query: "Find content related to the degrowth principle of sufficiency"

Expected Output Format:


```
json
{
"query": "degrowth principle of sufficiency",
"results": [
{
"document": {
"title": "WTO Trade Policy Review 2023",
"relevance_score": 0.85
},
"relevant_sections": [
{
"section": "Sustainable Development Goals",
"page": 45,
"relevance_score": 0.92,
"relevant_paragraphs": [
{
```

"text": "The WTO is committed to the right to food and right to water...",
"page": 46,
"relevance_score": 0.95
}
]
}
]
}
]
}

Tools Needed

- **Python Libraries**:
  - PyMuPDF (or pdfplumber) for PDF processing
  - Sentence-BERT for embeddings
  - FAISS for vector indexing
  - spaCy for text splitting and processing
- **Hardware**: Local machine with sufficient RAM for multiple indices

Expected Outcomes

- Hierarchical analysis of WTO documents showing relevance at multiple levels
- Ability to trace Degrowth concepts from document level down to specific sentences
- Contextual understanding of how Degrowth principles appear in different sections

Implementation Notes

1. **Text Chunking Strategy**
   - Use document structure (headers, sections) for logical splitting
   - Maintain minimum chunk sizes for meaningful embeddings
   - Preserve context in chunk metadata
2. **Index Management**
   - Implement efficient filtering between index levels
   - Cache frequently accessed results

- Consider using FAISS's hierarchical navigable small world (HNSW) index for larger datasets

3. **Result Presentation**
   - Provide context breadcrumbs for each match
   - Include relevance scores at each level
   - Enable drilling down from document to specific sentences

This enhanced approach allows for more precise and contextual analysis of how Degrowth principles appear in WTO documents, with the ability to trace concepts from broad documents down to specific sentences.