MSA 8395

SPECIAL TOPICS FOR ANALYTICS

Text Analysis: Extracting Licensor & Licensee

Hari Priya Avarampalayam Manoharan

Panther Id: 002711275

As part of this course, I worked on three problem statements.

**PROBLEM STATEMENT 1:**

Extracting Licensor and Licensee information from a set of License documents.

**Solution:**

Initially, Named Entity Recognition (NER) to retrieve Licensor and Licensee information from a collection of License documents, but the results were not satisfactory. Subsequently, experimented with the LLAMA2 model and discovered improved accuracy. Further refinement involved testing different temperatures (0.5, 0.8, and 0.9), with the optimal outcome achieved at a temperature setting of 0.8.

**Why LLAMA Model?**

Contextual Understanding: LLAMA, being a language model, has a strong capacity to understand the context of text. This is crucial for tasks like identifying Licensers and Licensees, as it often depends on the surrounding information and context within a document.

Flexibility: LLAMA's flexibility allows it to handle a wide range of text data and adapt to different types of information extraction tasks. It can generate responses based on user prompts, which makes it suitable for the conversational context in which the task is framed.

Potential for Improved Accuracy: LLAMA's ability to generate responses based on context may lead to more accurate information extraction compared to traditional NER models, which rely on predefined entity types. In summary, the decision to transition from NER models to the LLAMA model was driven by the need for improved accuracy and context-awareness in the task of extracting Licenser and Licensee information from text data. LLAMA's language generation capabilities and contextual understanding make it a promising choice for this specific information extraction task.

**Limitation:** We had limited access to the GPU.

**LLAMA Implementation:**

The below images show the implementation of LLAMA model .

A green and black text

Description automatically generated**A screenshot of a computer program

Description automatically generated**

A screenshot of a computer code

Description automatically generated

**Results:**

The below image shows the LLAMA giving the Licensor and Licensee details.

A screenshot of a computer

Description automatically generated

Finally, tried filtering the sentences containing words like (license, licensed, or licensing) and then tried feeding it to the model but it gave the same results. Also due to LLAMA limitation, we couldn’t proceed with the entire dataset.

**PROBLEM STATEMENT 2:**

Effectively identifying the technology utilized in each License document.

Experimented with NER, Turbo GPT-3.5 and LLAMA models. Additionally, tried LLAMA question/answer model and compared results with LLAMA prompt model.

**NER approach:**

Named Entity Recognition (NER) is a natural language processing (NLP) technique that focuses on identifying and classifying named entities in text into predefined categories.

**Limitation:**

A close up of text

Description automatically generatedA screenshot of a computer program

Description automatically generatedThe NER approach proved inadequate in accurately extracting technology information.

**Turbo GPT 3.5 model:**

GPT-3.5 API is a programming interface (API) that allows developers to access and use the GPT-3.5 language model from OpenAI.

**Limitation:**

This model had a limitation of limited access to the API.

**A screenshot of a computer

Description automatically generatedA screenshot of a computer code

Description automatically generated**

**LLAMA Prompt model:**

Llama Prompt Model is specifically designed for text generation tasks, and it can be used to generate different creative text formats, like poems, code, scripts, musical pieces, email, letters, etc. It is known for its ability to generate high-quality, creative, and grammatically correct text.

This model was able to extract the technology details accurately.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**LLAMA Question/Answering model:**

Llama Q/A model is specifically designed for question answering (QA) tasks, and it can be used to answer a wide variety of questions, including open ended, challenging, or strange questions, questions that require common sense reasoning, and questions that require multi-step reasoning. It is known for its ability to provide comprehensive and informative answers, even to complex and challenging questions.

**Example essay for comparing LLAMA prompt and Q/A model:**

A screenshot of a email

Description automatically generated

**Q/A model output:**

A screenshot of a computer program

Description automatically generated

**Prompt output:**



**Result:**

LLAMA prompt model gave better results compared to LLAMA Question/Answering model. Q/A model gave same name to both Licensor and Licensee details.

**PROBLEM STATEMENT 3:**

Implemented a methodology to extract similar license documents using Cosine and Fuzzywuzzy similarity scores.

**Solution:**

Engaged with data from the year 2011, conducted deduplication, and executed data preprocessing by eliminating special characters, stopwords, and converting text to lowercase. Applied cosine similarity and employed K-means clustering to group similar documents, experimenting with both 5 and 10 clusters.

However, it was observed that the Cosine Similarity method did not effectively identify similar documents, as evidenced by the word cloud below generated from Cluster 1, where dissimilar words were prevalent. The below word cloud shows how cosine similarity is not that effective.

A close-up of words

Description automatically generated

**Fuzzywuzzy:**

Leveraged the Fuzzywuzzy similarity score to identify comparable documents. Eliminated records that were 100% identical and filtered the remaining dataset to include only those with a similarity score greater than or equal to 80%.

A screenshot of a computer

Description automatically generated

**Learning from the course:**

I gained knowledge about recent models such as LLAMA and Turbo GPT 3.5, delving into the intricacies of generative AI architectures. Through hands-on experience with various models, I acquired insights into their respective limitations. The discovery of Fuzzywuzzy and its impressive results led me to recommend its application in a project at my workplace, specifically at LexisNexis Risk Solutions. Grateful for the opportunity, I extend my thanks to Professor Yusen Xia and Yaswanth Pothuru.