When SDMX Meets Ai: Leveraging Open Source LLMs to Make Official Statistics More Accessible and Discoverable

9th SDMX Global Conference

Alessandro Benedetti

29/10/2023

WHO AM I ?

ALESSANDRO BENEDETTI

- Born in Tarquinia (ancient Etruscan city in Italy)
- R&D Software Engineer
- Director
- Master degree in Computer Science
- PC member for ECIR, SIGIR and Desires
- Apache Lucene/Solr PMC member/committer
- Elasticsearch/OpenSearch expert
- Semantic search, NLP, Machine Learning technologies passionate
- Beach Volleyball player and Snowboarder

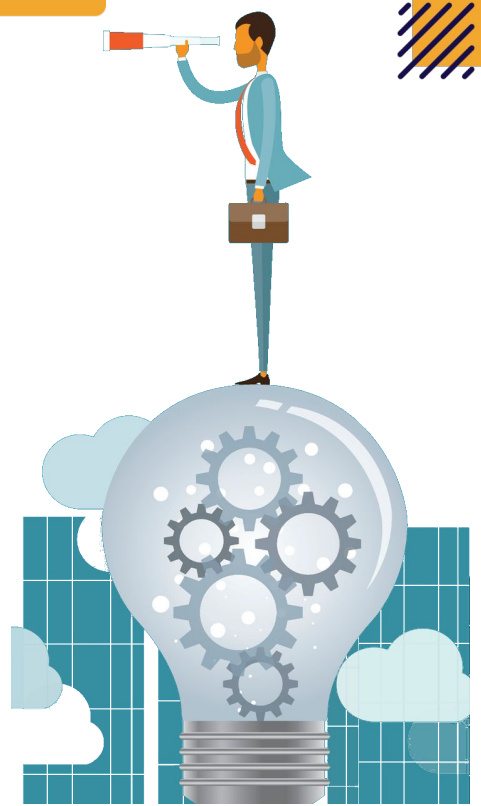
SEArch SErvices

www.sease.io

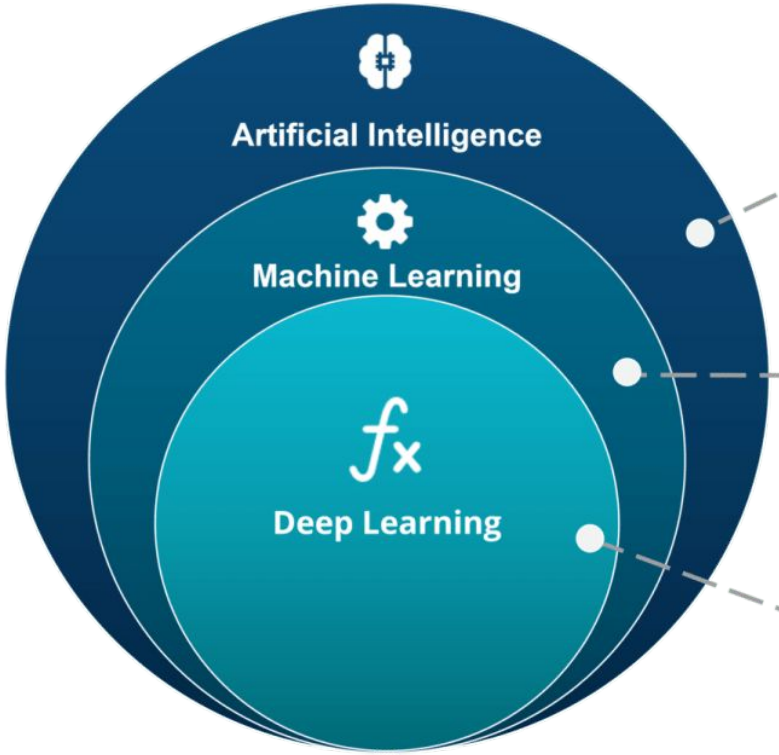
- Headquarter in London/distributed
- Open-source Enthusiasts
- Apache Lucene/Solr experts
- Elasticsearch/OpenSearch experts
- Community Contributors
- Active Researchers
- **Hot Trends** : Neural Search,
  Natural Language Processing
  Learning To Rank,
  Document Similarity,
  Search Quality Evaluation,
  Relevance Tuning

AGENDA

- LLMs and Open Source
- SIS-CC to enable AI applications with SDMX
- From Natural Language to structured queries
- Findings and future Works



AI, Machine learning and Deep Learning



ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

MACHINE LEARNING

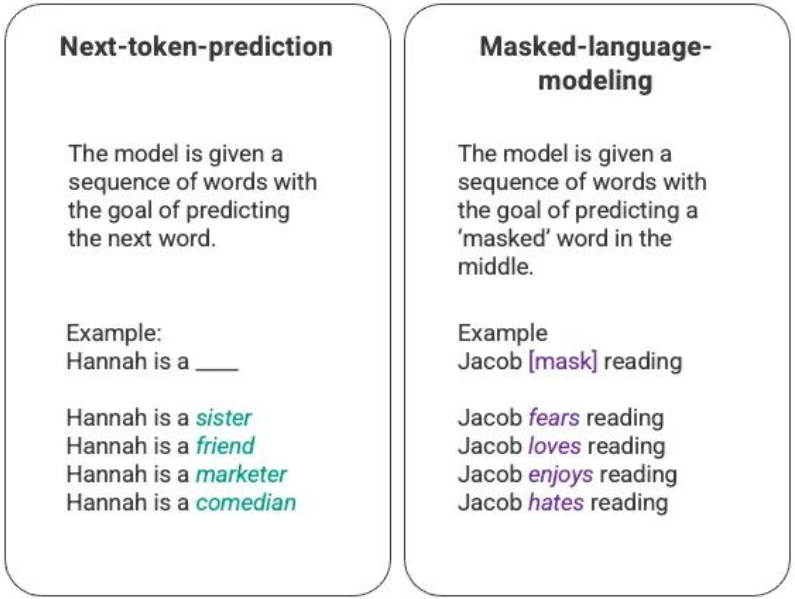
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible

WHAT IS A LARGE LANGUAGE MODEL?

- Transformers
- Next-token-prediction and masked-language-modeling
- estimate the likelihood of each possible word (in its vocabulary) given the previous sequence
- learn the statistical structure of language
- pre-trained on huge quantities of text



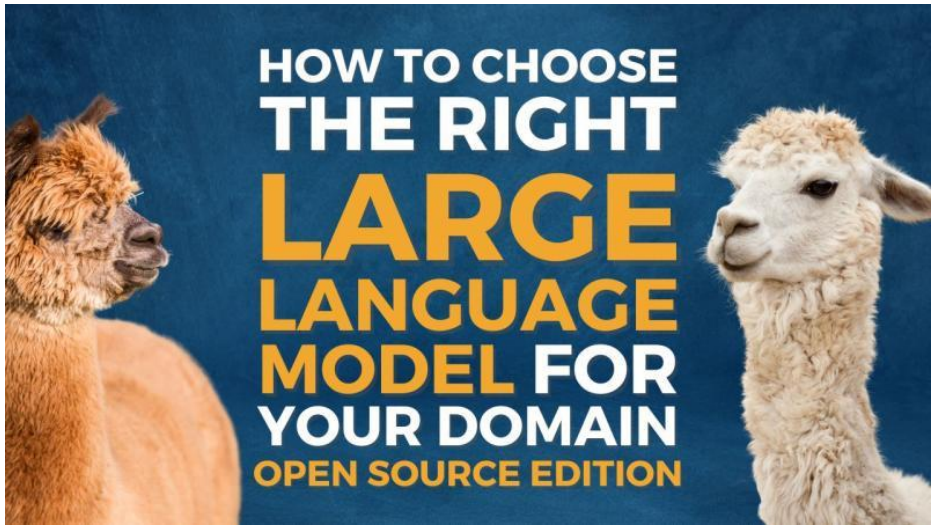
https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

OPEN SOURCE LARGE LANGUAGE MODELS

https://sease.io/2023/06/how-to-choose-the-right-large-language-model-for-your-domain-open-source-edition.html

- **Generalists**
  - Falcon
  - LLaMA
    - alpaca
    - vicuna

… many others!

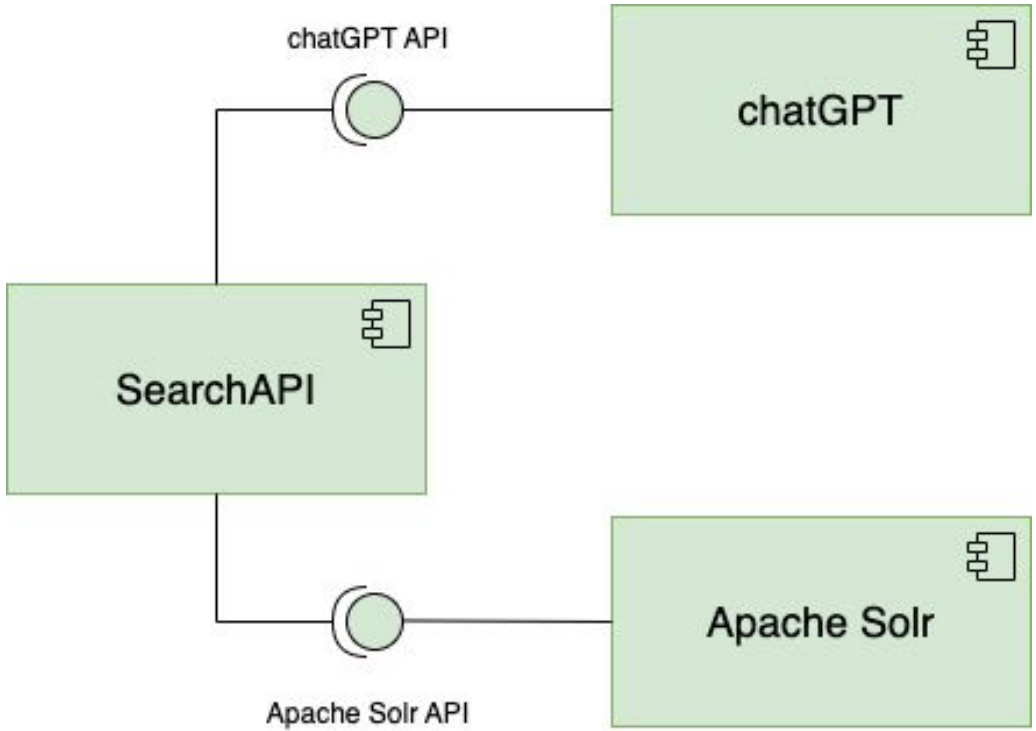


https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

SIS-CC TO ENABLE AI APPLICATIONS WITH SDMX

- OECD lead initiative (The Organisation for Economic Co-operation and Development)
- The Statistical Information System Collaboration Community
- .Stat Suite and Apache Solr https://siscc.org/developers/technology/

POC



FROM NATURAL LANGUAGE TO STRUCTURED QUERIES

Use case example

Natural language query:
What were the sulfur oxide emissions in Australia in 2013?

Search API

2 Requests to LLM

GENERATIVE

EXTRACTIVE

Statistic

```
{ "name":"Emissions of air pollutants",
"id":"ds-siscc-qa:DF_AIR_EMISSIONS"}] },
"Dimensions":[ "Country", "Pollutant", "Unit", "Year"]}, …
```

Filters

```
{
"Country": ["0|Australia#AUS#"],
"Pollutant": ["0|Sulphur Oxides#SOX#"],
"Unit Of measure": ["0|Total man-made emissions#TOT#"],
"Year": "2013"
}
```

IMPROVING OVER TIME

- How do you update this approach in time?
- New large language models?
- Better prompts?
- How to fit user interactions?

RESULTS: What were the sulfur oxide emissions in Australia in 2013

GPT Generative answer is:

['Sulfur dioxide emissions', 'Air pollution', 'Environmental impact', 'Fossil fuel combustion', 'Acid rain']

GPT Extractive answer is:

{'srQMgwl_en_ss': ['1|Environment#ENV#|Air and climate#ENV_AC#'], 'dimensions_en_ss': ['Time period', 'Reference area', 'Pollutant', 'Country']}

Dataflow retrieved is:

[{'id': 'ds-siscc-qa:DF_AIR_EMISSIONS', 'name': 'Emissions of air pollutants', 'description': ''}]

4 dimensions are available for the above dataflow:

['Country', 'Year', 'Pollutant', 'Variable']


```
{
    "dataflow": [
        {
            "description": "",
            "id": "ds-siscc-qa:DF_AIR_EMISSIONS",
            "name": "Emissions of air pollutants"
        }
    ],
    "filters": {
        "Country": "0|Australia#AUS#",
        "Pollutant": "0|Sulphur Oxides#SOX#",
        "Variable": "0|Total man-made emissions#TOT#",
        "Year": "2013"
    },
    "natural_language_query": "What were the sulfur oxide emissions in Australia in 2013"
}
```

FINDINGS

- **Promising! -** LLM are good in query expansion (generative or extractive)
- **gpt-3.5-turbo-instruct ->** new models can do much better!
- **4k tokens -** (for both prompt and response) is not enough
- **Mistakes -** Some times dimension values are associated to the wrong dimension, more prompt engineering!

THE ROAD TO PRODUCTION

- [Solr] Fine-tune the dimension retrieval Solr query
- [LLM] Select the best model to date - Explore the State of the Art (both commercially and Open source)
- [LLM] Refine the prompts according to the model
- [LLM] Implement integration tests with the most common failures -> LLM/prompt engineering to solve them
- [Solr] Finalising the dataflow retrieval Solr query
- [Performance] Stress test the solution
- [Quality] Set up queries/expected documents

FUTURE WORKS

- StatsBot - More conversation!
- Retrieval Augmented Generation
- Results Summarization