Portuguese Political Parties in Arquivo.pt

Information Processing and Retrieval, 2022

João Andrade, up201905589

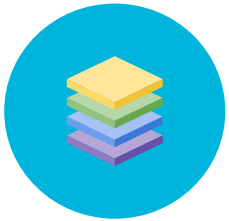
Maria Carneiro, up201907726

Miguel Azevedo, up201704590

Documents

Tool Selection

Since it is more text-oriented and our data is heavily textual, Solr met our information retrieval needs better



Document Definition

Added more parties (PCP, PAN, LIVRE, BE) and data since their respective website foundation. Also created a new party field and a title field in each document, and altered the date format to ISO8601

Indexing

stringParameter

link , type, party

textParameter

text, title

StandardTokenizerFactory
HTMLStripCharFilterFactory
ASCIIFoldingFilterFactory
LowerCaseFilterFactory
PortugueseStemFilterFactory

dateParameter

date

intParameter

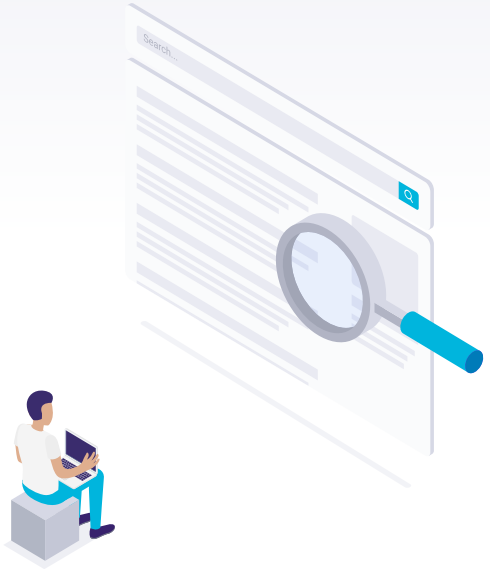
contentLength

All fields are indexed except for contentLength and link

Query Parser

Extended DisMax Parser

- **q** — represents the main query on which we are performing the search
- **q.op** — represents the operator used in the query expressions
- **qf** — list of fields whom importance gets determined by a boost factor associated with each one of them
- **fq** — defines a query that is used to restrict the superset of documents that can be returned
- **ps** — represents the phrase 'slop', which is the distance between the terms of the query while still considering it a phrase match



Queries

Grupo parlamentar do PSD em 2019

q.op=AND
query=grupo parlamentar
qf=title^5 text
fq=party:PSD
fq=date:[2019-01-01T00:00:00Z TO
2019-12-31T23:59:59Z]

Qual foi o candidato do Bloco de Esquerda às Eleições Presidenciais de 2016?

q.op=AND
query=candidato presidencial
qf=title^5 text
ps=100
fq=party:BLOCO
fq=date:[2015-01-01T00:00:00Z TO 2016-02-01T23:59:59Z]

Programa Eleitoral da Iniciativa Liberal para as Legislativas de 2019

q.op = AND
query = programa^3 eleitoral legislativas^3
qf=title^5 text
ps = 30
fq = party:IL
fq = date:[2019-01-01T00:00:00Z TO
2019-12-31T23:59:59Z]

Queries

Posição do PS sobre a eutanásia

q.op = AND
query = eutanásia
qf = title^10 text
fq = party:PS

Posição dos partidos sobre o aborto

q.op = AND
query = aborto
qf=title^10 text

Sampling

Given the nature of our dataset (sparse in terms of content), with a small sized sample, there were no relevant documents for most of the queries

Random Sample

1600 documents (200 per party)

Manually addition of relevant documents

Enriched Random Sample

1600+ documents (200+ per party)

By manually adding relevant documents we forego some of the randomness of our sample, however it was a necessary tradeoff

Complete Dataset

242890 documents

Impossible to know the exact number of relevant results for each query

Restrict the analysis to the first K results

242890

Not an exact approach, since there is no guarantee that after the kth result there won't be any more relevant results

Search Systems

- Non-Indexed Sample
  - No Boosts
    - Simple Queries
  - With Boosts
    - Complex Queries
- Indexed Sample
  - No Boosts
    - Simple Queries
  - With Boosts
    - Complex Queries

Evaluation

Performance Metrics for Simple Queries using Non-indexed System

Queries	Precision	P@10	AvP	F Measure	Recall
Q1	**0.057**	**0.4**	**1**	**0.108**	**1**
Q2	**0**	**0**	**0**	**0**	**0**
Q3	**0.384**	**0.4**	**0.6**	**0.526**	**0.833**
Q4	**0.065**	**0.6**	**0.837**	**0.123**	**1**
Q5	**0.333**	**0.1**	**1**	**0.5**	**1**

Evaluation

Performance Metrics for Complex Queries using Non-indexed System

Queries	Precision	P@10	AvP	F Measure	Recall
Q1	**0.2**	**0.4**	**1**	**0.333**	**1**
Q2	**0**	**0**	**0**	**0**	**0**
Q3	**0.454**	**0.5**	**0.722**	**0.588**	**0.833**
Q4	**0.065**	**0.6**	**0.916**	**0.123**	**1**
Q5	**0.333**	**0.1**	**1**	**0.5**	**1**

Evaluation

- Performance Metrics for Simple Queries using Indexed System

Queries	Precision	P@10	AvP	F Measure	Recall
Q1	**0.057**	**0.4**	**1**	**0.108**	**1**
Q2	**0.125**	**0.2**	**0.75**	**0.222**	**1**
Q3	**0.25**	**0.3**	**0.591**	**0.384**	**0.833**
Q4	**0.065**	**0.5**	**1**	**0.122**	**1**
Q5	**0.333**	**0.1**	**1**	**0.5**	**1**

Evaluation

Performance Metrics for Complex Queries using Indexed System

Queries	Precision	P@10	AvP	F Measure	Recall
Q1	**0.2**	**0.4**	**1**	**0.333**	**1**
Q2	**0.666**	**0.2**	**1**	**0.8**	**1**
Q3	**0.454**	**0.5**	**0.722**	**0.588**	**0.833**
Q4	**0.065**	**0.6**	**1**	**0.122**	**1**
Q5	**0.333**	**0.1**	**1**	**0.5**	**1**

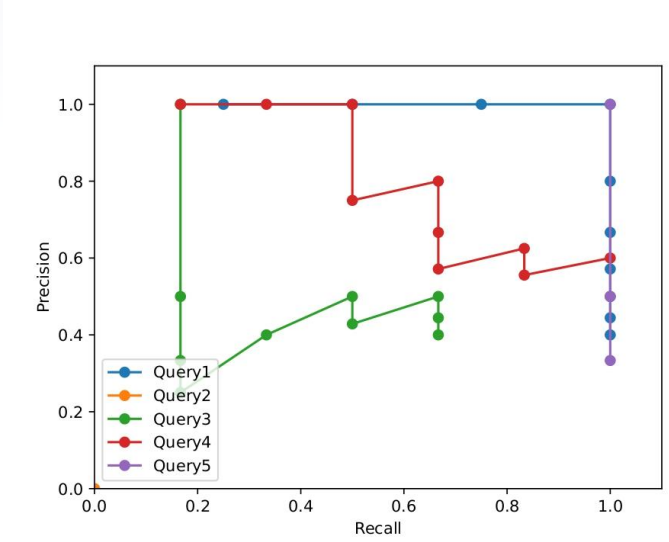
Evaluation

Mean Average Precision metrics for all systems

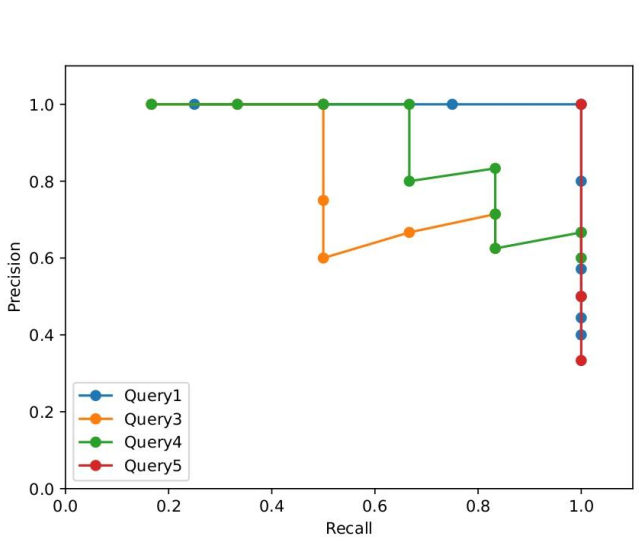
System	MAP
Non-Indexed Simple Queries	**0.687**
Non-Indexed Complex Queries	**0.759**
Indexed Simple Queries	**0.868**
Indexed Complex Queries	**0.944**

Evaluation

Non-Indexed Precision Recall Curves



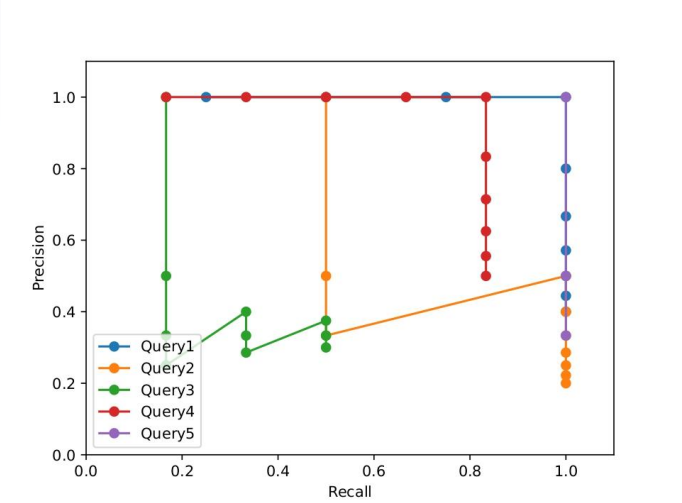
Simple Queries



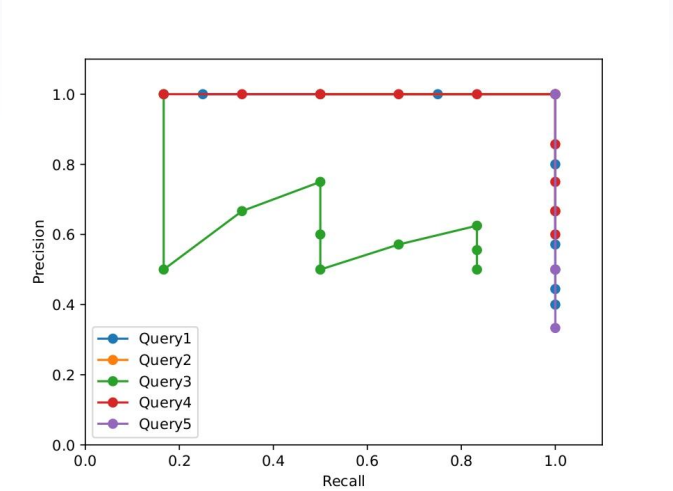
Complex Queries

Evaluation

Indexed Precision Recall Curves



Simple Queries



Complex Queries

Future Work

1. Check previous work to make sure everything is working as intended
2. Improve search system by adding more filters and query analyzers
3. Create final version of the search system