

The_Constitution_of_the_Italian_Republic

June 2, 2024

1 Installation and importing

```
[1]: %%capture
!pip install tika
!pip install langchain
!pip install -qU llama-index-llms-openai
!pip install langchain_community
!pip install langchain_core
!pip install langchain_openai
!pip install langchain_experimental
!pip install faiss-cpu
!pip install datasets
!pip install ragas

[2]: import os
import re
import getpass
import pandas as pd
from tika import parser
from datasets import load_dataset, Dataset
from ragas import evaluate
from ragas.metrics import (answer_relevancy, faithfulness, context_recall,
    ↪context_precision)
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain_community.vectorstores import FAISS
from langchain_openai.embeddings import OpenAIEmbeddings
from langchain_openai import ChatOpenAI
from langchain_core.prompts import ChatPromptTemplate
from langchain_core.runnables import RunnablePassthrough
from langchain_core.output_parsers import StrOutputParser
from langchain_experimental.text_splitter import SemanticChunker

[3]: os.environ["OPENAI_API_KEY"] =
    ↪"sk-proj-EpcuRNoHC6vZvieDWSR2T3B1bkFJVgs9EQBu6tx0C09VchbF"
```

2 Introduction

Our project aims to create a Question Answering (QA) model that delves into the foundation of human rights, exploring seminal legal documents like Justinian's Code and Magna Carta. To achieve this, we're developing a RAG (Retrieval-Augmented Generation) model that leverages the capabilities of Large Language Models (LLMs). Our RAG model will retrieve relevant passages from these historical texts and generate accurate answers to users' questions, providing insights into the evolution of human rights and their significance in modern times.

Our project will:

- Develop a RAG model that focuses on legal-human rights documents
- Train the model on a dataset comprising Justinian's Code, Magna Carta, and other relevant texts
- Evaluate the model's performance on a test set of questions
- Fine-tune the model for improved accuracy and relevance

Our RAG model builds upon the advancements in LLMs, which have demonstrated exceptional language understanding and generation capabilities. By integrating retrieval capabilities into our RAG model, we can ensure that the generated answers are grounded in the original texts, providing a more reliable and informative QA experience.

2.1 The Constitution of the Italian Republic

```
[4]: text_file = parser.from_file("The_Constitution_of_the_Italian_Republic.pdf")
     print(text_file['content'].replace('\n', ' ')[:100])
```

```
2024-06-02 17:44:26,655 [MainThread ] [WARNI] Failed to see startup log
message; retrying...
```

```
ART. 1 Italy is a Democratic Republic founded on labour. Sovereignty belongs to
the people, who exer
```

2.1.1 Chunking methods

Chunking is a crucial step in preprocessing text data for our RAG model. We explore three approaches: naive chunking, semantic chunking and manual chunking.

Recursive Character Text Splitter (also called naive chunking) - it is a fundamental tool in the LangChain suite for breaking down large texts into manageable, semantically coherent chunks. This method is particularly recommended for initial text processing due to its ability to maintain the contextual integrity of the text. It operates by recursively splitting text based on a list of user-defined characters, ensuring that related pieces of text remain adjacent to each other, thus preserving their semantic relationship.

```
[5]: naive_chunker = RecursiveCharacterTextSplitter(
     chunk_size=256,
     chunk_overlap=0,
     length_function=len,
     is_separator_regex=False
)
```

```
naive_chunks = naive_chunker.split_text(text_file['content'])
naive_chunks_df = pd.DataFrame(naive_chunks)
naive_chunks_df.head()
```

[5]: 0

0 ART. 1 \n\nItaly is a Democratic Republic foun...
1 The Republic acknowledges and guarantees the i...
2 economic and social solidarity be fulfilled. \\
3 It is the duty of the Republic to remove econo...
4 economic and social organization of the countr...

Semantic Chunking - it considers the relationships within the text. It divides the text into meaningful, semantically complete chunks. This approach ensures the information's integrity during retrieval, leading to a more accurate and contextually appropriate outcome. It is slower compared to the previous chunking strategy.

Interquartile chunking - in this method, the interquartile distance is used to split chunks.

```
[6]: interquartile_chunker = SemanticChunker(
      OpenAIEmbeddings(), breakpoint_threshold_type="interquartile"
    )

    interquartile_chunks = interquartile_chunker.
      create_documents([text_file['content']])
    interquartile_chunks_df = pd.DataFrame(interquartile_chunks)
    interquartile_chunks_df.head()
```

[illegible]

Percentile chunking - in this method, all differences between sentences are calculated, and then any difference greater than the X percentile is split.

```
[7]: percentile_chunker = SemanticChunker(
    OpenAIEmbeddings(), breakpoint_threshold_type="percentile"
)
```



```

User's Query:
{question}

Context:
{context}
"""

rag_prompt = ChatPromptTemplate.from_template(rag_template)

```

RAG models retrieve the context from vector databases. In this part, we will define 4 vector databases for each of our chunks. And to transfer our text file into numbers in the vector databases, we will use the “text-embedding-3-large” embedding model from OpenAI.

```

[10]: naive_vectorstore = FAISS.from_texts(naive_chunks,␣
      ↪embedding=OpenAIEmbeddings(model="text-embedding-3-large"))
naive_retriever = naive_vectorstore.as_retriever(search_kwargs={"k" : 4})

```

```

[11]: interquartile_vectorstore = FAISS.from_documents(interquartile_chunks,␣
      ↪embedding=OpenAIEmbeddings(model="text-embedding-3-large"))
interquartile_retriever = interquartile_vectorstore.
      ↪as_retriever(search_kwargs={"k" : 4})

```

```

[12]: percentile_vectorstore = FAISS.from_documents(percentile_chunks,␣
      ↪embedding=OpenAIEmbeddings(model="text-embedding-3-large"))
percentile_retriever = percentile_vectorstore.as_retriever(search_kwargs={"k" :␣
      ↪4})

```

```

[13]: manual_vectorstore = FAISS.from_texts(manual_chunks,␣
      ↪embedding=OpenAIEmbeddings(model="text-embedding-3-large"))
manual_retriever = manual_vectorstore.as_retriever(search_kwargs={"k" : 4})

```

Our base model for question answering system will be ChatOpenAI.

```

[14]: base_model = ChatOpenAI()

```

These are the main bodies of our models. We will first define the context by setting it to relevant retriever from our vectorbase and in the same way, we will define the question by setting it to the user defined question. Then these two will go through our pre-defined prompt. As we know from earlier, the prompt will retrieve the context and augment it with the question before feeding them to our base model. When the base model receives these two, it will generate an answer. The answer will solely be related to our database.

```

[15]: naive_rag_chain = (
      {"context" : naive_retriever, "question" : RunnablePassthrough() }
      | rag_prompt
      | base_model
      | StrOutputParser()
    )

```

```
[16]: interquartile_rag_chain = (
    {"context" : interquartile_retriever, "question" : RunnablePassthrough()}
    | rag_prompt
    | base_model
    | StrOutputParser()
)
```

```
[17]: percentile_rag_chain = (
    {"context" : percentile_retriever, "question" : RunnablePassthrough()}
    | rag_prompt
    | base_model
    | StrOutputParser()
)
```

```
[18]: manual_rag_chain = (
    {"context" : manual_retriever, "question" : RunnablePassthrough()}
    | rag_prompt
    | base_model
    | StrOutputParser()
)
```

2.1.3 Testing with random questions

For comparing the different chunking methods and in general, for comparing the models, we will create some random questions based on our text and get answers from these models. We will also create some questions that are not in the document to see if it outputs “I don’t know”.

```
[19]: # Test 1
question = "What is prescribed for admission to and graduation from the various_
↪school levels and grades and to qualify for a profession?"

naive_answer_1 = naive_rag_chain.invoke(question)
interquartile_answer_1 = interquartile_rag_chain.invoke(question)
percentile_answer_1 = percentile_rag_chain.invoke(question)
manual_answer_1 = manual_rag_chain.invoke(question)

print(f"The question is: {question}")
print(f"\nThe answer of naive chunking is: {naive_answer_1}")
print(f"\nThe answer of interquartile chunking is: {interquartile_answer_1}")
print(f"\nThe answer of percentile chunking is: {percentile_answer_1}")
print(f"\nThe answer of manual chunking is: {manual_answer_1}")
```

The question is: What is prescribed for admission to and graduation from the various school levels and grades and to qualify for a profession?

The answer of naive chunking is: A state examination is prescribed for admission to and graduation from the various school levels and grades and to qualify for a profession.

The answer of interquartile chunking is: A state examination is prescribed for admission to and graduation from the various school levels and grades and to qualify for a profession.

The answer of percentile chunking is: A state examination is prescribed for admission to and graduation from the various school levels and grades and to qualify for a profession.

The answer of manual chunking is: A state examination is prescribed for admission to and graduation from the various school levels and grades and to qualify for a profession.

```
[20]: # Test 2
question = "Who represents the unity of the Nation?"

naive_answer_2 = naive_rag_chain.invoke(question)
interquartile_answer_2 = interquartile_rag_chain.invoke(question)
percentile_answer_2 = percentile_rag_chain.invoke(question)
manual_answer_2 = manual_rag_chain.invoke(question)

print(f"The question is: {question}")
print(f"\nThe answer of naive chunking is: {naive_answer_2}")
print(f"\nThe answer of interquartile chunking is: {interquartile_answer_2}")
print(f"\nThe answer of percentile chunking is: {percentile_answer_2}")
print(f"\nThe answer of manual chunking is: {manual_answer_2}")
```

The question is: Who represents the unity of the Nation?

The answer of naive chunking is: The President of the Republic represents the unity of the Nation.

The answer of interquartile chunking is: The President of the Republic represents the unity of the Nation.

The answer of percentile chunking is: The President of the Republic represents the unity of the Nation.

The answer of manual chunking is: The President of the Republic represents the unity of the Nation.

2.1.4 Model Evaluation

In the final part, we will ask the OpenAI model to generate questions (to behave like a teacher preparing questions for the students) from the document in order to assess our model's ability to answer questions.

In order to achieve this, we'll first create synthetic documents from our original data by chunking them "naively". Next, we'll create our question prompt for the OpenAI model, which will take

this prompt and generate questions based on the synthetic documents (it will behave like a teacher trying to prepare questions for an exam). Then in order to be able to compare the generated answers with the real answers, we'll create also a ground truth prompt for the OpenAI model. This ground truth will always give the right answers because we extract the context from which the question was generated from and feed it to the ground truth prompt.

```
[21]: synthetic_data_splitter = RecursiveCharacterTextSplitter(
    chunk_size=256,
    chunk_overlap=0,
    length_function=len,
    is_separator_regex=False
)

synthetic_data_chunks = synthetic_data_splitter.
    ↪create_documents([text_file['content']])
```

```
[22]: questions = []
ground_truths_semantic = []
contexts = []
answers = []
```

We'll start with percentile chunking.

```
[23]: question_prompt = """\
You are a teacher preparing a test. Please create a question that can be
    ↪answered by referencing the following context.

Context:
{context}
"""

ground_truth_prompt = """\
Use the following context and question to answer this question using *only* the
    ↪provided context.

Question:
{question}

Context:
{context}
"""

question_prompt = ChatPromptTemplate.from_template(question_prompt)
ground_truth_prompt = ChatPromptTemplate.from_template(ground_truth_prompt)

question_chain = question_prompt | ChatOpenAI(model="gpt-3.5-turbo") |
    ↪StrOutputParser()
```



```

ground_truth_chain = ground_truth_prompt |
↳ ChatOpenAI(model="gpt-4-turbo-preview") | StrOutputParser()

for chunk in synthetic_data_chunks[10:20]:
    questions.append(question_chain.invoke({"context" : chunk.page_content}))
    contexts.append([chunk.page_content])
    ground_truths_semantic.append(ground_truth_chain.invoke({"question" :
↳ questions[-1], "context" : contexts[-1]}))
    answers.append(percentile_rag_chain.invoke(questions[-1]))

```

```

[24]: qagc_list = []

for question, answer, context, ground_truth in zip(questions, answers,
↳ contexts, ground_truths_semantic):
    qagc_list.append({
        "question" : question,
        "answer" : answer,
        "contexts" : context,
        "ground_truth" : ground_truth
    })

eval_dataset = Dataset.from_list(qagc_list)

```

```

[25]: result = evaluate(eval_dataset, metrics=[context_precision,
                                             faithfulness,
                                             answer_relevancy,
                                             context_recall]);

```

Evaluating: 0% | 0/40 [00:00<?, ?it/s]

No statements were generated from the answer.

```

[26]: result_df = result.to_pandas()
result_df

```

```

[26]:
question \
0 In what way does the Italian legal system regu...
1 Question: According to the context provided, w...
2 Question: How does the Italian legal system ha...
3 Question: How are the legal status of foreigne...
4 Question: According to the Italian Constitutio...
5 Question: According to the context provided, w...
6 Question: According to the provided context, w...
7 Question: According to the context provided, w...
8 Question: In what circumstances is it permissi...
9 Question: In what circumstances may law enforc...

answer \

```

0 The Italian legal system regulates the relatio...
 1 The primary purpose of technical research in r...
 2 I don't know.
 3 The legal status of foreigners is regulated by...
 4 Foreigners who, in their own country, are deni...
 5 According to the context provided, Article 11 ...
 6 According to the provided context, Italy must ...
 7 The flag of the Republic of Italy is the Itali...
 8 Personal freedom can be restricted in exceptio...
 9 In exceptional cases of necessity and urgency,...

contexts \
 0 [conflict with the Italian legal system. \n\nT...
 1 [technical research. \n\nIt shall safeguard th...
 2 [interest of future generations. State law sha...
 3 [The legal status of foreigners shall be regul...
 4 [democratic freedoms guaranteed by the Italian...
 5 [The extradition of a foreigner for political ...
 6 [of other peoples and as a means of settling i...
 7 [promote and encourage international organizat...
 8 [No form of detention, inspection or personal ...
 9 [in the manner provided for by law. \n\nIn exc...

	ground_truth	context_precision	\
0	The Italian legal system regulates the relatio...	1.0	
1	The primary purpose of technical research in r...	1.0	
2	The Italian legal system handles the safeguard...	1.0	
3	The legal status of foreigners is regulated by...	1.0	
4	Those who are guaranteed democratic freedoms b...	1.0	
5	According to the context provided, Article 11 ...	0.0	
6	According to the provided context, Italy must ...	1.0	
7	The flag of the Republic of Italy is the Itali...	1.0	
8	According to the provided context, it is permi...	1.0	
9	In exceptional cases of necessity and urgency,...	0.0	

	faithfulness	answer_relevancy	context_recall
0	0.066667	0.984892	1.0
1	0.333333	0.974306	1.0
2	NaN	0.000000	0.5
3	1.000000	0.936707	1.0
4	1.000000	0.967791	1.0
5	0.500000	0.896462	1.0
6	0.500000	0.939094	1.0
7	1.000000	0.949101	1.0
8	0.142857	0.940497	1.0
9	1.000000	0.952111	0.5

```
[27]: for metric, score in result.items():
        print(f"Metric: {metric}, score: {score}", end='\n')
```

```
Metric: context_precision, score: 0.79999999992
Metric: faithfulness, score: 0.6158730158730159
Metric: answer_relevancy, score: 0.854096123550281
Metric: context_recall, score: 0.9
```

Let's do the same things for manual chunking.

```
[28]: answers_manual = []

for question in questions:
    answers_manual.append(manual_rag_chain.invoke(question))
```

```
[29]: qagc_manual = []

for question, answer, context, ground_truth in zip(questions, answers_manual,
↪ contexts, ground_truths_semantic):
    qagc_manual.append({
        "question" : question,
        "answer" : answer,
        "contexts" : context,
        "ground_truth" : ground_truth
    })

eval_dataset_manual = Dataset.from_list(qagc_manual)
```

```
[30]: result_manual = evaluate(eval_dataset_manual, metrics=[context_precision,
                                                                faithfulness,
                                                                answer_relevancy,
                                                                context_recall])
```

```
Evaluating:  0%|          | 0/40 [00:00<?, ?it/s]
```

No statements were generated from the answer.

```
[31]: result_manual_df = result_manual.to_pandas()
result_manual_df
```

```
[31]:                                     question \
0  In what way does the Italian legal system regu...
1  Question: According to the context provided, w...
2  Question: How does the Italian legal system ha...
3  Question: How are the legal status of foreigne...
4  Question: According to the Italian Constitutio...
5  Question: According to the context provided, w...
6  Question: According to the provided context, w...
7  Question: According to the context provided, w...
```

8 Question: In what circumstances is it permissi...
 9 Question: In what circumstances may law enforc...

answer \
 0 The Italian legal system regulates the relatio...
 1 The primary purpose of technical research in r...
 2 I don't know.
 3 The legal status of foreigners in Italy is reg...
 4 According to the Italian Constitution, foreign...
 5 According to the context provided, Article 11 ...
 6 In order to ensure peace and justice among nat...
 7 According to the context provided, the flag of...
 8 In exceptional cases of necessity and urgency,...
 9 Law enforcement may adopt temporary measures i...

contexts \
 0 [conflict with the Italian legal system. \n\nT...
 1 [technical research. \n\nIt shall safeguard th...
 2 [interest of future generations. State law sha...
 3 [The legal status of foreigners shall be regul...
 4 [democratic freedoms guaranteed by the Italian...
 5 [The extradition of a foreigner for political ...
 6 [of other peoples and as a means of settling i...
 7 [promote and encourage international organizat...
 8 [No form of detention, inspection or personal ...
 9 [in the manner provided for by law. \n\nIn exc...

	ground_truth	context_precision	\
0	The Italian legal system regulates the relatio...	0.0	
1	The primary purpose of technical research in r...	1.0	
2	The Italian legal system handles the safeguard...	1.0	
3	The legal status of foreigners is regulated by...	1.0	
4	Those who are guaranteed democratic freedoms b...	1.0	
5	According to the context provided, Article 11 ...	0.0	
6	According to the provided context, Italy must ...	1.0	
7	The flag of the Republic of Italy is the Itali...	1.0	
8	According to the provided context, it is permi...	1.0	
9	In exceptional cases of necessity and urgency,...	0.0	

	faithfulness	answer_relevancy	context_recall
0	0.000000	0.919698	1.0
1	0.600000	0.974306	1.0
2	NaN	0.000000	1.0
3	0.142857	0.901217	1.0
4	0.000000	0.948034	1.0
5	1.000000	0.967560	1.0
6	0.750000	0.929063	1.0

7	1.000000	0.949101	1.0
8	1.000000	0.961056	1.0
9	1.000000	0.953554	1.0

```
[32]: for metric, score in result_manual.items():
      print(f"Metric: {metric}, score: {score}", end='\n')
```

```
Metric: context_precision, score: 0.69999999993
Metric: faithfulness, score: 0.6103174603174604
Metric: answer_relevancy, score: 0.8503588634599286
Metric: context_recall, score: 1.0
```

```
[33]: result_manual_df.query('faithfulness < 1')
```

```
[33]:                                     question \
```

```
0 In what way does the Italian legal system regu...
1 Question: According to the context provided, w...
3 Question: How are the legal status of foreigne...
4 Question: According to the Italian Constitutio...
6 Question: According to the provided context, w...
```

```
                                     answer \
```

```
0 The Italian legal system regulates the relatio...
1 The primary purpose of technical research in r...
3 The legal status of foreigners in Italy is reg...
4 According to the Italian Constitution, foreign...
6 In order to ensure peace and justice among nat...
```

```
                                     contexts \
```

```
0 [conflict with the Italian legal system. \n\nT...
1 [technical research. \n\nIt shall safeguard th...
3 [The legal status of foreigners shall be regul...
4 [democratic freedoms guaranteed by the Italian...
6 [of other peoples and as a means of settling i...
```

```
                                     ground_truth context_precision \
```

```
0 The Italian legal system regulates the relatio...      0.0
1 The primary purpose of technical research in r...      1.0
3 The legal status of foreigners is regulated by...      1.0
4 Those who are guaranteed democratic freedoms b...      1.0
6 According to the provided context, Italy must ...      1.0
```

	faithfulness	answer_relevancy	context_recall
0	0.000000	0.919698	1.0
1	0.600000	0.974306	1.0
3	0.142857	0.901217	1.0
4	0.000000	0.948034	1.0
6	0.750000	0.929063	1.0

```
[34]: for _, row in result_manual_df.query('faithfulness < 1').iterrows():
        print(row.question, end='\n')
        print("Answer: ", row.answer, end='\n')
        print("Ground truth: ", row.ground_truth, end='\n')
        print("Faithfulness: ", row.faithfulness, end='\n')
        print('-----')
```

In what way does the Italian legal system regulate the relations between the State and their respective representatives according to the context provided?
 Answer: The Italian legal system regulates the relations between the State and their respective representatives by vesting legislative powers in the State and the Regions in compliance with the Constitution and international obligations. The State has exclusive legislative powers in various subject matters, including foreign policy, immigration, defense, currency, public order, and more. Additionally, the President of the Council of Ministers and the Ministers are subject to ordinary courts for crimes committed in the exercise of their duties.
 Ground truth: The Italian legal system regulates the relations between the State and their respective representatives by requiring that these relations be governed by laws that are based on agreements with their respective representatives.
 Faithfulness: 0.0

Question: According to the context provided, what is the primary purpose of technical research in relation to safeguarding the nation's natural beauties and heritage?
 Answer: The primary purpose of technical research in relation to safeguarding the nation's natural beauties and heritage is to promote the development of culture, scientific knowledge, and to safeguard the natural beauties, historical, and artistic heritage of the Nation.
 Ground truth: The primary purpose of technical research in relation to safeguarding the nation's natural beauties and heritage is to safeguard the natural beauties, the historical and artistic heritage of the nation, and the environment, biodiversity, and ecosystems.
 Faithfulness: 0.6

Question: How are the legal status of foreigners regulated in compliance with international provisions and treaties?
 Answer: The legal status of foreigners in Italy is regulated by law in compliance with international provisions and treaties. Foreigners who are denied democratic freedoms in their own country may have the right to asylum in Italy. Extradition of foreigners for political offenses is not permitted, and extradition of citizens is only allowed in cases provided for in international conventions. Citizens have the right to move and reside freely in Italy, with restrictions only allowed for health or security reasons, not for political reasons.
 Ground truth: The legal status of foreigners is regulated by laws that are in compliance with international provisions and treaties.
 Faithfulness: 0.14285714285714285

Question: According to the Italian Constitution, who shall have the right of asylum in the territory of the Italian Republic?

Answer: According to the Italian Constitution, foreigners who are denied the actual exercise of democratic freedoms guaranteed by the Italian Constitution in their own country shall have the right of asylum in the territory of the Italian Republic.

Ground truth: Those who are guaranteed democratic freedoms by the Italian Constitution shall have the right of asylum in the territory of the Italian Republic, in accordance with the conditions set forth by law.

Faithfulness: 0.0

Question: According to the provided context, what conditions must Italy consent to in order to ensure peace and justice among nations?

Answer: In order to ensure peace and justice among nations, Italy must consent to sovereignty limitations required for a world order that ensures peace and justice among Nations, on conditions of equality with other States. Italy shall also promote and encourage international organizations furthering such ends.

Ground truth: According to the provided context, Italy must consent to sovereignty limitations required for a world order that ensures peace and justice among nations, on conditions of equality with other States.

Faithfulness: 0.75

[]: