SHL Assessment Search System: Technical Approach

Overview

This system provides semantic search capabilities for SHL assessment data, allowing users to find relevant assessments based on natural language queries with additional filtering by job level, language, and duration. The solution consists of three main components:

1. **Data Collection Pipeline**: Web scraper to extract assessment data from SHL's website
2. **Data Ingestion Pipeline**: Processes raw data and indexes it in a vector database
3. **Search API**: FastAPI service enabling semantic and filtered search through the data

Data Collection

The system uses Selenium to scrape SHL's product catalog, extracting:

- Assessment titles and URLs
- Descriptions
- Job levels (e.g., Entry Level, Mid Level)
- Supported languages
- Assessment duration

The scraper navigates through paginated results and visits each assessment's detail page, storing results in both CSV and JSON formats for subsequent processing.

Data Ingestion Pipeline

The ingestion pipeline transforms raw assessment data into a searchable index:

1. **Metadata Extraction**: Uses Groq's llama3-70b-8192 LLM to parse and normalize:
    - Job levels into standardized categories
    - Languages (removing regional indicators)
    - Duration in minutes from text descriptions
2. **Embedding Generation**: Converts assessment descriptions to vector embeddings using the BAAI/bge-small-en-v1.5 sentence transformer model.

3. **Vector Database Storage**: Stores processed data in Qdrant vector database via LlamaIndex, enabling:
    - Semantic similarity search
    - Structured metadata filtering
    - Hybrid search capabilities

Search API

The FastAPI service provides two main endpoints:

- POST /query: Advanced query interface with structured request body

Key features:

- **Natural Language Understanding**: Extracts implied filters from user queries using LLM
- **Hybrid Search**: Combines vector similarity with metadata filtering
- **Metadata Filtering**: Supports filtering by:
    - Job levels (case and hyphen insensitive)
    - Programming/natural languages
    - Duration (minimum and maximum)
- **Result Ranking**: Returns results sorted by relevance score

Technical Stack

- **Web Scraping**: Selenium, Python
- **Vector Embeddings**: Sentence Transformers (all-MiniLM-L6-v2)
- **Vector Database**: Qdrant
- **LLM Integration**: Groq API (Mixtral model)
- **Search Framework**: LlamaIndex
- **API Framework**: FastAPI with Pydantic models
- **Data Processing**: Pandas, NumPy

This architecture creates a robust, semantically-aware search system that helps users find the most appropriate SHL assessments based on their specific requirements and natural language descriptions.