TEXAS
The University of Texas at Austin

C S 329E
Elements of Data Analytics

Spring 2023
Prof. Kuhlmann (Section 52150)

Quiz 15-01

Take 5 Minutes now

- Complete individually
- **No access code**

Week 15 - Lecture 1

Monday, April 17, 2023

- Natural Language Processing (NLP)
- Embeddings
- Autoencoders
- AI Safety

Use Cases

C S 329E
Spring 2023 (52150)

TEXAS
The University of Texas at Austin

Named Entity Recognition

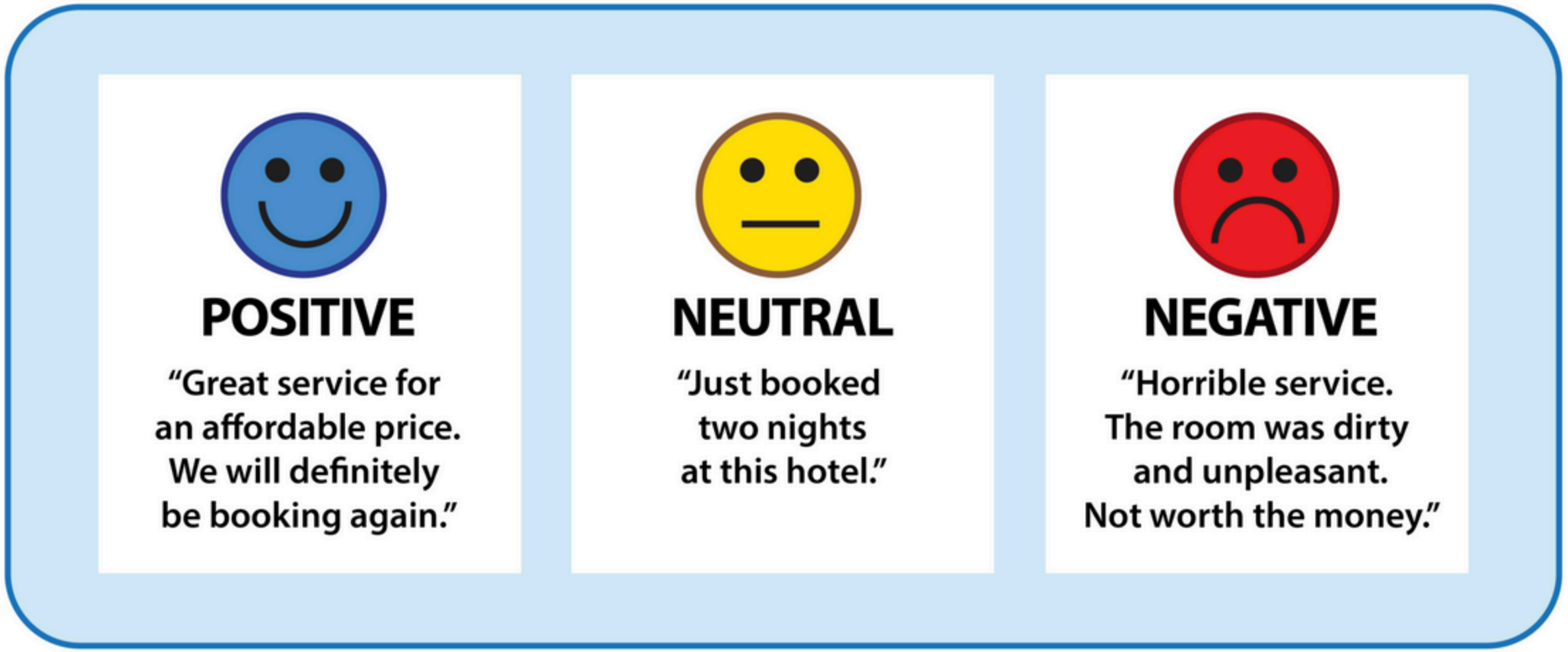
Extract entities from text into pre-defined categories

- Résumé parsing
- Customer support
- Knowledge graphs

- **NORP (nationality or religious political group)**
- **GPE (geopolitical entity)**

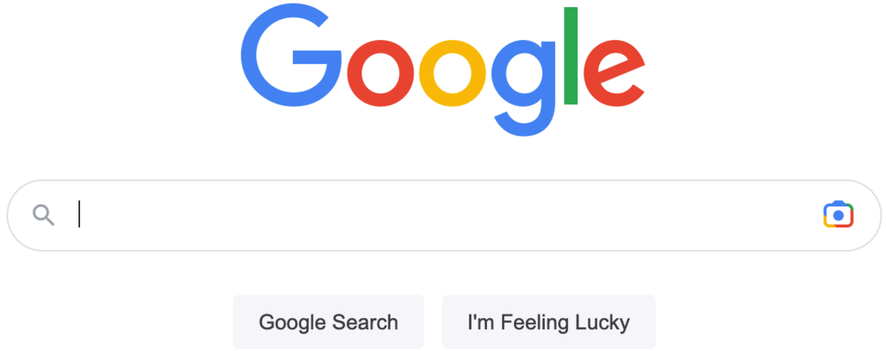
Sentiment Analysis

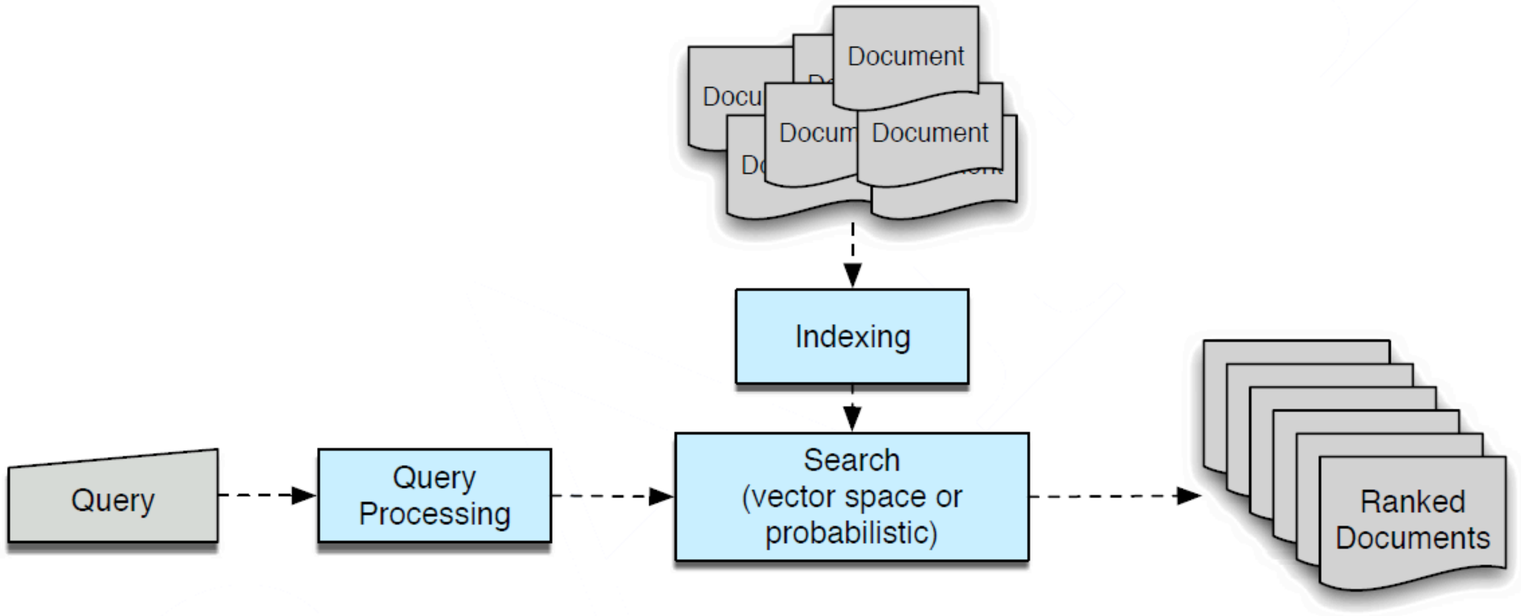
Given text, classify its emotional quality



Information Retrieval

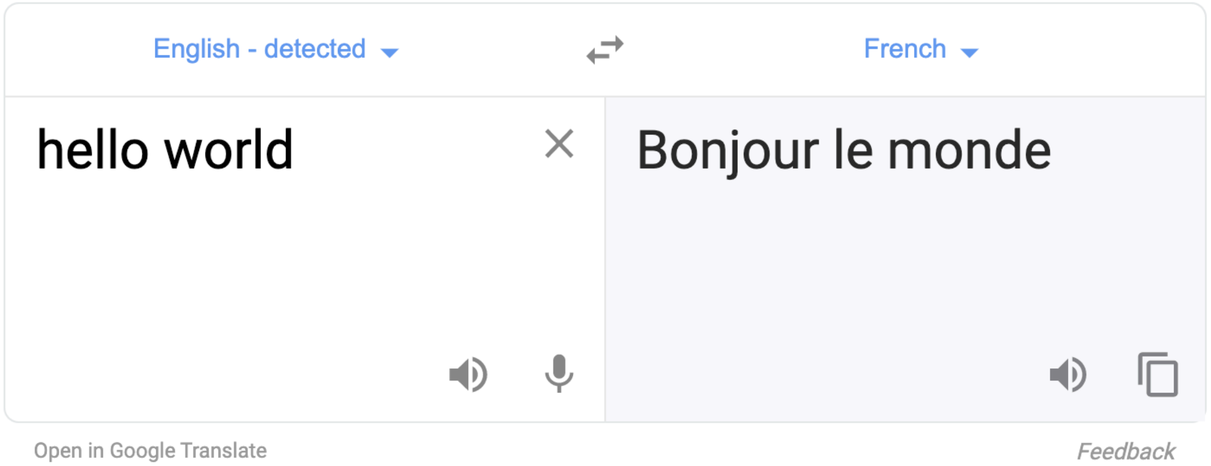
Find documents most relevant to a query

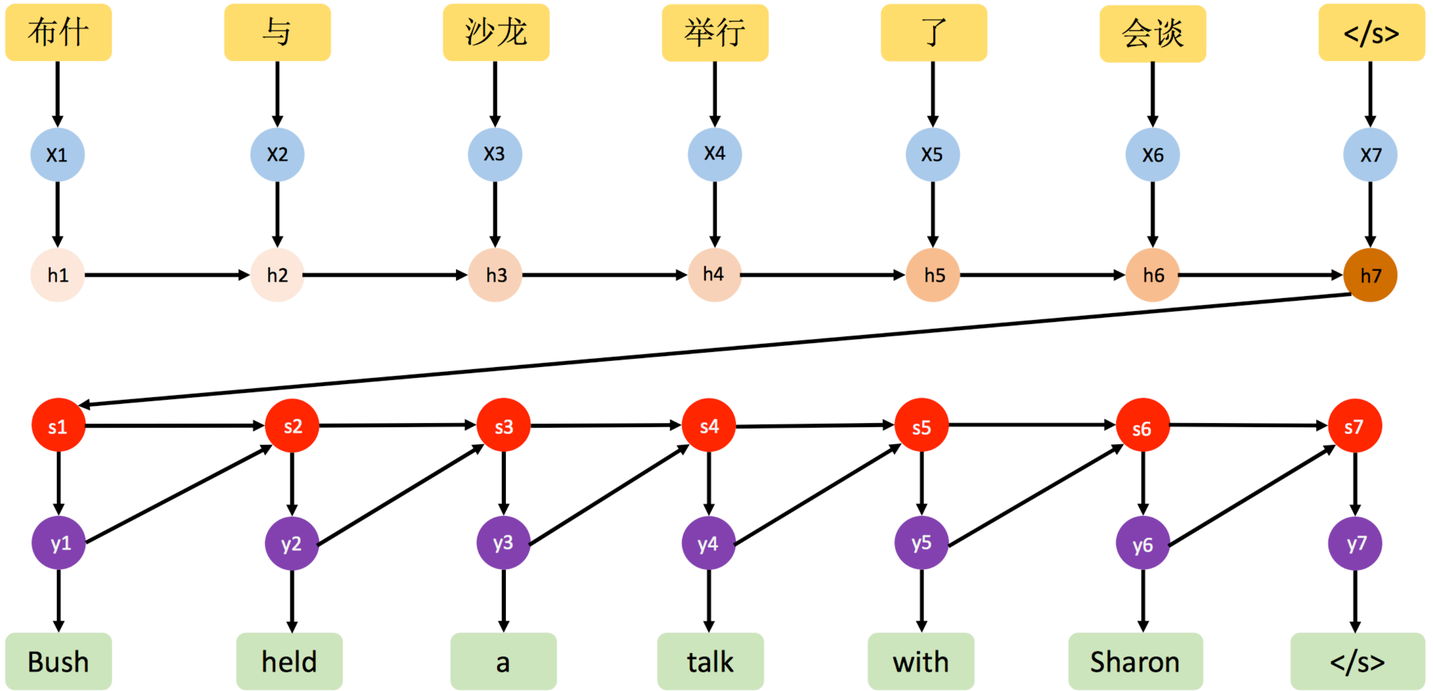




Machine Translation

Translate text to another language





Generative AI

ChatGPT

Why did the chicken cross the road?

The phrase "Why did the chicken cross the road?" is a classic example of a riddle joke, dating back to the 19th century. The punchline is often "To get to the other side." The humor in this joke derives from its anti-humor or anti-climax, as it sets up the expectation of a clever or funny answer,

Jasper Features Company Support Pricing Business API

MEET JASPER 👋

Create amazing sales emails 10X faster with AI.

DALL·E 2

Midjourney /stable-diffusion-2-1

GitHub Copilot

NLP Basics

C S 329E
Spring 2023 (52150)

TEXAS
The University of Texas at Austin

```
import nltk
```

Data Preprocessing

Text often transformed prior to task to improve performance

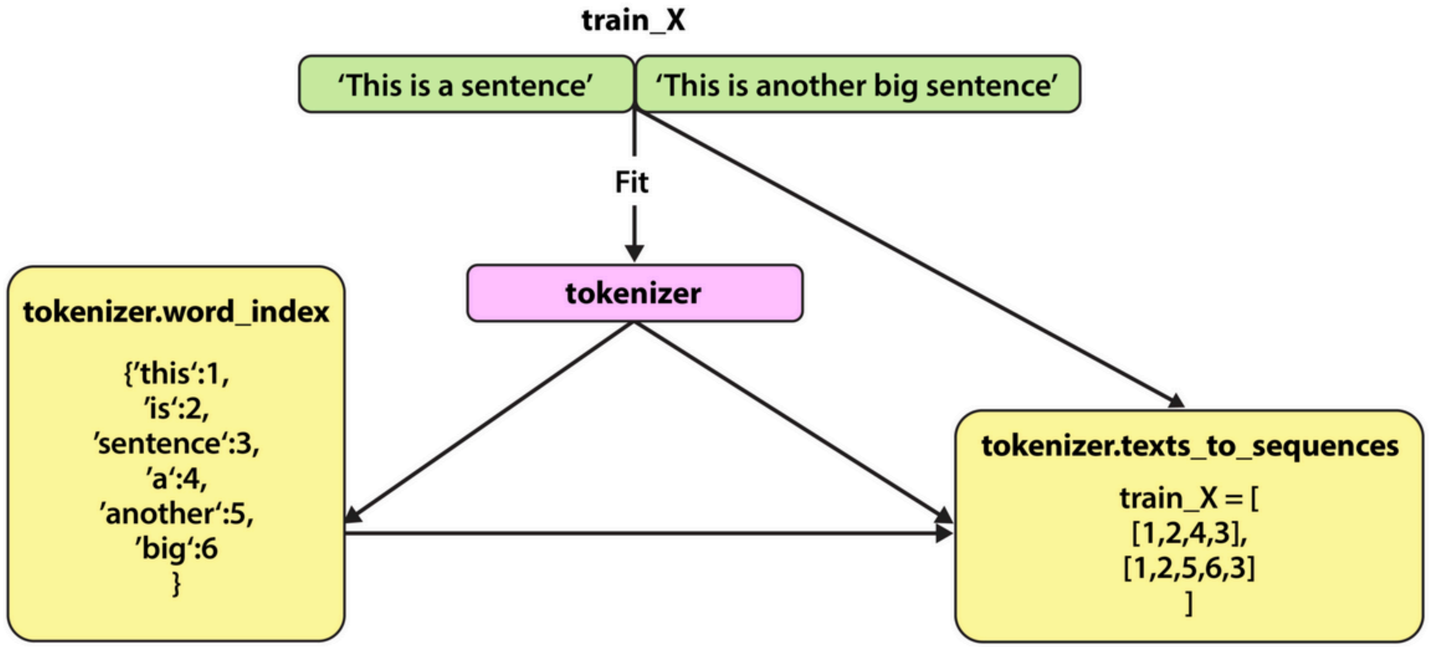
- Sentence segmentation: break up document into sentences
- Stemming: (closed, closely, closing) → close
- Stop word removal

Type	Examples
Determiners	a, the, etc.
Conjunctions	is, am, are, etc.
Pronouns	he, him, this, that, they, them, etc.
Interrogative	what, where, when, how, etc.

- Tokenization

Tokenization

Given a vocabulary, map every word (or sub-word) to an index



- Every document becomes a sequence of indices

ChatGPT

- 1 token ~= 4 chars in English
- 1 token ~= ¾ words
- 100 tokens ~= 75 words

My favorite color is red.

Text | Token IDs

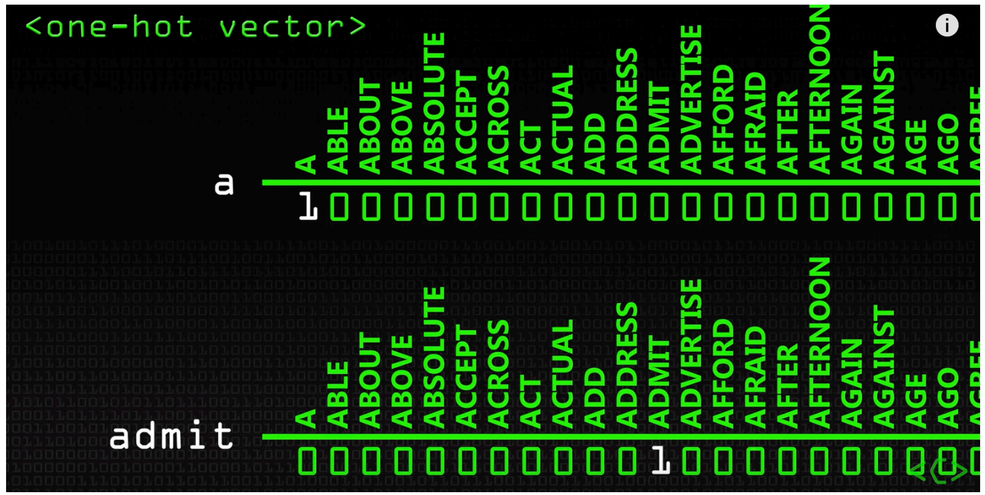
[3666, 4004, 3124, 318, 2266, 13]

Text | Token IDs

Feature Extraction

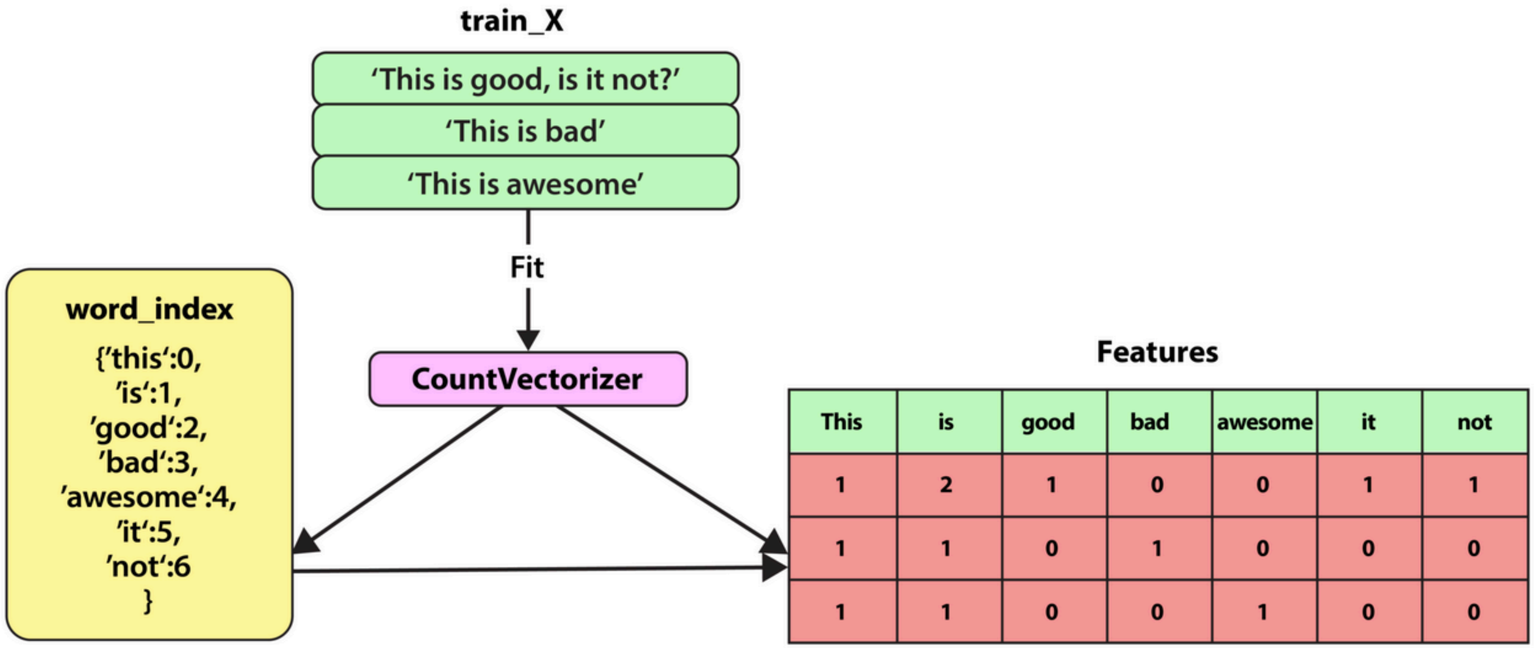
How do we represent text to a model?

- Documents have different length
- ‘car’ and ‘cat’ are lexically similar but semantically very different
- Number of possible words is huge



Feature Extraction: Bag of Words

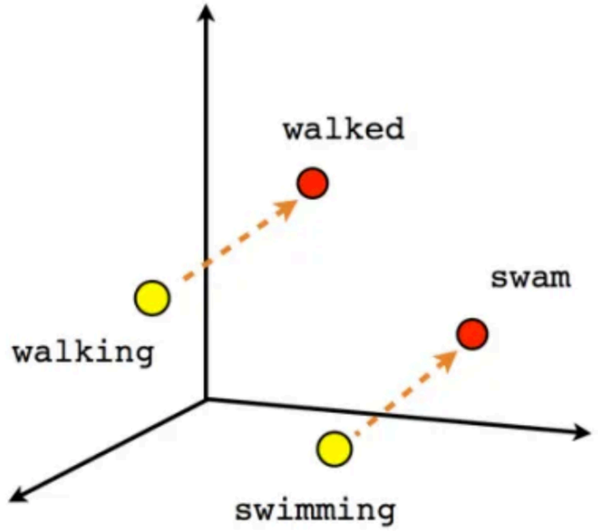
Represent a document by the number of times each word occurs



- TF-IDF: weight word's count by how "rare" the word is across documents

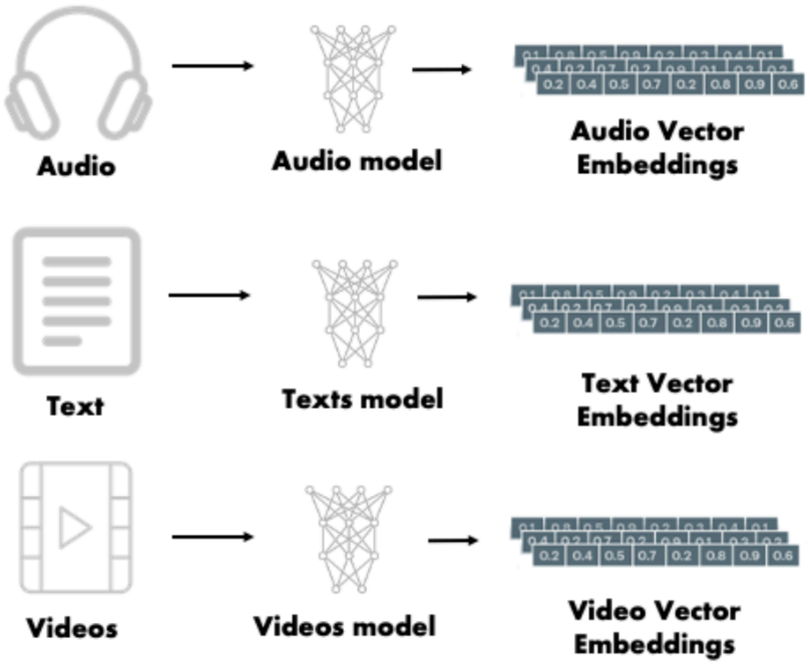
What are the shortcomings of this approach?

Embeddings

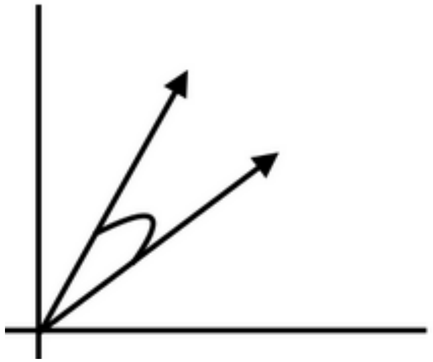


What is an embedding?

An embedding is a numerical representation, often in a lower-dimensional space, that captures the relationships and structure of complex data, such as words, images, or nodes in a network, enabling efficient analysis and computation.



Desirable properties

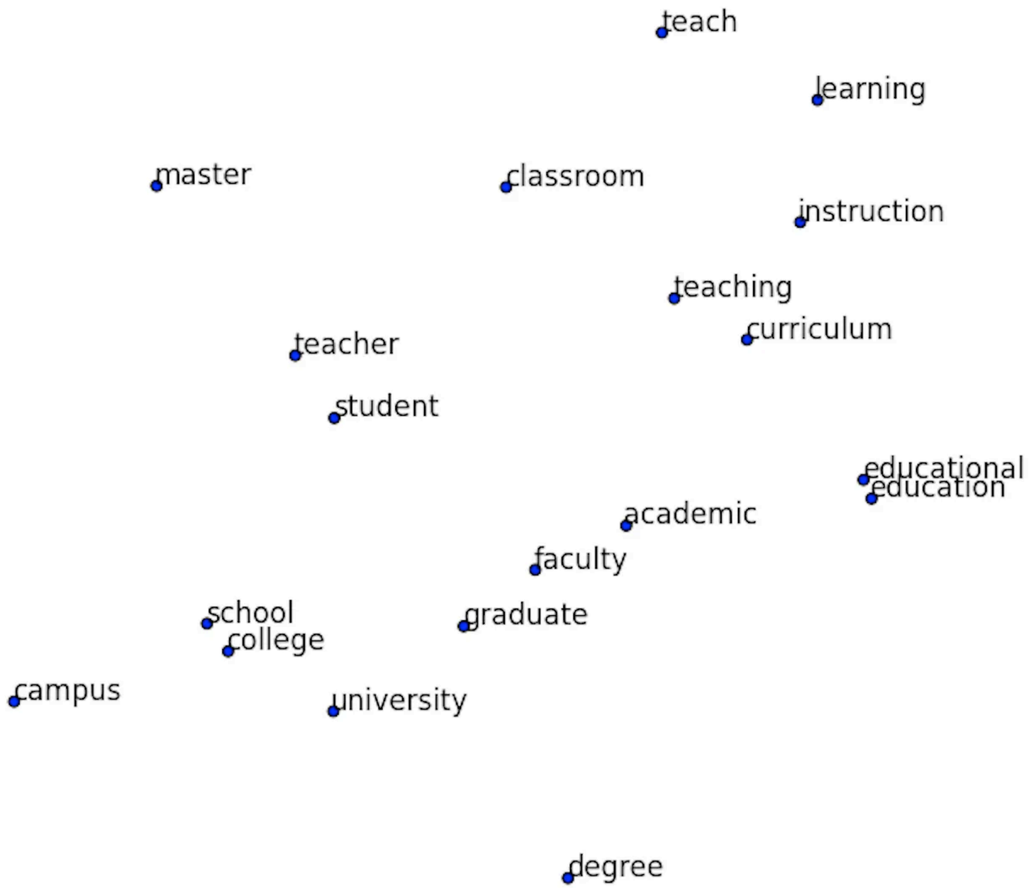


- Compression: dimensionality reduction
- Semantically similar inputs produce similar vectors
  - for some metric, typically cosine similarity (or euclidean distance)
- Dissimilar inputs produce far away vectors

- Meaning of “similar” depends on how embedding model is trained
  - songs that appear together in playlists
  - words that appear in the context similar other words in documents

Word Embeddings

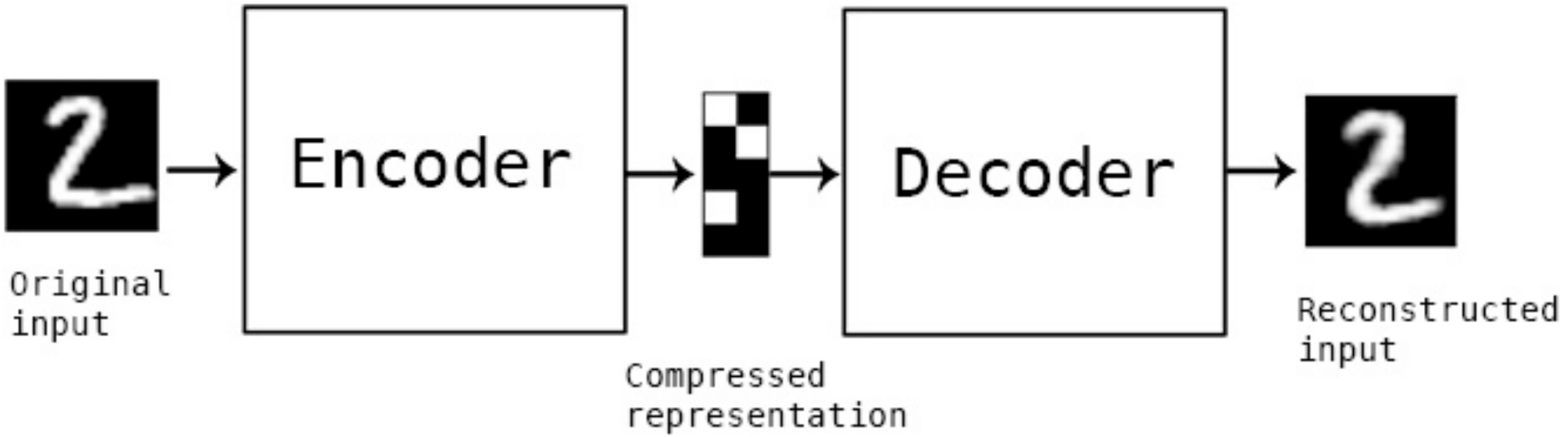
Each word is a point (vector) in n-dimensional space



Two-dimensional visualization

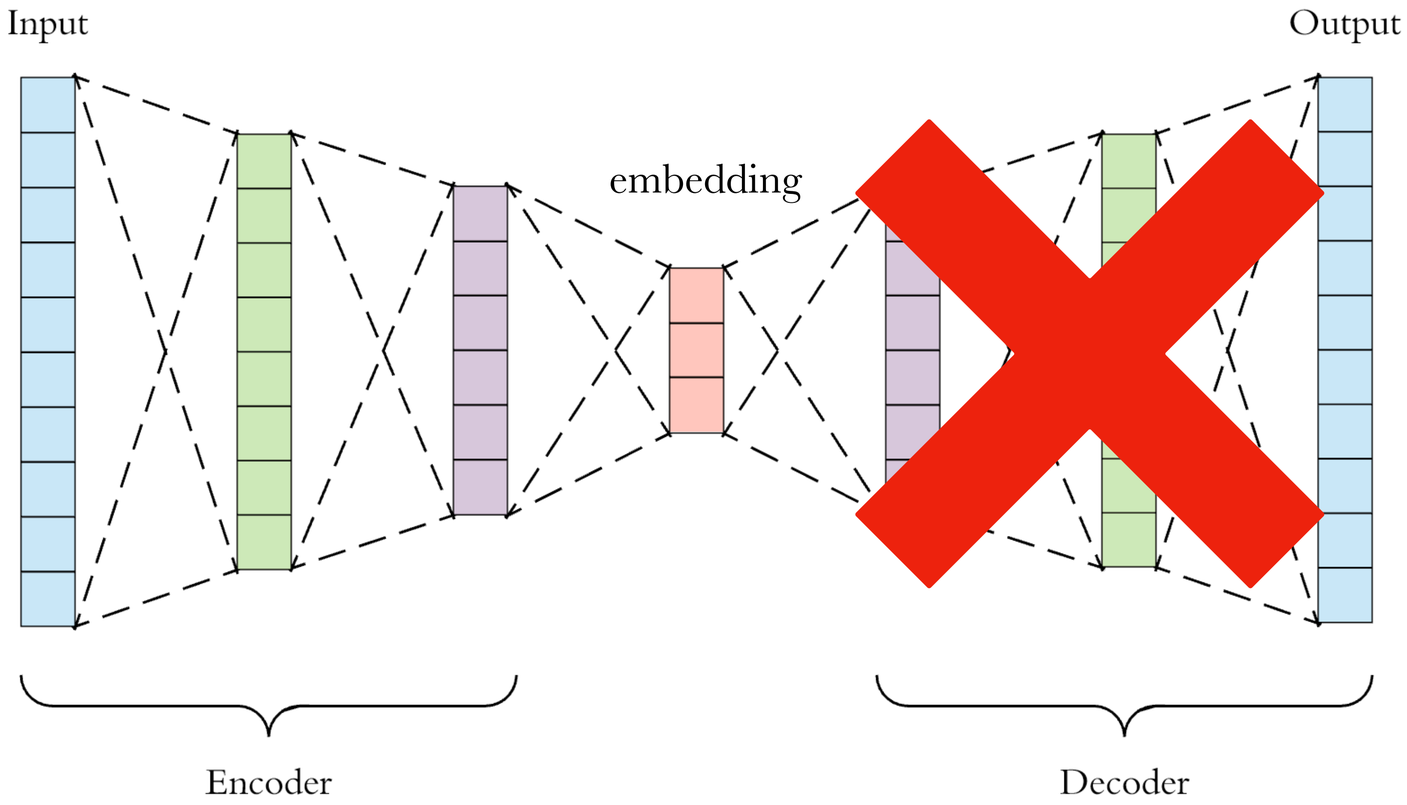
Code

Generating Embeddings for MNIST images



Autoencoders

Unsupervised way to generate embeddings

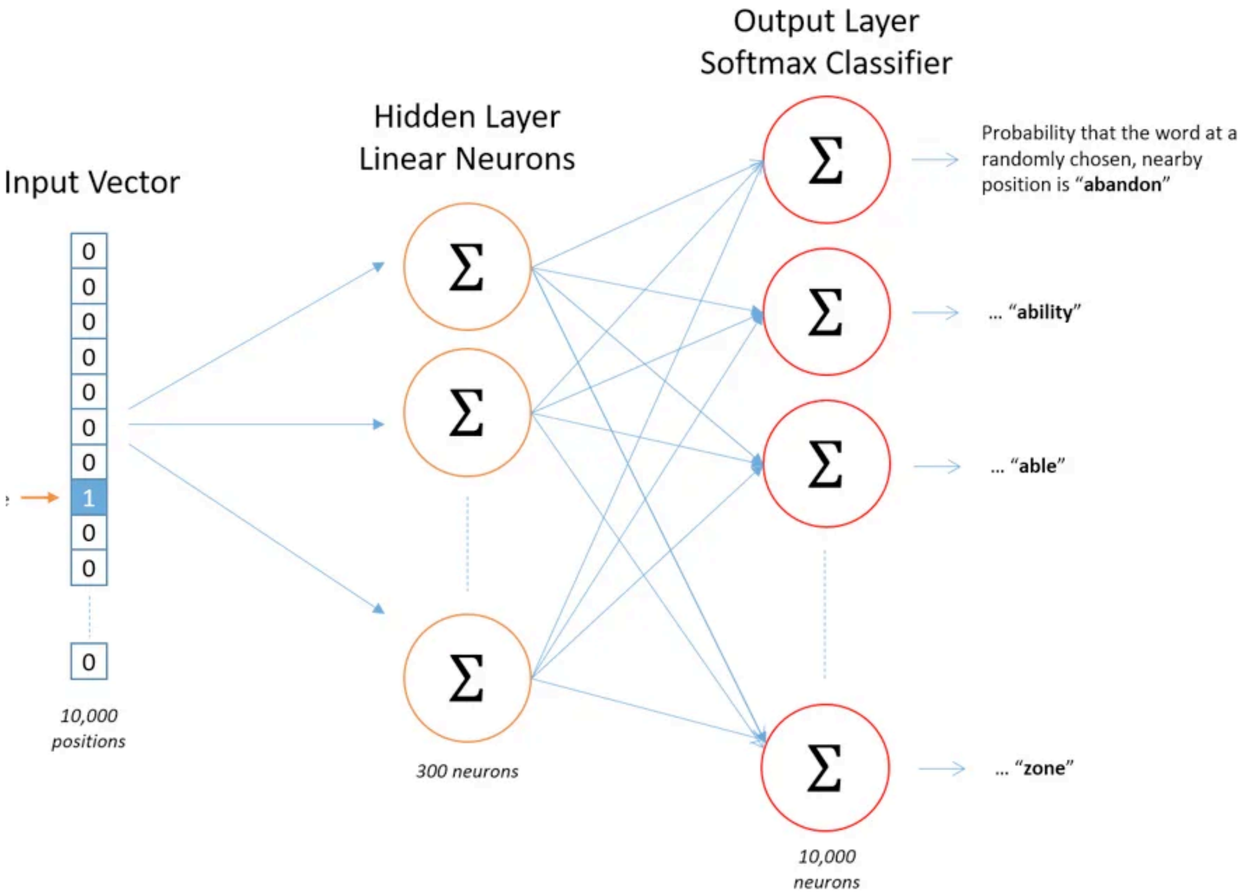


- NN with small hidden layer
- Essential info must pass through bottleneck
- Train network to reproduce each given input
  - Present input
  - Backprop loss function: difference between output and input
- In the end we can discard decoder

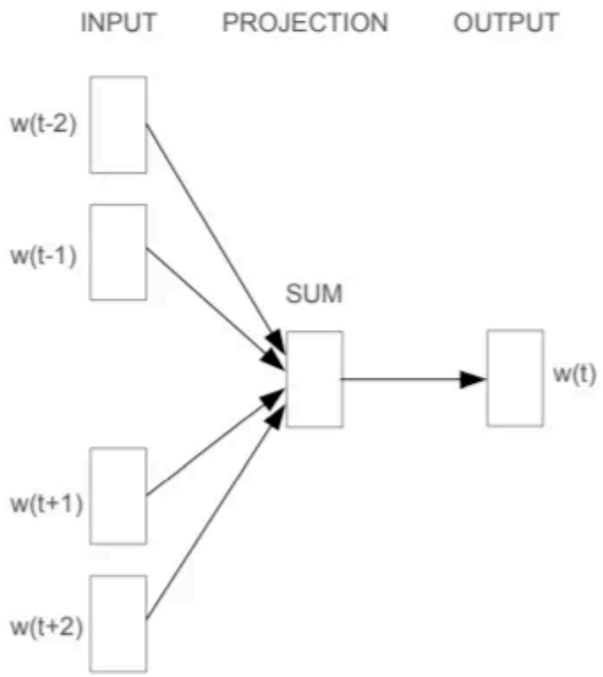
Word2Vec

Predict surrounding words, given an input word

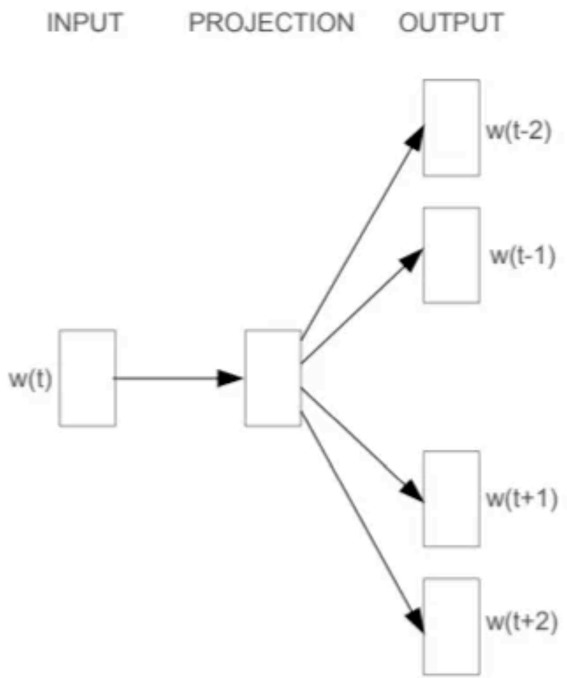
Source Text	Training Samples generated from source text			
I will have orange juice and eggs for breakfast	(will, I)	(will, have)	(will, orange)	
I will have orange juice and eggs for breakfast	(have, I)	(have, will)	(have, orange)	(have, juice)
I will have orange juice and eggs for breakfast	(orange, will)	(orange, have)	(orange, juice)	(orange, and)
I will have orange juice and eggs for breakfast	(juice, have)	(juice, orange)	(juice, and)	(juice, eggs)
I will have orange juice and eggs for breakfast	(and, orange)	(and, juice)	(and, eggs)	(and, for)
I will have orange juice and eggs for breakfast	(eggs, juice)	(eggs, and)	(eggs, for)	(eggs, breakfast)
I will have orange juice and eggs for breakfast	(for, and)	(for, eggs)	(for, breakfast)	



Word2Vec



CBOW



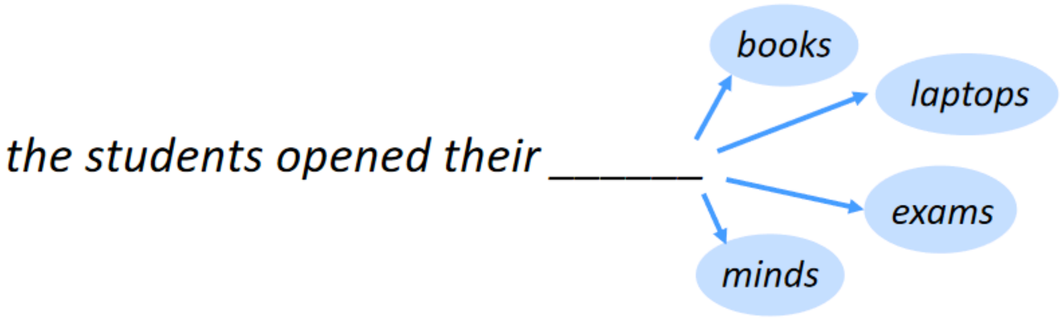
Skip-gram

- **Skip-gram**
  Performs well even on small, sparse data set
- **Continuous Bag-of-Words**
  Faster to train

BERT, GPT, …

Large Language Models

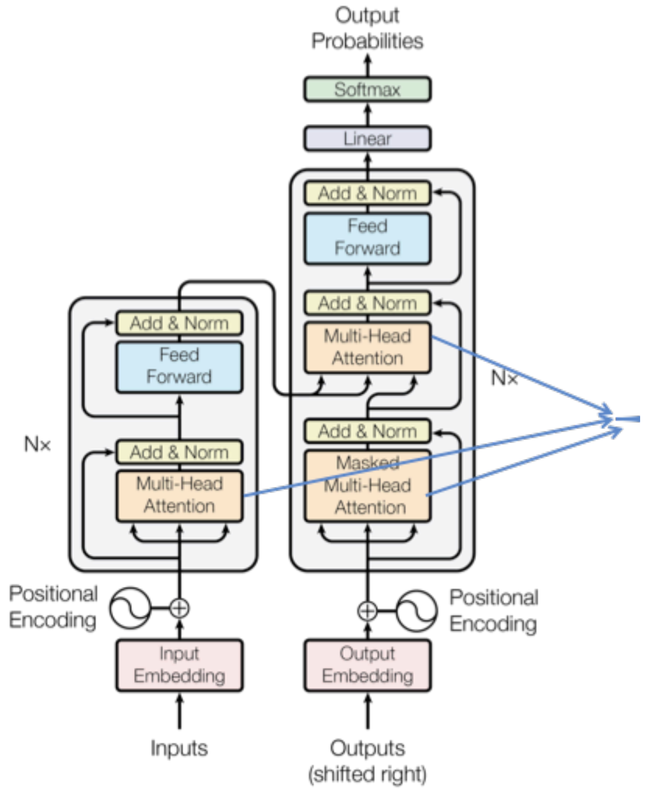
- Next-token prediction



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

- NN architecture: “transformers”

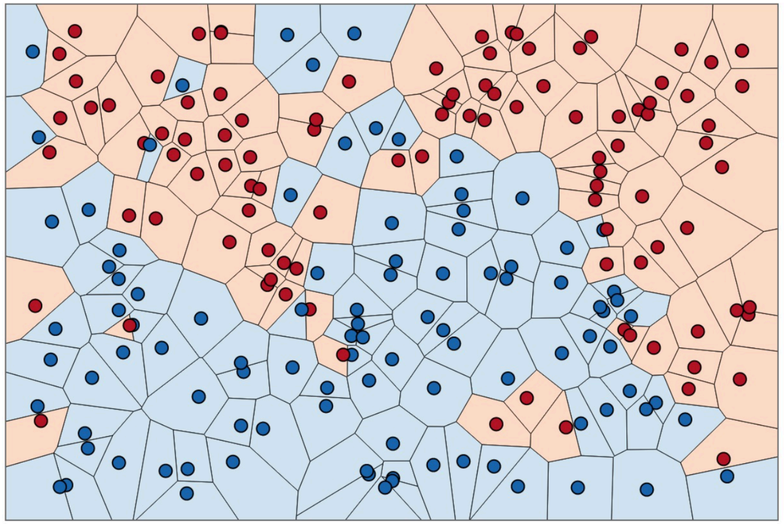


https://course.fast.ai/

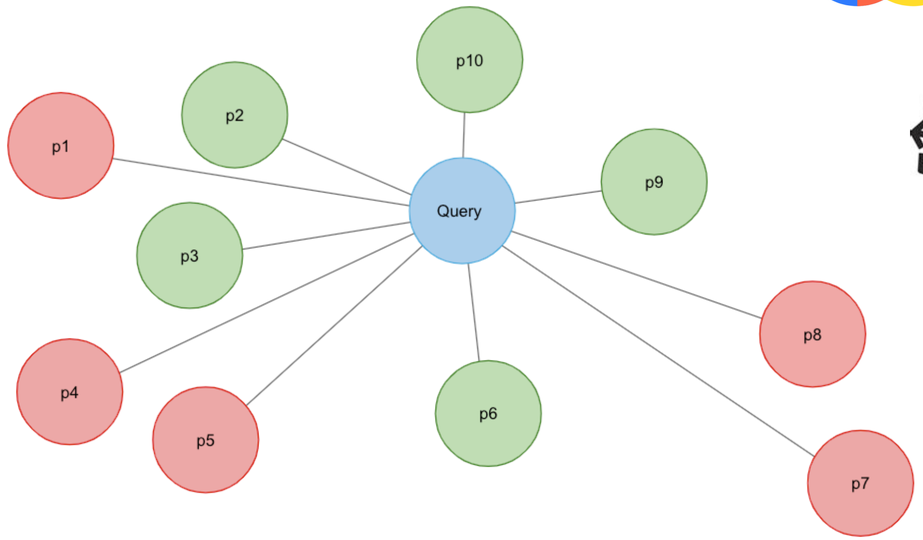
Vector Databases

Databases optimized for Similarity Search

- k-NN Search



- Approximate Nearest Neighbors (ANN)



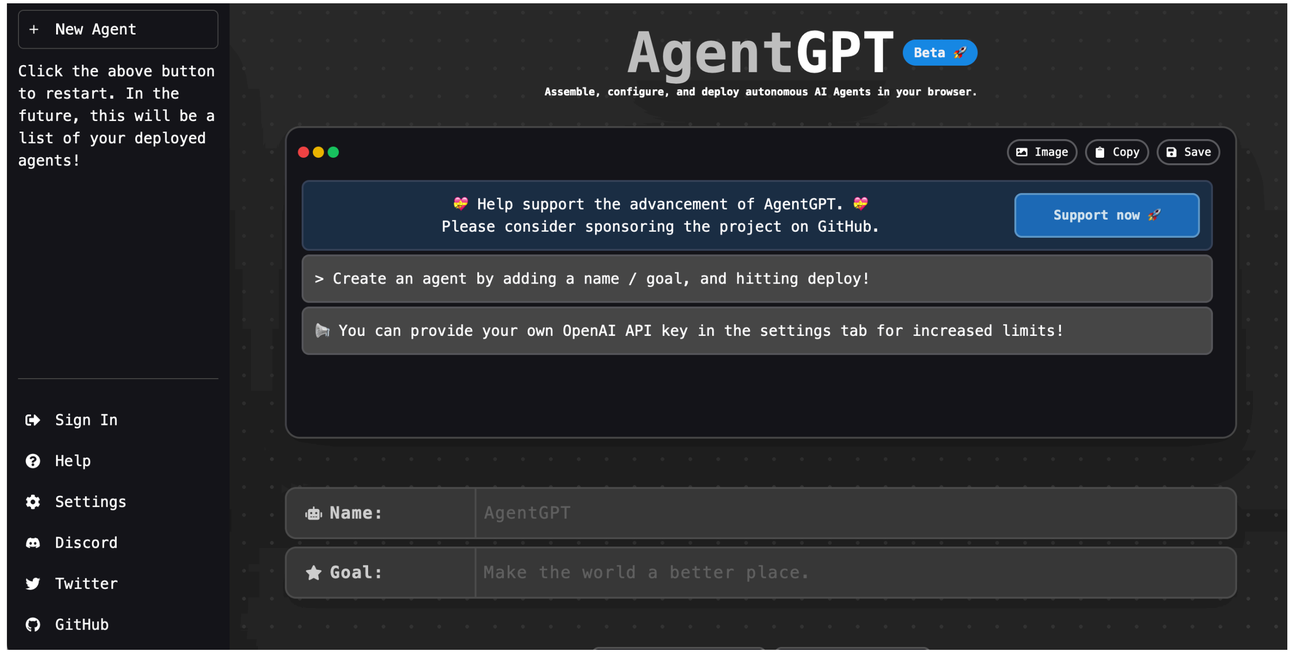
zilliz

milvus

Chroma

Pinecone

Agents



Auto-GPT

https://github.com/Significant-Gravitas/Auto-GPT

The BabyAGI Revolution: The Dawn of Autonomous AI Agents

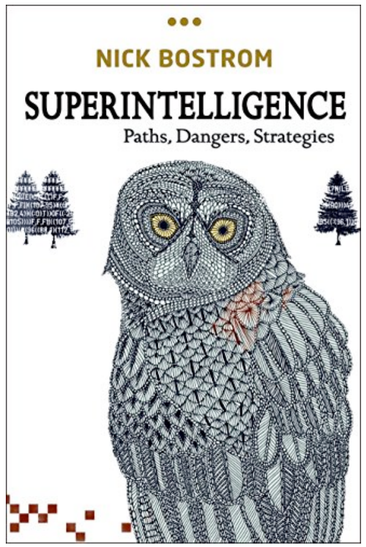
AI Risk Poll

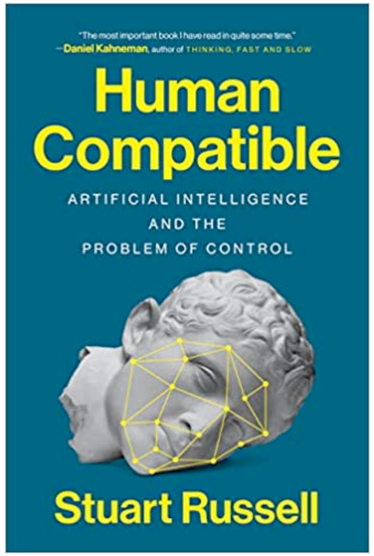
- Go to **menti.com** and use the code **3292 4666**

AI Alignment & Existential Risk

Resources

- Scott Aaronson Talks AI Safety - Effective Altruism at UT (Nov 2022)
- Existential Risk from Power-Seeking AI - Carlsmith (March 2023)
- Recent podcasts with Lex Fridman: Eliezer Yudkowsky, Max Tegmark





TECHNOLOGY

An open letter signed by tech leaders, researchers proposes delaying AI development

March 29, 2023 · 6:17 PM ET

Heard on All Things Considered

By Linah Mohammad, Patrick Jarenwattananon, Juana Summers

4-Minute Listen + PLAYLIST

Prof. Peter Stone

https://course.mlsafety.org/

https://www.agisafetyfundamentals.com/

Up Next

- HW12 due Wed at midnight
- Lecture 15-02: Exam 3 Review