Scalable and Cost-Effective Deduplication: Leveraging Algorithms and LLMs

Team: Rohan Chaudhary, Sarmad Kahut, Valentina Torres

Sponsor Motion

OVERVIEW

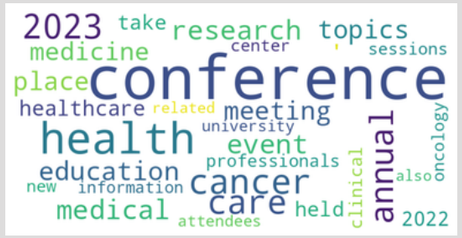
- The company operates an automated web scraping and AI-powered filtering system to maintain a comprehensive database of events in the US
- Their mission is to ensure easy discoverability of events and facilitate effective sponsor-event connections

BUSINESS OBJECTIVES

1. Scalability:
   - Establishing a fully automated and cost-effective data quality control process to identify duplicate records
   - Optimizing the costs of post-scrape filtering and processing of event records
2. Additional goals:
   - Smarter characterization
   - Event recommendations

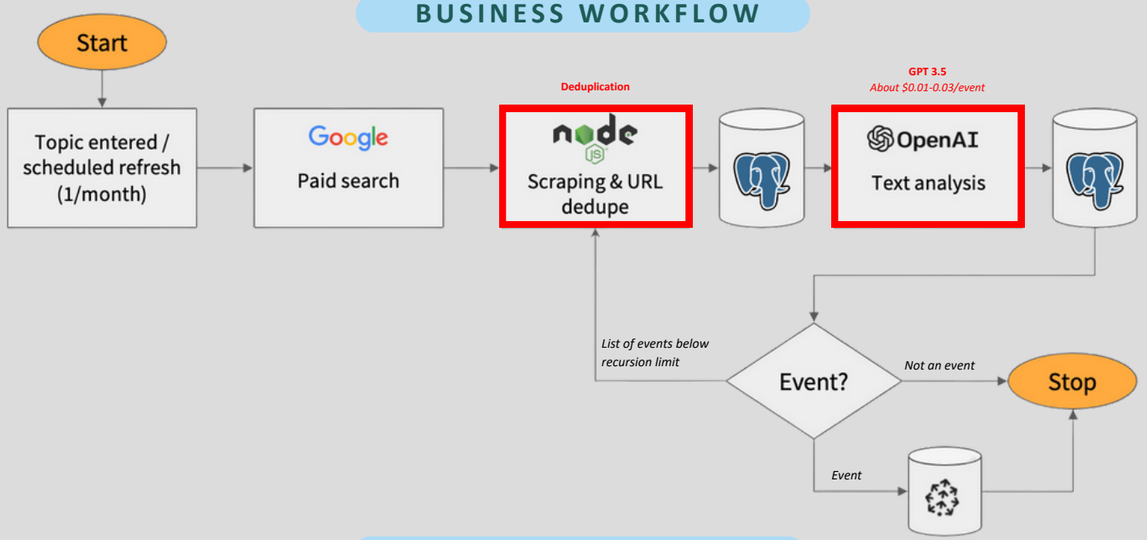
DATA ANALYSIS /PREPROCESSING

- Selecting the most relevant columns
- Data Cleaning (Missing Values)
- Dates Formatting
- Summary Column Clean-up (Stemming)

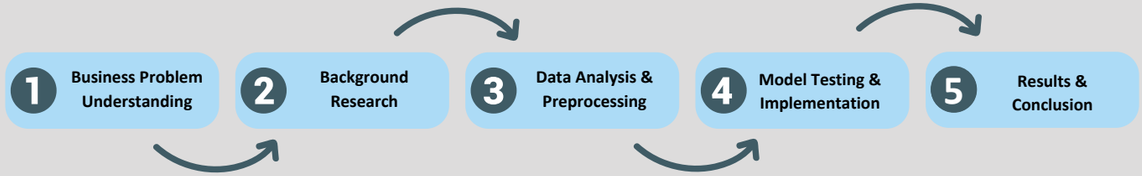


Word cloud of the most frequent and generic words in the text (summary) column

BUSINESS WORKFLOW

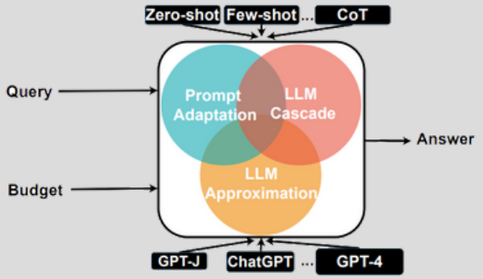


METHODOLOGY



BACKGROUND RESEARCH

1. Large Language Models:
   - Functioning and architecture
   - Cost Structure: capabilities and price points
   - Cost Optimization: Frugal GPT
     - Adaptively triage different queries in the dataset to different combinations of LLMs (LLM cascading)
   - Data Extraction/preprocessing using LLMs
2. Functioning and types of text similarity detection algorithms
3. Word Embedding, Vectorization and Sentence transformation



(b) Proposed FrugalGPT

Proposed FrugalGPT - LLM Cascading

Reference: https://arxiv.org/abs/2305.05176

DEDUPLICATION PROBLEM

1. Goal: finding a right balance of the false positives and false negatives such that the algorithm can be implemented on a larger scale
2. Models Considered:
   - Semantic text similarity:
     - TF-IDF
     - Cosine Similarity
     - Word2Vec
     - BERT with cosine distance
3. Irregularity in duplicate identification
4. Different techniques gave varying results for various thresholds of similarity tested

CHALLENGES

- Evolving GPTs are unstable technologies and are not fully understood yet
- Having bad data as a result of weak deduplication
- Defining the threshold for duplicate detection

FUTURE STEPS

- Bucketing data by using sorting algorithm
- Implementing Fuzzy Matching on event names that contains only few words
- Cost optimization of the analysis and filtering step using LLMs and automate the duplicate data identification

REPORT AND REFERENCES