

Scalable and Cost-Effective Deduplication: Leveraging Algorithms and LLMs

SponsorMotion

Discover • **Connect** • Act

Rohan Chaudhary, Sarmad Kahut, Valentina Torres

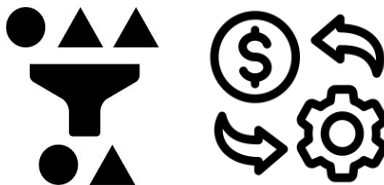
Mission

- Comprehensive database of healthcare related events in the US
- Make these events easily discoverable
- Connect the right sponsors with the right events



Business Problem

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

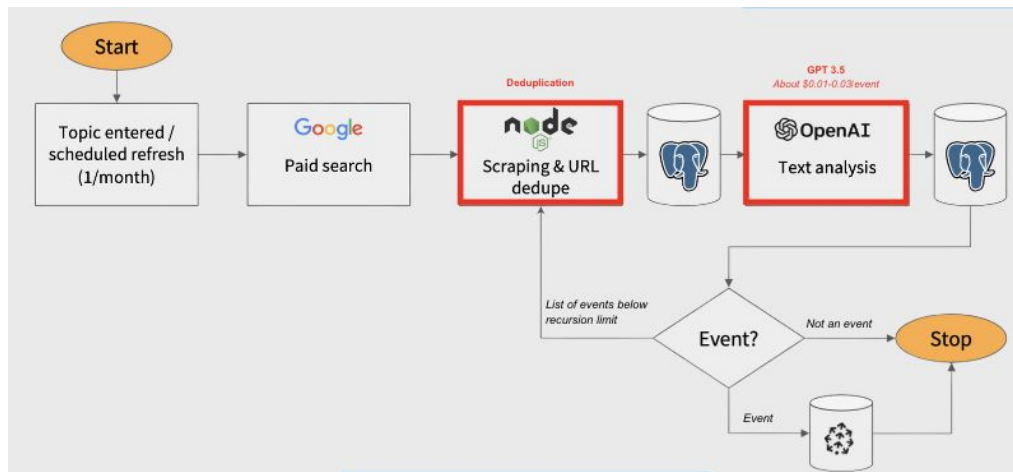


Data Analysis/Preprocessing

- Selected the most relevant columns
- Data cleaning (missing values)
- Start date formatting
- Summary column clean-up (Stemming)





Business Workflow



Research

1. Large Language Models:
 - Functioning
 - Cost Structure
 - Cost Optimization: Frugal GPT
 - Data Extraction/preprocessing
2. Text Similarity Detection Algorithms
3. Word Embedding and Vectorization

Duplicate Data of events

 NUTRITION 2023	Jul 22, 2023	Jul 25, 2023	MA	77%
 NUTRITION 2023	Jul 22, 2023	Jul 25, 2023	MA	76%

Work Done and Findings

1. Goal: Find the right balance between false positives and false negatives
2. Models:
 - Semantic text similarity (Summary)
 - a. TF-IDF
 - b. Cosine Similarity
 - Word2Vec
 - BERT
3. Irregularity in duplicate identification
4. Different techniques gave varying results



Challenges

- Evolving GPT: new and not completely understood
- Bad data as a result of weak deduplication
- Defining the threshold for duplicate detection



Future Steps

- Sorting algorithm
- Fuzzy matching of event names
- Algorithmic cost optimization

