Information Search System for Versioned Portuguese News Articles about Technology

Information Processing and Retrieval course project

FEUP

16th November 2021

João Romão

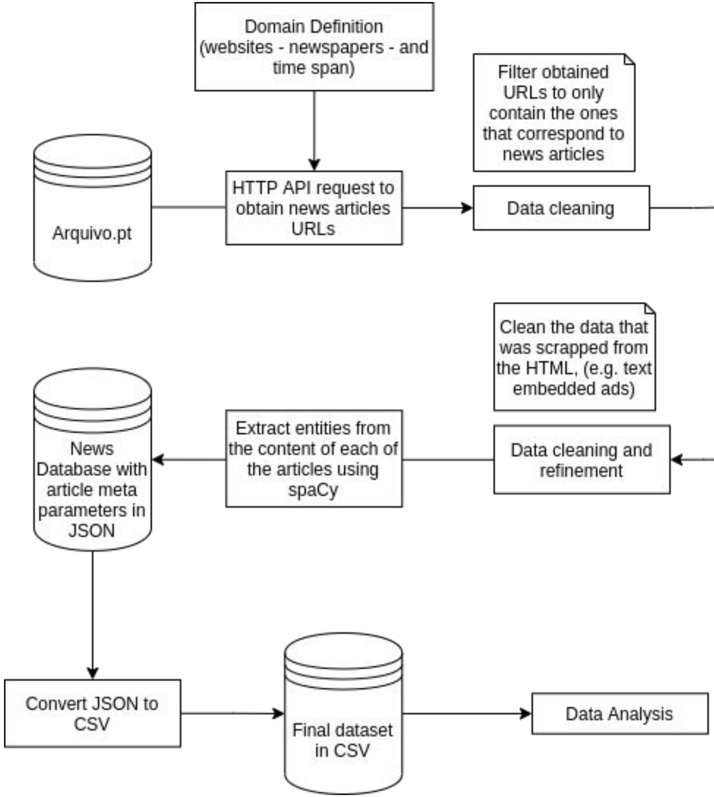
Rafael Cristino

Xavier Pisco

Arquivo.pt Dataset

- Arquivo.pt was created by *Fundação para a Ciência e Tecnologia* (FCT).
- Arquivo.pt has been crawling the Portuguese web and storing the webpages it finds since 1996 (over time storing multiple versions of each webpage).
- This means that it contains a HUGE amount of data.
- We decided to collect our dataset from a smaller scope of data: news articles about technology that were indexed by Arquivo.pt in 2021, from the 1st of January to the 1st of November, and that were published by one of 3 portuguese news media companies: *Notícias ao Minuto*, *Jornal de Notícias*, and *Exame Informática*.
- This resulted in a 183.5MB dataset, with 19580 unique articles, and in total, counting with all of the different versions of each article, with 83838 entries.

Data Processing and Retrieval Pipeline



Domain Definition (websites - newspapers - and time span)

Arquivo.pt

HTTP API request to obtain news articles URLs

Filter obtained URLs to only contain the ones that correspond to news articles

Data cleaning

HTTP API request to obtain the HTML of each article

Scrapping each article's HTML content with Beautiful Soup to obtain article information (e.g. author, title, content, published date, etc)

Clean the data that was scrapped from the HTML, (e.g. text embedded ads)

Data cleaning and refinement

Extract entities from the content of each of the articles using spaCy

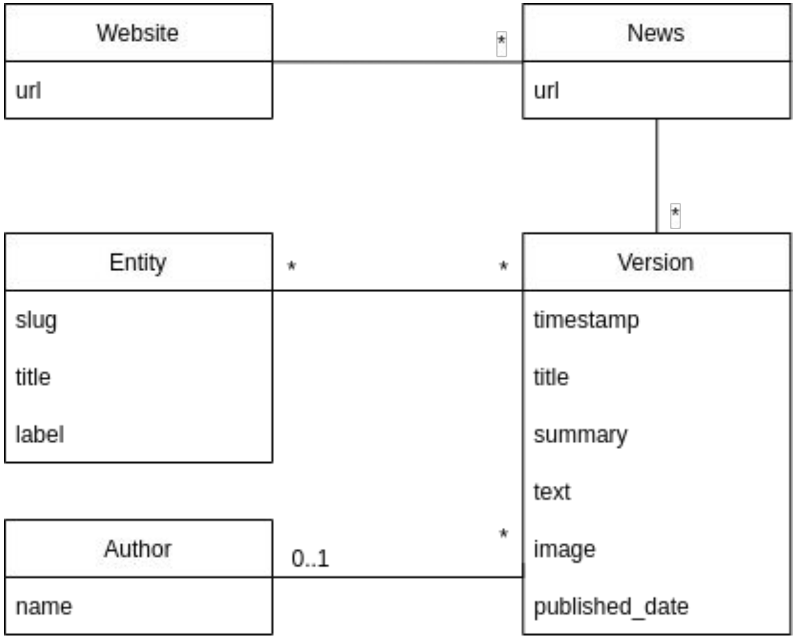
News Database with article meta parameters in JSON

Convert JSON to CSV

Final dataset in CSV

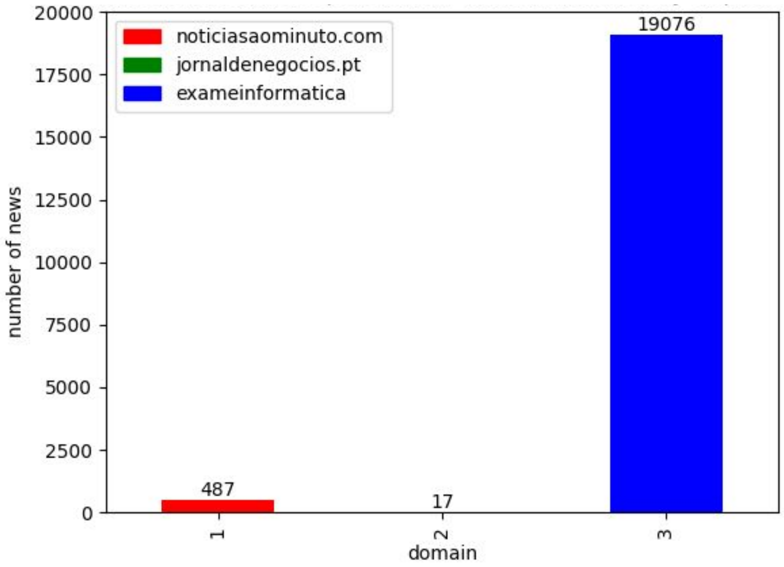
Data Analysis

Conceptual Data Model

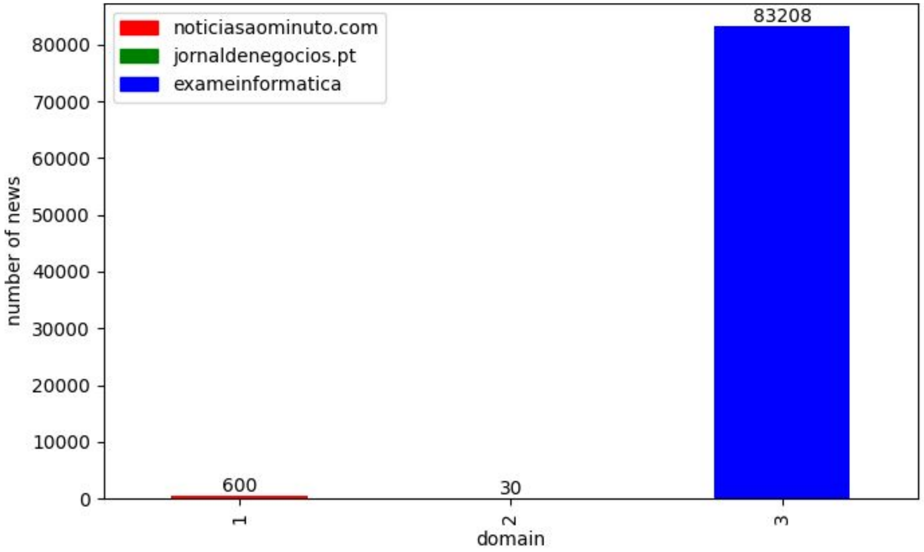


Data Characterization

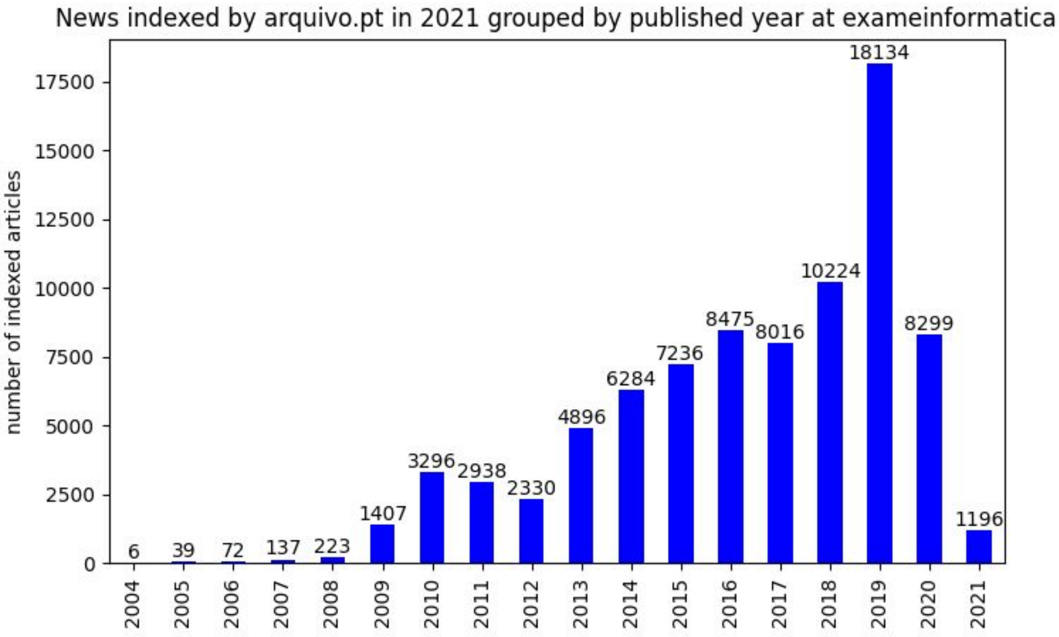
Total number of unique articles per domain indexed by arquivo.pt in 2021



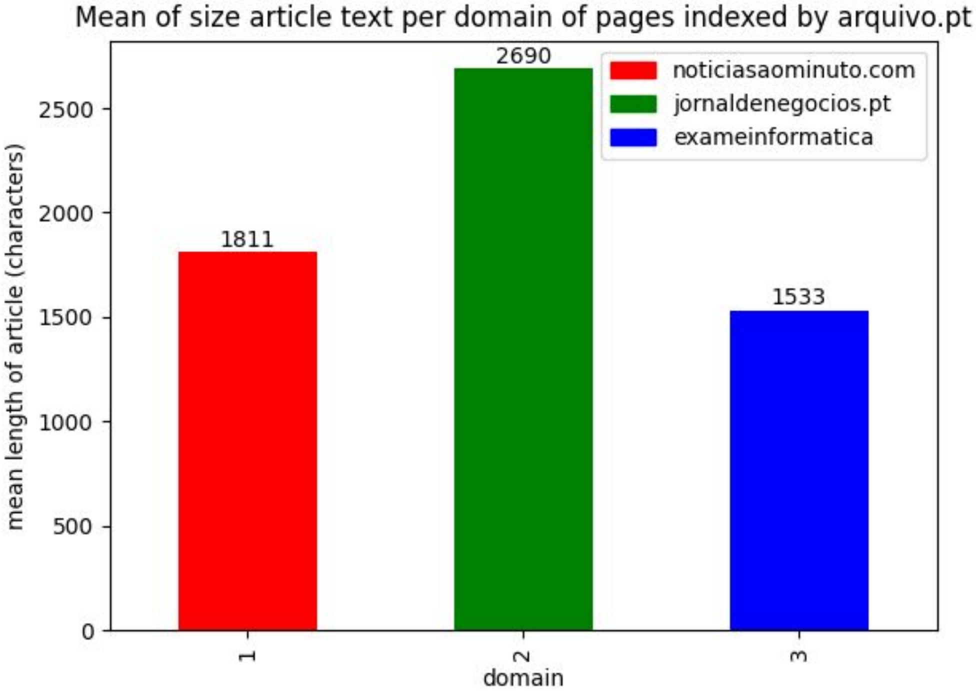
Total number of article versions per domain indexed by arquivo.pt in 2021



Data Characterization

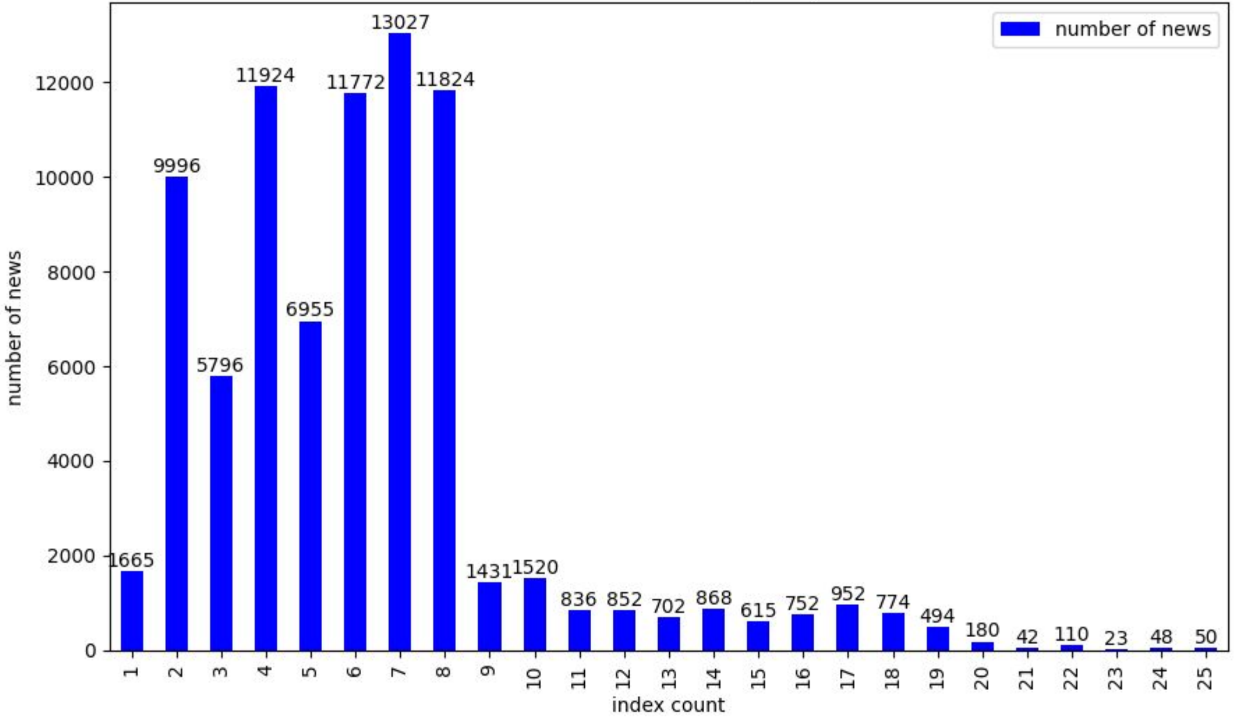


Data Characterization

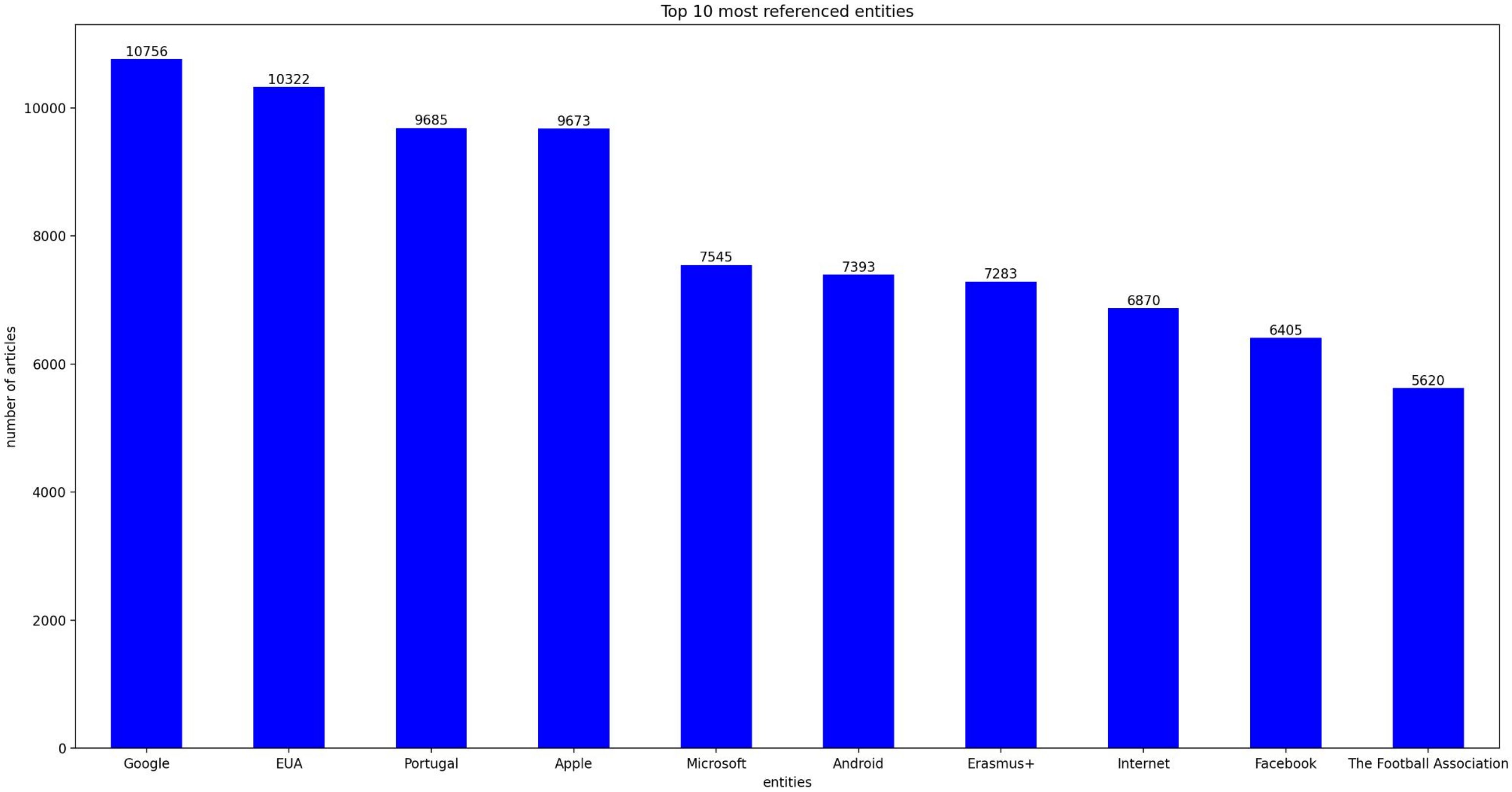


Data Characterization

Distribution of news articles by the amount of indexations by Arquivo.pt



Data Characterization



Possible Search Queries

- Search for a specific webpage to obtain the contents of previous versions.
  - Example search query: “https://visao.sapo.pt/exameinformatica/noticias-ei/insolitos/2019-01-28-soma-pipe-o-sintetizador-que-se-sopra/”
  - The result would be the version entries that correspond to this page, if any.
- Search for any topic (as in a search engine).
  - Example search query: “*convenção aeroespacial*”
  - The result would be the most relevant articles, if any.
- Search (or filter) by articles that were published in a certain date or in a range of dates.
  - Example search query: “published in 19/01/2019 - 25/02/2019”
  - The result would be the articles that were published between the given dates.

Possible Search Queries

- Search (or filter) by articles that were written by a specific author.
  - Example search query: “written by Miguel Patinha Dias”
  - The result would be the articles that were authored by the given author.
- Search (or filter) by the most or the least indexed news.
  - Example search query: “indexed more than 10 times”
  - The result would be the articles that were indexed more than 10 times.
- Search for articles whose versions have text content differences between them (which means that they were edited after publishing by the news publisher)
  - Example search query: “articles that were edited”
  - The result would be the articles that have text differences between their versions.

Future Improvements

- Maybe choose newspapers that are indexed at a similar rate (in the newspapers we chose, *Exame Informática* is indexed at a higher rate than the others - probably because they publish more frequently).
- Improve entity cleaning methods (sometimes entities are erroneously classified).
- Maybe we don’t need to store information about each of the news’ versions if their content is not different from one to the other. In that case they could share the information.