AIQuest: Navigating My Data Cosmos

A Personal Search Engine

Project Overview:

The semantic search engine project aims to create a personalized search interface that allows users to search through their personal data takeouts efficiently. Additionally, users can execute targeted queries to retrieve specific information from their data collections. By leveraging embeddings generated using the OpenAI Embeddings API and utilizing the Pinecone vector similarity search service, users will be able to find relevant information and execute custom queries on their collected data.

Project Scope:

The project will cover the following aspects:

- Integration with personal data sources, including Gmail, LinkedIn,Instagram and other supported services.
- Text-based search engine focusing on email content, messages, and posts.
- User-friendly search interface accessible via a web application.
- Custom query functionality to retrieve specific data based on user-defined criteria.

User Stories and Use Cases:

1. As an end user, I want to search my emails for keywords to quickly find relevant information.
2. As an end user, I want to retrieve past LinkedIn messages related to a particular topic.
3. As a user, I want to retrieve data based on some other specific attributes (e.g. Retrieve all connection requests received on LinkedIn on a specific date).
4. As an administrator, I want to monitor system performance and ensure data privacy.

Data Collection and Preprocessing:

Personal data will be obtained through data takeouts from supported services (e.g., Gmail, LinkedIn). Data preprocessing includes:

- Cleaning and removing irrelevant content.
- Tokenization and stemming for text data.

- Removing anything other that ascii characters.

Embeddings Generation:

The OpenAI Embeddings API will be used to convert preprocessed text data into dense vector embeddings. The following parameters will be used:

- Model: GPT-3.5-turbo
- Max Tokens: 512
- Temperature: 0.7

Pinecone Integration:

Pinecone will be used to index and search vector embeddings efficiently. Embeddings will be batched and upserted to Pinecone for indexing and semantic search.

Query Processing:

1. User queries will be converted into embeddings using the Embeddings API.
2. These embeddings will be sent to Pinecone for semantic search.
3. Top results from Pinecone will be retrieved and displayed to the user.

Custom Query Functionality:

1. Users will have the option to execute custom queries for specific data.
2. Queries can include filters based on date, sender, keywords, and other parameters.
3. Custom queries will be processed by sending queries to data via postman and filtering results based on user criteria.

Performance and Scalability:

- Regular performance testing will be conducted to ensure fast response times.
- Pinecone's scalability features will be leveraged to handle increased user queries.

Testing:

- Unit testing for data preprocessing and embeddings generation.
- Integration testing for Pinecone indexing and semantic search.
- API endpoints are tested using Postman.

Future Enhancements:

- Interactive User Interface including filters.
- Integration with additional data sources.
- Support for multimedia content search.

- Natural language processing improvements (depending on OpenAI).
- Enhanced user customization of search parameters and custom queries.

Conclusion:

The semantic search engine project aims to empower users with a powerful search tool for efficiently retrieving valuable information from their personal data. By leveraging embeddings and vector similarity search, along with custom query capabilities, the project provides an innovative solution to the challenge of organizing and accessing personal data from various sources while enabling users to execute specific data retrieval tasks.