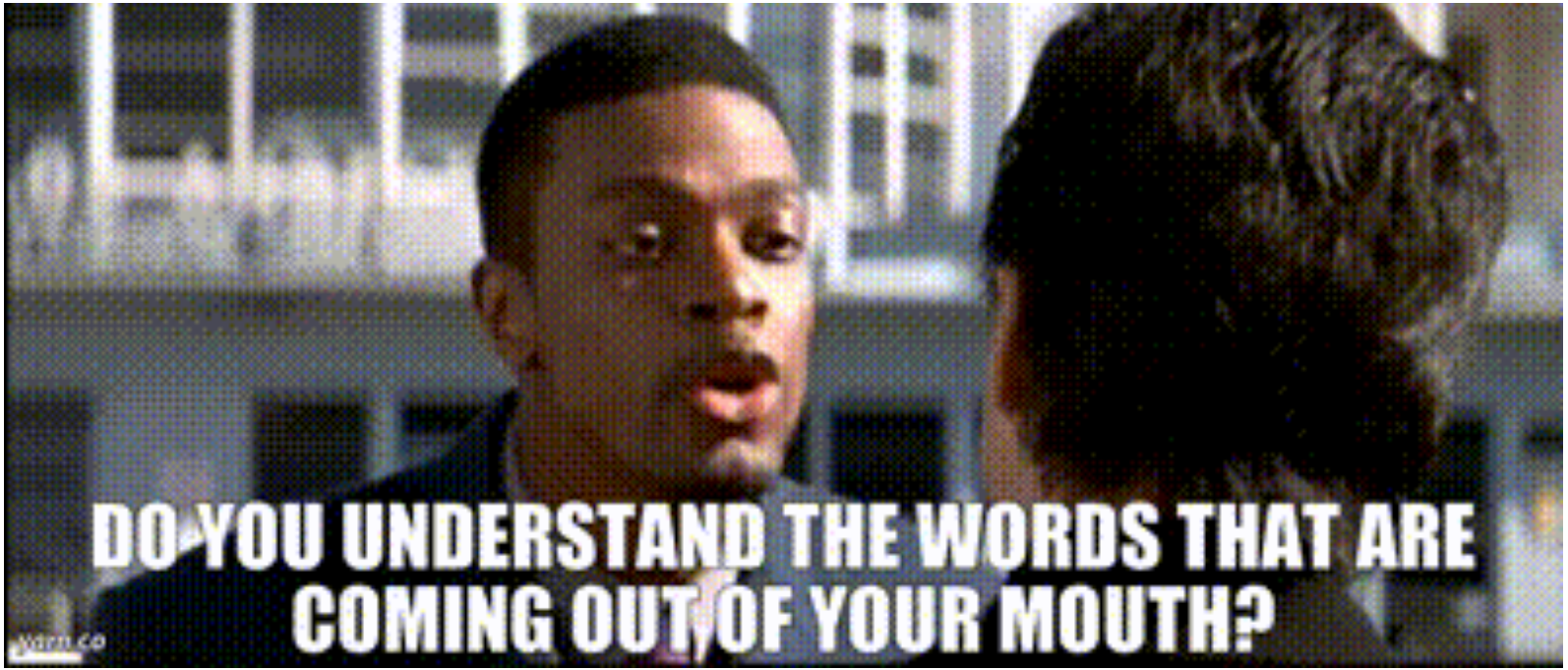
CrossComm

Computers That “Get It”

Using LLMs to Give Software Semantic Understanding

Will computers ever understand what we're saying?



Natural Language Processing (NLP) has been historically difficult

- Humans use a lot of words and expressions to mean the same thing, even when speaking the “same” language:
  - I’m excited.
  - I’m fired up
  - I’m beside oneself
  - I’m worked up

Natural Language Processing (NLP) has been historically difficult

- And humans sometimes use a single word/expression to mean many things:
  - “bow”

Natural Language Processing (NLP) has been historically difficult

- And humans sometimes embed sentiment that communicates as much (if not more) than the literal words:
  - “Oh, sure, because I have nothing better to do than help you with that.”

If computers could understand semantic meaning…

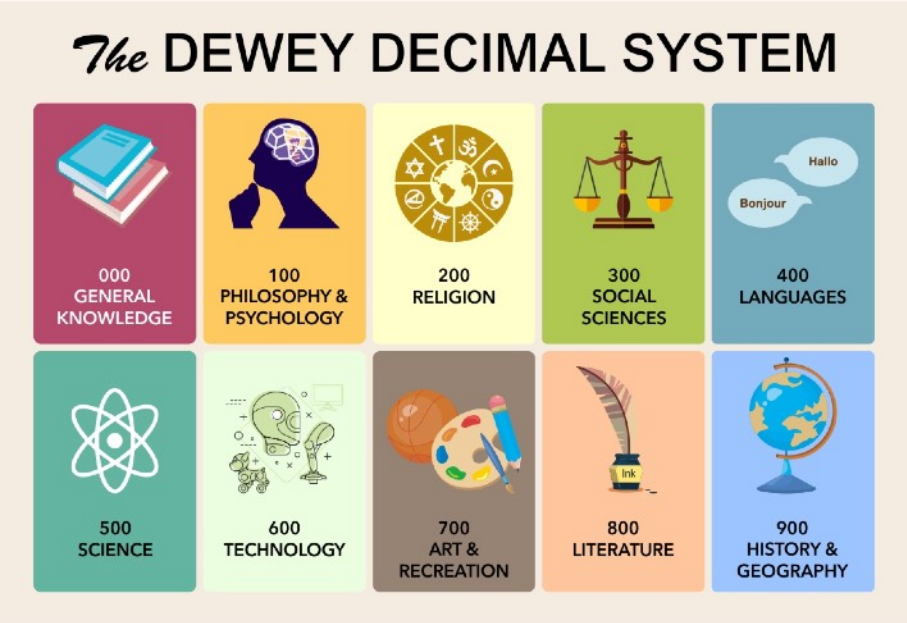
- They could search for information of semantic relevance to a given topic, regardless of differences or ambiguities in expression
- They could matchmake/pair people or objects of semantically similar attributes
- They could understand the underlying sentiment of an expression or statement

Vector Databases

Mapping by similar semantic meaning

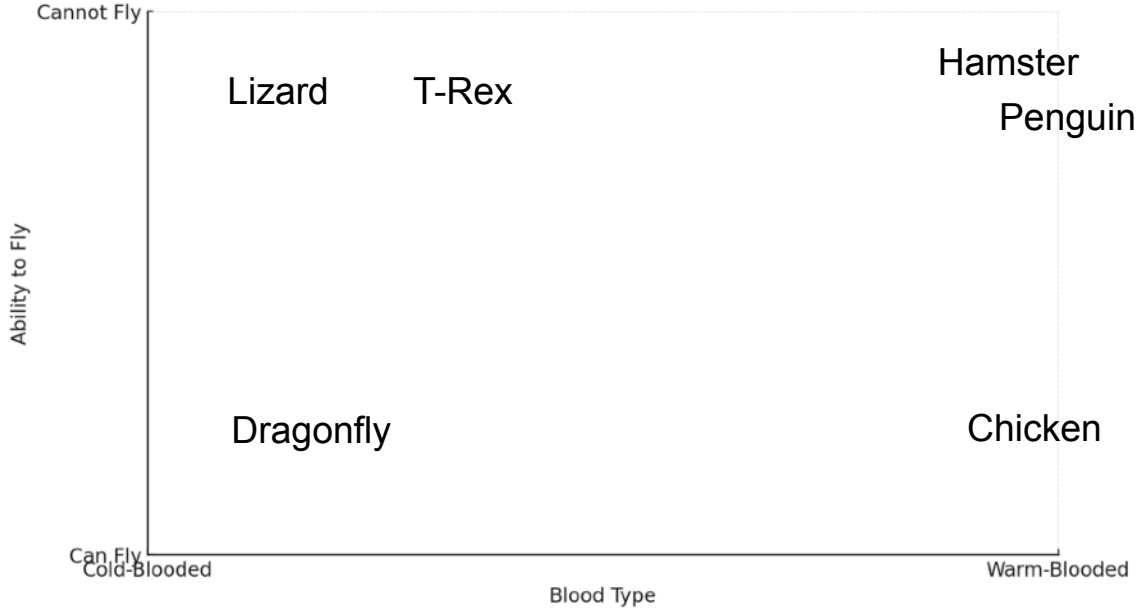
How to map semantic meaning

- Imagine mapping the semantic meaning and context of all the knowledge of the world along one axis where similar meanings are in close proximity to each other. What would that look like?



How to map semantic meaning

- What if we graphed all semantic meanings within a 2-axis graph?



How to map semantic meaning

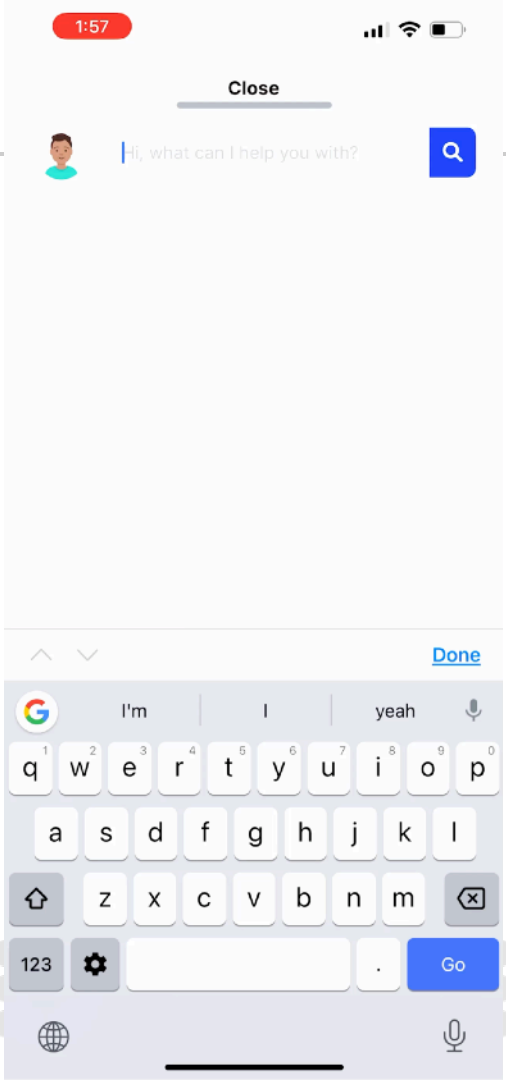
- What about 3 dimensions, or 8 dimensions… n-dimensions...
- LLMs can represent semantic meanings as embeddings – high dimensional (dense) vectors – according to semantic similarity so that nearest neighbors in the n-dimension space have similar meaning and context
- An LLM text prompt can also be transformed into a vector so that semantically similar entries in a vector database can be retrieved by proximity

Some Examples

Made possible by understanding semantic meaning

Semantic Search/Recommendations

- Semantic search finds content similar in meaning
- Hybrid approach merges traditional keyword results with semantic results
- Related content recommendations generated by semantic meaning

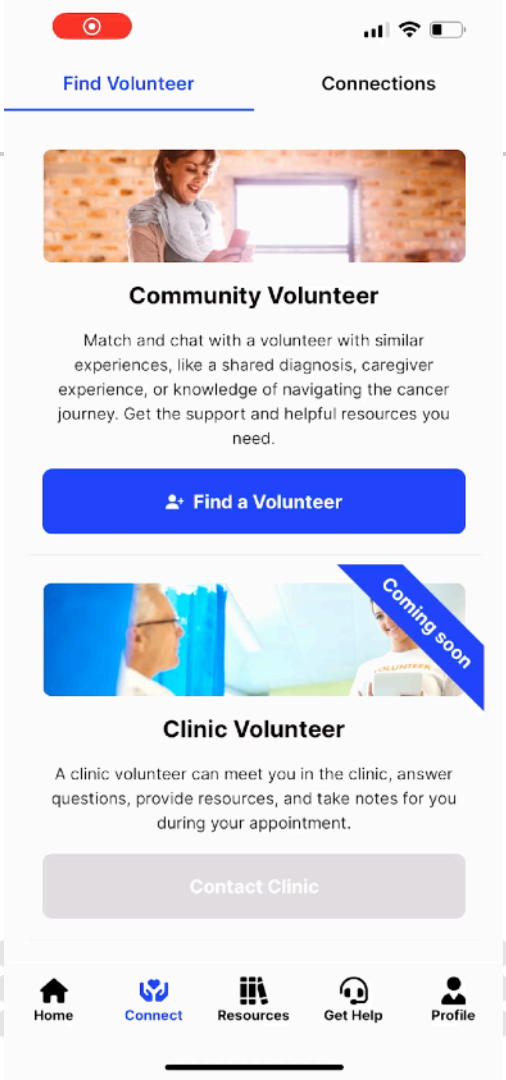


Semantic Search/Recommendations: Why not RAG?

- Sometimes less can be more.
  - Scenarios where hallucination cannot be tolerated
  - Scenarios where provenance and sourcing are just as important as the content itself
  - Scenarios where the content is better experienced than summarized (e.g. non-text media)
  - Can be faster.

Pairing/Matchmaking

- Matchmaking of patients and volunteers based on background story provided by users
- Hybrid approach merges traditional and semantic similarity results



Sentiment Analysis

- Two inquiries:
  - “Hi! I’d like to go to Maui”
  - “I'd love to book my trip to Maui sometime this century, please.”
- If sentiments could be quantified, how should you (or your software) handle those requests differently?
- Sentiments can serve as dimensions of semantic meaning as well.

Helpful Tools

Components for Handling Semantic Meaning

Let’s start with the LLM

- all-MiniLM-L6-v2
  - Sentence-transformer (a.k.a. SBERT) model: Maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.
  - “Our model is intended to be used as a sentence and short paragraph encoder. Given an input text, it outputs a vector which captures the semantic information. The sentence vector may be used for information retrieval, clustering or sentence similarity tasks.”
  - (relatively) computationally fast

Vector Databases

- Chroma vector database is free (Apache 2.0 license)
  - Decent performance and scalability
  - Straightfoward integration
- Pinecone
  - Managed service, commercial support

No-code Services

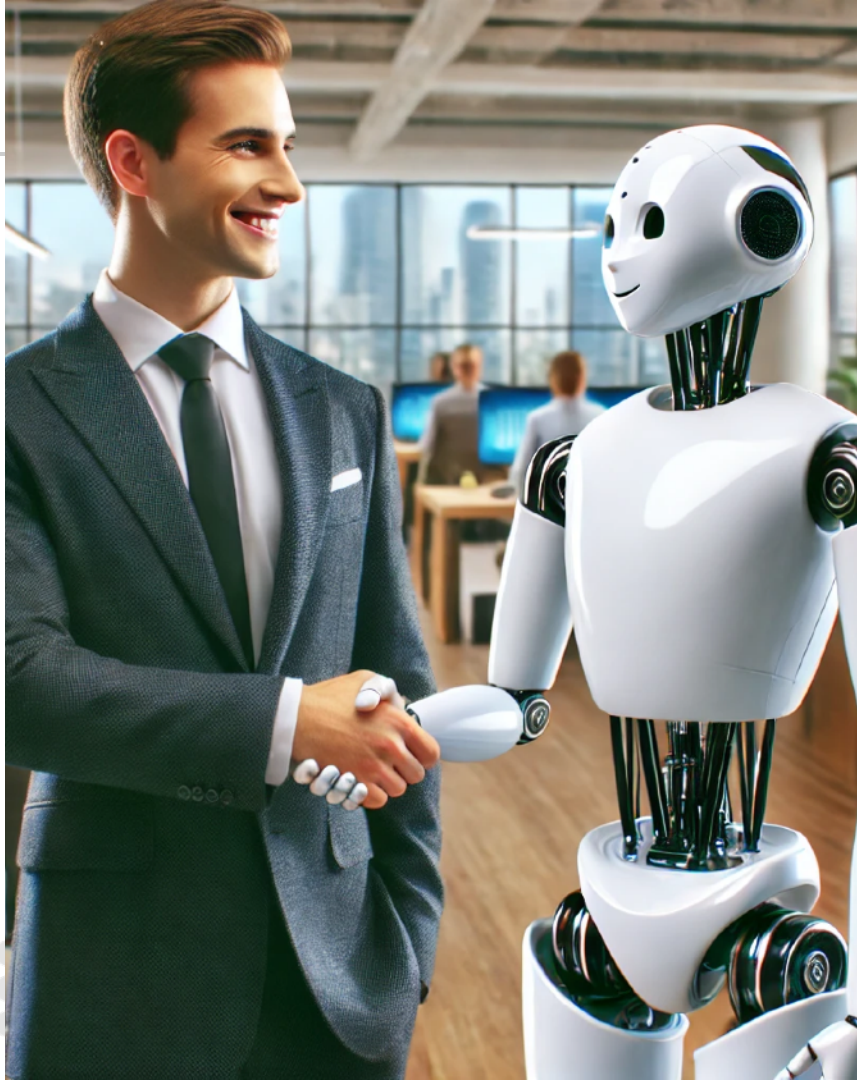
- H20GPT – LLM of your choice, Vector DB of your choice, LangChain integration, and a nice Gradio UI that supports document ingestion.
  - Available as hosted service of open source stack/ecosystem with option to self-host
  - Provides RAG (Retrieval Augmented Generation) capability

Wrapping Up

Computers Can Finally “Get It”

Wrapping Up

- A sentence transformer model can map semantic meaning for words, sentences, and paragraphs to multi-dimensional vectors, i.e. plots in an n-dimension space
- Similar semantic meanings are in close proximity within an n-dimension space.
- LLMs are helping computers to finally understand our language- and understand us.



Thank You

Any Questions?

Don Shin Don.Shin@crosscomm.com

LinkedIn: https://www.linkedin.com/in/donshin1/

Twitter: @donshin