Information Search System for Versioned Portuguese News Articles about Technology

Information Processing and Retrieval course project

FEUP

14th December 2021

João Romão

Rafael Cristino

Xavier Pisco

Collection

- Solr search platform was chosen.
- Transformed previous .json file that contained our dataset into a collection of documents.
- Dates were converted into Solr format: YYYY-MM-DDThh:mm:ssZ.
- Each document contains the following information: urlkey, timestamp, URL, article, newspaper.
- The article field contains:
  - title
  - summary
  - url to cover image
  - published date
  - author name
  - content
  - entities

Indexation

- We created indexes on the fields we would be using the most for our queries:
  - url
  - article title
  - article summary
  - article content
  - timestamp
  - article entities
  - article authors
  - article publish date
  - urlkey

Indexation

- For the text fields we:
  - Tokenized them with solr.StandardTokenizerFactory
  - Filtered them with solr.ASCIIFoldingFilterFactory, solr.LowerCaseFilterFactory and solr.PortugueseStemFilterFactory
- For the date fields we treated them as solr.DatePointField.
- The same was applied to the queries using those fields.

Information Retrieval

Search for different versions of an article based on a URL

- q: url:"https://www.noticiasaominuto.com/tech/1125833/sonda-da-nasa-ja-aterrou-em-marte"

Date filtered search

- q: article.publish_date: [2021-05-20T00:00:00Z TO 2021-08-15T00:00:00Z]
- fq: {!collapse field="urlkey" sort='timestamp desc'}

Information Retrieval

Search for the number of times a page was indexed in an interval of time

- q: url:"https://visao.sapo.pt/exameinformatica/noticias-ei/2010-04-26-microsoft-touch-pack-para-windows-7-ja-disponivel/" AND timestamp:[2021-03-01T00:00:00Z TO 2021-03-31T23:59:59Z]
- Raw Parameter Queries: group=true & group.field=urlkey & group.sort=timestamp desc

Information Retrieval

Text search

- q:
  - article.text:"aterragem em Marte" OR
  - article.title:"aterragem em Marte" OR
  - article.entities.title:"aterragem em Marte" OR
  - article.summary:"aterragem em Marte" OR
  - article.authors:"aterragem em Marte"
- defType: disMax
- qf: article.title³ article.entities.title³ article.text article.summary article.publish_date

Information Retrieval

Combination of multiple parameters

- "Search for an article that has *Sistema Operativo Android* in its text, was authored by someone called *Pedro* and published in the *Exame Informática* newspaper before *20-05-2011*, and has the entity titled *Google* associated with it"
- q:
  - article.entities.title:"Google" AND
  - newspaper:exameinformatica AND
  - article.publish_date:[* TO 2011-05-20T00:00:00Z] AND
  - article.authors:Pedro AND
  - article.text:Sistema Operativo Android
- fq: {!collapse field="urlkey" sort='timestamp desc'}

Information Retrieval

Search for articles that have differences in their texts in each indexed version (proposed search query that is not included)

- q: url:"https://visao.sapo.pt/exameinformatica/noticias-ei/software/2019-08-08-direcoes-em-realidade-aumentada-chegam-ao-google-maps/"
- fq: {!collapse field="urlkey" sort='timestamp desc'}
- Raw Parameter Queries: expand=true & expand.field=article.text

Evaluation

- For the evaluation of our information retrieval system we used 2 different text queries:
  - “aterragem em Marte”
  - “Microsoft Teams”
- We tested both of the queries in 3 different scenarios:
  - Schemaless and no attribute weighting
  - With a custom schema and no weights
  - With a custom schema and weights

Aterragem em Marte

Schemaless and weightless:

k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	R	R	N	R	N	R	R
P@k	1	1	1	1	1	0.83	0.86	0.75	0.77	0.8
R@k	0.09	0.18	0.27	0.36	0.45	0.45	0.54	0.54	0.63	0.73

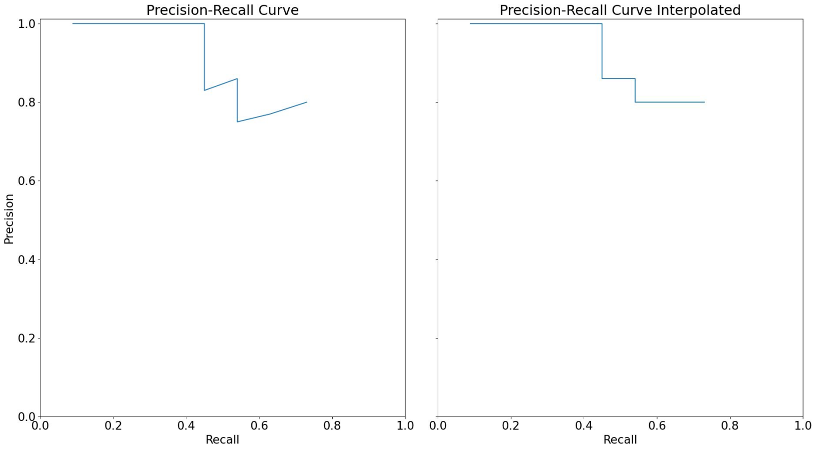
Schema and weightless:

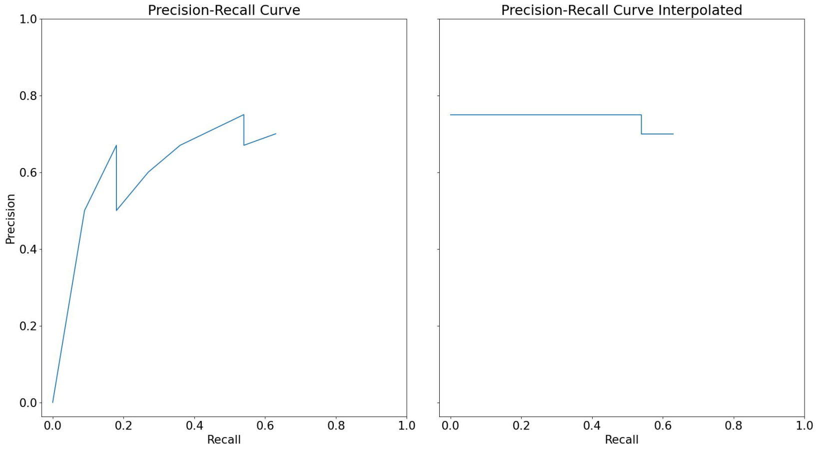
k	1	2	3	4	5	6	7	8	9	10
Relevant	R	N	N	R	N	N	N	N	R	N
P@k	0	0.50	0.67	0.50	0.60	0.67	0.71	0.75	0.67	0.7
R@k	0	0.09	0.18	0.18	0.27	0.36	0.45	0.54	0.54	0.63

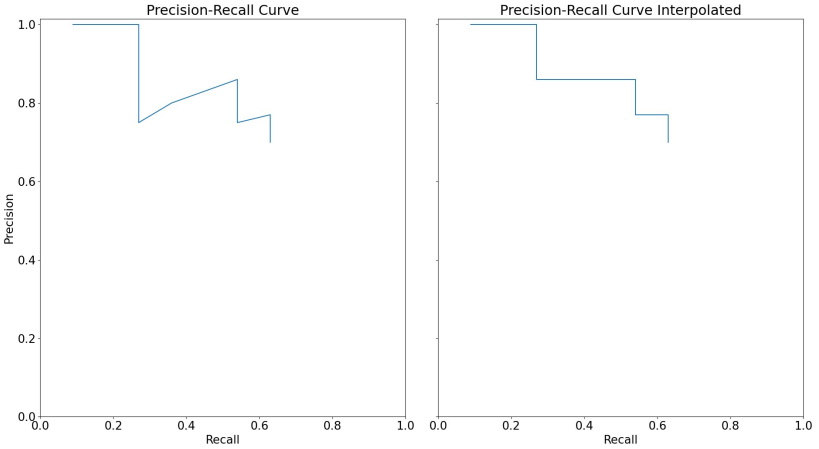
Schema and weights:

k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	N	R	R	R	N	R	N
P@k	1	1	1	0.75	0.80	0.83	0.86	0.75	0.77	0.7
R@k	0.09	0.18	0.27	0.27	0.36	0.45	0.54	0.54	0.63	0.63

Aterragem em Marte







Microsoft Teams

Schemaless and weightless:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	N	N	N	N	N	N	N	N	R
P@k	0	0	0	0	0	0	0	0	0	0.1
R@k	0	0	0	0	0	0	0	0	0	0.1

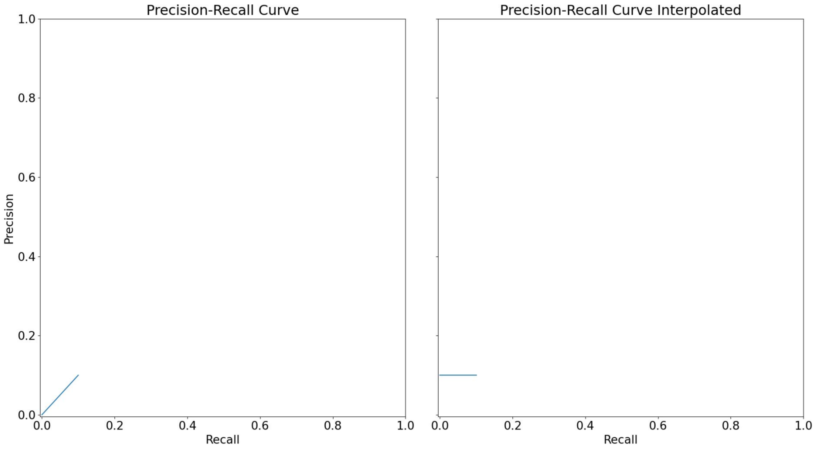
Schema and weightless:

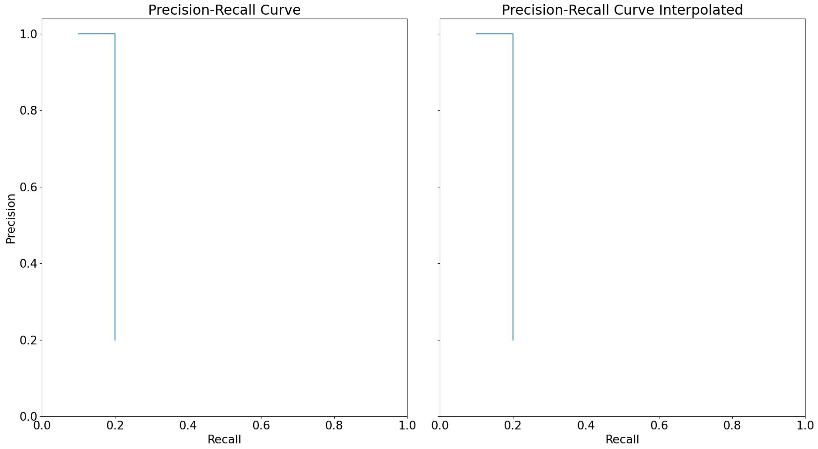
k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	N	N	N	N	N	N	N	N
P@k	1	1	0.67	0.5	0.4	0.33	0.29	0.25	0.22	0.2
R@k	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

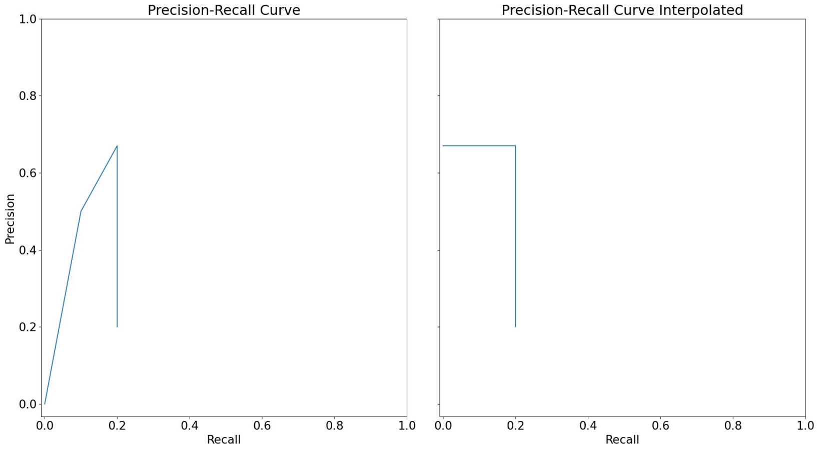
Schema and weights:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	R	R	N	N	N	N	N	N	N
P@k	0	0.5	0.67	0.5	0.4	0.33	0.29	0.25	0.22	0.2
R@k	0	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

Microsoft Teams







Future Improvements

- Improve the results of the queries with better weights.