SMART INDIA HACKATHON - 2024

INQUEST.ai

Semantic Search in Unstructured and Semi-Structured Data

- **Problem Statement ID:** SIH1600
- **Problem Statement Title:** Student Innovation from
- **Organization:** Ministry of Education's Innovation Cell (MIC)
- **Theme:** Smart Automation
- **PS Category:** Software
- **Team ID:** 16841
- **Team Name:** SLASH 6

SLASH6

SLASH6

INQUEST.ai

SMART INDIA HACKATHON 2024

Semantic Search in Unstructured and Semi-Structured Data

PROPOSED SOLUTION

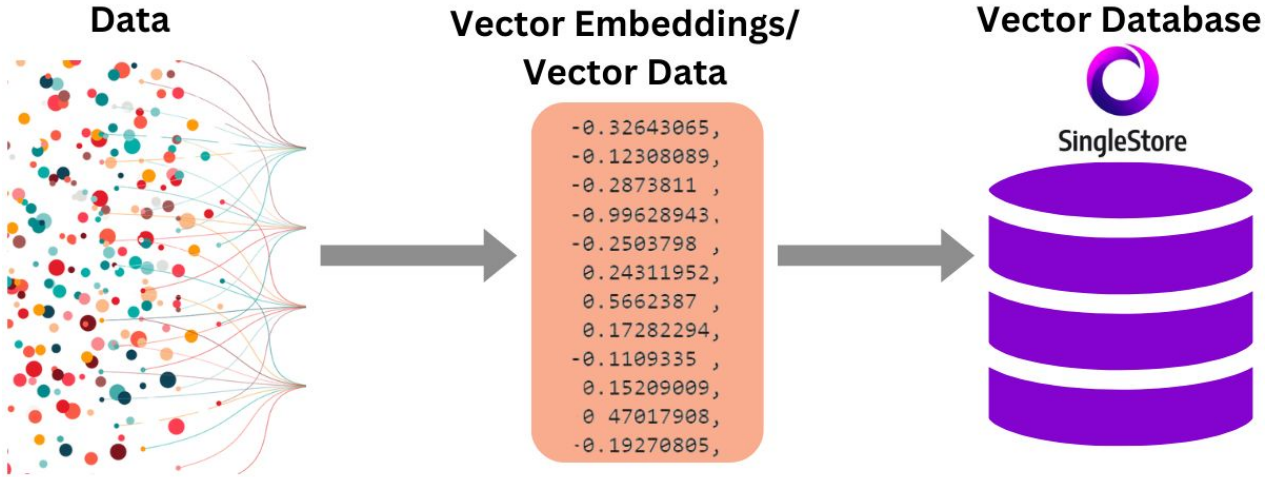
- To build a generalized AI-driven research engine that can handle both structured and unstructured data and convert it into vector embeddings, allowing for powerful semantic search capabilities across domains.
- In order to attain contextually relevant search we intend on using advanced technologies like FAISS or ElasticSearch.

PROTOTYPE

- The prototype integrates raw data and builds a vector database.
- It leverages FAISS/ElasticSearch for semantic search, ensuring fast and efficient retrieval of results based on embeddings.
- The user interface allows for natural language querying, returning relevant and ranked results in real time.
- This research engine will offer fast and accurate search results, making it a versatile tool for research in any field.

UNIQUENESS IN APPROACH?

- Domain - Agnostic and usage across diverse databases
- Contextually relevant and Semantic Search
- Dynamic Schema of Databases
- Scalable
- User-Friendly and Customizable
- Self Learning and Improving



TECHNICAL APPROACH

TECHNOLOGIES USED:

- Primary Language will be Python with C++ build dependency for GPU (CUDA) support.
- Hardware component like GPU for vector embedding with CUDA support.
- Libraries such as PyTorch, Ultralytics, Torchvision .
- Databases such as PostgreSQL, MongoDB and ChromaDB.
- Frameworks to build Interfaces Django, FastAPI and Streamlit.
- Other core python dependencies and libraries such as NumPy, Pandas, SpaCy, NLTK etc

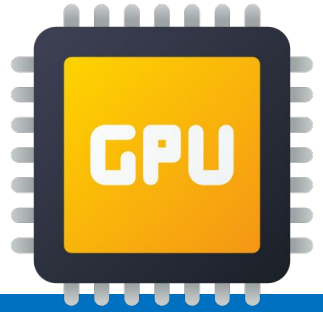
WORKFLOW

- Initial data setup from various structured and unstructured data sources.
- Data is ingested and preprocessing is done and it is converted to a Vector Database.
- Heuristic Configuration and Setup of domain specific custom preferences.
- Execution of contextual search to obtain semantically relevant results.
- Display results and refinement and filtering of desired results.

FEASIBILITY AND VIABILITY

FEASIBILITY ANALYSIS

- Data Agnosticism
- HIghly Scalable Mechanism
- Computationally affordable
- Improved research and referencing
- Saves TIme and Manual Labour



POTENTIAL CHALLENGES

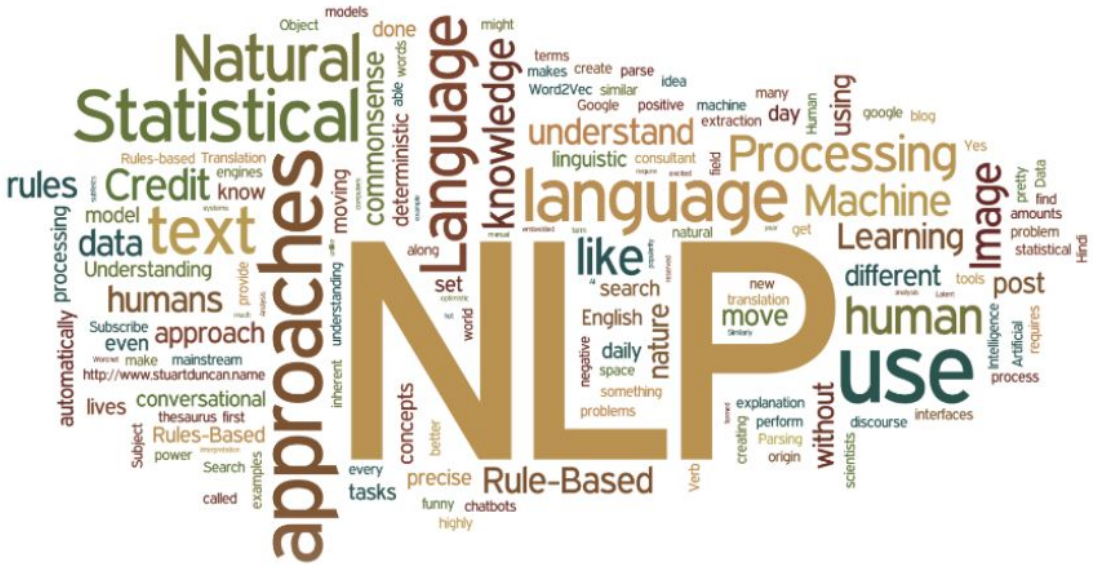
- Availability of quality, consistent and not noisy data
- Multilingual Support
- User Adoption
- Data Privacy
- Security threat handling with rigorous testing

SLASH6

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2024

- ➔ Domain Independent Application and Usecases
- ➔ Enhanced Efficiency
- ➔ Scalable and Inexpensive
- ➔ Improved Decision Making
- ➔ Widespread Adoption
- ➔ Reduces manual work of citing, referencing and researching
- ➔ Highly Adaptable and Versatile
- ➔ Continuous Learning and Improvement of search results



RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

The following resources were referred:

- Research paper titled "*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*" - [link]
- An open-sourced resource by *Facebook Research*- [link]
- An open-sourced resource by *elastic.co* - Team of elastic search - [link]
- Open Source templates for Interface and UI/UX Designs
- Documentations of various technologies such as ChromaDB, FAISS, PostgreSQL etc

SLASH6