Data Mining – Jeopardy Project

In the thrilling world of game shows, few are as captivating as Jeopardy!. But can the computer beat us? The secret is found in a simple tactic: to use the vast collection of human knowledge stored inside the Wikipedia pages.

Chris Cărpineanu, Victor Măcinic, Horațiu Pop

Maria Sîntea, Alexandra Tudorescu

Indexing and Retrieval Process

Using **Apache Lucene** and **CoreNLP** for the indexing and retrieval methods.

Lemmatization

Grammar-based preprocessing, ensuring accuracy as opposed to stemming.

Parallelized Index Building

To efficiently write index to file and be able to reuse it.

QueryParser & Searcher

Efficient information retrieval.

Building the Query

Using the category and all words in the clue (except for "Alex") → maximizing relevance

Challenges & Highlights

Challenges

Inconsistent File Formatting:

"[[Titles]]", "[[Publishers]]" - which is which??

→ Titles are identified by checking if a line starts with "[[" and ends with "]]".

Empty Document Handling:

Empty documents, disrupting processing → skip these pages during processing.

Highlights

Multiple relevant results

Once the system runs out of pages similar to the content of the query, the other results are related to the category of the question.

e.g.: for a question related to the Nile and Cairo, the system returned as a first answer Cairo (which is the correct answer), and then was followed by the "Geography of Egypt" and "Demographics of Egypt", and then results related to African cities category.

Deep Learning Approach for Enhanced Retrieval

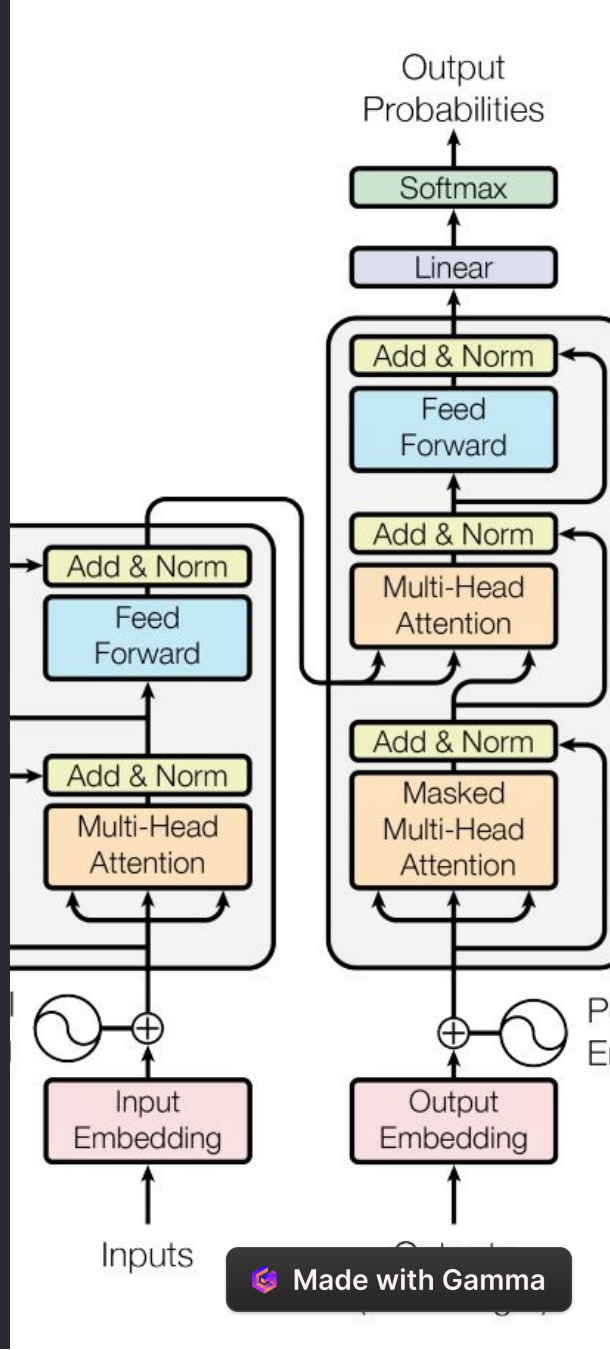
Utilizing a specialized pre-trained NLP model to optimize performance and enhance question retrieval.

1 Embedding Indexing

Creating a vector database (***Qdrant***) for efficient and accurate information retrieval.

2 Similarity Inference

Utilizing cosine similarity metrics to retrieve the most similar embeddings.



Measuring Performance and Error Analysis

Metric	Lucene Index Retrieval	Deep Learning
MRR	23.18%	25.84%
P@1	17%	23%
P@3	24%	28%

Error Analysis

"The Naples Museum of Art" → should be Florida

Our system → list of museums (Maryhill Museum of Art, Colin Center for the Art)

Why? **Clue**! "We'll give you the museum. You give us the state."