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.

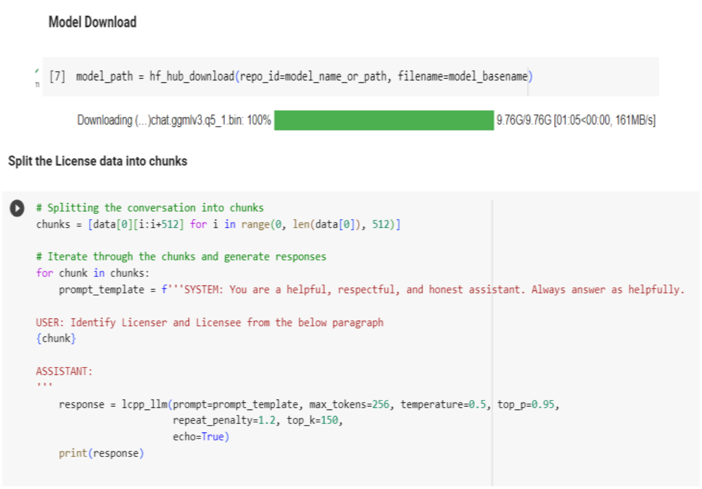
Flexibility: LLAMA's flexibility allows it to handle a wide range of text data and adapt to different types of information extraction tasks. It can generate responses based on user prompts, which makes it suitable for the conversational context in which the task is framed.

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 .

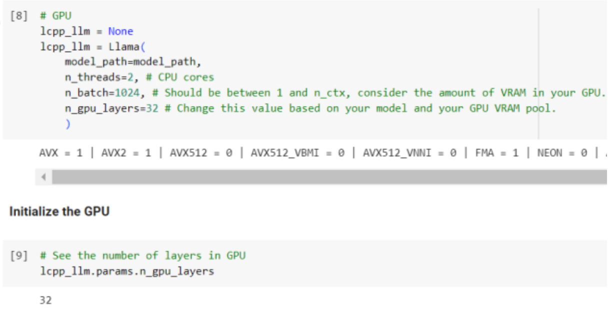


Model Download

```
[7]  model_path = hf_hub_download(repo_id=model_name_or_path, filename=model_basename)

      Downloading (…)chat.ggmlv3.q5_1.bin: 100%  9.76G/9.76G [01:05<00:00, 161MB/s]
```

Split the License data into chunks

```
# Splitting the conversation into chunks
chunks = [data[0][i:i+512] for i in range(0, len(data[0]), 512)]

# Iterate through the chunks and generate responses
for chunk in chunks:
    prompt_template = f'''SYSTEM: You are a helpful, respectful, and honest assistant. Always answer as helpfully.

USER: Identify Licenser and Licensee from the below paragraph
{chunk}

ASSISTANT:
'''
    response = lcpp_llm(prompt=prompt_template, max_tokens=256, temperature=0.5, top_p=0.95,
                        repeat_penalty=1.2, top_k=150,
                        echo=True)
    print(response)
```



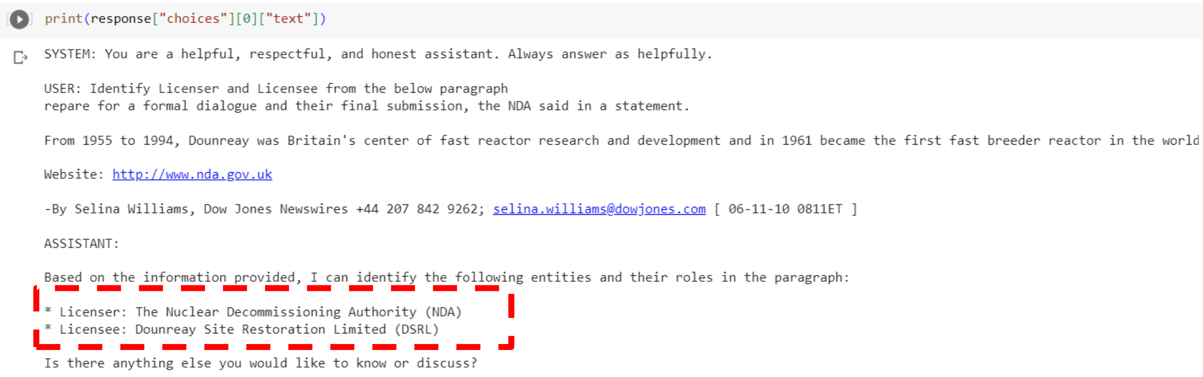
```
[8]  # GPU
     lcpp_llm = None
     lcpp_llm = Llama(
         model_path=model_path,
         n_threads=2, # CPU cores
         n_batch=1024, # Should be between 1 and n_ctx, consider the amount of VRAM in your GPU.
         n_gpu_layers=32 # Change this value based on your model and your GPU VRAM pool.
         )

     AVX = 1 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 |
```

Initialize the GPU

```
[9]  # See the number of layers in GPU
     lcpp_llm.params.n_gpu_layers

     32
```

Results:

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



```
print(response["choices"][0]["text"])
```

```
SYSTEM: You are a helpful, respectful, and honest assistant. Always answer as helpfully.

USER: Identify Licenser and Licensee from the below paragraph
repare for a formal dialogue and their final submission, the NDA said in a statement.

From 1955 to 1994, Dounreay was Britain's center of fast reactor research and development and in 1961 became the first fast breeder reactor in the world

Website: http://www.nda.gov.uk

-By Selina Williams, Dow Jones Newswires +44 207 842 9262; selina.williams@dowjones.com [ 06-11-10 0811ET ]

ASSISTANT:

Based on the information provided, I can identify the following entities and their roles in the paragraph:

* Licenser: The Nuclear Decommissioning Authority (NDA)
* Licensee: Dounreay Site Restoration Limited (DSRL)

Is there anything else you would like to know or discuss?
```

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.

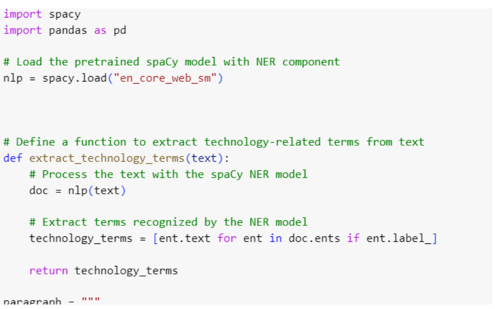
Experimented with NER, Turbo GPT-3.5 and LLAMA models. Additionally, tried LLAMA question/answer model and compared results with LLAMA prompt model.

NER approach:

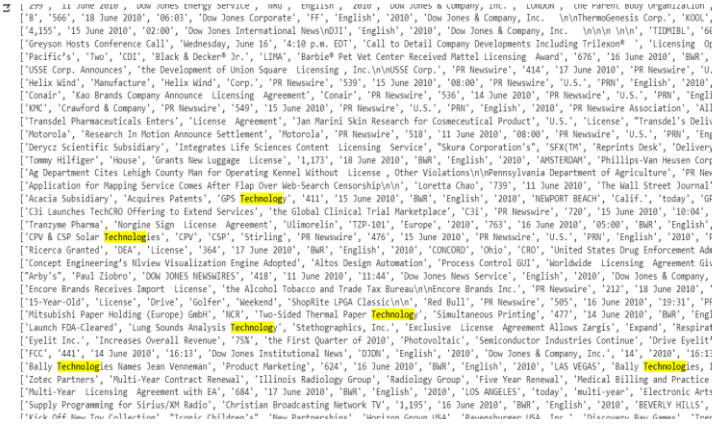
Named Entity Recognition (NER) is a natural language processing (NLP) technique that focuses on identifying and classifying named entities in text into predefined categories.

Limitation:

The NER approach proved inadequate in accurately extracting technology information.



```
import spacy
import pandas as pd

# Load the pretrained spaCy model with NER component
nlp = spacy.load("en_core_web_sm")


# Define a function to extract technology-related terms from text
def extract_technology_terms(text):
    # Process the text with the spaCy NER model
    doc = nlp(text)

    # Extract terms recognized by the NER model
    technology_terms = [ent.text for ent in doc.ents if ent.label_]

    return technology_terms

paragraph = """
```



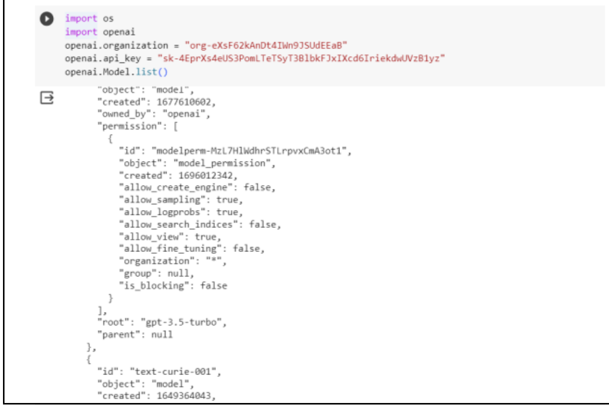
```
['299', '11 June 2010', 'Dow Jones Energy Service', 'NRG', 'English', '2010', 'Dow Jones & Company, Inc.', 'LONDON', 'the Parent Body Organization',
['8', '566', '18 June 2010', '06:03', 'Dow Jones Corporate', 'FF', 'English', '2010', 'Dow Jones & Company, Inc.   \n\nThermoGenesis Corp.', 'KOOL',
['4,155', '15 June 2010', '02:00', 'Dow Jones International News\nDJI', 'English', '2010', 'Dow Jones & Company, Inc.   \n\n\n \n\n', 'TIDMIBL', '66
['Greyson Hosts Conference Call', 'Wednesday, June 16', '4:10 p.m. EDT', 'Call to Detail Company Developments Including Trilexon®  ', 'Licensing  Op
['Pacific's', 'Two', 'CDI', 'Black & Decker® Jr.', 'LIMA', 'Barbie® Pet Vet Center Received Mattel Licensing  Award', '676', '16 June 2010', 'BWR',
['USSE Corp. Announces', 'the Development of Union Square  Licensing , Inc.\n\nUSSE Corp.', 'PR Newswire', '414', '17 June 2010', 'PR Newswire', 'U.
['Helix Wind', 'Manufacture', 'Helix Wind', 'Corp.', 'PR Newswire', '539', '15 June 2010', '08:00', 'PR Newswire', 'U.S.', 'PRN', 'English', '2010',
['Conair', 'Kao Brands Company Announce  Licensing  Agreement', 'Conair', 'PR Newswire', '536', '14 June 2010', 'PR Newswire', 'U.S.', 'PRN', 'Engli
['KMC', 'Crawford & Company', 'PR Newswire', '549', '15 June 2010', 'PR Newswire', 'U.S.', 'PRN', 'English', '2010', 'PR Newswire Association', 'All
['Transdel Pharmaceuticals Enters', 'License  Agreement', 'Jan Marini Skin Research for Cosmeceutical Product', 'U.S.', 'License', "Transdel's Deli
['Motorola', 'Research In Motion Announce Settlement', 'Motorola', 'PR Newswire', '518', '11 June 2010', '08:00', 'PR Newswire', 'U.S.', 'PRN', 'Eng
['Derycz Scientific Subsidiary', 'Integrates Life Sciences Content  Licensing  Service', "Skura Corporation's", 'SFX(TM', 'Reprints Desk', 'Delivery
['Tommy Hilfiger', 'House', 'Grants New Luggage  License', '1,173', '18 June 2010', 'BWR', 'English', '2010', 'AMSTERDAM', 'Phillips-Van Heusen Corp
['Ag Department Cites Lehigh County Man for Operating Kennel Without  License , Other Violations\n\nPennsylvania Department of Agriculture', 'PR New
['Application for Mapping Service Comes After Flap Over Web-Search Censorship\n\n', 'Loretta Chao', '739', '11 June 2010', 'The Wall Street Journal'
['Acacia Subsidiary', 'Acquires Patents', 'GPS Technology', '411', '15 June 2010', 'BWR', 'English', '2010', 'NEWPORT BEACH', 'Calif.', 'today', 'GF
['C3i Launches TechCRO Offering to Extend Services', 'the Global Clinical Trial Marketplace', 'C3i', 'PR Newswire', '720', '15 June 2010', '10:04',
['Tranzyme Pharma', 'Norgine Sign  License  Agreement', 'Ulimorelin', 'TZP-101', 'Europe', '2010', '763', '16 June 2010', '05:00', 'BWR', 'English',
['CPV & CSP Solar Technologies', 'CPV', 'CSP', 'Stirling', 'PR Newswire', '476', '15 June 2010', 'PR Newswire', 'U.S.', 'PRN', 'English', '2010', 'F
['Ricerca Granted', 'DEA', 'License', '364', '17 June 2010', 'BWR', 'English', '2010', 'CONCORD', 'Ohio', 'CRO', 'United States Drug Enforcement Adm
['Concept Engineering's Nlview Visualization Engine Adopted', 'Altos Design Automation', 'Process Control GUI', 'Worldwide  Licensing  Agreement Giv
["Arby's", 'Paul Ziobro', 'DOW JONES NEWSWIRES', '418', '11 June 2010', '11:44', 'Dow Jones News Service', 'English', '2010', 'Dow Jones & Company,
['Encore Brands Receives Import  License', 'the Alcohol Tobacco and Trade Tax Bureau\n\nEncore Brands Inc.', 'PR Newswire', '212', '18 June 2010', '
['15-Year-Old', 'License', 'Drive', 'Golfer', 'Weekend', 'ShopRite LPGA Classic\n\n', 'Red Bull', 'PR Newswire', '505', '16 June 2010', '19:31', 'PF
['Mitsubishi Paper Holding (Europe) GmbH', 'NCR', 'Two-Sided Thermal Paper Technology', 'Simultaneous Printing', '477', '14 June 2010', 'BWR', 'Engl
['Launch FDA-Cleared', 'Lung Sounds Analysis Technology', 'Stethographics, Inc.', 'Exclusive  License  Agreement Allows Zargis', 'Expand', 'Respirat
['Eyelit Inc.', 'Increases Overall Revenue', '75%', 'the First Quarter of 2010', 'Photovoltaic', 'Semiconductor Industries Continue', 'Drive Eyelit'
['FCC', '441', '14 June 2010', '16:13', 'Dow Jones Institutional News', 'DJDN', 'English', '2010', 'Dow Jones & Company, Inc.', '14', '2010', '16:13
['Bally Technologies Names Jean Venneman', 'Product Marketing', '624', '16 June 2010', 'BWR', 'English', '2010', 'LAS VEGAS', 'Bally Technologies, I
['Zotec Partners', 'Multi-Year Contract Renewal', 'Illinois Radiology Group', 'Radiology Group', 'Five Year Renewal', 'Medical Billing and Practice
['Multi-Year  Licensing  Agreement with EA', '684', '17 June 2010', 'BWR', 'English', '2010', 'LOS ANGELES', 'today', 'multi-year', 'Electronic Arts
['Supply Programming for Sirius/XM Radio', 'Christian Broadcasting Network TV', '1,195', '16 June 2010', 'BWR', 'English', '2010', 'BEVERLY HILLS',
['Kick Off New Toy Collection', "Iconic Children's", 'New Partnerships', 'Horizon Group USA', 'Ravensburger USA, Inc.', 'Discovery Bay Games', 'Tren
```

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.

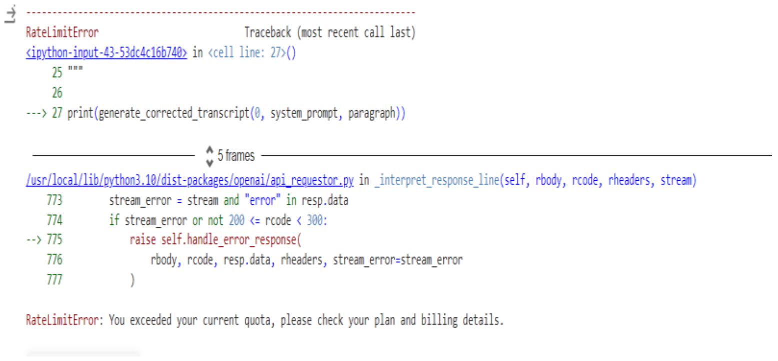
Limitation:

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



```
import os
import openai
openai.organization = "org-eXsF62kAnDt4IWn9JSUdEEaB"
openai.api_key = "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"
openai.Model.list()
```

```
    "object": "model",
    "created": 1677610602,
    "owned_by": "openai",
    "permission": [
      {
        "id": "modelperm-MzL7HlWdhrSTLrpvxCmA3ot1",
        "object": "model_permission",
        "created": 1696012342,
        "allow_create_engine": false,
        "allow_sampling": true,
        "allow_logprobs": true,
        "allow_search_indices": false,
        "allow_view": true,
        "allow_fine_tuning": false,
        "organization": "*",
        "group": null,
        "is_blocking": false
      }
    ],
    "root": "gpt-3.5-turbo",
    "parent": null
  },
  {
    "id": "text-curie-001",
    "object": "model",
    "created": 1649364043,
```

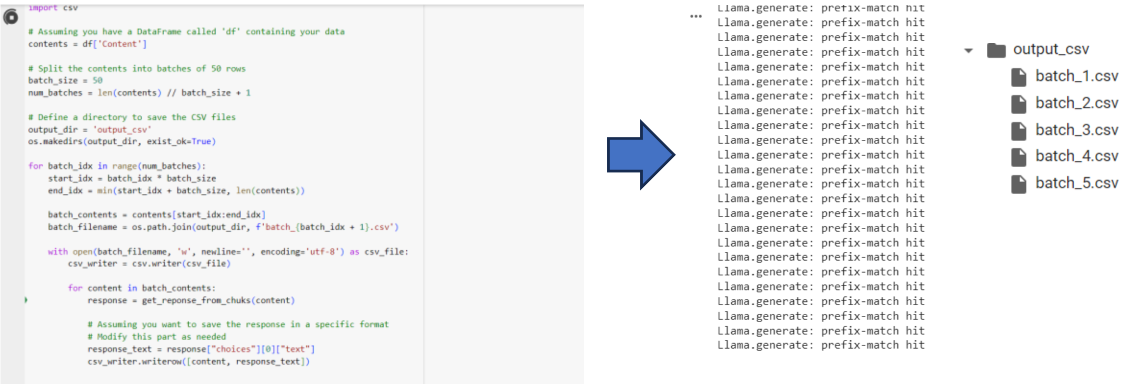


```
---------------------------------------------------------------------------
RateLimitError                            Traceback (most recent call last)
<ipython-input-43-53dc4c16b740> in <cell line: 27>()
     25 """
     26
---> 27 print(generate_corrected_transcript(0, system_prompt, paragraph))

5 frames
/usr/local/lib/python3.10/dist-packages/openai/api_requestor.py in _interpret_response_line(self, rbody, rcode, rheaders, stream)
    773         stream_error = stream and "error" in resp.data
    774         if stream_error or not 200 <= rcode < 300:
--> 775             raise self.handle_error_response(
    776                 rbody, rcode, resp.data, rheaders, stream_error=stream_error
    777             )

RateLimitError: You exceeded your current quota, please check your plan and billing details.
```

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.



```
import csv

# Assuming you have a DataFrame called 'df' containing your data
contents = df['Content']

# Split the contents into batches of 50 rows
batch_size = 50
num_batches = len(contents) // batch_size + 1

# Define a directory to save the CSV files
output_dir = 'output_csv'
os.makedirs(output_dir, exist_ok=True)

for batch_idx in range(num_batches):
    start_idx = batch_idx * batch_size
    end_idx = min(start_idx + batch_size, len(contents))

    batch_contents = contents[start_idx:end_idx]
    batch_filename = os.path.join(output_dir, f'batch_{batch_idx + 1}.csv')

    with open(batch_filename, 'w', newline='', encoding='utf-8') as csv_file:
        csv_writer = csv.writer(csv_file)

        for content in batch_contents:
            response = get_reponse_from_chuks(content)

            # Assuming you want to save the response in a specific format
            # Modify this part as needed
            response_text = response["choices"][0]["text"]
            csv_writer.writerow([content, response_text])
```

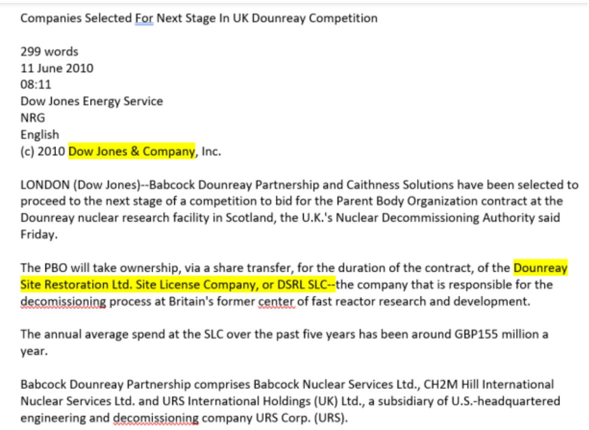
```
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
Llama.generate: prefix-match hit
```

```
output_csv
    batch_1.csv
    batch_2.csv
    batch_3.csv
    batch_4.csv
    batch_5.csv
```

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:



Companies Selected For Next Stage In UK Dounreay Competition

299 words
11 June 2010
08:11
Dow Jones Energy Service
NRG
English
(c) 2010 Dow Jones & Company, Inc.

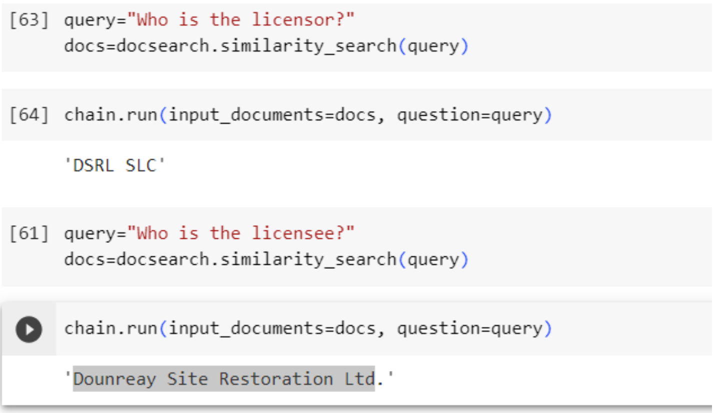
LONDON (Dow Jones)--Babcock Dounreay Partnership and Caithness Solutions have been selected to proceed to the next stage of a competition to bid for the Parent Body Organization contract at the Dounreay nuclear research facility in Scotland, the U.K.'s Nuclear Decommissioning Authority said Friday.

The PBO will take ownership, via a share transfer, for the duration of the contract, of the Dounreay Site Restoration Ltd. Site License Company, or DSRL SLC--the company that is responsible for the decomissioning process at Britain's former center of fast reactor research and development.

The annual average spend at the SLC over the past five years has been around GBP155 million a year.

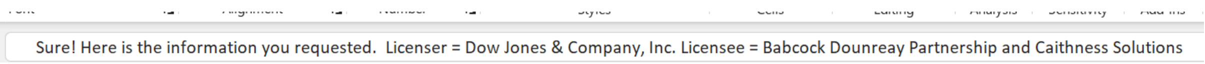
Babcock Dounreay Partnership comprises Babcock Nuclear Services Ltd., CH2M Hill International Nuclear Services Ltd. and URS International Holdings (UK) Ltd., a subsidiary of U.S.-headquartered engineering and decomissioning company URS Corp. (URS).

Q/A model output:



```
[63] query="Who is the licensor?"
     docs=docsearch.similarity_search(query)

[64] chain.run(input_documents=docs, question=query)

     'DSRL SLC'

[61] query="Who is the licensee?"
     docs=docsearch.similarity_search(query)

     chain.run(input_documents=docs, question=query)

     'Dounreay Site Restoration Ltd.'
```

Prompt output:



Sure! Here is the information you requested. Licenser = Dow Jones & Company, Inc. Licensee = Babcock Dounreay Partnership and Caithness Solutions

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.

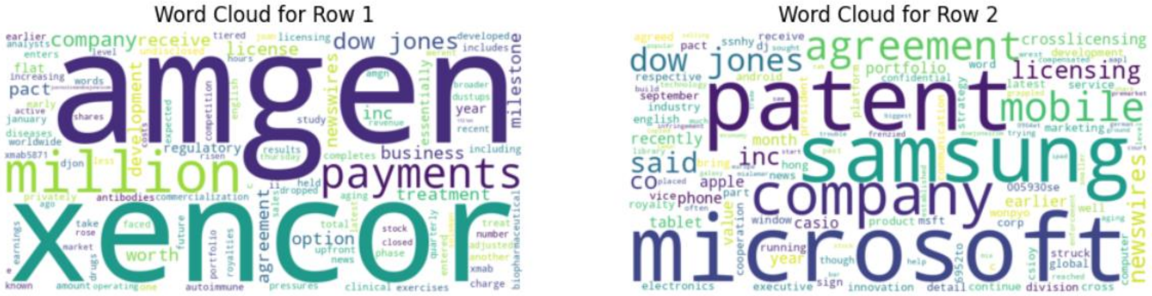
PROBLEM STATEMENT 3:

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.

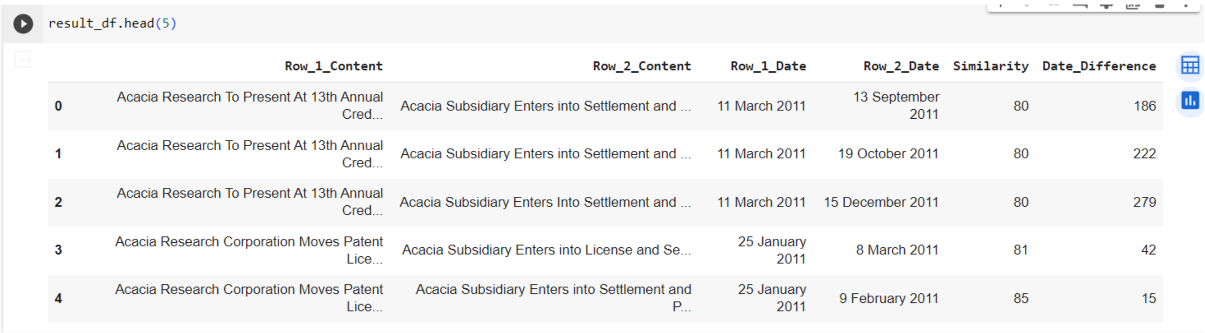


Word Cloud for Row 1

Word Cloud for Row 2

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%.



```
result_df.head(5)
```

	Row_1_Content	Row_2_Content	Row_1_Date	Row_2_Date	Similarity	Date_Difference
0	Acacia Research To Present At 13th Annual Cred...	Acacia Subsidiary Enters into Settlement and ...	11 March 2011	13 September 2011	80	186
1	Acacia Research To Present At 13th Annual Cred...	Acacia Subsidiary Enters into Settlement and ...	11 March 2011	19 October 2011	80	222
2	Acacia Research To Present At 13th Annual Cred...	Acacia Subsidiary Enters Into Settlement and ...	11 March 2011	15 December 2011	80	279
3	Acacia Research Corporation Moves Patent Lice...	Acacia Subsidiary Enters into License and Se...	25 January 2011	8 March 2011	81	42
4	Acacia Research Corporation Moves Patent Lice...	Acacia Subsidiary Enters into Settlement and P...	25 January 2011	9 February 2011	85	15

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.