BOSTON UNIVERSITY

Scalable and Cost-Effective Deduplication: Leveraging Algorithms and LLMs

SponsorMotion

Rohan Chaudhary, Sarmad Kahut, Valentina Torres

Table of Contents

Business Problem, Data and Initial Analysis

Data

- 48k rows
- 21 columns
- Duplicates

Business Problems

- **Deduplication**
- **Cost Optimization**

Initial Analysis

- **Word 2 Vec**
- **TF-IDF**

Deduplication Algorithm; Python Notebook Workflow

- Dropped all rows with no name, no state, no start date. Data went from 48k to 45k
- Removed special characters and blank spaces, all names were changed to lowercase
- Picked 6 states with most records to test our approach (CA, MA, NY, TX, FL, and NV)
- Found exact duplicates, with same name and same start date (+/- 1 day)
- Kept duplicate records with longest summaries and dropped the rest
- Applied fuzzy matching to remaining records and applied 3 thresholds (75, 80, and 85)
- Added ‘Human Verification’ column to show rows needing attention (Yes; similar name but different start dates, No; similar names and date)

Final Approach and Method

Skip to Main Content
All Events
SGO ANNUAL MEETING ON WOMEN‚ÄôS CANCER
SGO Annual Meeting on Women‚Äôs Cancer
Registration
Hotel Information
Schedule
Accreditation Information
Exhibits And Sponsorships
Exhibitor Information
Key Concepts
Destination Information & Activities
Industry Supported Symposia
Promotional Resources
Request Meeting Space
Conflict Of Interest
Media & Press
Mar 25 - Mar 28, 2023 Tampa, Florida
Tampa Florida
SGO 2023 ANNUAL MEETING ON WOMEN‚ÄôS CANCER
Join us March 25 ‚Äì 28, 2023, at the Tampa Convention Center in Tampa, FL.
The SGO Annual Meeting on Women‚Äôs Cancer¬Æ is the premier educational and scientific event for those who treat and care for women with gynecologic cancer.
Since 1969, women‚Äôs health care professionals have convened at the Society of Gynecologic Oncology (SGO)‚Äôs Annual Meeting on Women‚Äôs Cancer to discuss the latest science in the field, receive educational programming, and to network.
Members of the entire gynecologic cancer care team who provide treatment and care in the areas of chemotherapy, radiation therapy, surgery, and palliative care attend the SGO Annual Meeting. Gynecologic oncologists make up a large population of the attendees, along with medical oncologists, pathologists, radiation oncologists, hematologists, surgical oncologists, obstetrician/gynecologists, nurses, physician assistants, fellows in training, residents, and pharmacists.
The SGO Annual Meeting also brings together exhibitors from a variety of medical device and service companies. An array of state-of-the-art products and services geared towards members of the gynecologic cancer care team are on display in the exhibition area. There are numerous sponsorship opportunities as well for various events, services and takeaway items at the SGO Annual Meeting.
*Full abstract details posted the day of scheduled presentation.
REGISTER NOW BOOK HOTEL
VIEW SCHEDULE & ABSTRACTS
Abstract titles can be found in the full schedule by clicking into the sessions. Full abstract details posted the day of scheduled presentation.
2023 Program Committee
The Program Committee for the SGO 2023 Annual Meeting on Women‚Äôs Cancer is co-chaired by Dineo Khabele, MD, from Washington University in St. Louis, and Paul DiSilvestro, MD, from the Women & Infants Hospital of Brown University.
CO-CHAIRS
Dineo Khabele, MD
Paul DiSilvestro, MD
STEERING COMMITTEE
Floor Backes, MD
Leslie Boyd, MD
Lee-May Chen, MD
Bill Cliby, MD
Bradley Corr, MD
Andrea Hagemann, MD
Emily Hill, MD
Neil Horowitz, MD
Adrianne Mallen, MD
Alexander Olawaiye, MD
Dmitriy Zamarin, MD, PhD
PROGRAM COMMITTEE
Victoria Bae-Jump, MD, PhD
Joyce Barlin, MD
Evelyn Cantillo, MD, MPH
Allan Covens, MD, FRCSC
Keiichi Fujiwara, MD, PhD
Stephanie Gaillard, MD, PhD
Sophia George, PhD
Lindsay Hasselwander, MSN
Gloria Huang, MD
Marilyn Huang, MD
Tilley Jenkins Vogel, MD
Andrea Jewell, MD
Emily Ko, MD, MSCR
Elizabeth Lokich, MD
Jyoti Mayadev, MD
Alexander Melamed, MD, MPH
Navya Nair, MD
Michaela Onstad Grinsfelder, MD, MPH
Neil Phippen, MD
Monica Prasad Hayes, MD
Angelica Rodrigues, PhD
Mian Shahzad, MD, PhD
Kristy Ward, MD
Annelise Wilhite, MD
Israel Zighelboim, MD
Annie Ellis (Patient Advocate)
Jubilee Brown, MD (Board Liason)
THANK YOU FOR YOUR INTEREST IN EXHIBITING AT THE SGO 2023 ANNUAL MEETING!
More than 2,000 leaders in the field of gynecologic oncology convene every year at the SGO Annual Meeting on Women‚Äôs Cancer.¬† Attendees come from the entire gynecologic cancer care team who provide treatment and care to cancer patients, including gynecologic oncologists, medical oncologists, pathologists, radiation oncologists, hematologists, surgical oncologists, obstetrician/gynecologists, nurses, physician assistants, fellows in training, residents and social workers.
There is no better time to reach the key decision-makers in gynecologic oncology care than at the SGO Annual Meeting. Interested in Sponsoring or Exhibiting? Please contact Jessie Par¥©, Manager, Industry Relations and Business Development.
Not an SGO member?
Sign up to receive notifications about SGO meetings, education and membership.
Fields marked with an * are required
FIRST NAME *
LAST NAME *
EMAIL *
PROFESSION
SIGN UP
Advertisement
Advertise With Us
Future Dates
MARCH 16 ‚Äì MARCH 19, 2024
San Diego, CA
MARCH 15 ‚Äì MARCH 18, 2025
Seattle, WA
Meetings & Events
SGO is the premier source for gynecologic oncology education. Several meetings are hosted by SGO throughout the year for professional development.
VIEW UPCOMING

Text Analysis using LLM

Data Columns

Name

SGO Annual Meeting on Women

Start Date

3/25/2023 12:00:00 AM

Summary

The SGO Annual Meeting on Women‚Äôs Cancer¬Æ is the premier educational and scientific event for those who treat and care for women with gynecologic cancer. Members of the entire gynecologic cancer care team who provide treatment and care in the areas of chemotherapy, radiation therapy, surgery, and palliative care attend the SGO Annual Meeting. The conference also includes exhibitors from a variety of medical device and service companies.

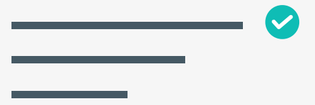
State

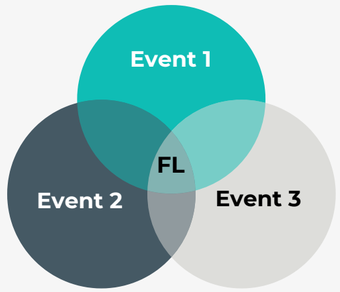
FL

Post Preprocessing

sgoannualmeetingonwomen

2023/03/25





Key Findings: Deduplication

Performance Assessment

- Used 3 threshold level
- Six states with the highest number of records
- Actual duplicates = Duplicates identified + FN - FP
- % of Actual duplicates identified = Duplicates identified / Actual duplicates

Accuracy Results

- Higher thresholds = more false negatives
- **Cost function:** estimates the cost associated with each threshold
- **Test dataset:**
  - 15% actual duplicate records to be identified
  - 75% threshold identified 12% of those duplicates

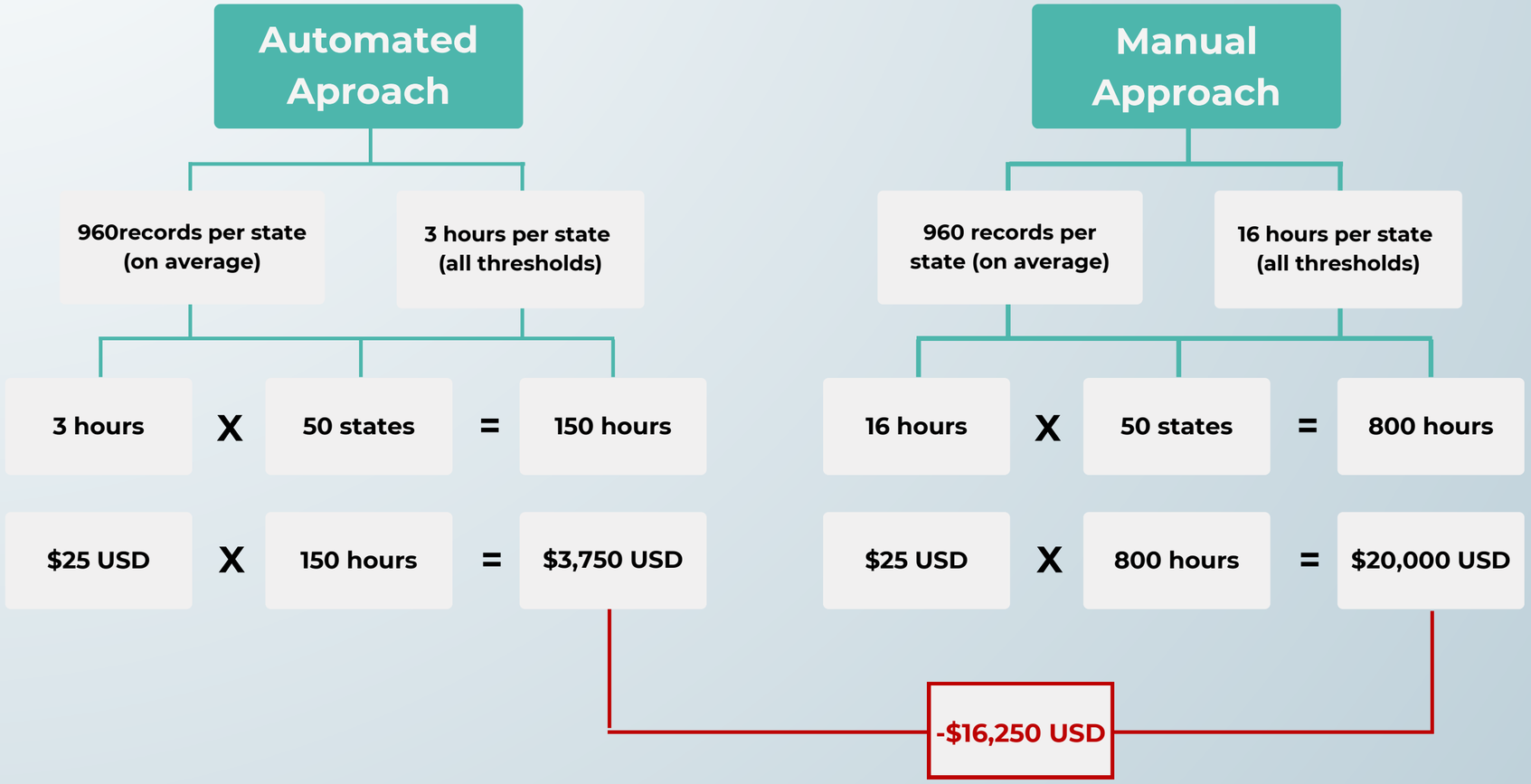
Similarity Threshold	Total Records	Duplicates Identified	False Negatives	False Positives	Actual Duplicates	% of Actual Duplicates Identified
75	5646	680	223	65	838	81.15%
80	5646	530	288	35	783	67.69%
85	5646	426	336	26	736	57.88%

Cost (fn) = 3 x FN + 1 x FP

- Cost (Threshold: **75**) = 3 x 223 + 1 x 65= 669 + 65 = **734**
- Cost (Threshold: **80**) = 3 x 288 + 1 x 35 = 864 + 35 = **899**
- Cost (Threshold: **85**) = 3 x 336 + 1 x 26 = 1008 + 26 = **1034**

Where
FN = Number of False Negatives
FP = Number of False Positives

Cost Optimization



Dynamic Model Selection

- Choose LLMs (GPT-J, GPT-4, ChatGPT) based on:
  - Query needs
  - Balancing accuracy
  - Budget for data extraction
- Pilot Project: Compare GPT-3.5 and frugal LLMs (GPT-J, ChatGPT) on 100 healthcare URLs for accuracy, cost, and data quality insights.

Business Impact & Implications

Deduplication

Automation

Efficient preprocessing and automated fuzzy matching reduce manual effort, streamline deduplication, and enhance productivity

Scalability

Scalable strategy handles expanding healthcare data, maintains accuracy, and supports database growth seamlessly

Cost Savings

Cost reduction, efficiency boost, data integrity improvement, and enhanced business outcomes

LLM Optimization

Faster Processing

Speedy processing, rapid analysis, enhanced efficiency for textual content tasks

Enhanced Scalability

Affordable scaling, process large data volumes, expand analytics for small businesses

Cost Savings & Efficiency

Enhanced cost-effectiveness, more analysis within budget with pay-per-token model

Limitations and Avenues for Further Improvement

Enhancing Data Quality for Accurate Analysis

- Implement filters to highlight ".org" URLs
- Optimize web scraping by eliminating null values in key columns

www.ncchc.org
Correctional Health Care

Advancing User Experience through Taxonomy Labeling

- **Enhanced Categorization** of events (e.g., oncology, cardiology)
- **Improve user experience** through quick access and filtering based on categories

Balancing Business Benefits with Model Selection

- **Strategic Consideration:** Evaluate trade-offs in accuracy, response time, and comprehensiveness when choosing LLM sizes
- **Workflow Integration:** Integrate LLM cascading with minimal workflow disruption while assessing task suitability

Summary

Our Path of Achievements, Insights and Learning

Automated Data Quality Control and Optimized Post-Scrape Filtering

Preprocessed textual data and explored similarity detection techniques

Achieved 81.15% Accuracy at 75% Threshold

Developed a cost function to assess results and estimated savings through Deduplication Process Automation

Informed Decision-Making: Pilot Project for LLM Comparison: GPT-3.5, GPT-J, ChatGPT

Future Vision: Streamlined Operations and Enhanced Navigation

Business impact and deliverables aligning with the objective of the project

Balanced Trade-offs: accuracy,response time, ccomprehensiveness in LLM Usage

the Final Project Report

Thank You for your time!