AI-Driven Analysis of BC Post-Secondary Institution Strategy Documents

Helping consultants and policymakers extract strategic insights from institutional documents using AI

Current Scope (MVP Focus)

Analyzing institutional strategies manually is time-consuming. Our AI solution automates insight extraction

To build a focused and high-impact MVP, we targeted four key PDF documents per institution:

- Strategic Plan
- Financial Statement
- Government Mandate Letter
- Courses List

Total: 40 documents across 10 institutions

Each document was parsed page-by-page and tagged with standardized metadata for institution name, document type, and source

Decision-makers can identify trends across institutions through an interactive user interface

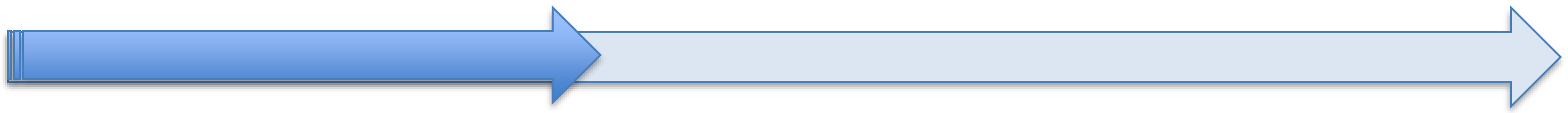
Progress

PHASE 1 – INGESTION PIPELINE

PHASE 2 – BUILD KNOWLEDGE GRAPH

PHASE 3 – SEMANTIC SEARCH & EMBEDDING

PHASE 4 – CONSULTANT-FACING SEMANTIC INTERFACE



Phase 1 – Ingestion Pipeline (Completed)

PDFs loaded using LangChain’s PyPDFLoader

Each page treated as an individual content unit

Custom metadata tagging applied:

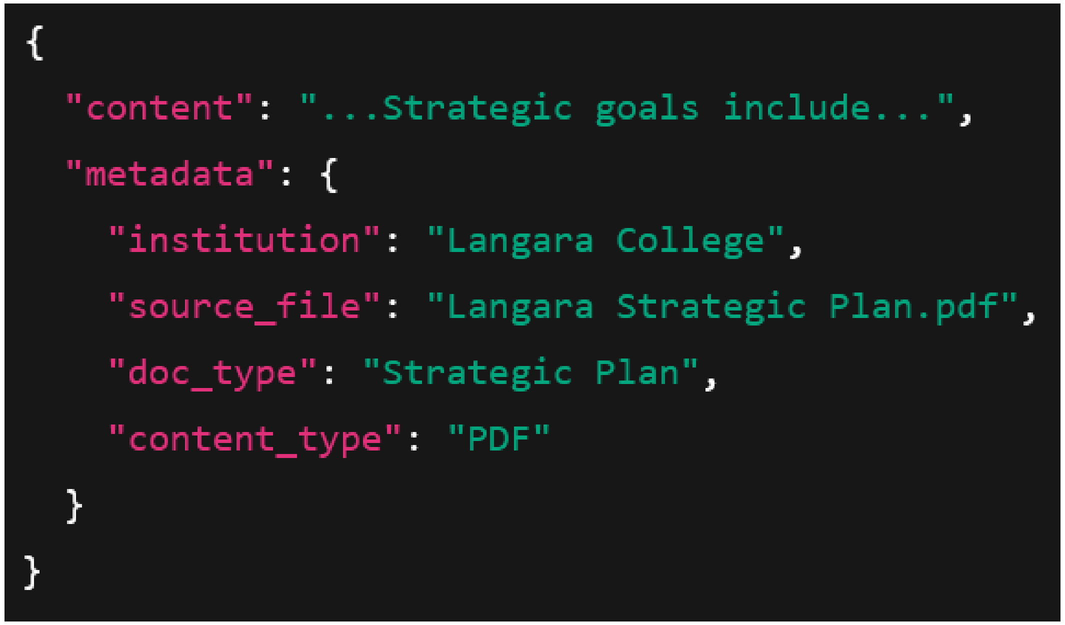
- institution
- source_file
- doc_type
- content_type

documents.json – full text and metadata per page

pdf_metadata.json – summary metadata only

Low-value pages are skipped using a hybrid approach: text density thresholds and keyword-based filtering, balancing **efficiency and accuracy** while ensuring valuable content isn't lost.

Sample Data Snapshot



```
{
  "content": "...Strategic goals include...",
  "metadata": {
    "institution": "Langara College",
    "source_file": "Langara Strategic Plan.pdf",
    "doc_type": "Strategic Plan",
    "content_type": "PDF"
  }
}
```

- Clean and structured for downstream use
- Enables fast search, QA, or filtering
- Supports graph and semantic search applications

Phase 2 – Knowledge Graph (In Progress)

Goal:

- Extract (subject, predicate, object) triples from raw document text using OpenAI GPT3.5 Turbo

Examples:

- Selkirk College, offers, Renewable Energy Diploma
- BCIT, partnered_with, Indigenous Communities

Tools:

- spaCy for named entity extraction (NER)
- OpenAI / OpenRouter LLMs for triple generation
- Output saved as triples.jsonl

Quality Assurance:

- evaluating knowledge graph accuracy by comparing extracted triples against a reference dataset

LLM Pipeline (In Progress)

- We use OpenAI GPT3.5 Turbo for relationship extraction
- Each page is chunked (up to ~2000 characters) and sent to the LLM with structured prompts
- Extracted triples are returned in JSON format and tagged with source metadata
- Deloitte has provided $62 worth of OpenRouter tokens, but access was unsuccessful for unknown security reasons, so OpenAI token used instead.
- Efficient use of prompts is required to stay within budget
  - Skipping short or low-value pages

Phase 3 – Semantic Search & Embedding (Next)

Planned Next Steps:

- Chunk pages into ~500-token segments
- Embed using OpenAI or Mistral models
- Store in FAISS or Pinecone vector database

FAISS vs. Pinecone:

- FAISS for local efficiency
- Pinecone for managed scalability

Purpose:

- Enable consultants to ask natural questions like:
- *“Which colleges offer AI-related diplomas?”*
- *“What sustainability goals exist across institutions?”*

Phase 4 – Consultant-Facing Semantic Interface

Streamlit-Powered QA Interface:

Simple UI to type questions about BC institutions

LLM-backed retrieval pipeline fetches best answers

Includes citations to source PDFs and institutions

Inlcudes a simple feedback mechanism (e.g., thumbs-up/down for answers)

Benefits:

Fast, explainable answers

Document traceability

Strategic insight engine for RFPs and client work

Technical Stack Summary

- Tools Used So Far:
  - LangChain (PDF loading, RAG pipeline)
  - spaCy (Named Entity Recognition)
  - OpenRouter (LLM access)
  - Python 3.10 (venv, CLI)
  - JSON (document storage, triplet output)
  - Ready for Neo4j, FAISS, or Streamlit to be decided for UI