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

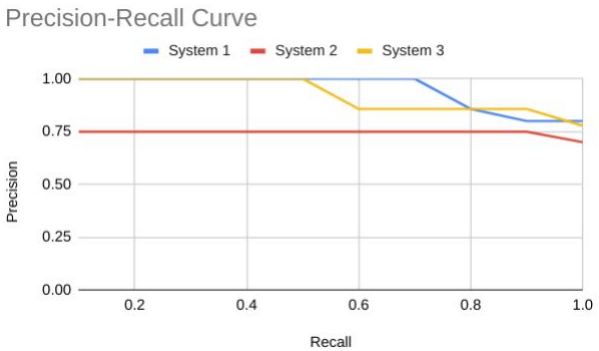
Milestone 3

- Revisions to the last milestone
  - Corrected precision-recall curves
- Improvement of dataset
  - Further cleaning of entities data
- Study of weights for the system
- Implementation of synonyms
- Implementation of facets
- Front-end implementation

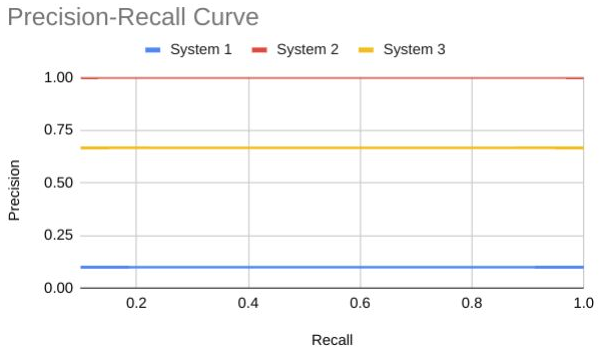
Revisions to the last milestone

- Recalls were being calculated incorrectly.
- As a result, the graphics were incorrect.
- Conclusions drawn from graphics analysis were revised.

‘aterragem em Marte’ query



‘Microsoft Teams’ query



Improvement of dataset

- Our dataset is imbalanced, most of our news are from 1 website.
- Some websites are not properly indexed by “Arquivo.pt”.
- Some websites have no difference in URLs between tech news and other news.
- We found a possible addition to our dataset “tek.sapo.pt”.
    - All the available websites are Wireframes, which meant we couldn’t get the news from it.
- With no increase in the number of documents, we only fixed some errors we found as we were working.

Weights

- System 1: High importance to the title of the news
  - qf: article.title^10 article.entities.title article.text article.summary
- System 2: High importance to the entities in the text
  - article.title article.entities.title^10 article.text article.summary
- System 3: Overall importance
  - article.title^5 article.entities.title^5 article.text article.summary^3

Microsoft Teams

System 1:

k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	R	R	R	N	N	N	N
P@k	1	1	1	1	1	1	0.86	0.75	0.67	0.6
R@k	0.17	0.33	0.5	0.67	0.83	1	1	1	1	1

System 2:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	N	R	N	N	N	N	N	N	N
P@k	0	0	0.33	0.25	0.2	0.17	0.14	0.13	0.11	0.1
R@k	0	0	1	1	1	1	1	1	1	1

System 3:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	R	R	R	R	R	R	N	N	R
P@k	0	0.5	0.67	0.75	0.8	0.83	0.86	0.75	0.67	0.7
R@k	0	0.14	0.29	0.43	0.57	0.71	0.86	0.86	0.86	1

Prémio Nobel

System 1:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	R	R	R	R	R	R	R	N	N
P@k	0	0.5	0.67	0.75	0.8	0.83	0.86	0.88	0.78	0.7
R@k	0	0.14	0.29	0.43	0.57	0.71	0.86	1	1	1

System 2:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	N	R	N	N	N	N	N	N	R
P@k	0	0	0.33	0.25	0.2	0.17	0.14	0.13	0.11	0.2
R@k	0	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1

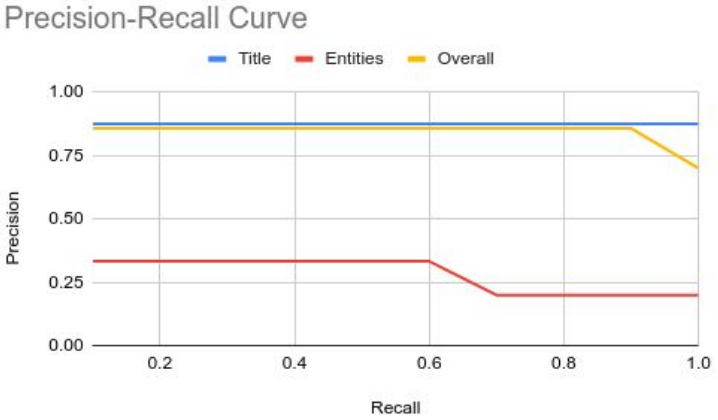
System 3:

k	1	2	3	4	5	6	7	8	9	10
Relevant	N	R	R	R	R	R	R	N	N	R
P@k	0	0.5	0.67	0.75	0.8	0.83	0.86	0.75	0.67	0.7
R@k	0	0.14	0.29	0.43	0.57	0.71	0.86	0.86	0.86	1

Precision-Recall curves

Microsoft Teams

Prémio Nobel



Synonyms

- We wanted a more robust system that would search for documents with synonyms of the words we used on the query.
- Since our dataset is focused in news related to technologies, our synonyms list has terms associated with technology:
    - TV, televisão, ecrã, plasma
    - App, aplicação, software
    - aterrar, chegar, pousar
- In Solr we used **solr.SynonymGraphFilterFactory** filter with a synonyms file.

App Televisão

Without synonyms:

k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	R	N	N	N	N	N	N
P@k	1	1	1	1	0.8	0.67	0.57	0.5	0.44	0.4
R@k	0.25	0.5	0.75	1	1	1	1	1	1	1

With synonyms:

k	1	2	3	4	5	6	7	8	9	10
Relevant	R	N	R	R	N	N	R	R	N	N
P@k	1	0.5	0.67	0.75	0.6	0.5	0.57	0.63	0.56	0.5
R@k	0.2	0.2	0.4	0.6	0.6	0.6	0.8	1	1	1

Aterragem em Marte

Without synonyms:

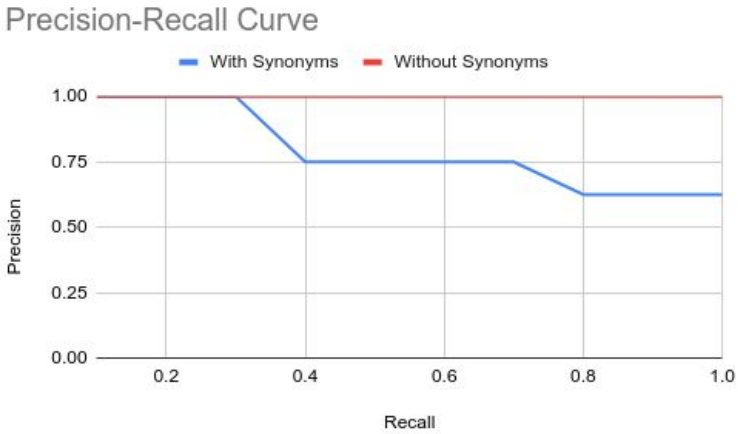
k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	R	N	N	N	N	N	N
P@k	1	1	1	1	0.8	0.67	0.57	0.5	0.44	0.4
R@k	0.25	0.5	0.75	1	1	1	1	1	1	1

With synonyms:

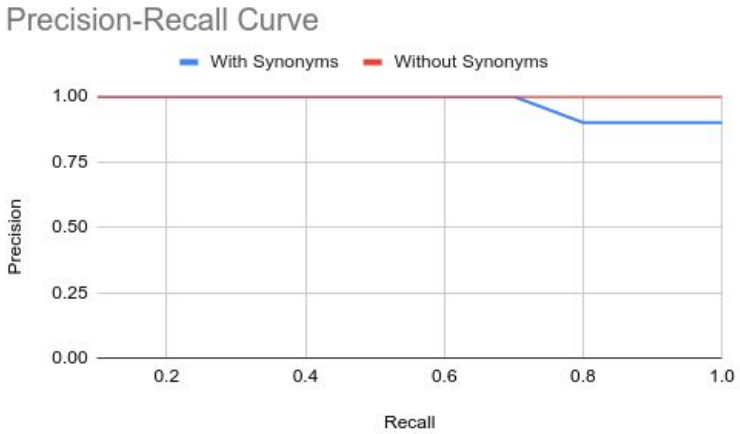
k	1	2	3	4	5	6	7	8	9	10
Relevant	R	R	R	R	R	R	N	R	R	R
P@k	1	1	1	1	1	1	0.86	0.88	0.89	0.9
R@k	0.11	0.22	0.33	0.44	0.56	0.67	0.67	0.78	0.89	1

Precision-Recall Curves

App Televisão:



Aterragem em Marte:



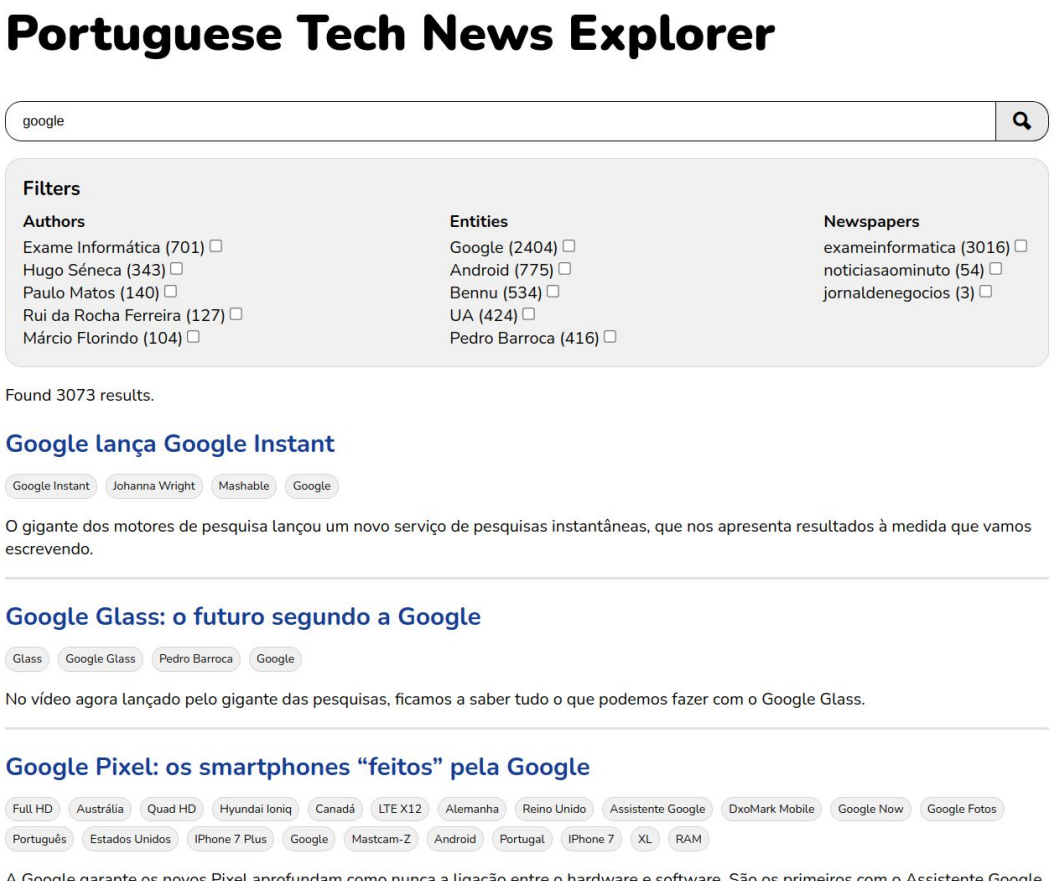
Facets

- To allow the end user to filter the results provided by the search system, the facets utility provided by solr was used.
- With ‘facet=true’ the query returns the number of documents that are categorized in each of the facet values.
- Facets introduced:
  - Article entities’ title
  - Article authors
  - Newspaper

Facets

- Since the *PortugueseStemFilterFactory* is being applied, the facet results were, at first, not the expected ones.
- The *KeywordRepeatFilterFactory* was added as documentation states:
    - ‘If placed before a stemmer, the result will be that you will get the unstemmed token preserved on the same position as the stemmed one.’
- Even though the obtained results were better, they were not the intended ones.
- Final solution was to duplicate the fields and store them as strings.

Frontend


Portuguese Tech News Explorer
google
Filters
Authors
Exame Informática (701)
Hugo Séneca (343)
Paulo Matos (140)
Rui da Rocha Ferreira (127)
Márcio Florindo (104)
Entities
Google (2404)
Android (775)
Bennu (534)
UA (424)
Pedro Barroca (416)
Newspapers
exameinformatica (3016)
noticiasaominuto (54)
jornaldenegocios (3)
Found 3073 results.
Google lança Google Instant
Google Instant
Johanna Wright
Mashable
Google
O gigante dos motores de pesquisa lançou um novo serviço de pesquisas instantâneas, que nos apresenta resultados à medida que vamos escrevendo.
Google Glass: o futuro segundo a Google
Glass
Google Glass
Pedro Barroca
Google
No vídeo agora lançado pelo gigante das pesquisas, ficamos a saber tudo o que podemos fazer com o Google Glass.
Google Pixel: os smartphones “feitos” pela Google
Full HD
Austrália
Quad HD
Hyundai Ioniq
Canadá
LTE X12
Alemanha
Reino Unido
Assistente Google
DxoMark Mobile
Google Now
Google Fotos
Português
Estados Unidos
IPhone 7 Plus
Google
Mastcam-Z
Android
Portugal
IPhone 7
XL
RAM

Future Improvements

- Improve frontend for a better experience - allow navigating between different same-article versions.
- Search results ranking based on user behaviour.