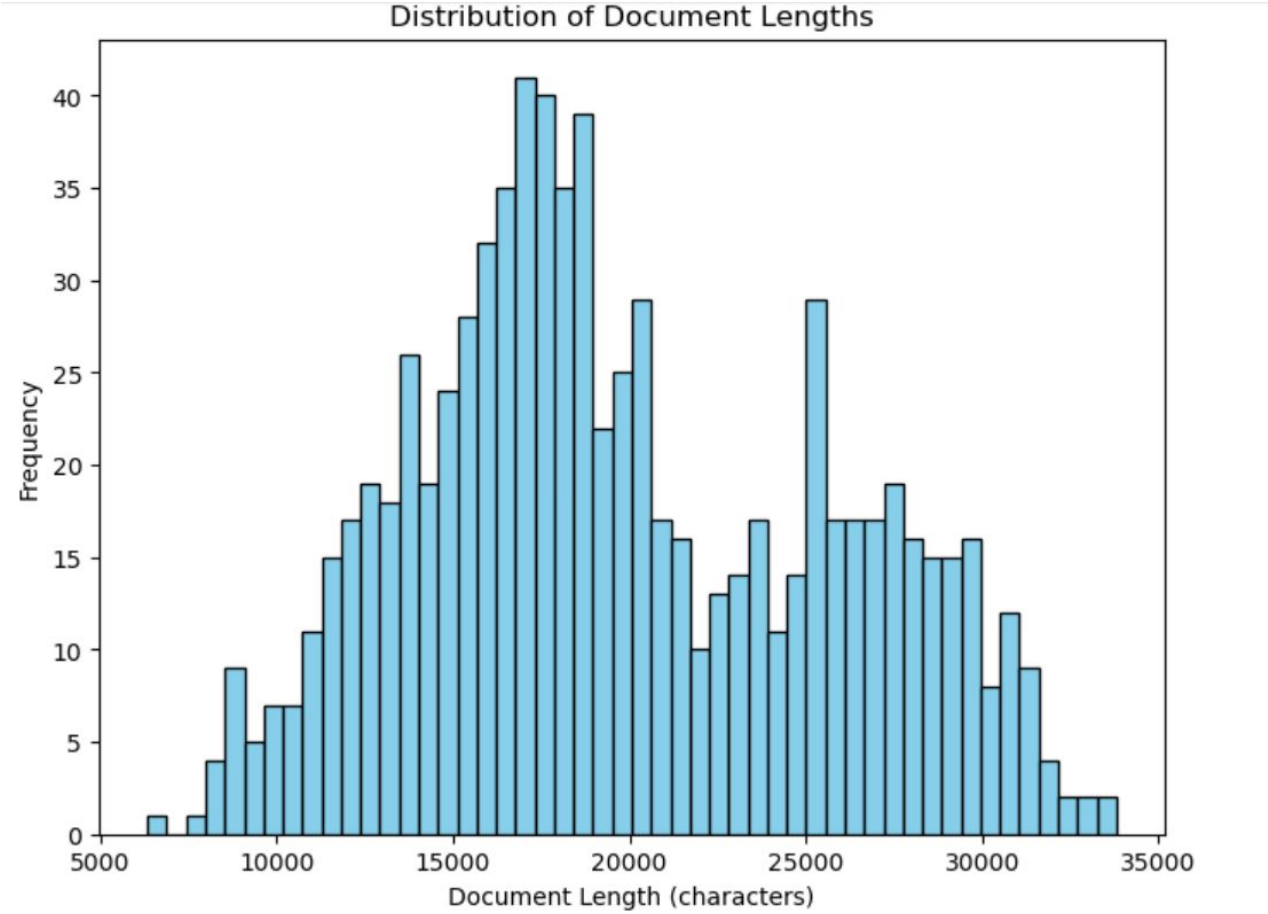
RAG PROJECT

JARIAN DEL VALLE

JAVIER DASTAS

EXPLORATORY DATA ANALYSIS (EDA)



EXPLORATORY DATA ANALYSIS (EDA)

- Could present challenges due to content length variety, may exceed allowed token limit when combining chunks
- Suitable data set for query retrieval due to chunked format

ETL

- Recursive Text Splitter, chunk size 1000 and chunk overlap 200 (within recommended 10 – 20% overlap)
- Iterate through the text file to perform splitting
- Batch insert into ChromaDB

EMBEDDINGS

- Initial trial: Hugging Face embeddings with model- "sentence-transformers/all-MiniLM-L6-v2“

```
query = "What is the economic impact of hurricanes in Puerto Rico?"

sentence_transformer_vector = embedding_model.embed_query(user_query)

print(f"SentenceTransformer Embedding Vector: {sentence_transformer_vector[0:30]}")
```

```
SentenceTransformer Embedding Vector: [0.06662759929895401, 0.013993947766721249, 0.009934347122907639, 0.1170114278793335, 0.012308242730796337, -0.0163
47035765647888, -0.04397964105010033, -0.023287108168005943, -0.022707929834723473, -0.0012624622322618961, 0.031106624752283096, 0.008816748857498169, -
0.010537777096033096, -0.016138488426804543, 0.010201184079051018, 0.0077844359911978245, -0.02792900614440441, -0.03427750989794731, 0.01525187026709318
2, 0.006409345660358667, 0.010645732283592224, -0.01788138970732689, -0.09436211735010147, -0.0017113065114244819, -0.009319481439888477, -0.015047031454
741955, -0.019765324890613556, 0.03430965542793274, -0.06701117008924484, -0.0362224280834198]
```

- Second trial: Open AI Embeddings with model- "text-embedding-ada-002"

```
# Test query output as vectors
user_query = "What industry contributes the most to the Puerto Rican economy?"

text_embedding_ada_vector = open_ai_embedding_model.embed_query(user_query)

print(f"Text Embedding ADA 002 Embedding Vector: {text_embedding_ada_vector[0:30]}")
```

```
Text Embedding ADA 002 Embedding Vector: [-9.969390521291643e-05, -0.02118777669966221, 0.011862811632454395, -0.02057608962059021, -0.02207276970148086
5, -0.01704913191497326, 0.00205956120043993, 0.020940497517585754, -0.002923405496403575, -0.039121899753808975, 0.006897740066051483, 0.020901454612612
724, 0.01643744483590126, 0.0009126491495408118, -0.008459492586553097, 0.0015715134795755148, 0.012324830517172813, -0.010496278293430805, 0.00142021873
03453684, -0.01594289019703865, -0.03412429243326187, 0.031755633652210236, 0.004568126052618027, -0.016580605879426003, 0.02582097426056862, 0.031443282
9618454, 0.019495876505970955, -0.003575762500986457, -0.006660223938524723, -0.014146874658763409]
```

RETRIEVAL

- Retrieval with ChromaDB-Hugging Face Embedding with Sentence Transformers

```
Document 2:
Source: 19240126_1.txt
Content: were only two decent solutions to the political problem of Puerto Rico and the only "bles- don frti'WbY of a cultured and civilized people: the
State or Independence...

Document 3:
Source: Unknown
Content: and sovereignty which the National Congress granted to the Legislature of Puerto Rico it had "sta and has now the right and power to effect the
greatest and most permanent liberation of the country and of all its people, abolishing the horrendous physical and mental misery which Onychia numerous
com...

Document 4:
Source: 19220930_1.txt
Content: and sovereignty which the National Congress granted to the Legislature of Puerto Rico it had "sta and has now the right and power to effect the
greatest and most permanent liberation of the country and of all its people, abolishing the horrendous physical and mental misery which Onychia numerous
com...
```

RETRIEVAL

- Retrieval with ChromaDB- OpenAI Embedding with Text Embedding ADA 002

```
Document 2:
Source: 19471025_1.txt
Content: . The sudden drop in exports found the industry operating at an accelerated volume, and with considerable reserves in the possession of its cont
inental distributors. Although there is hope that a favorable change in the present situation may arise, there is little optimism among Puerto Rican indu
str...

Document 3:
Source: 19500121_1.txt
Content: . 1 There are, however, two or three factors of much greater importance to the island economy than to the national one. "The size of the cane cr
op and the amount of cane that can be sold at a remunerative price is of great significance to our economy. In this sector of the island economy there is
un...

Document 4:
Source: 19470426_1.txt
Content: . market if the problem facing our industry is not resolved quickly. On the other hand, it was reported to EL MUNDO that the Island runs the ris
k of losing its ascendancy in the U.S. market, with more serious damage* to the industry, "I trafficking with advantage* the ions of Cuba and the Virgin
Isl...

Document 5:
Source: 19220415_1.txt
Content: . A people without its industry is like an unattractive woman or like a wasteland in complete solitude and utter neglect. Puerto Rico industries
and it is an imperious duty to defend and maintain them, if it wishes to have a personality of its own, a life of its own and a wealth of its own that wi
ll...
```

MODEL EVALUATION- QUERY

1. Basic Fact Checking and Accuracy

- Question- Who is the current govenor of Puerto Rico?
- Question- What are the main insudtries in Puerto Rico?

2. Citation and Source Attribution

- Question- Where in the documentation does it say that Puerto Rico has a strong pharmaceutical industry?

3. Edge Cases and Robustness

- Question- Does Puerto Rico export more coffee or electronics?

4. Bias and ethical testing

- Question: Is Puerto Rico better off as a U.S. Territory or as an independent country?

5. Formatting and response structure

- Question: List three key takeaways from Puerto Rico's latest financial report.

6. Hallucination Detection

- Question: Provide information about Puerto Rico's space program

RAG DEPLOYMENT

http://dsmlbootcamp.org

Implementation

- Linode Cloud Service
- Ubuntu 24.0 Server OS
- Back-end: Flask (API) + NGINX
- Front-end: Flask-Template

Requirements

- flask 3.1.0
- openai 1.61.1
- langchain_chroma 0.2.1
- langchain_huggingface 0.1.2
- langchain_openai 0.3.4

THANK YOU

JARIAN DEL VALLE

JAVIER DASTAS