Laptop-sized ML for Text, with Open Source

Nick Burch

Berlin Buzzwords 2023

Nick Burch

Director of Engineering

FLEC

Nick Burch

@Gagravarr

@nick@social.earth.li

Code, Scripts, Slides

github.com/Gagravarr/BBuzz23-LaptopML

All code in slides is taken from here!

Our talk was going to be...

- LLMs - Large Language Models
- Why you can't have one
- Simpler vector-space language models
- Word2vec, GloVe, ELMo and BERT
- How they differ from traditional text relevancy like TF-IDF
- How to run them with Open Source libraries
- How to tweak them to get better relevance or speed or memory
- What you can do even though you don't have a LLM

But then…

Llamas happened

LLaMA

Large Language Model

from Facebook (they're still around!)

LLaMA

ai.facebook.com/blog/large-language-model-llama-meta-ai/

https://arxiv.org/abs/2302.13971v1

LLaMa changed my talk, and the ML-for-text world!

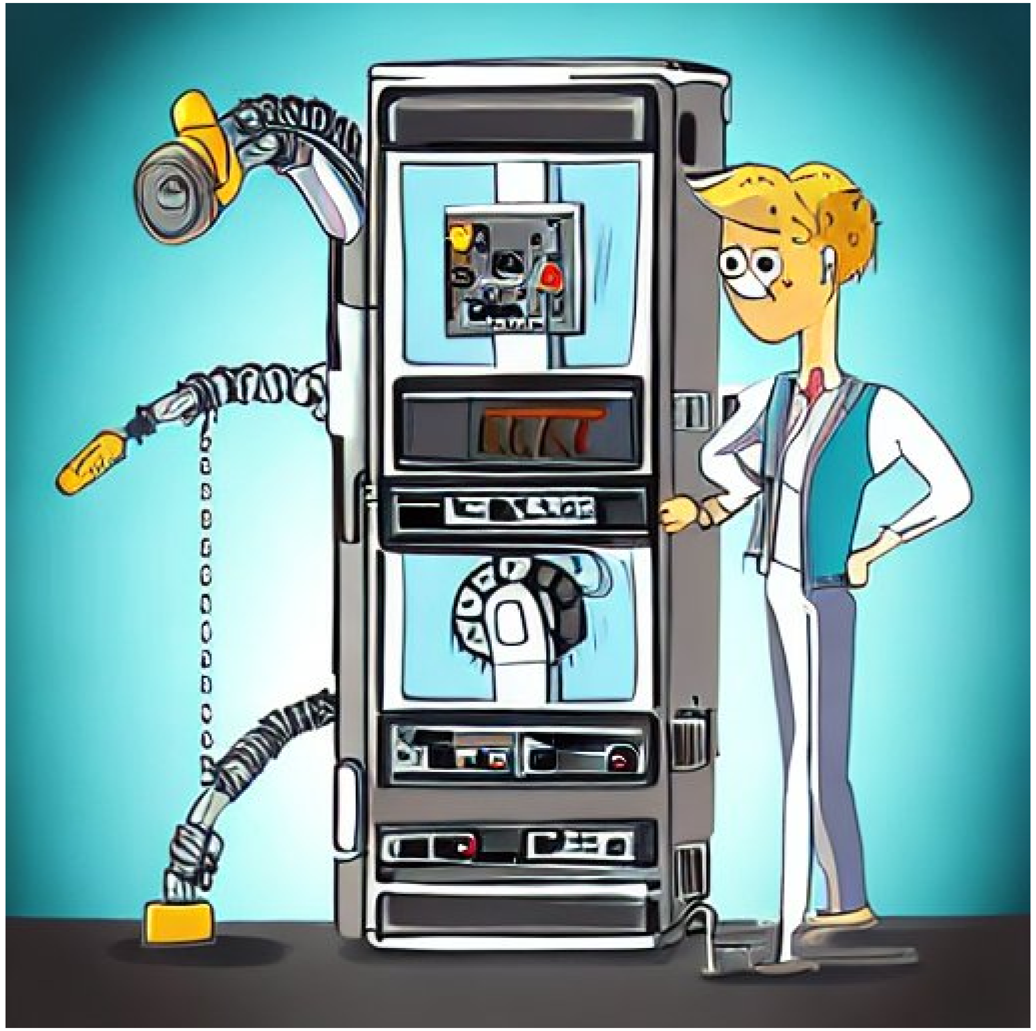
All images in this talk are from Stable Diffusion

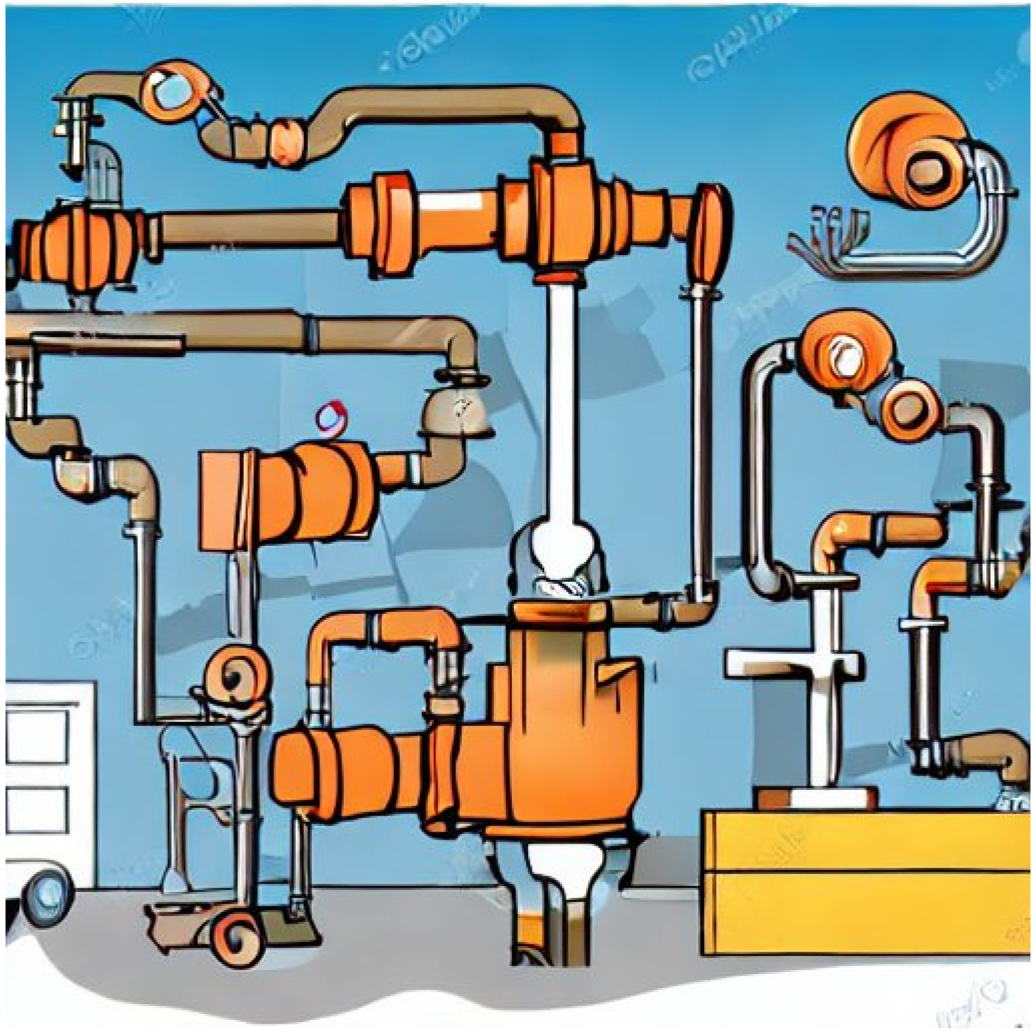
But this talk is all about text

So we'll skip over all of that crazy-fun area of AI/ML!

stablediffusionweb.com

So what is an LLM, and/or a LLaMa?





LLM - Large Language Model

LLaMA - Facebook's (slightly open) LLM

LLaMa - Slightly open?

Code is open, model is not

But more on that later!

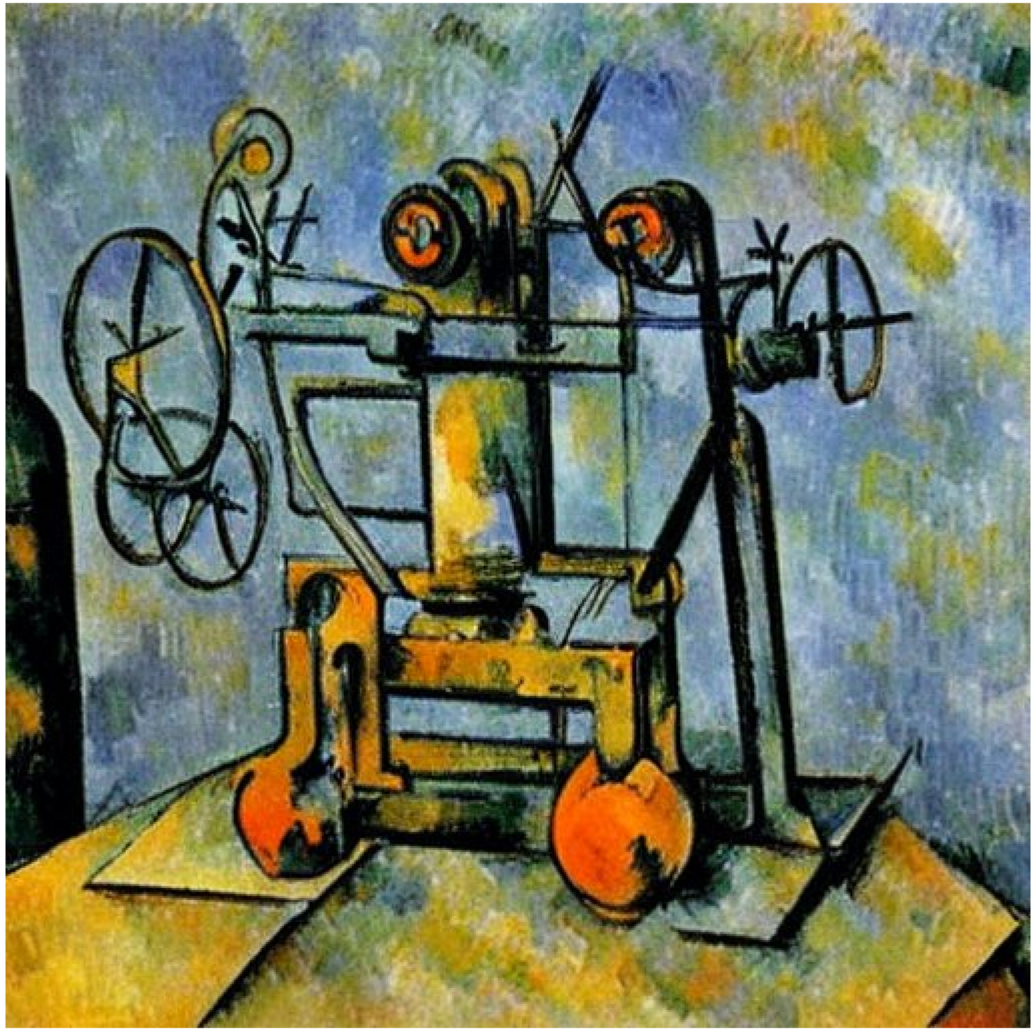
LLM - Large Language Model?

A text-focused neural network with billions of parameters, trained on large amounts of text.

LLMs are deep-learning models, typically general purpose, which excel at a wide range of text-related tasks.

Some level of encoding on syntax and sematics of human language, and some level of general knowledge / facts.

Bigger models (both number of parameters and training corpus) will tend to show more knowledge and better semantics.



Don't ignore the old stuff!

Everything you could do with BERT or ELMo, you can do with a LLM

- Embeddings? ✅
- Vector Search? ✅
- Similar Words? ✅
- Semantic Relationships? ✅

So look back through the Buzzwords 2022 and 2021 presentations

(They're all online on youtube)

If it mentioned BERT, you can try it with an LLM, and it'll (probably) be better!

But a LLM on a laptop?

Live demo time!

llama.cpp, Facebook's LLaMa 7B (7 billion parameters) model, and a few of your questions

Some ML + Information Retrieval terms to know

Tokenization

Breaking the input text into chunks, and possibly some simple transformations

eg *Hello there! Welcome to my talk.*

Could become

`hello there welcome to my talk`

Or... `hello there [!] welcome to my talk [end]`

Embeddings

Both IR systems like Lucene, and ML techniques, need to work on numeric representations.

Lucene has a *term dictionary*, eg `1=hello 142=there`

ML techniques have *embeddings*, a vector-space representation / projection, eg `[0, 0, 0.9, 0, 0.1, 0.8]`

Embeddings

If this is all new to you, or you're a bit hazy on the details...

What are embeddings

by *Vicki Boykis*

github.com/veekaybee/what_are_embeddings

What could the "old stuff" like BERT do?

Similar Words (semantically similar)

What other words are similar to X?

How similar are the words X and Y?

Similar tokens to: linux

- Cosine sim=0.849: unix
- Cosine sim=0.793: open-source
- Cosine sim=0.778: kernel

Similar tokens to: raise

- Cosine sim=0.890: raising
- Cosine sim=0.871: pay
- Cosine sim=0.848: benefit

Difference between raise and risen is 56

Difference between raise and above is 52

Difference between raise and below is 54

Difference between raise and shine is 30

Difference between raise and linux is 11

Word relationships

What is the equivalent of X to Y, for Z?

The analogy of X - Y for Z = ???

berlin - germany for paris = france

madrid - spain for lisbon = portugal

man - boy for woman = girl

Warning - it can be wrong, it can be sexist!

The analogy of X - Y for Z = ???

spain - madrid for portugal = spain

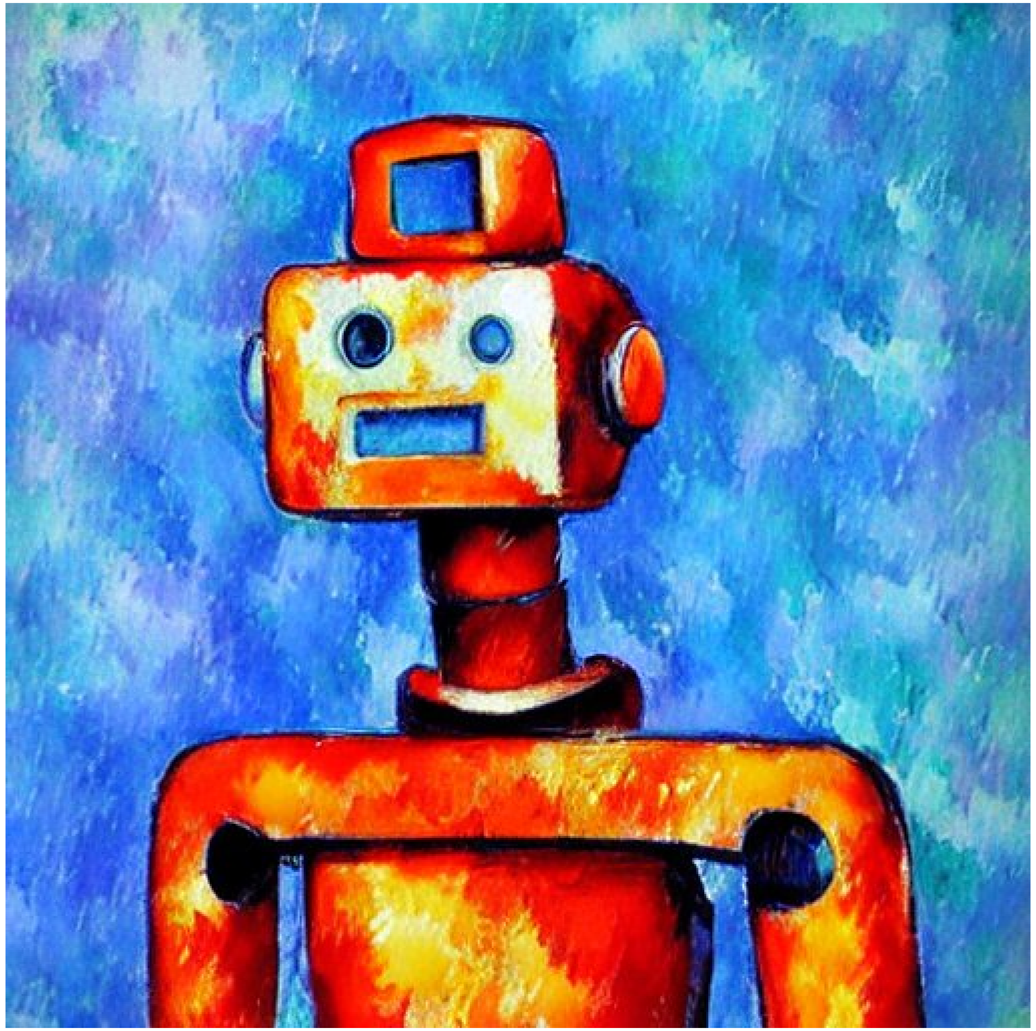
doctor - man for nurse = woman

Embeddings - sentence and/or document

Used in Vector Search (supported by Lucene, elastic etc)

- Transform the documents into the embedding vector space
- Transform the query into the embedding vector space
- Find documents similar to the query

Combination of embedding vector search and normal search can work well, see previous talks for details!



Rest to follow...

Code, Scripts, Slides

github.com/Gagravarr/BBuzz23-LaptopML

All code in slides is taken from here!