Applying NLP to the ICDS Knowledge Base

Justin Petucci, Ph.D.

9/11/2023

Contents

- Project Team
- Project Details
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

ICDS NLP@KB team

- RISE: Justin Petucci
- i-ASK: Lindsay Wells, Emery Etter, Mohammad Moeini
- Leadership: Amit Amritkar

- Project Team
- **Project Details**
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

Project Aims

Aim 1: Improve support for users of the ICDS research computing systems

Aim 2: Reduce the 'support burden' on client facing ICDS teams (i-ASK, RISE, OPs, etc.)

Aim 3: Improve access and usability of internal KB (cross-team knowledge sharing)

Project Deliverables

Consolidated Knowledge Base: i-ASK (WHD, SN), website, BookStack, etc.

Vector Database: Elements of the KB will stored in numerical form (text embedding generation)

Internal KB semantic search tool and chatbot: An internal tool that provides semantic search and conversation using the vector database

External chatbots: A client facing Chatbot that has knowledge/access to the external portion of ICDS KB

Project Requirements/Constraints

Open Source: Get as close to OpenAI models (Ada and GPT4) as possible with open source models and auxiliary libraries

Easily Extensible/Updat(e)able: Ability to extend/update KB, use new text embedding models, LLMs, etc.

Low resource requirements: Does not require 100s of GB of GPU RAM to run reasonable inference

Follow Responsible AI principles

- Project Team
- Project Details
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

ICDS Knowledge Base

Components:

- Help Desk Tickets
  - Legacy SolarWinds WHD + Service Now
- ICDS Website
  - User-guide, services, events, etc.
- Internal documentation
  - gitlab (software stack), confluence, BookStack, Slurm, software user-guides, etc.

ICDS Knowledge Base - 2

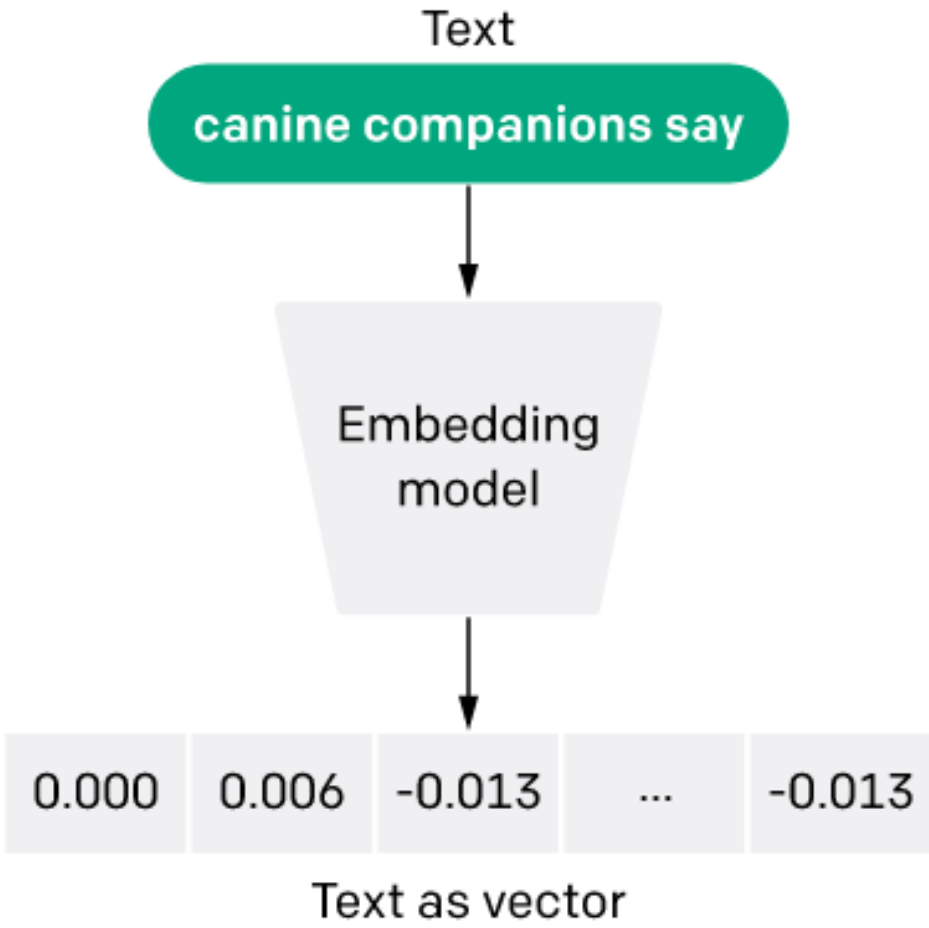
Task list:

- Collect data
  - Web scraper (icds.psu.edu), legacy WHD, SNow, BookStack transition
- Create database
  - define metadata, structure, accessibility, etc.
- Update procedure
- Data use restrictions, internal vs external delineation

- Project Team
- Project Details
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

Text Embeddings

- Numerical/vector representation of words, sentences, paragraphs, documents, etc.
- encodes semantic information
- meaning $\Leftrightarrow$ location
- SOTA models are transformer based



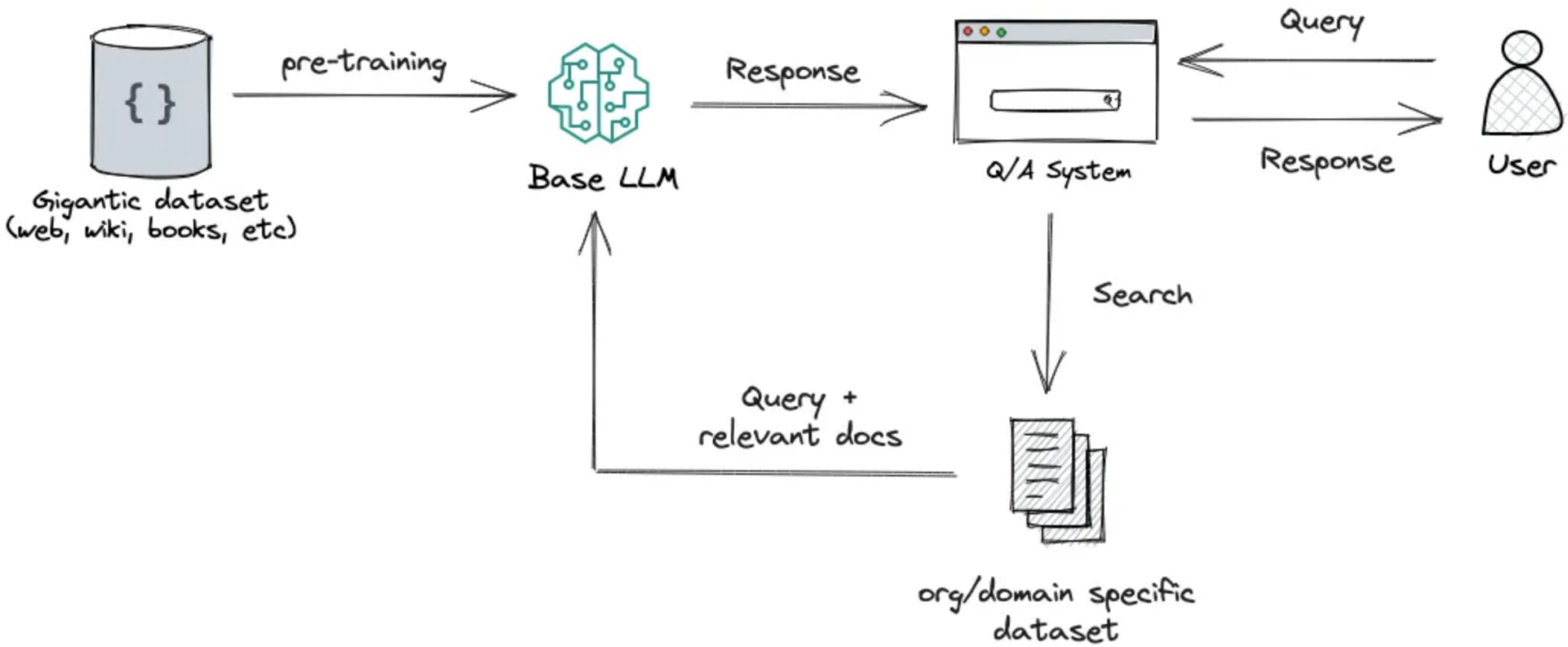
Text Embeddings - 2

Task list:

- Format KB for model ingestion
  - content splitting/chunking (context length limit) - use GPT4?
- Evaluate pretrained text embedding models - Massive Text Embedding Benchmark (MTEB)
- Explore fine tuning embedding models
- Generate and store embeddings
  - dedicated vector DB (Milvus, marqo, etc) ?

- Project Team
- Project Details
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

Retrieval Augmented Generation (RAG)



Source: Heiko Hotz

Retrieval Augmented Generation (RAG) - 2

Task list:

- LLM model exploration
  - off the shelf, full fine tuning $$, LORA, context length limits/expansion, model quantization
- Similarity search
  - cosine similarity, FAISS (Facebook AI Similarity Search), BM25 supplement, Re-ranker model
- Query/Prompt Engineering
  - combining top KBs to fit in model context window, minimize hallucinations, etc.
- User interface: Gradio web app
- Compare best open model to GPT4

- Project Team
- Project Details
- The ICDS KB
- Vector Database
- Internal search and chatbot
- External chatbot

Client facing chatbot

Task list:

- Use only approved externally facing KB, all other back-end components the same
- Integrate into i-ASK for Level 0-1 user self service
- Frontend - beyond Gradio
- Backend - scale to many? users