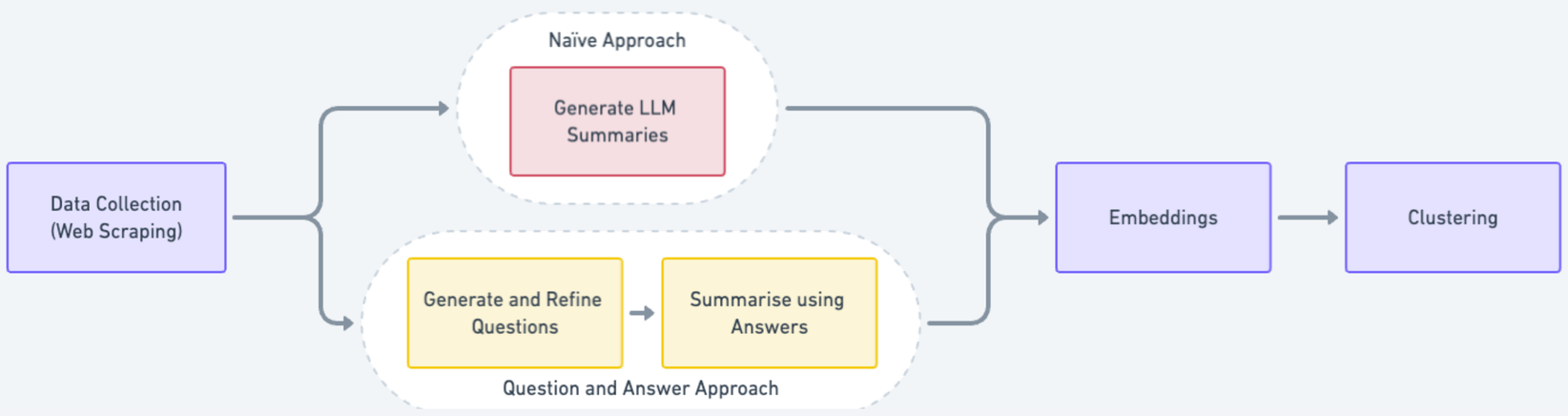
Potential Customer Segmentation using LLMs

[1]Koushik Sai Veerella , [1]Nitish Satya Sai Gedela , [1]Siva Ram Kottapalli , [1]Kevin Patel
[1] Michigan State University
[2]Jian Yang , [2] Michael Dessauer
[2] Westlake Chemical

Objective

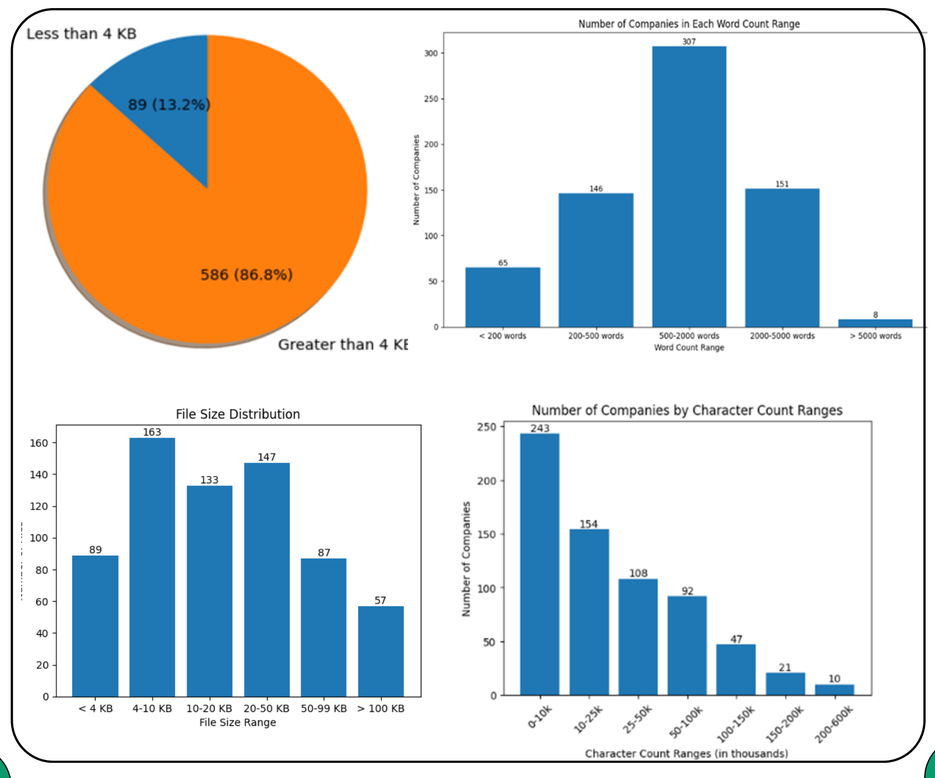
Leverage Large Language Models (LLMs) & Clustering techniques on Web Scraped data to predict potential customers for targeted marketing strategies.

Project Work Flow



Data Collection

- Total companies Count: **722**
- Invalid/Duplicate Companies: **47**
- Companies Scraped: **675**
- Constraints used for Scraping:
  - Depth crawled per website: **5**
  - Max URLs per company: **150**
- **Approach**:
  - Scrap company websites using **Selenium**.
  - Save Scraped data to PDFs (1 per company)



Statistics

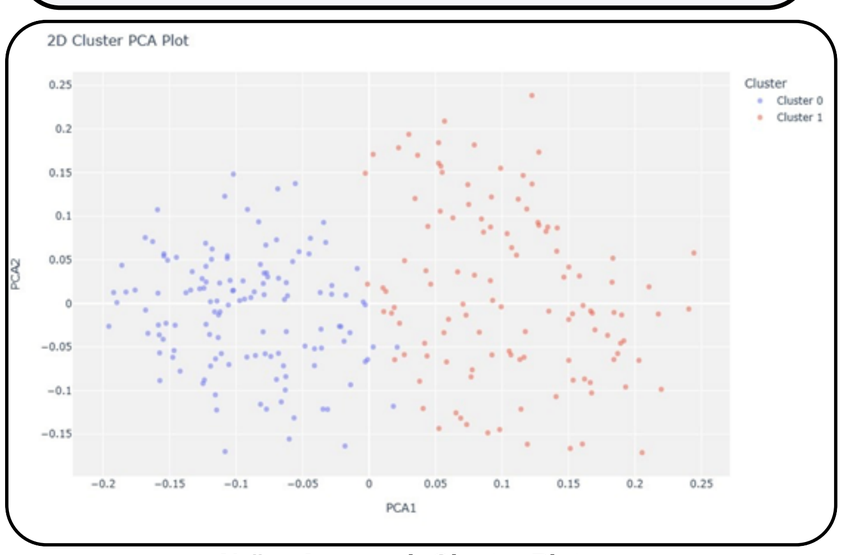
- Total files: **675** (1 per company)
- Companies with < 4 KB data: **89**
- Companies with > 4 KB data: **586**
- Mean File Size: **33.13** KB
- Mean Word Count (per PDF): **1302**
- Mean Char Count (per PDF): **39628**
- Mean Char Count (per Page): **2741**

Naïve Approach

Input: PDFs from Web Scraped data
Output: Summarized Company data of ~500 words

Approach:

- In-depth study and experimentation of Text Splitters and chain_types (Stuffing, Map-Reduce, Refine, Map-Rerank) for PDF Text Summarisation.
- Leverage LangChain's 'load_summarize_chain' with the 'map_reduce' option and employ prompt engineering to derive summaries from input PDF



Naïve Approach Cluster Plot

Embeddings

Input: Summarized/Structured Company data from Naive / Q&A Approaches.
Output: Vector Embeddings of Summaries / Structured data.

Approach:

- Use **OpenAIEmbeddings** (text-embedding-ada-002) to convert character chunks into embeddings.
- Use **FAISS** DB to store the embedded chunks.

Clustering

Input: Vector Embeddings of Summaries / Structured Answer Set Matrix (14 x N).
Output: Clusters of companies.

Approach:

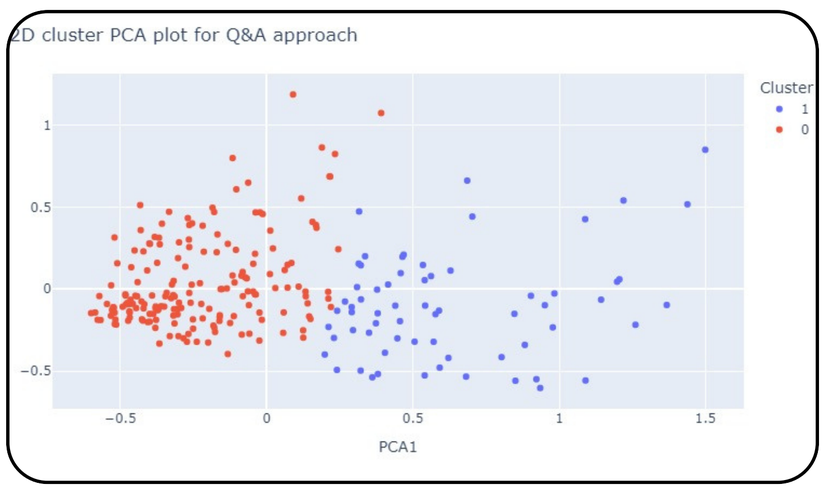
- Unsupervised clustering of Vector Embeddings of Summaries/Structured Answer set Matrix .
- Utilize clustering techniques - K-Means, DBScan - to generate clusters of semantically similar companies.
- Use Metrics (Silhoutte Score) to evaluate performance of Clustering.

Q&A Approach

Input: PDFs from Web Scraped data
Output: Structured Representation of Company Data (14xN (14 Questions. N dimensions for embedding model)).

Approach:

- Generate Questions from PDFs using LangChain's load_summarize_chain (question_prompt).
- Refine Questions using Clustering & Statistics.
- Use RAG (Retrieval Augmented Generation) approach to generate Answers for Select Questions.
- Generate Answers & Embeddings for Select Questions.



Q&A Approach Cluster Plot

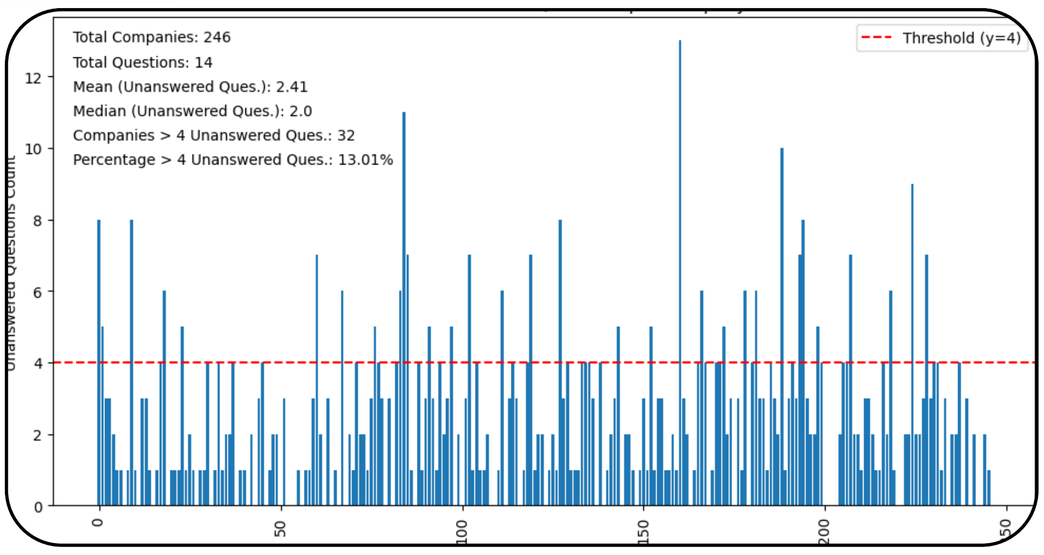
Conclusion & Way Forward

Conclusion:

- Leveraging LLMs can definitely help gain valuable insights from data.
- Q&A approach has improved the overall clustering performance and is a more structured approach than Naive Summaries Approach.

Way Forward:

- Use more open and available sources to generate & leverage information of potential customer profiles.
- Use supervised clustering (with existing clients data) techniques to boost the objective of predicting potential customers using LLMs.



Q&A Approach Evaluation

MICHIGAN STATE UNIVERSITY

Westlake