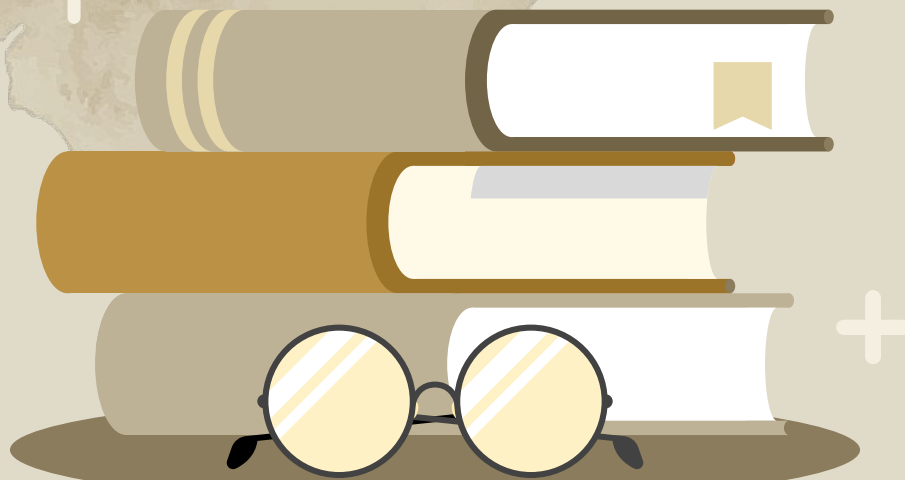


U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Historical Conflicts Search Engine



Information Retrieval

Dinis Sousa (up202006303)
Gabriel Ferreira (up201906072)
João Matos (up202006280)
João Pinheiro (up202008133)

Table of contents

01. Introduction

04. Indexing

07. Conclusions

02. Data

05. Parameters

03. Information
Retrieval Tool

06. Evaluation

01

Introduction



Introduction



In this milestone, we will explore the information retrieval processing we have conducted with the data collected and processed in the first milestone.



02

Dataset



Wikidata

The conflicts are
extracted from wikidata,
as well as structured
attributes

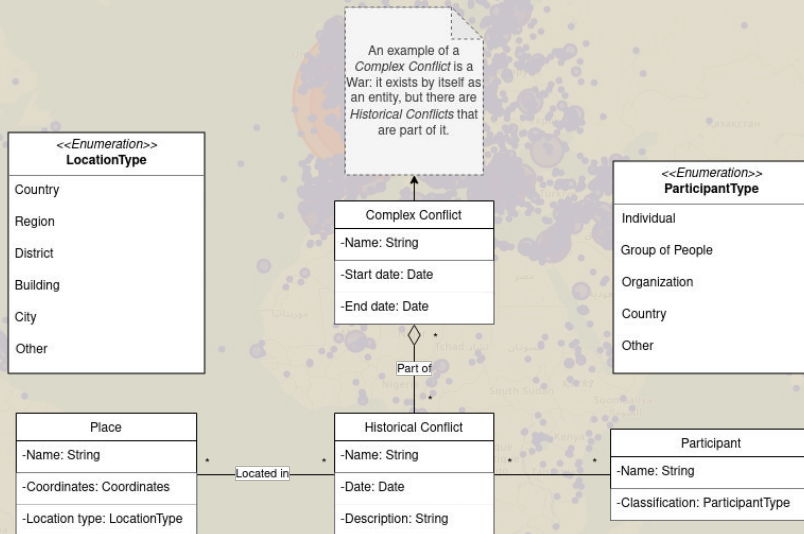


Wikipedia

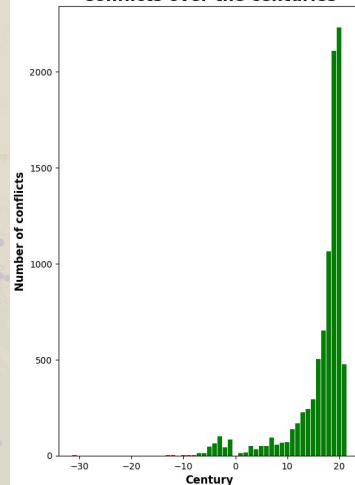
We use the summaries
from Wikipedia as the rich
text source

Dataset

Remembering our dataset...



Conflicts over the centuries

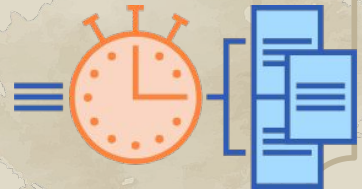


Information Retrieval Tool



Information Retrieval Tool

- Advanced Full-Text Search Capabilities
- Flexible and Adaptable
- Advanced Configurable Text Analysis
- Near Real-Time Indexing



+
Solr

Uses and
extends

APACHE
LUCENE

Indexing

04



Indexing

We indexed the fields that will be used for filtering, search and tokenization.

There was the need to create new types for the schema, as seen below.

| Type | Filters |
|-----------------------|--|
| caseInsensitiveString | ASCIIFoldingFilterFactory LowerCaseFilterFactory |
| urlString | ASCIIFoldingFilterFactory LowerCaseFilterFactory |
| richText | StopFilterFactory ASCIIFoldingFilterFactory EnglishMinimalStemFilterFactory LowerCaseFilterFactory SynonymGraphFilterFactory |

| Type | Field | Indexed |
|-----------------------|----------------------------|---------|
| caseInsensitiveString | country | true |
| | instance_of | true |
| | location | true |
| location | coordinate_location | true |
| pdate | date | true |
| | end_time | false |
| | inception | false |
| | point_in_time | true |
| | start_time | false |
| pint | participants_count | true |
| richText | participants | true |
| | part_of | true |
| | summary | true |
| string | article | false |
| | day_in_year_for_... | false |
| | destroyed | false |
| | located_in_on_physical_... | false |
| | significant_person | true |
| | end_time | false |
| | facet_of | false |
| | followed_by | false |
| | follows | false |
| | has_cause | false |
| | has_effect | false |
| | image | false |
| | in_opposition_to | false |
| | present_in_work | false |
| | time_period | false |
| | topics_main_category | false |
| | labeled | false |
| | text | false |
| urlString | event | true |
| text_en | label | true |

Parameters



Parameters

Parameter Selection Process

- Parameters were chosen based on their relevance to information needs and their impact on search results. Considerations included the balance between precision and recall, as well as the specific requirements of each query.


Parameter Examples

- "q: Portugal": Focuses on documents mentioning Portugal.
- "fq: date:[1300-01-01T00:00:00Z TO 1799-12-31T23:59:00Z]": Filters results within a specified date range.
- "(Boosted system only) bq: allied 3^3 participants_count:[3 TO *]^7": Boosts documents with many participants.



06

Evaluation





Base vs Enhanced System

Our information retrieval system underwent a comparative evaluation between the base and enhanced (boosted) configurations.

Comparison Criteria

Metrics:

- Precision at 10 ($P@10$): The proportion of relevant results among the top 10 retrieved documents.
 - Average Precision (AvP): The mean precision across different recall levels.
 - Precision-Recall Curves (P-R curves): Visual representation of precision at varying recall levels.
- 




Base vs Enhanced System

Query-Specific Performance

- Each information need was assessed independently to capture nuanced performance differences.

Results Overview

- **Base System:**
Overview of performance metrics for each query.
 - **Enhanced (Boosted) System:**
Comparative analysis, emphasizing improvements and modifications.
- 

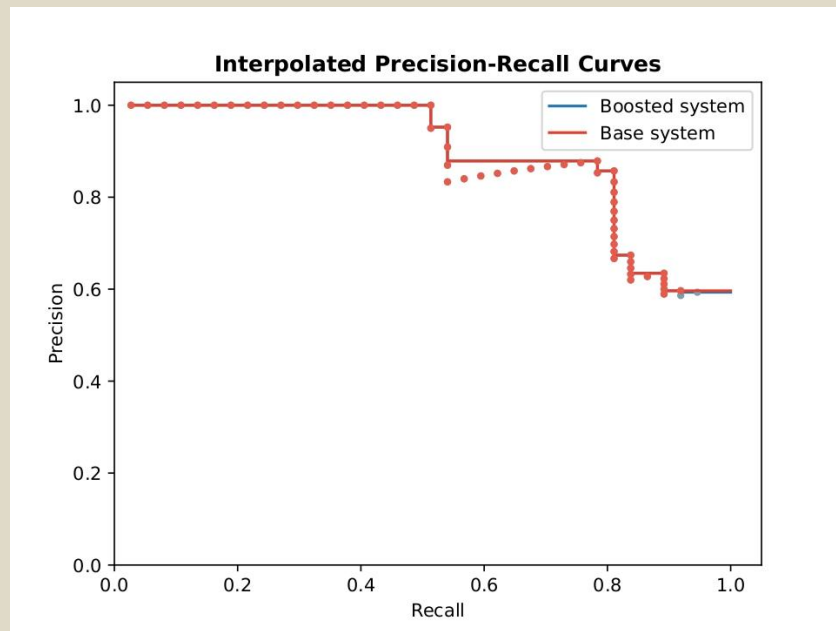
1

Evaluation

What battles took place near
a **river** between **1700** and
1775?

| Simple | Boosted |
|---|--|
| q:river | q:river |
| qf:summary location label participants | qf:summary participants^2 location^2 label^2" |
| fq:[date:[start date, end date]] | fq:[date:[start date, end date]] |

| | AvP | P@10 |
|----------------|----------|----------|
| Base System | 0.905253 | 1.000000 |
| Boosted system | 0.905253 | 1.000000 |



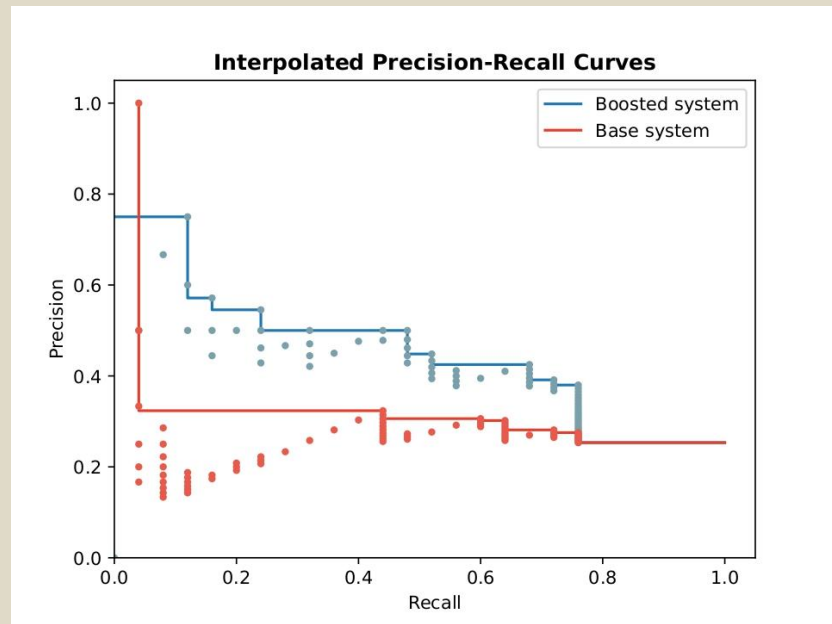
2

Evaluation

What were the most **destructive** battles in **Europe** during **World War I**?

| Simple | Boosted |
|--|--|
| q:destruction ruins bomb* devastat* destroy* damaging | q:destruction ruins bomb* devastat* destroy* damaging |
| qf:summary location label participants | qf:summary participants^2 location^2 label^2 |
| fq:[date:[start date, end date], part_of:("war 1"), summary:(europe italy germany france ...) location: (europe italy germany france britain...)] | fq:[date:[start date, end date], part_of:("war 1"), summary:(europe italy germany france ...) location: (europe italy germany france britain...)] |
| | bq:fort^3 bombard^5 bridge^5 city^5 terrain^4 village^5 devastated^7 ruins^7 severely^2 enormous^2 |

| | AvP | P@10 |
|----------------|----------|----------|
| Base System | 0.303171 | 0.200000 |
| Boosted system | 0.488829 | 0.500000 |



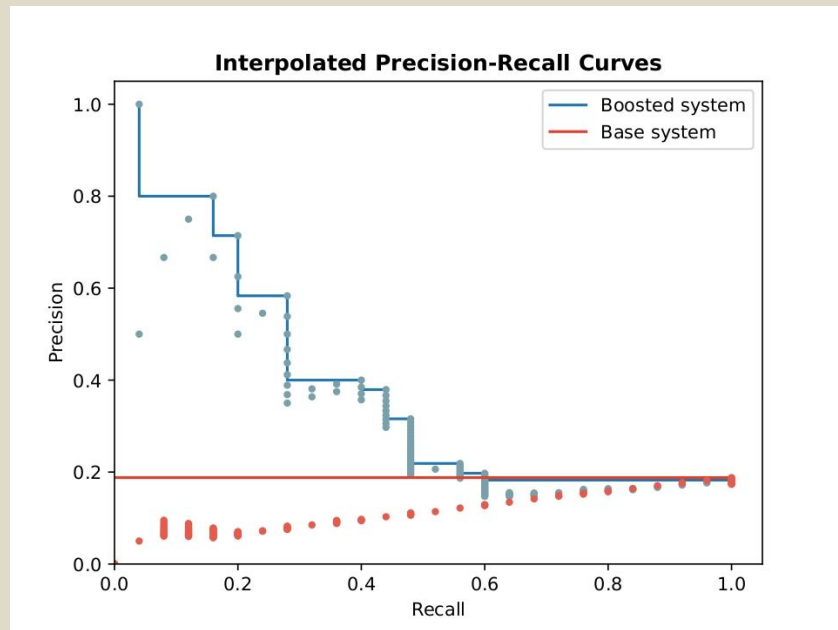
3

Evaluation

In which battles did the **Portuguese** participate between **1300 and 1800**, where they were not the main participant?

| Simple | Boosted |
|---|---|
| q:portugal | q:portugal |
| qf:summary, location, label, participants | qf:summary participants^2 location^2 label^2" |
| fq:[date:[start date, end date]] | fq:[date:[start date, end date]] |
| | bq:allied~3^3 participantes count:[3 TO *]^7 |

| | AvP | P@10 |
|----------------|----------|----------|
| Base System | 0.120492 | 0.000000 |
| Boosted system | 0.368246 | 0.500000 |



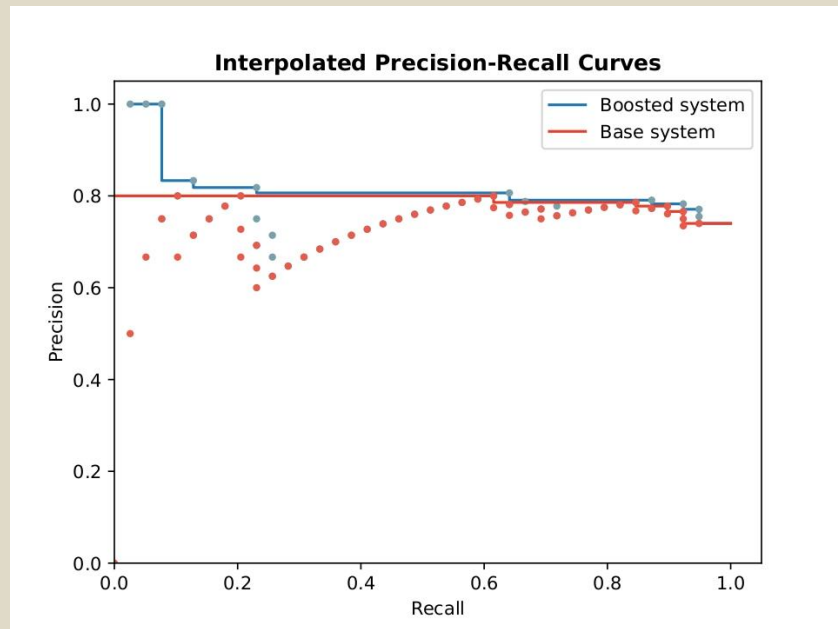
4

Evaluation

What battles, that are part of revolutions, had economics consequences?

| Simple | Boosted |
|---|---|
| q:economy | q:economy |
| qf:summary participants location label | qf:summary participants^2 location^2 label^2 |
| fq:[label:revolution* or summary:revolution* or part_of:revolution* or part_of:Revolution*] | fq:[label:revolution* or summary:revolution* or part_of:revolution* or part_of:Revolution*] |
| | bq:part_of:Revolution^3 instance_of:revolution^3 label:Revolution^3 summary: consequence |

| | AvP | P@10 |
|----------------|----------|----------|
| Base System | 0.740397 | 0.800000 |
| Boosted system | 0.782353 | 0.800000 |

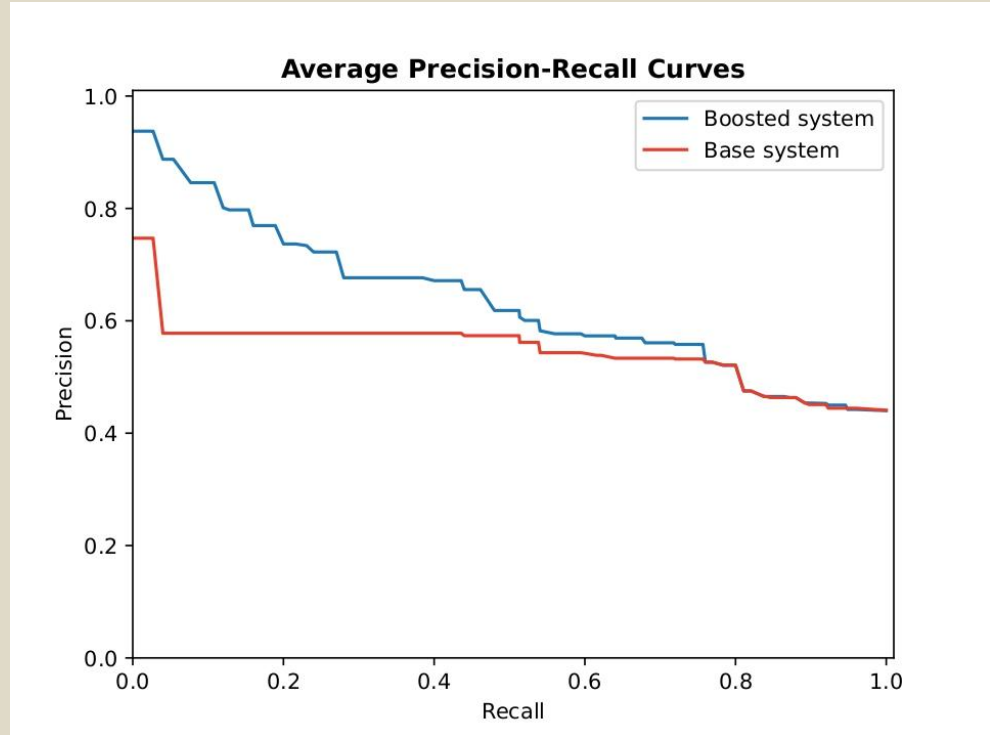


Evaluation

We ended the evaluation by comparing the overall average of the precision-recall curves for the 4 information needs.

The boosted system is consistently better than the base system.

| | mAvP |
|----------------|------------|
| Base System | 0.51732825 |
| Boosted system | 0.63617025 |



Conclusions



Conclusion

Solr's Strength in Information Retrieval

- Solr stands out as a powerful Information Retrieval tool, providing a robust foundation for our data exploration.

Critical Role of Evaluation

- Emphasizing the importance of result evaluation:

Focus Metrics:



Objective:

Precision at 10 (P@10).

Average Precision (AvP).

Precision-Recall Curves (P-R curves).

Understanding user information needs

Identifying areas for improvement.

Diverse Query Challenges

Conclusion

Diverse Query Challenges

- Not all information needs are equal:
 - Some queries are more straightforward.
 - Varying levels of complexity in finding relevant documents.

Consistent Boosted System Performance

- Across all four information needs:
 - The boosted system consistently outperformed the base system.
 - Improved precision, recall, and overall retrieval effectiveness.

Future Directions

- Leveraging Solr's capabilities for continued improvement.
- Expanding query capabilities for diverse historical research questions.
- Incorporating advanced NLP techniques for enhanced understanding of user queries.



Thanks

Do you have any questions?

