U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Art Analysis

Dataset Characterization

MIEIC 2020/2021

Descrição, Armazenamento e Pesquisa de Informação

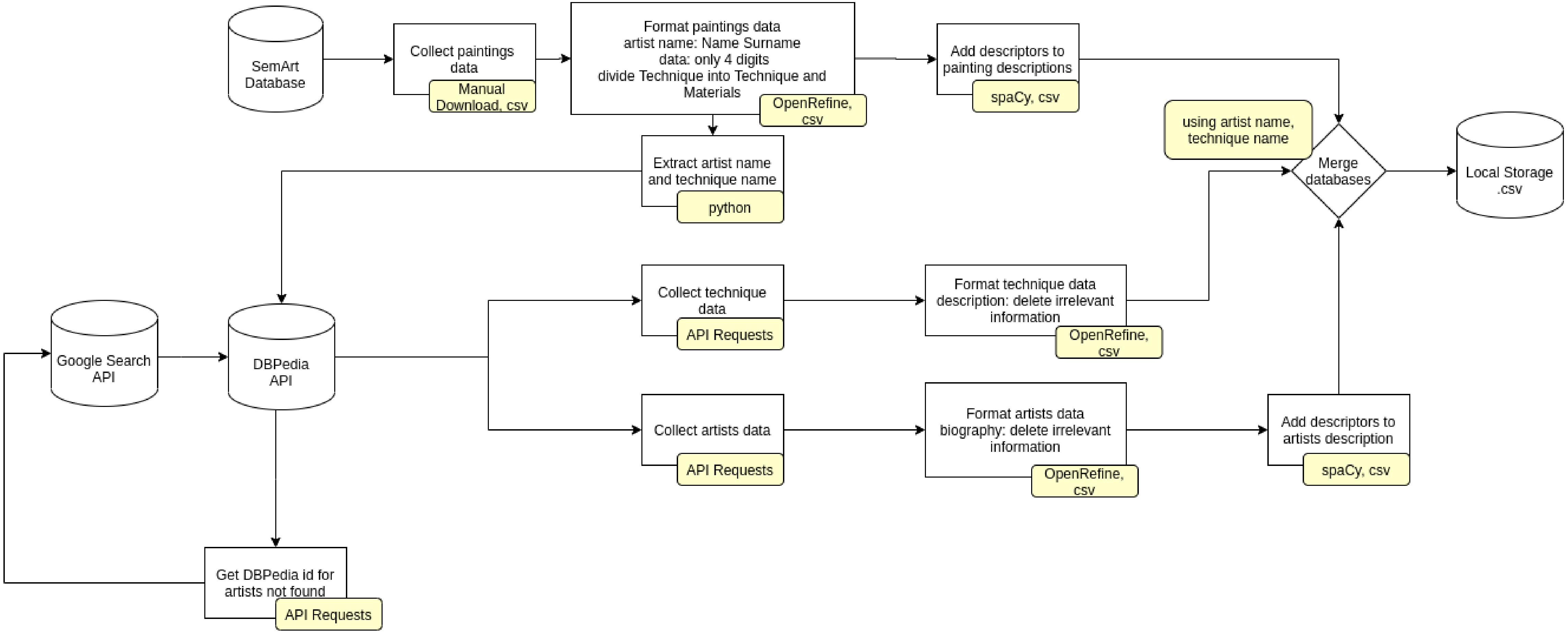
Ana Silva, up201604105

Fábio Araújo, up201607944

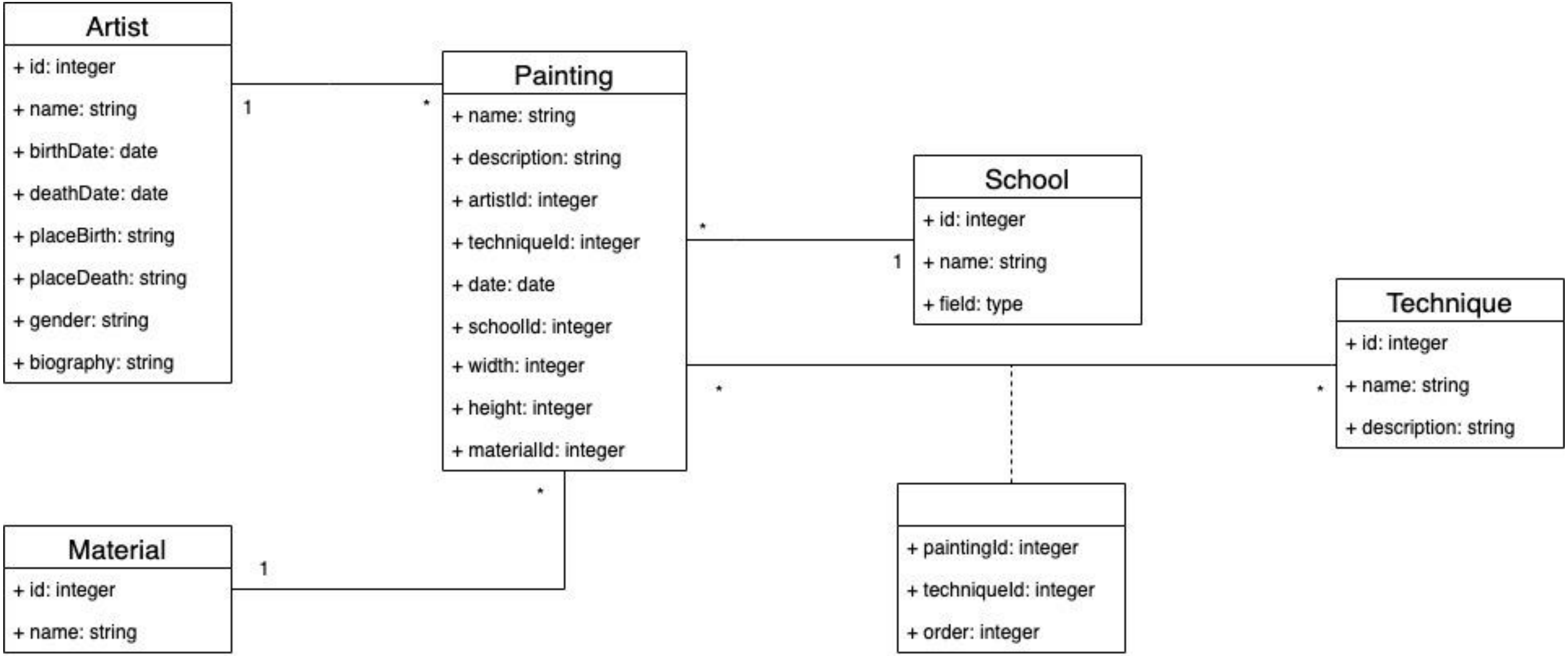
Gonçalo Santos, up201603265

Susana Lima, up201603634

Data Collection



Conceptual Model



Artist

+ id: integer
+ name: string
+ birthDate: date
+ deathDate: date
+ placeBirth: string
+ placeDeath: string
+ gender: string
+ biography: string

Painting

+ name: string
+ description: string
+ artistId: integer
+ techniqueId: integer
+ date: date
+ schoolId: integer
+ width: integer
+ height: integer
+ materialId: integer

School

+ id: integer
+ name: string
+ field: type

Technique

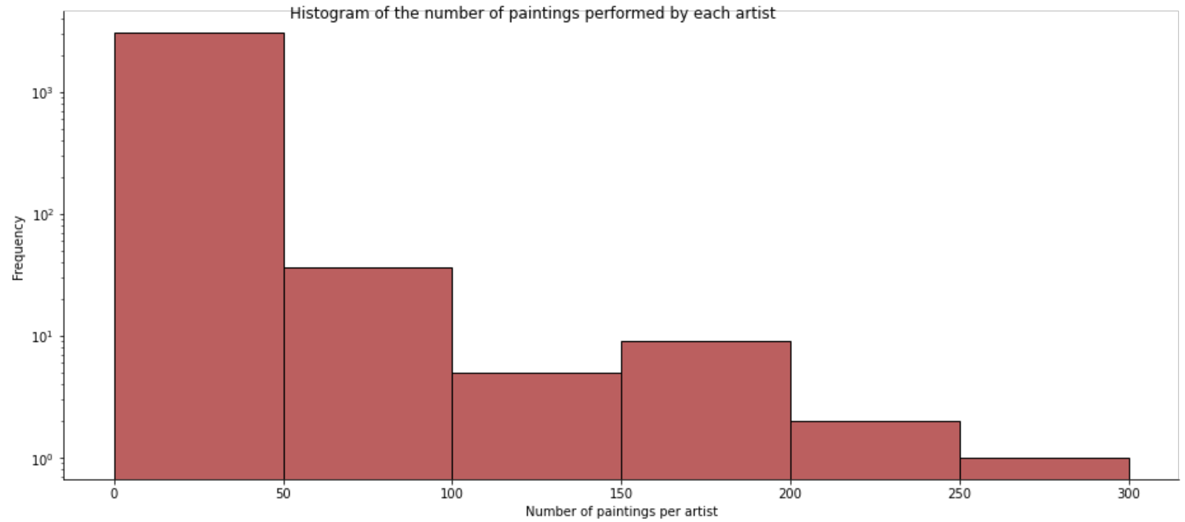
+ id: integer
+ name: string
+ description: string

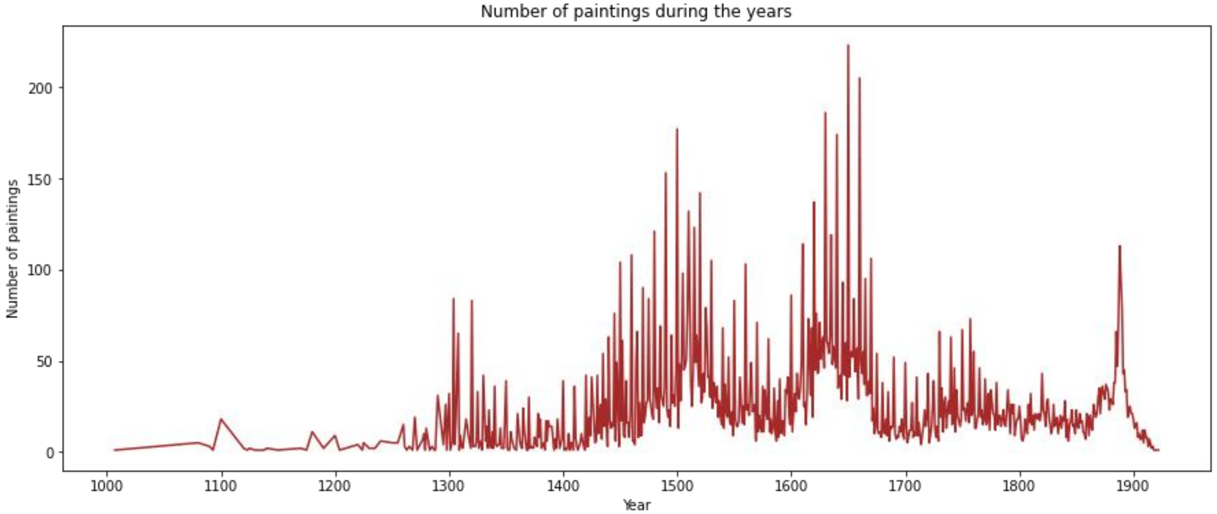
Material

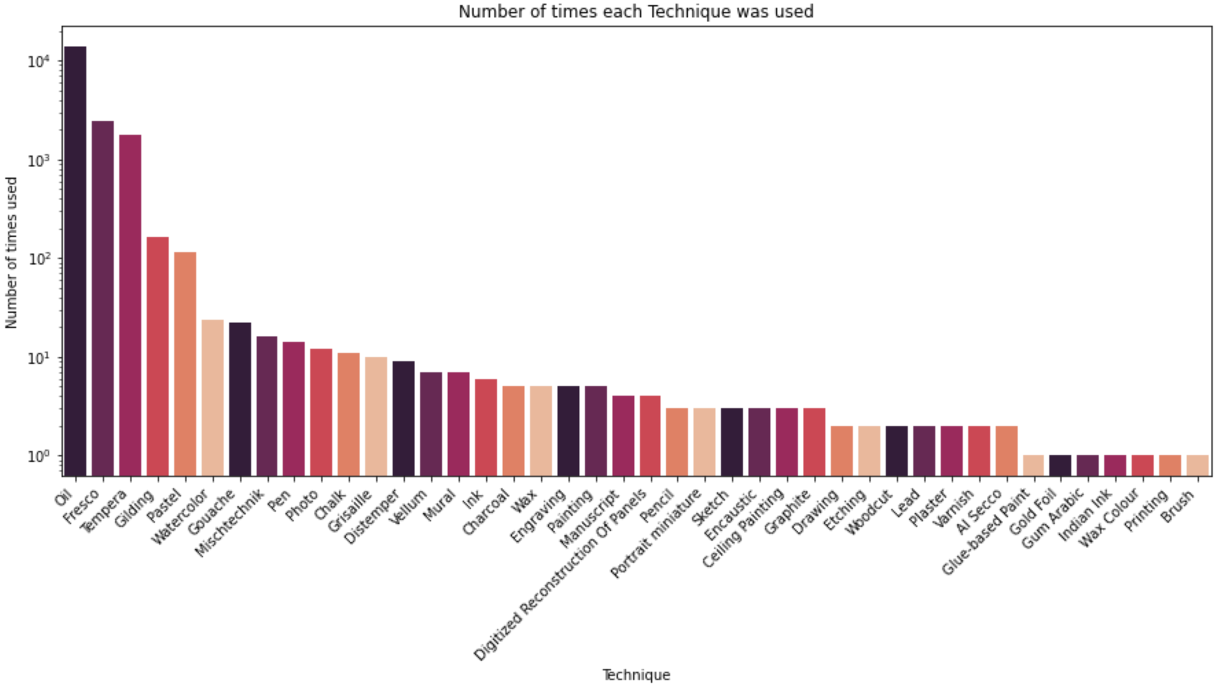
+ id: integer
+ name: string

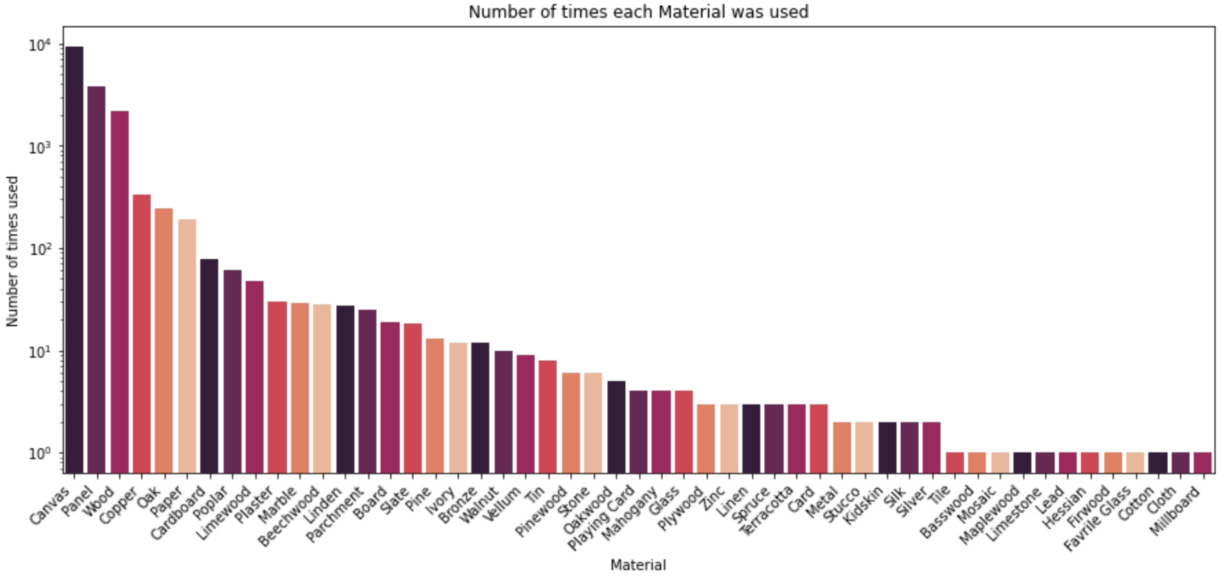
+ paintingId: integer
+ techniqueId: integer
+ order: integer

Data Characterization

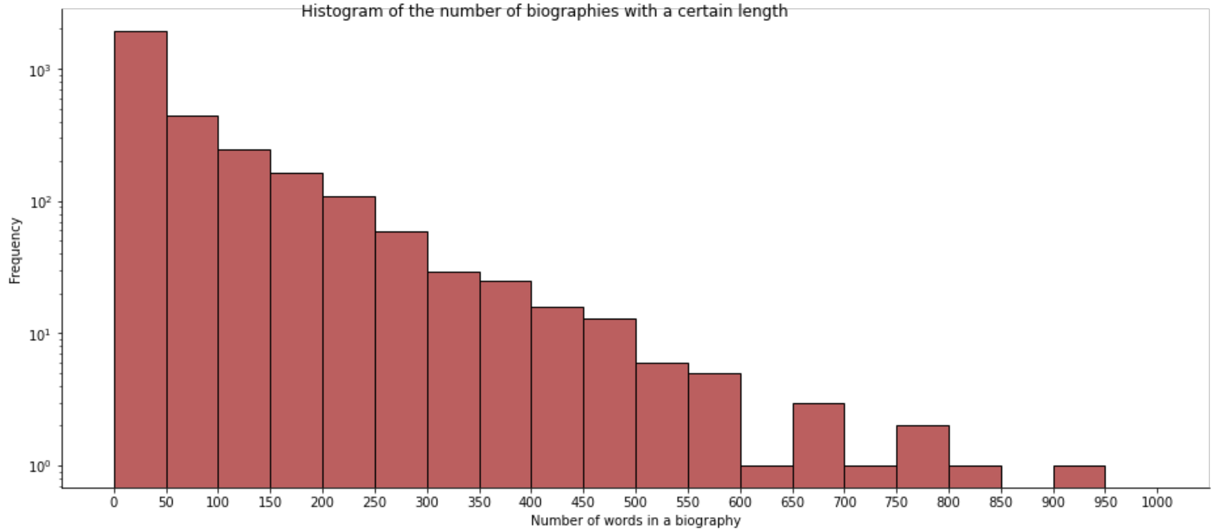


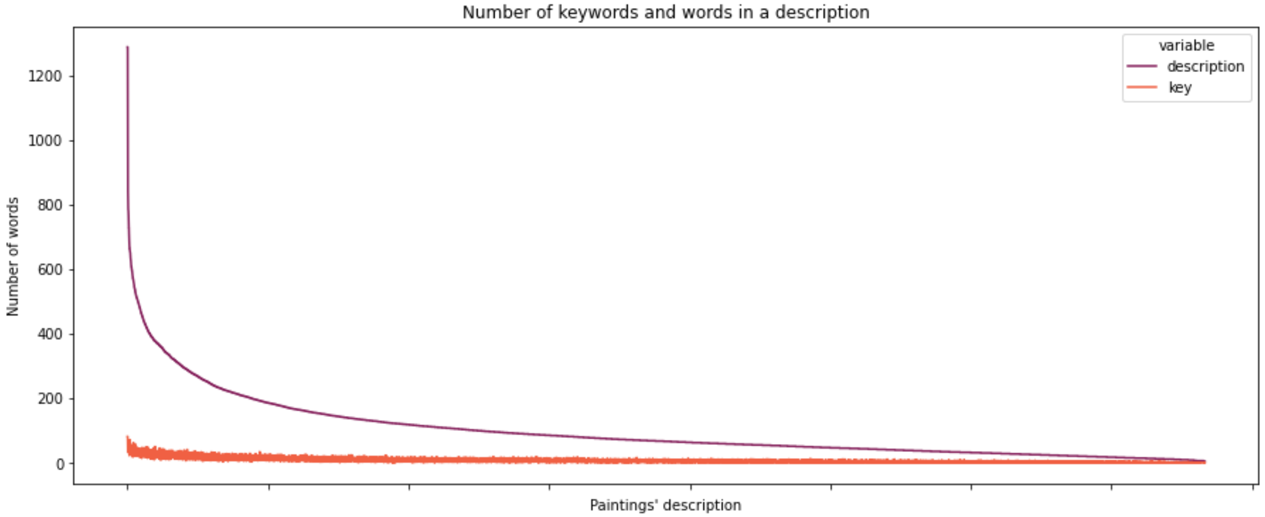


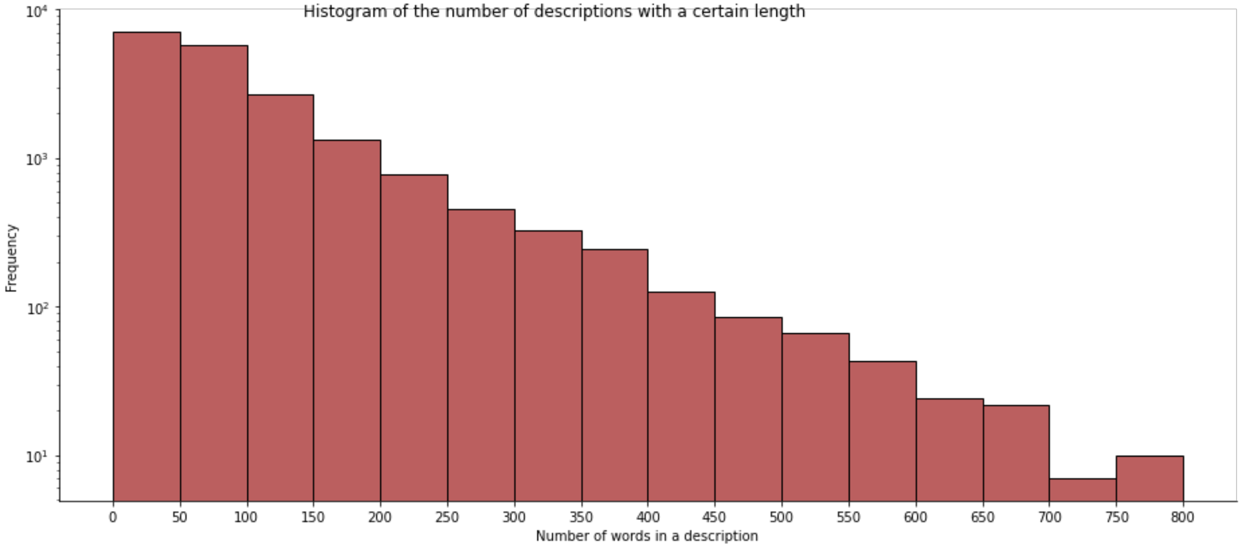


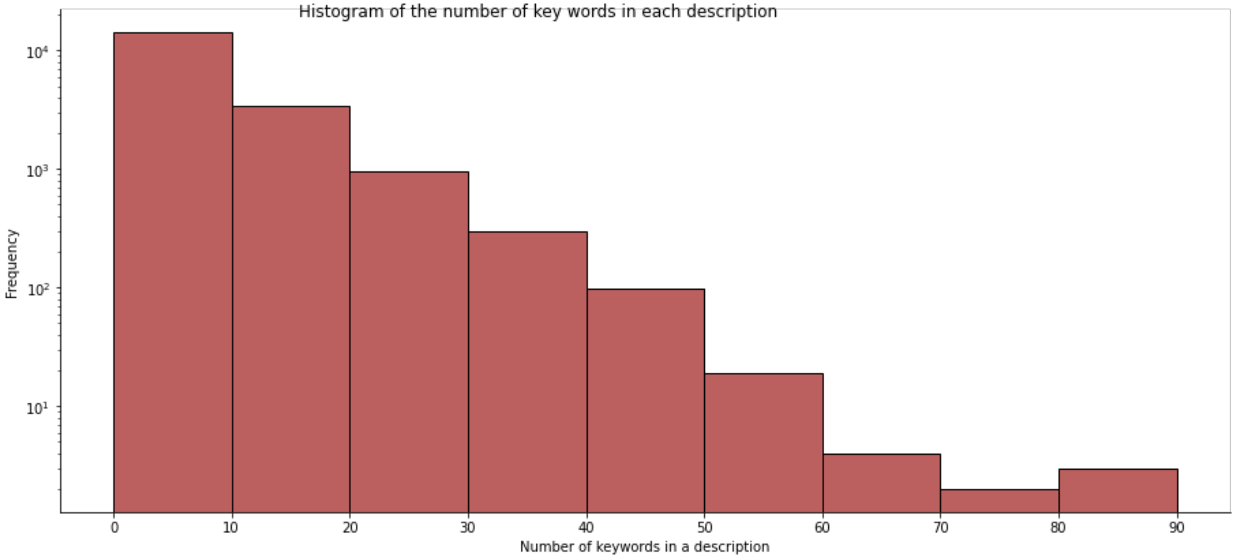


Data Characterization


Histogram of the number of biographies with a certain length


Number of keywords and words in a description


Histogram of the number of descriptions with a certain length


Histogram of the number of key words in each description

STEAM GAMES

Milestone #1 - Data Preparation

DAPI - October 2020

Ângelo Teixeira	up201606516
Duarte Frazão	up201605658
Mariana Aguiar	up201605904
Pedro Costa	up201605339

DATASETS

1. GAMES

- .csv format
- 50 MB
- 27k rows

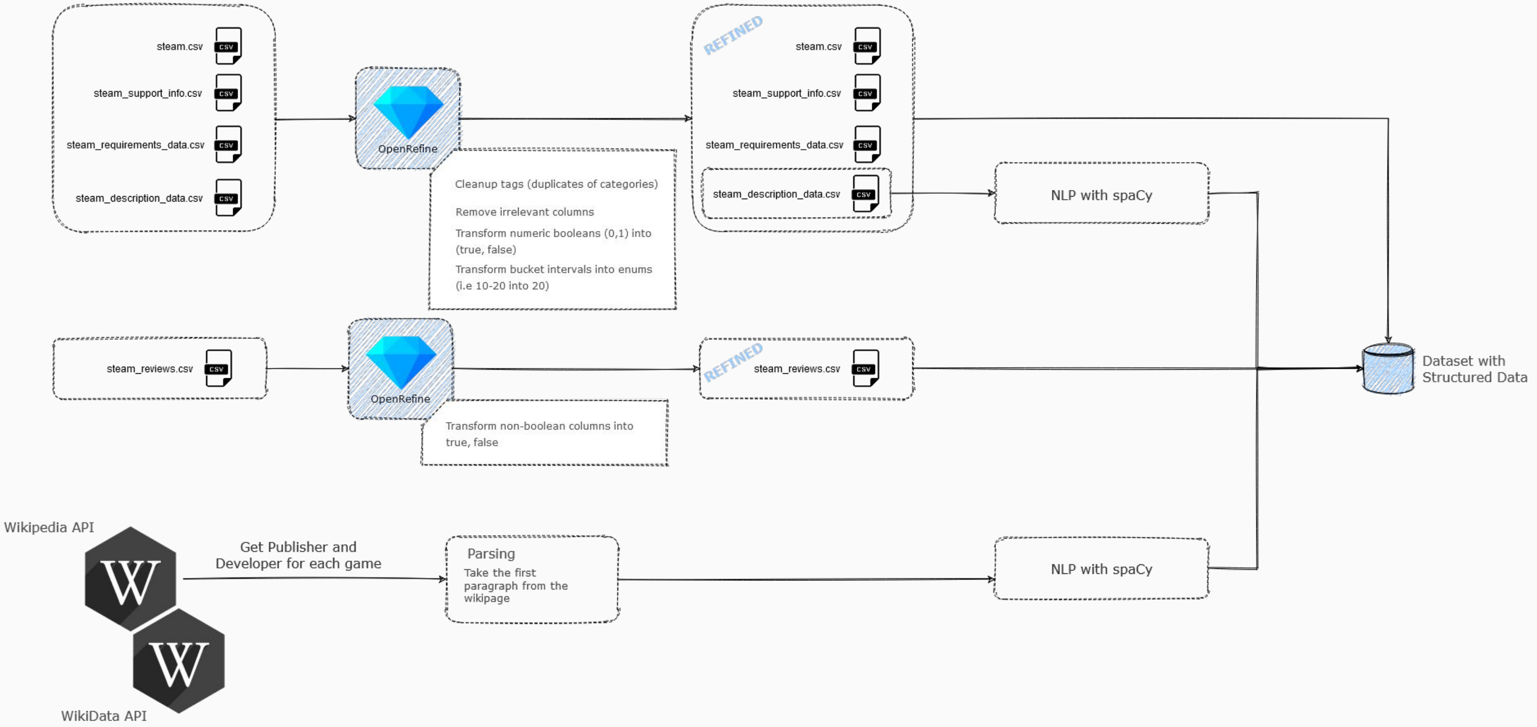
- **Steam** website
- Pre-collected data from **Kaggle**
- Sample of the whole domain

2. REVIEWS

- .csv format
- 2 GB
- 6.4M rows

- **Steam** website
- Pre-collected data from **Kaggle**
- Sample of the whole domain

REFINEMENT AND ENRICHMENT



steam.csv

steam_support_info.csv

steam_requirements_data.csv

steam_description_data.csv

OpenRefine

Cleanup tags (duplicates of categories)

Remove irrelevant columns

Transform numeric booleans (0,1) into (true, false)

Transform bucket intervals into enums (i.e 10-20 into 20)

REFINED

steam.csv

steam_support_info.csv

steam_requirements_data.csv

steam_description_data.csv

NLP with spaCy

steam_reviews.csv

OpenRefine

Transform non-boolean columns into true, false

REFINED

steam_reviews.csv

Dataset with Structured Data

Wikipedia API

WikiData API

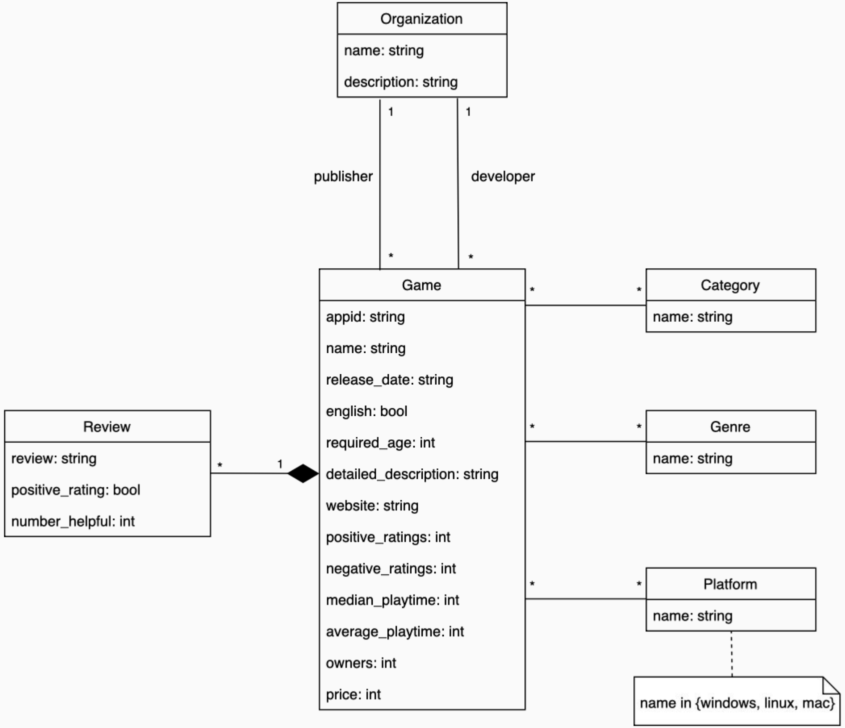
Get Publisher and Developer for each game

Parsing

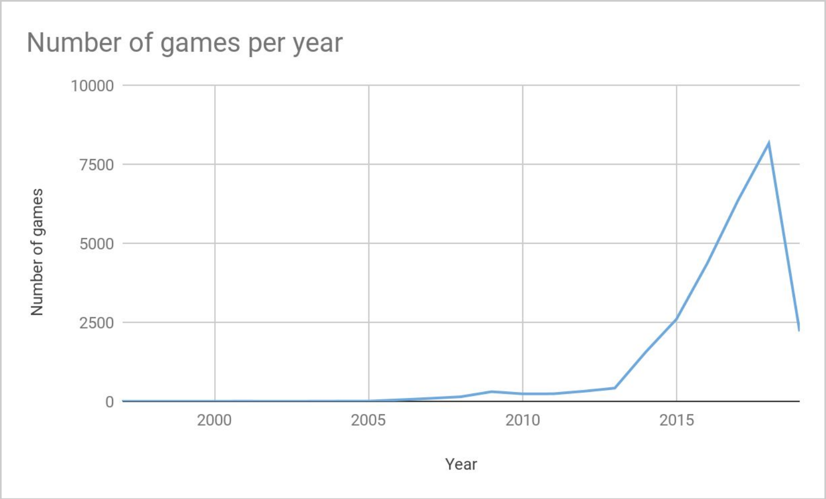
Take the first paragraph from the wikipage

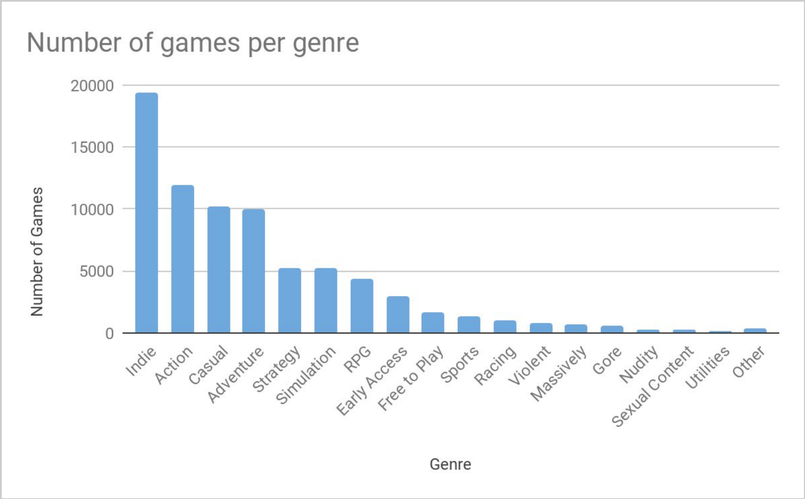
NLP with spaCy

CONCEPTUAL MODEL



DATA CHARACTERIZATION - DOCUMENTS



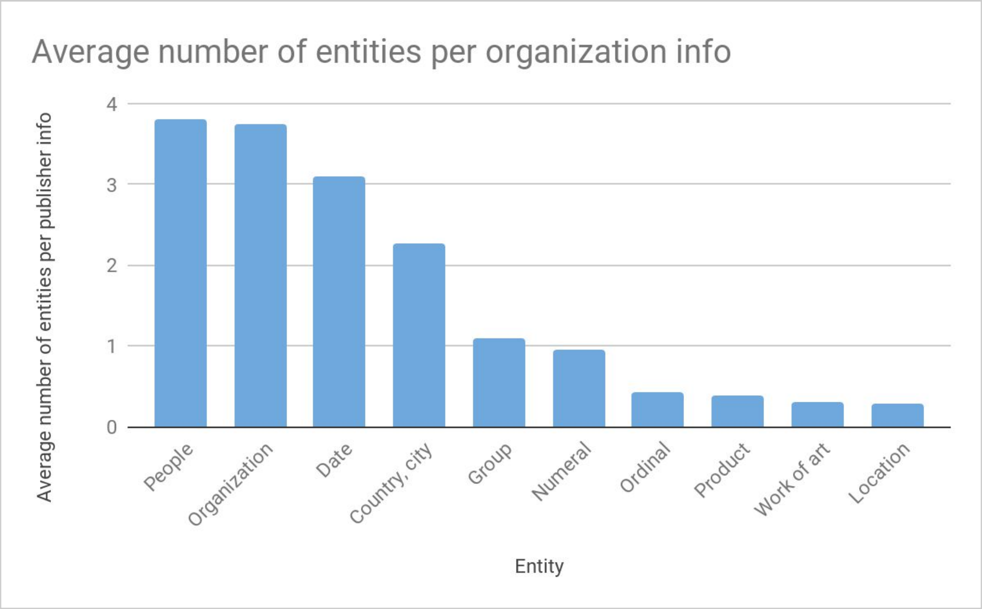


DATA CHARACTERIZATION - TEXT

		LENGTH				
	Total number	Average	p25	p50	p75	p95
Organization descriptions	1011	729	308	526	983	1975
Reviews	6417105	304	30	104	310	1237
Game descriptions	27334	1634	837	1303	2026	3907

- Games with organizations without descriptions: **85%**
- Games with reviews: **8980**

DATA CHARACTERIZATION



POSSIBLE SEARCH QUERIES

- Search for Steam Games
- Search for games’ publishers and developers
- Search for Steam Games Categories and Genres
- Search for games’ reviews
- Search games from a specific organization (publisher/developer)
- Top reviews of a game
- Related games (via tags/category/same organizations/etc)

Goodreads Books and Reviews

DAPI 2020/21 - Group 3

Bruno Sousa - up201604145
Filipa Durão - up201606640
Miguel Duarte - up201606298
Rui Alves - up201606746

Datasets

Books dataset:

- Retrieved from the goodbooks-10k GitHub repository (in *CSV* format)
- Built by scrapping Goodreads pages.

Reviews dataset:

- Retrieved from the USCD Book Graph website (in *JSON* format)
- Built by scrapping Goodreads pages.

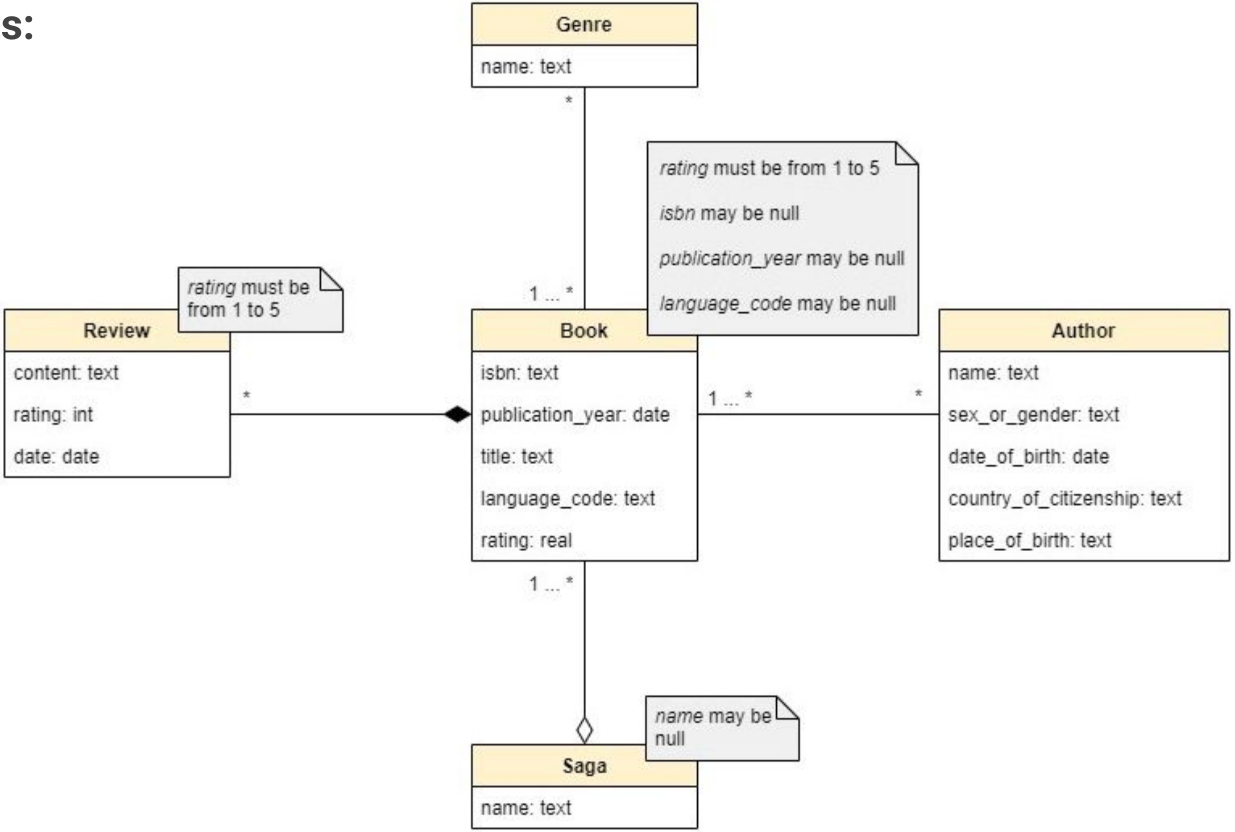
Authors dataset:

- Built by the team using the WikiData website
- Populated by performing an API request for each author in the **Books dataset**, using the wptools CLI tool
- Each input author entry is in *JSON* format

Domain Conceptual Model

Retrieval Units:

- Book
- Review
- Author



Dataset Preparation - Books

Original Data

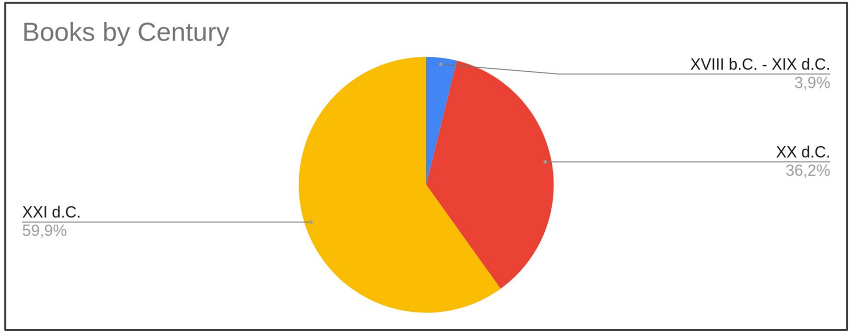
- **10,000** book entries
- *CSV* format

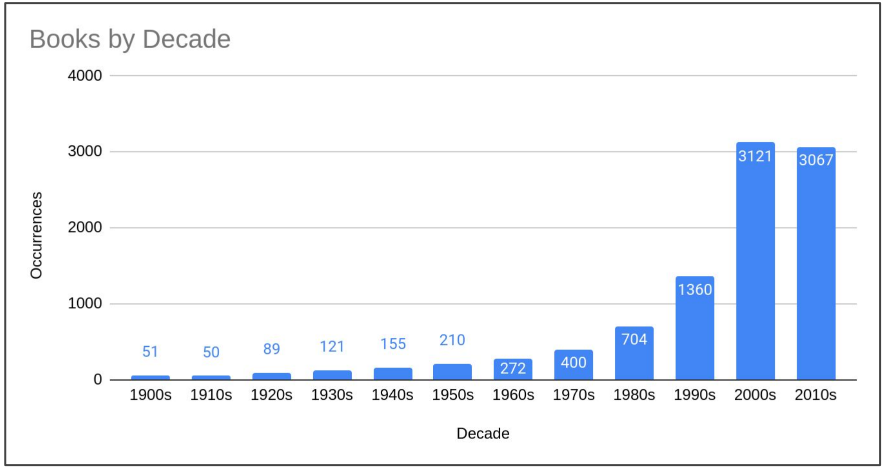
Preparation operations

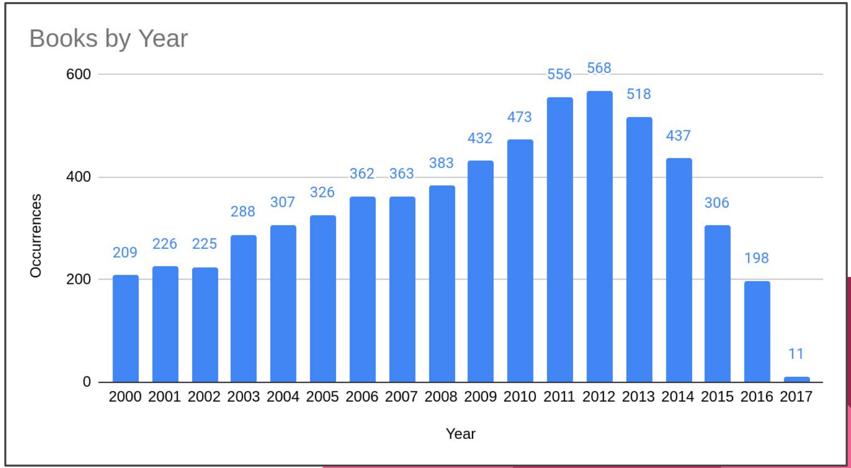
- Removing duplicate entries
- Null/empty fields normalization
- Whitespace trimming
- Attributes selection

Dataset Characterization - Books (timeline)

- About **96%** of books are from the past two centuries
- The majority of books are quite recent (from the past two decades)

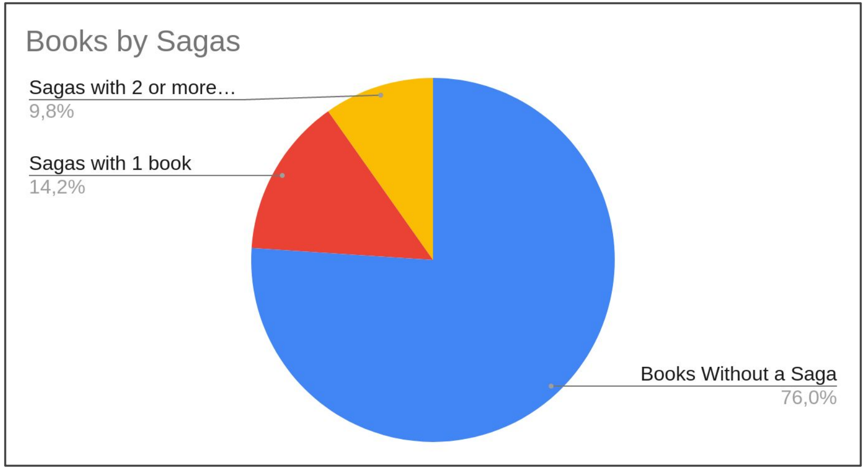


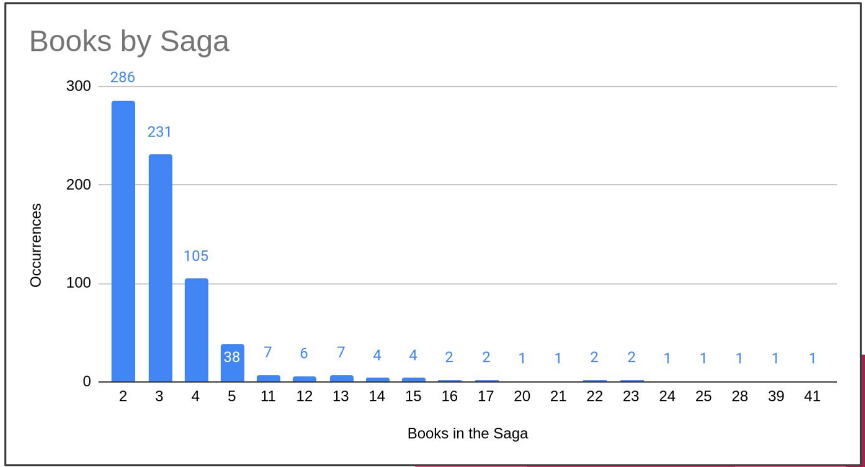




Dataset Characterization - Books (sagas)

- About three quarters of Books do not belong to a Saga
- **14%** of Books belong to a Saga that has only 1 book
- The remaining **10%** of Books belong to larger sagas with 2 or more books





Dataset Preparation - Reviews

Original Data

- More than **15 million** review entries
- Separated in **8 different categories** (Romance, Fantasy, …)
- *JSON* format

Preparation operations

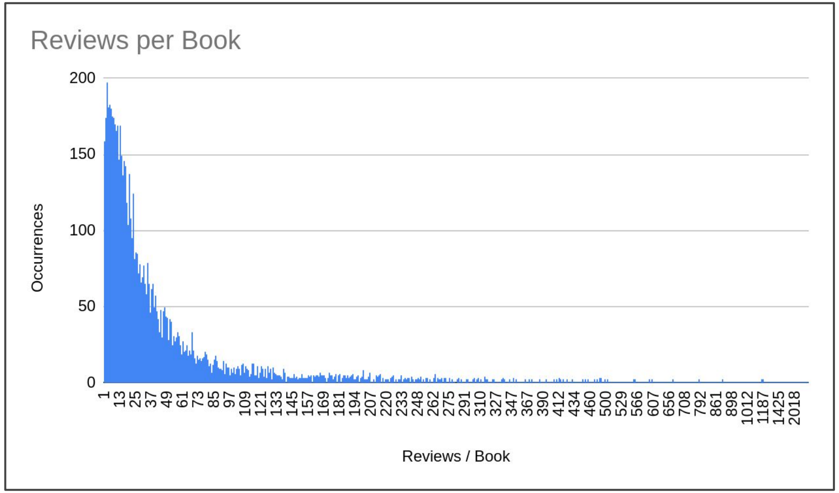
- Filtering by presence in the Books dataset
- Filtering reviews from 2016 to the present
- Null/empty fields normalization and whitespace trimming
- Attributes selection

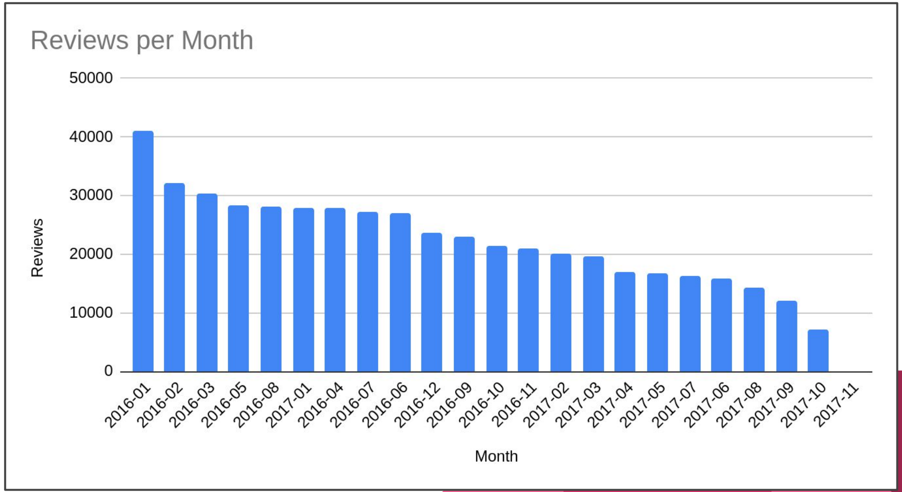
Results

- **500,000** review entries
- **70%** of books have at least one review (dataset intersection)

Dataset Characterization - Reviews (timeline)

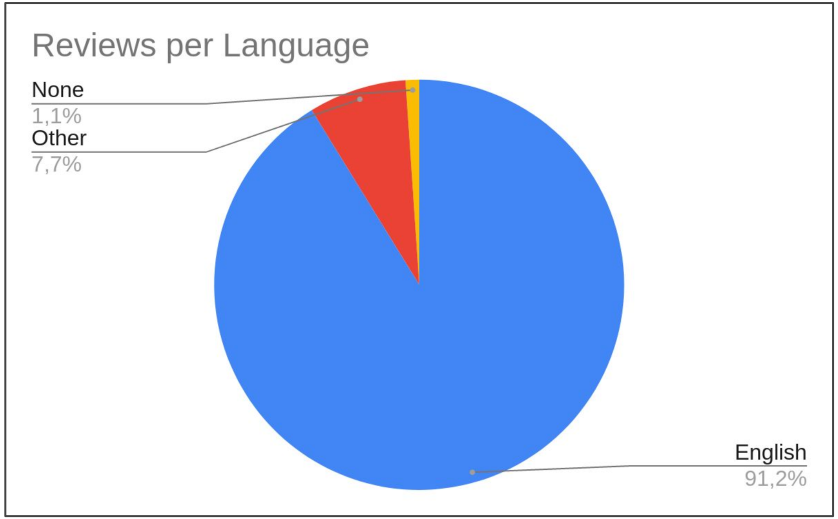
- The vast majority of books have less than 50 reviews
- The reviews time distribution from January 2016 to the present is somewhat linear, although the number of reviews is slightly descending with time

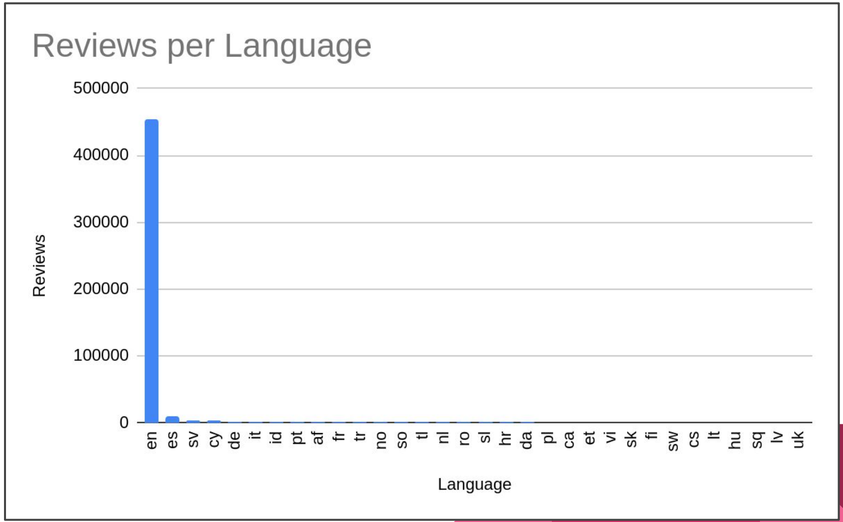

Reviews per Book


Reviews per Month

Dataset Characterization - Reviews (language)

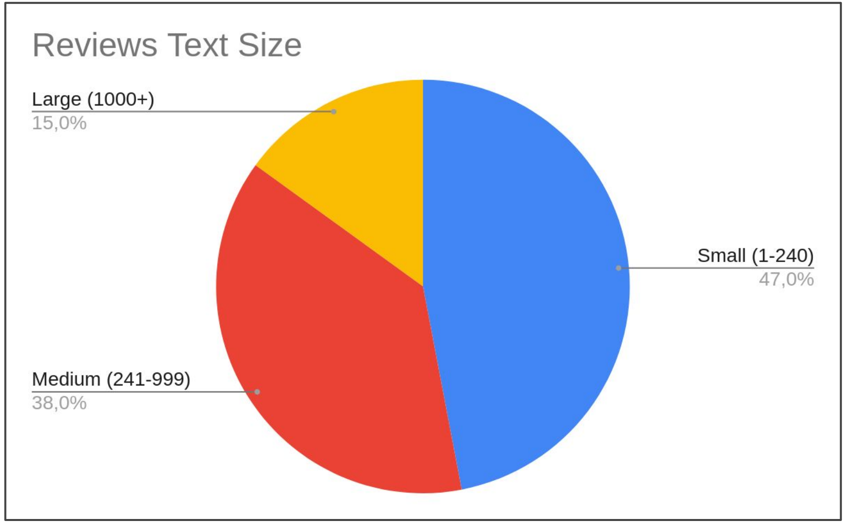
- The vast majority of reviews are written in **English**
- Only **1%** of reviews were in an unclassifiable language

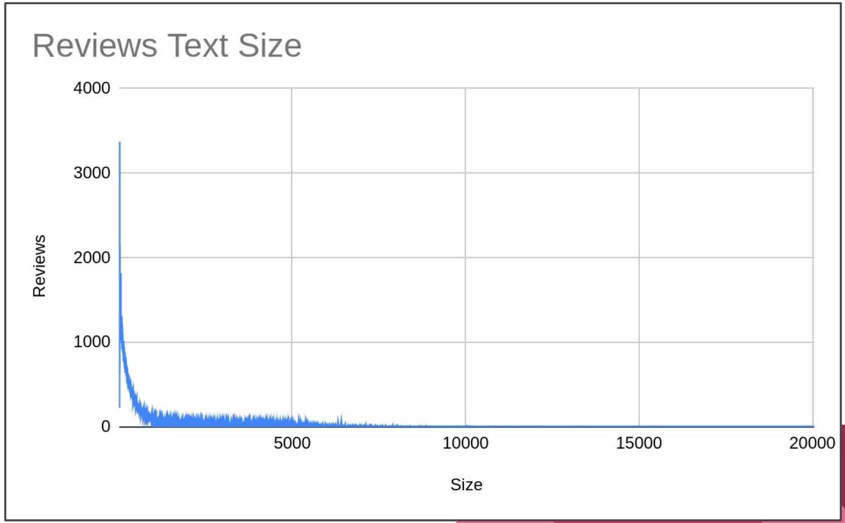




Dataset Characterization - Reviews (content size)

- About half of the reviews are quite **small** (using Twitter's post size for reference)
- About 15% of the reviews are quite extensive, featuring a lot of content





Dataset Preparation - Authors

Building the raw Dataset

- Extracted list of unique authors from Books dataset
- For each unique author, get the author entry from **WikiData** API

Preparation operations

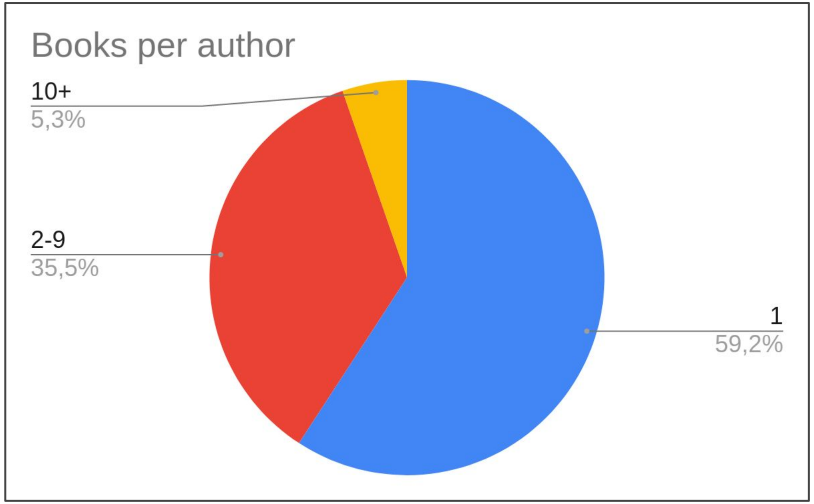
- Null/empty fields normalization
- Whitespace trimming
- Attributes selection

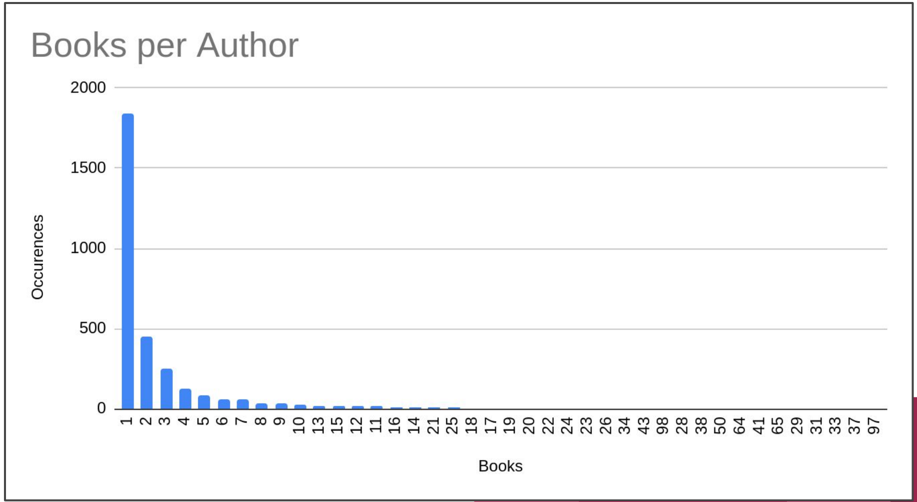
Results

- **3,100** author entries
- **76%** of books have information regarding their authors (intersection percentage)

Dataset Characterization - Authors

- The majority of authors have written only **one book**
- More than **40%** of authors have written multiple books





Retrieval Tasks

- Search for books rated over **R**, filtered by genre **G**
- Search for books that were co-authored by authors **A_1** and **A_2**
- Search for reviews of the most well rated book in the saga **S**
- Search for reviews between dates **D_1** and **D_2** of books that were authored by **A**
- Search for medium-sized reviews in books written by authored **A** that are not from genre **G**
- Search for authors that published over **N** books, filtered by their country of citizenship **C**
- Search for authors who have written at least **N** books rated over **R**

Information Description, Storage and Retrieval

Popular movies and streaming

Milestone 1 - Group 4

Carlos Gomes - up201603404
Eduardo Silva - up201603135
Joana Silva - up201208979
Joana Ramos - up201605017

Obtaining and Preparing the Datasets

kaggle

Streaming Dataset

Structured dataset in .csv format with information regarding the streaming platform in which a movie is available.

kaggle

IMDb Scraped Dataset

Structured dataset with movie information retrieved through scraping of IMDb's website.

IMDb

IMDb Official Dataset

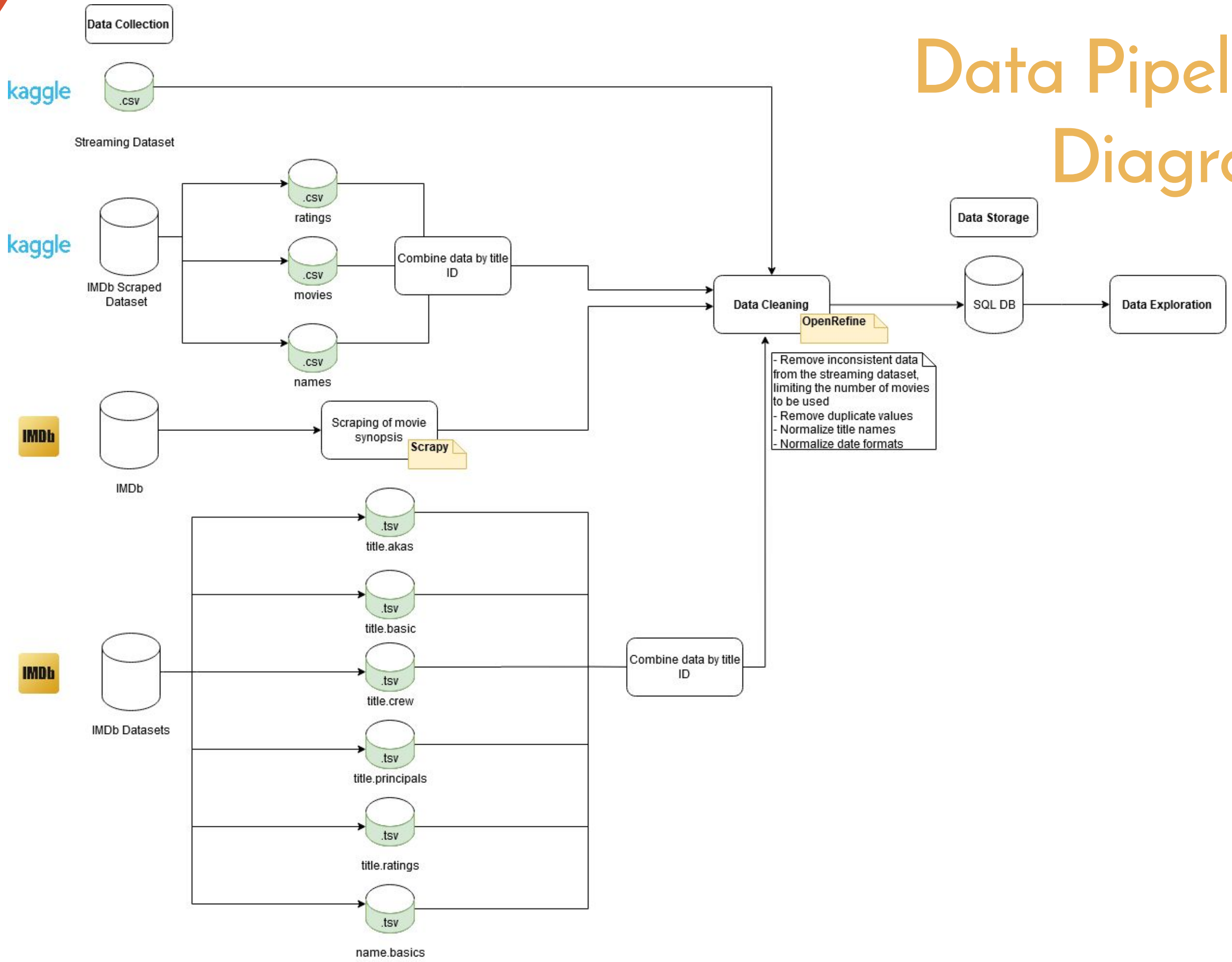
Structured dataset in TVS format with IMDb's website information regarding movies.

IMDb

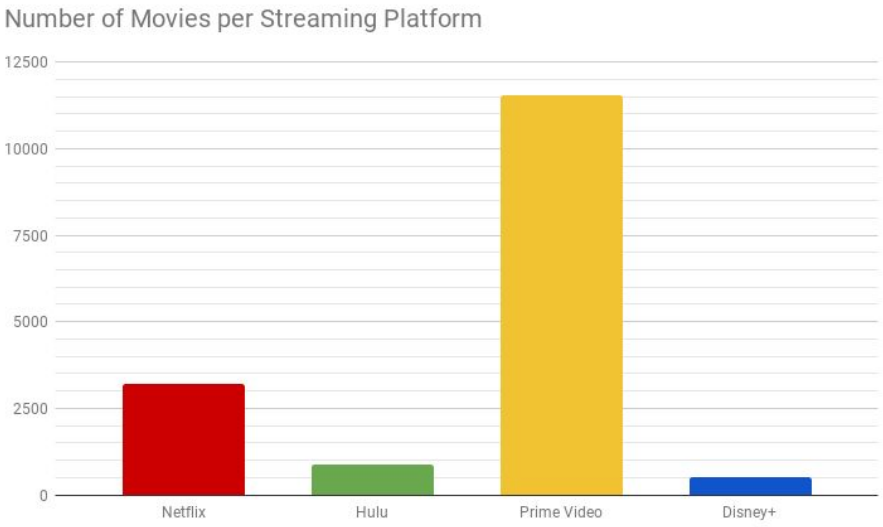
IMDb Dataset

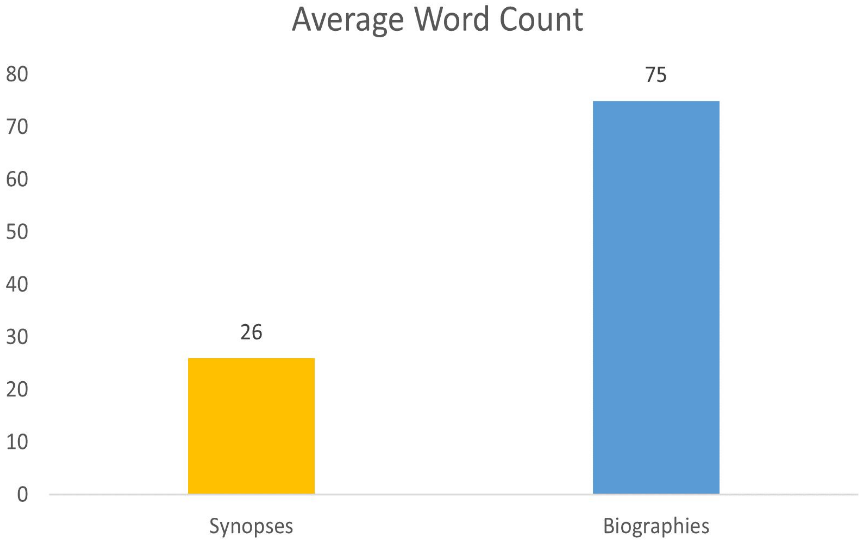
Unstructured data (movie synopsis) obtained by the scraping of IMDB's movie pages.

Data Pipeline Diagram

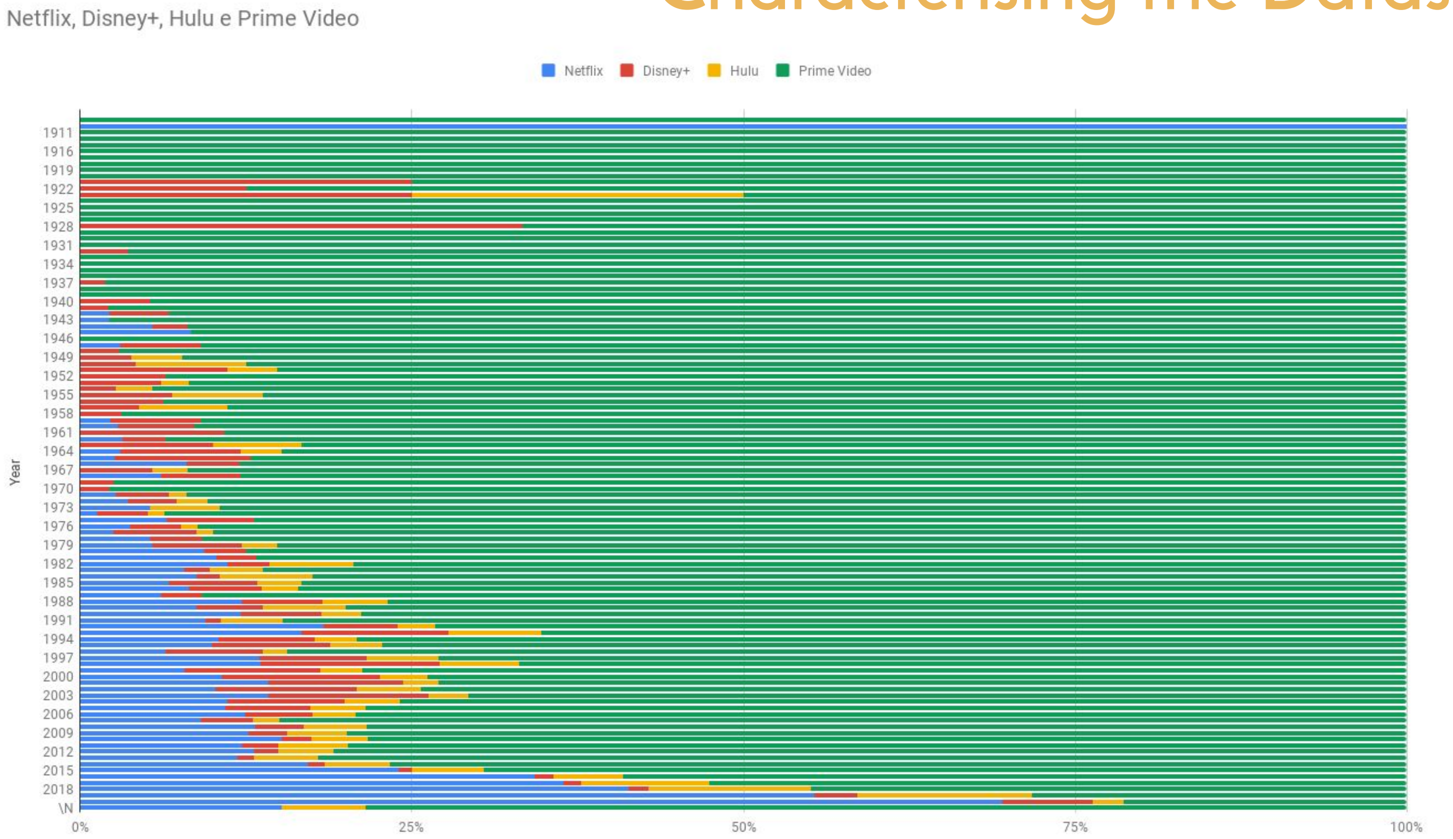


Characterising the Datasets

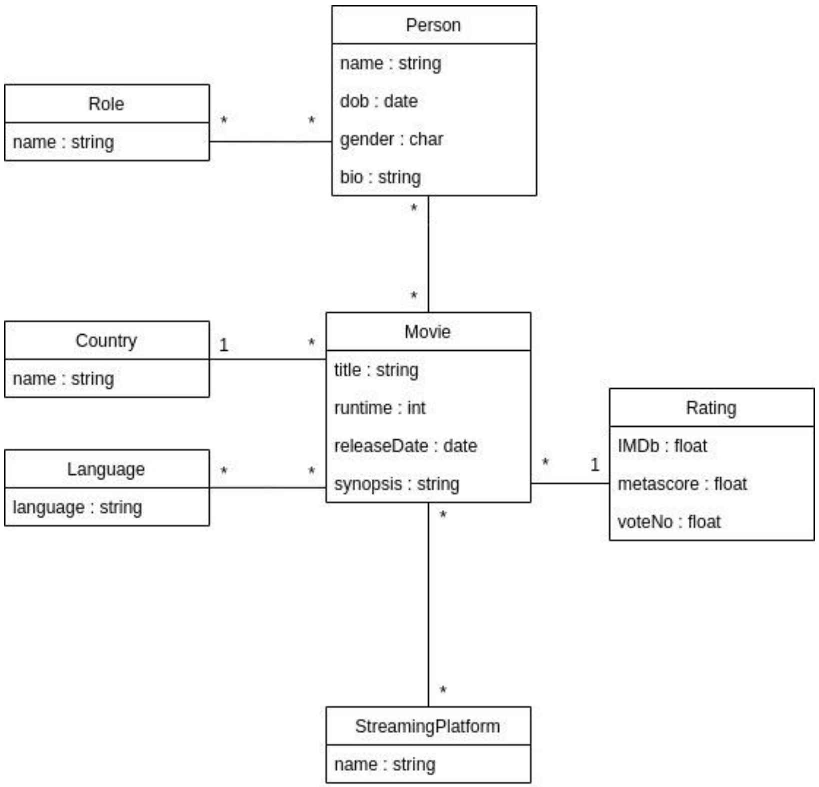




Characterising the Datasets



Conceptual Model of the Domain



The main entities of our domain are movies and people involved in them (actors, directors, writers etc.).

Data Retrieval Tasks

The results of our queries focus on the two main entities: **Person** and **Movie**.

Here are some possible queries:

- Retrieve all movies in which an actor appears;
- Retrieve all movies by director;
- Retrieve high rated movies for each genre;
- Retrieve high rated movies for each year;
- Retrieve high rated movies for each language;
- Retrieve movies that fit into a text description (e.g. “second world war movies”);
- Retrieve people that fit into a text description (e.g. “young puerto rican actor”).

European Parliament Data

Dataset Preparation

What is our project about?

705 MEPs

27 Countries

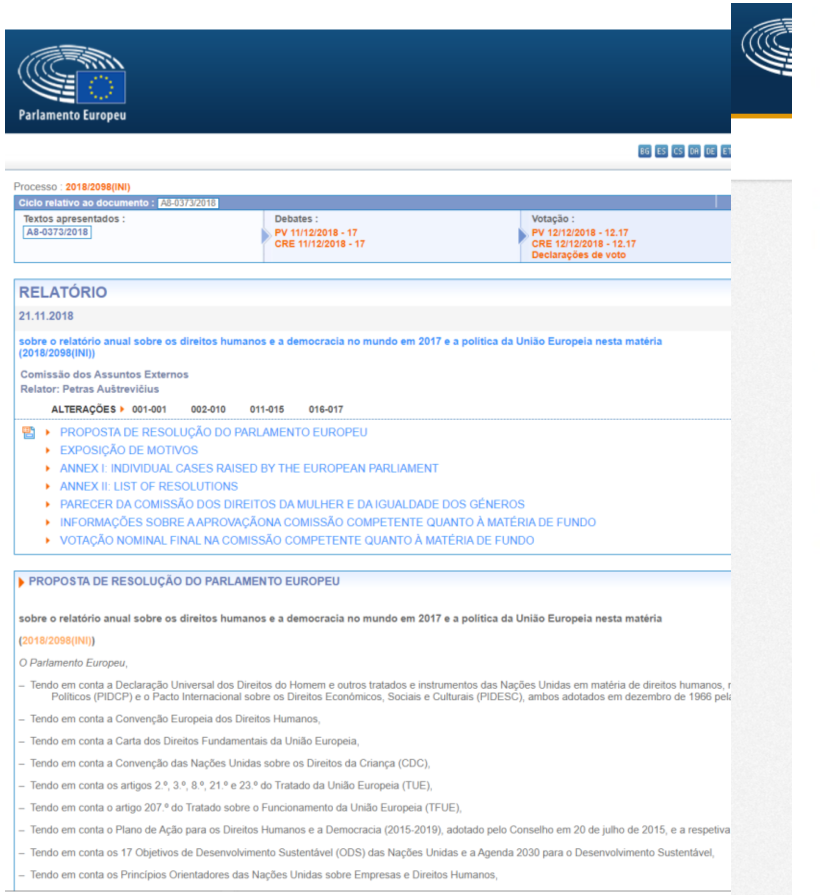
7 Political groups

22 Committees

~200 National parties

6000 Voting sessions per year

Search tasks source



Parlamento Europeu

BG ES CS DA DE ET

Processo : 2018/2098(INI)

Ciclo relativo ao documento : A8-0373/2018

Textos apresentados :	Debates :	Votação :
A8-0373/2018	PV 11/12/2018 - 17 CRE 11/12/2018 - 17	PV 12/12/2018 - 12.17 CRE 12/12/2018 - 12.17 Declarações de voto

RELATÓRIO

21.11.2018

sobre o relatório anual sobre os direitos humanos e a democracia no mundo em 2017 e a política da União Europeia nesta matéria (2018/2098(INI))

Comissão dos Assuntos Externos

Relator: Petras Auštrevičius

ALTERAÇÕES ▸ 001-001 002-010 011-015 016-017

- PROPOSTA DE RESOLUÇÃO DO PARLAMENTO EUROPEU
- EXPOSIÇÃO DE MOTIVOS
- ANNEX I: INDIVIDUAL CASES RAISED BY THE EUROPEAN PARLIAMENT
- ANNEX II: LIST OF RESOLUTIONS
- PARECER DA COMISSÃO DOS DIREITOS DA MULHER E DA IGUALDADE DOS GÉNEROS
- INFORMAÇÕES SOBRE A APROVAÇÃONA COMISSÃO COMPETENTE QUANTO À MATÉRIA DE FUNDO
- VOTAÇÃO NOMINAL FINAL NA COMISSÃO COMPETENTE QUANTO À MATÉRIA DE FUNDO

PROPOSTA DE RESOLUÇÃO DO PARLAMENTO EUROPEU

sobre o relatório anual sobre os direitos humanos e a democracia no mundo em 2017 e a política da União Europeia nesta matéria

(2018/2098(INI))

O Parlamento Europeu,

– Tendo em conta a Declaração Universal dos Direitos do Homem e outros tratados e instrumentos das Nações Unidas em matéria de direitos humanos, r
Políticos (PIDCP) e o Pacto Internacional sobre os Direitos Económicos, Sociais e Culturais (PIDESC), ambos adotados em dezembro de 1966 pela

– Tendo em conta a Convenção Europeia dos Direitos Humanos,

– Tendo em conta a Carta dos Direitos Fundamentais da União Europeia,

– Tendo em conta a Convenção das Nações Unidas sobre os Direitos da Criança (CDC),

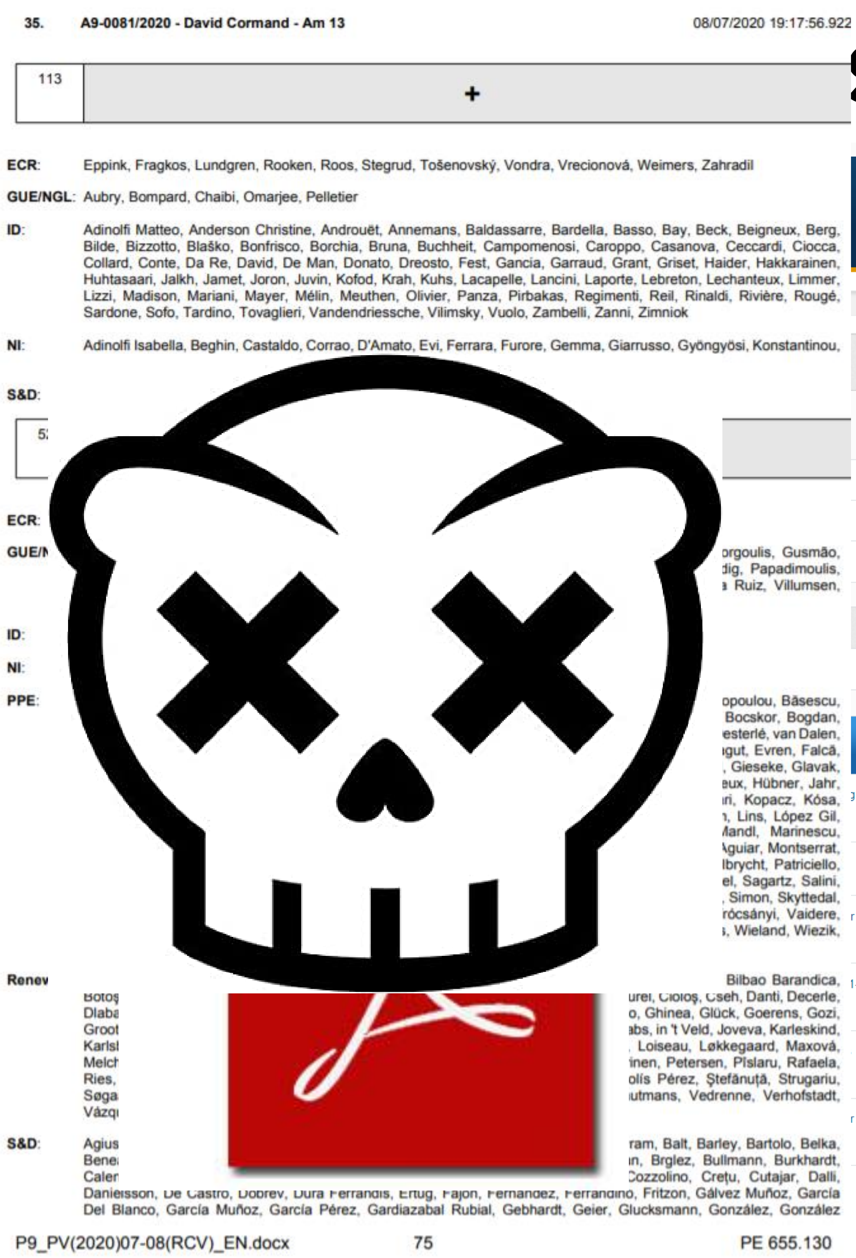
– Tendo em conta os artigos 2.º, 3.º, 8.º, 21.º e 23.º do Tratado da União Europeia (TUE),

– Tendo em conta o artigo 207.º do Tratado sobre o Funcionamento da União Europeia (TFUE),

– Tendo em conta o Plano de Ação para os Direitos Humanos e a Democracia (2015-2019), adotado pelo Conselho em 20 de julho de 2015, e a respetiva

– Tendo em conta os 17 Objetivos de Desenvolvimento Sustentável (ODS) das Nações Unidas e a Agenda 2030 para o Desenvolvimento Sustentável,

– Tendo em conta os Princípios Orientadores das Nações Unidas sobre Empresas e Direitos Humanos,



35. A9-0081/2020 - David Cormand - Am 13 08/07/2020 19:17:56.922

113	+

ECR: Eppink, Fragkos, Lundgren, Rooken, Roos, Stegrud, Tošenovský, Vondra, Vrecionová, Weimers, Zahradil

GUE/NGL: Aubry, Bompard, Chaibi, Omarjee, Pelletier

ID: Adinolfi Matteo, Anderson Christine, Androuët, Annemans, Baldassarre, Bardella, Basso, Bay, Beck, Beigneux, Berg, Bilde, Bizzotto, Blaško, Bonfrisco, Borchia, Bruna, Buchheit, Campomenosi, Caroppo, Casanova, Ceccardi, Ciocca, Collard, Conte, Da Re, David, De Man, Donato, Dreosto, Fest, Gancia, Garraud, Grant, Griset, Haider, Hakkarainen, Huhtasaari, Jalkh, Jamet, Joron, Juvin, Kofod, Krah, Kuhs, Lacapelle, Lancini, Laporte, Lebreton, Lechanteux, Limmer, Lizzi, Madison, Mariani, Mayer, Mélin, Meuthen, Olivier, Panza, Pirbakas, Regimenti, Reil, Rinaldi, Rivière, Rougé, Sardone, Sofo, Tardino, Tovaglieri, Vandendriessche, Vilimsky, Vuolo, Zambelli, Zanni, Zimniok

NI: Adinolfi Isabella, Beghin, Castaldo, Corrao, D'Amato, Evi, Ferrara, Furore, Gemma, Giarrusso, Gyöngyösi, Konstantinou,

S&D:

5

ECR:

GUE/N orgoulis, Gusmão,
dig, Papadimoulis,
a Ruiz, Villumsen,

ID:

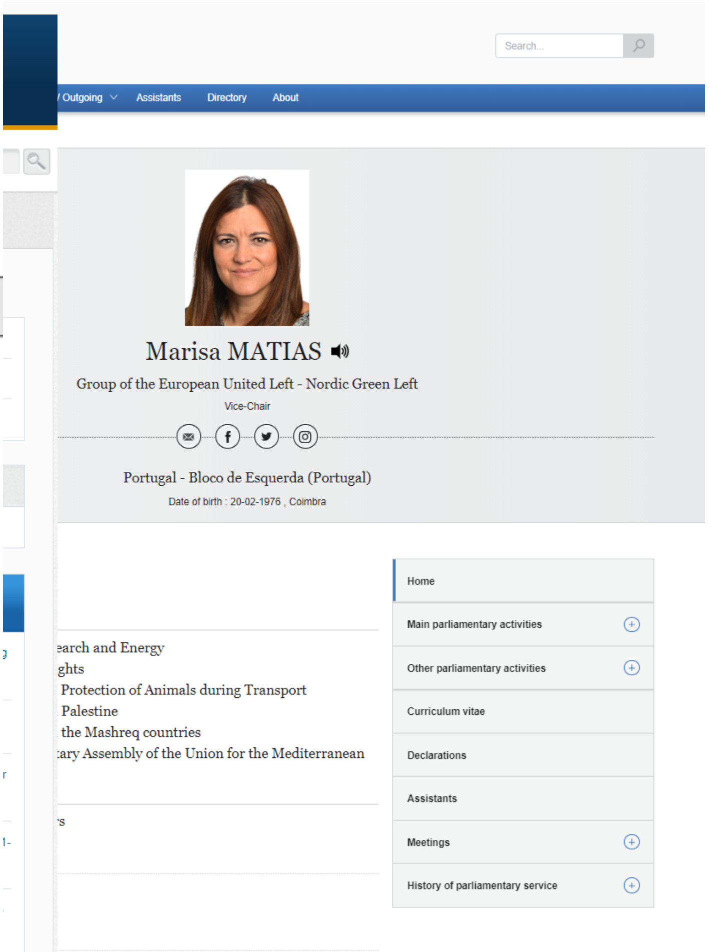
NI:

PPE: opoulou, Băsescu,
Bocskor, Bogdan,
esterlé, van Dalen,
gut, Evren, Falcă,
, Gieseke, Glavak,
eux, Hübner, Jahr,
ri, Kopacz, Kósa,
, Lins, López Gil,
Mandl, Marinescu,
Aguiar, Montserrat,
lbrycht, Patriciello,
el, Sagartz, Salini,
, Simon, Skyttedal,
rócsányi, Vaidere,
s, Wieland, Wiezik,

Renev Bilbao Barandica,
Botoş urel, Cioloş, Cseh, Danti, Decerle,
Dlaba o, Ghinea, Glück, Goerens, Gozi,
Groot abs, in 't Veld, Joveva, Karleskind,
Karlsl , Loiseau, Løkkegaard, Maxová,
Melch inen, Petersen, Pîslaru, Rafaela,
Ries, olís Pérez, Ştefănuță, Strugariu,
Søga utmans, Vedrenne, Verhofstadt,
Vázq

S&D: Agius ram, Balt, Barley, Bartolo, Belka,
Bene n, Brglez, Bullmann, Burkhardt,
Calen Cozzolino, Crețu, Cutajar, Dalli,
Danielsson, De Castro, Dobrev, Dura Ferrandis, Ertug, Fajon, Fernandez, Ferrandino, Fritzon, Gálvez Muñoz, García Del Blanco, García Muñoz, García Pérez, Gardiazabal Rubial, Gebhardt, Geier, Glucksmann, González, González

P9_PV(2020)07-08(RCV)_EN.docx 75 PE 655.130



Search...

/ Outgoing ⌄ Assistants Directory About

Marisa MATIAS 🔊

Group of the European United Left - Nordic Green Left

Vice-Chair

✉ f 🐦 ◎

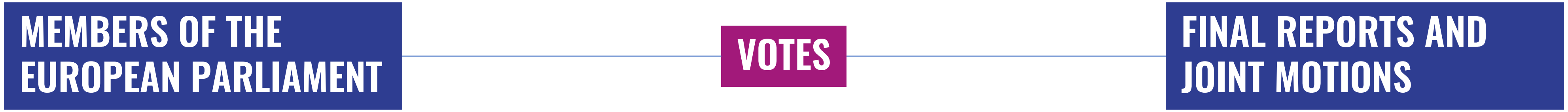
Portugal - Bloco de Esquerda (Portugal)

Date of birth : 20-02-1976 , Coimbra

earch and Energy
ghts
Protection of Animals during Transport
Palestine
the Mashreq countries
tary Assembly of the Union for the Mediterranean
's

- Home
- Main parliamentary activities ⊕
- Other parliamentary activities ⊕
- Curriculum vitae
- Declarations
- Assistants
- Meetings ⊕
- History of parliamentary service ⊕

The dataset



MEMBERS OF THE EUROPEAN PARLIAMENT

VOTES

FINAL REPORTS AND JOINT MOTIONS

- Name
- Country
- Age
- Political Group(s)
- National Party
- Committees
- Social Media

+ In favor
- Against
0 Abstention

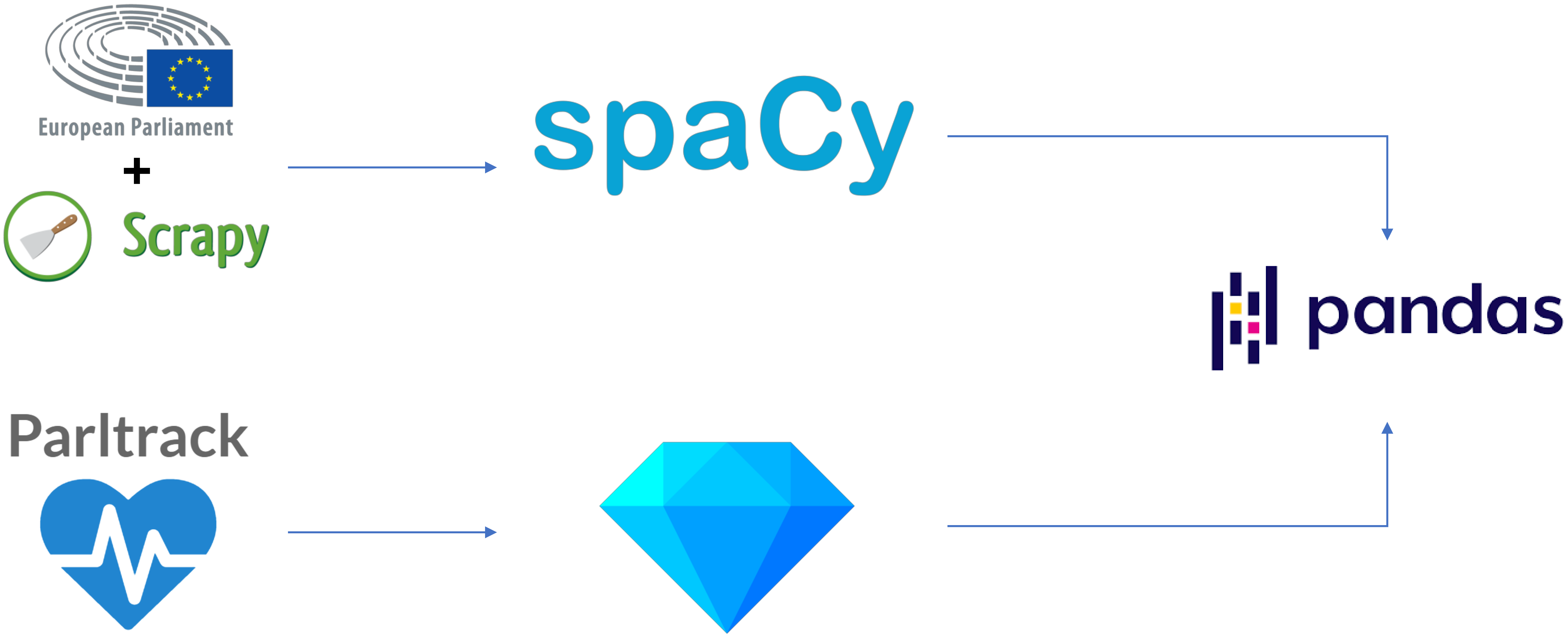
- Title
- (A lot of) Text
- Rapporteur(s)
- (Committee)

Data sources

European Parliament

Parltrack

Data preparation



Possible search tasks

- Search for a report
- Search for a MEP
- Search for a committee
- Get the votes casted by a MEP
- Get the votes on a specific report
- Get reports of a committee

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

2020/2021

Diseases, Symptoms and Treatments

Information Description, Storage and Retrieval

Group 6:

- André Esteves - up201606673
- Francisco Filipe - up201604601
- Helena Montenegro - up201604184
- Juliana Marques - up201605568

Introduction

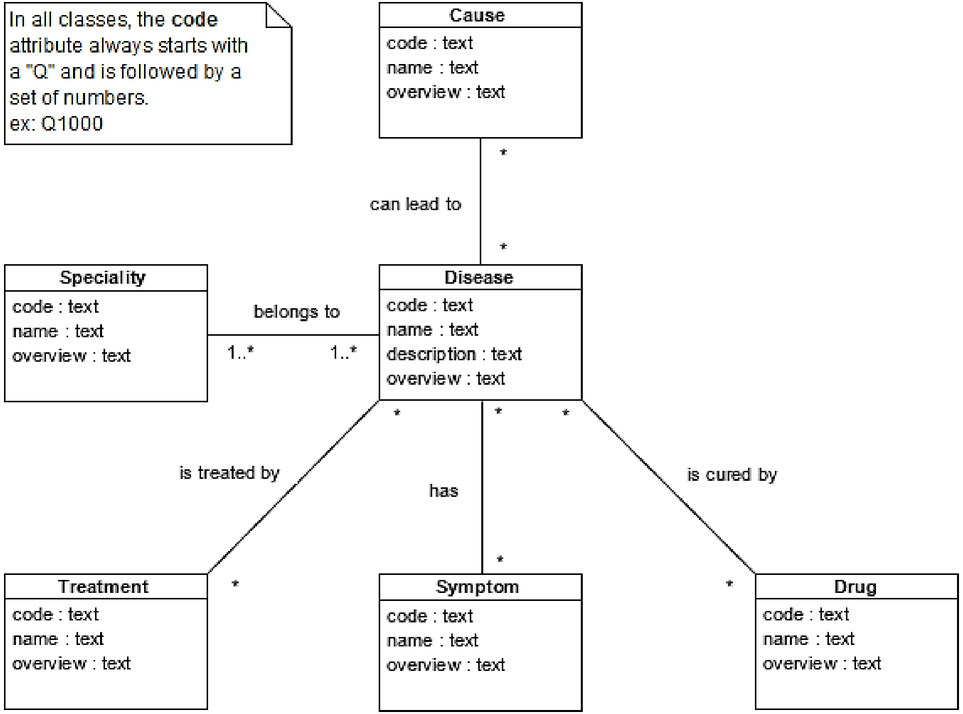
Problem:

- Documents shown by current search mechanisms may not be reliable.
- Misleading and exaggerated information may lead to panic.
- Lack of search mechanism focused on health matters.

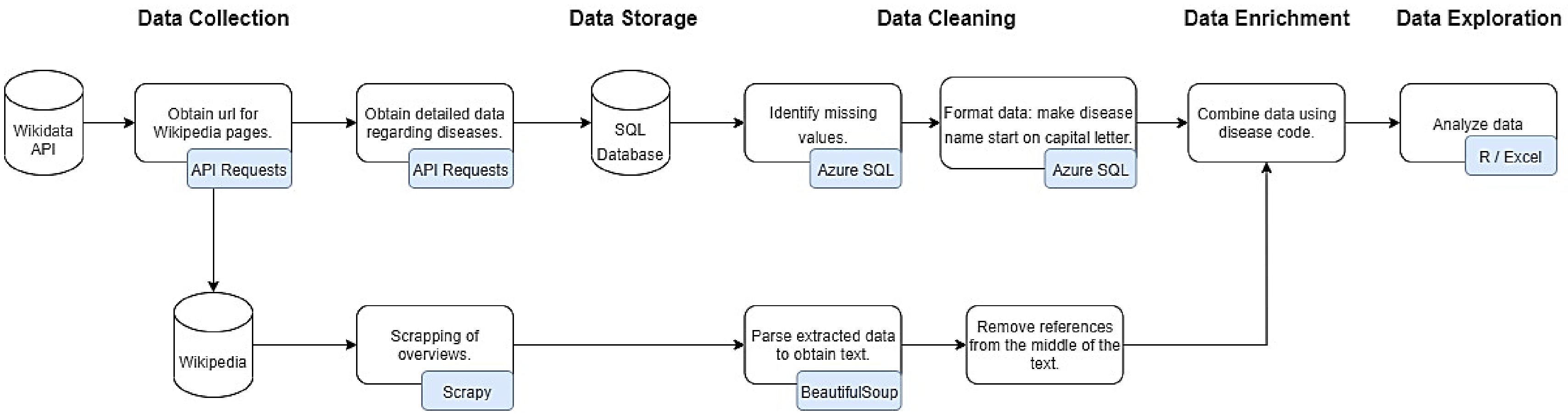
Goal of the project:

- Develop search mechanism focused on diseases, treatments and symptoms.

Conceptual Model



Data Pipeline


Data Collection
Data Storage
Data Cleaning
Data Enrichment
Data Exploration
Wikidata API
Obtain url for Wikipedia pages.
API Requests
Obtain detailed data regarding diseases.
API Requests
SQL Database
Identify missing values.
Azure SQL
Format data: make disease name start on capital letter.
Azure SQL
Combine data using disease code.
Analyze data
R / Excel
Wikipedia
Scrapping of overviews.
Scrapy
Parse extracted data to obtain text.
BeautifulSoup
Remove references from the middle of the text.

Data Collection

Wikidata

Structural data obtained through API requests.

The information is:

- Unreliable.
- Incomplete.

License:

- *Creative Commons Public Domain Dedication 1.0*
- Free to modify and share, even for commercial purposes.

Wikipedia

Textual data obtained through scraping (with Scrapy).

The information is:

- Verifiable against authoritative sources.

License:

- *Creative Commons Attribution-ShareAlike 3.0 Unported*
- Free to modify and share, even for commercial purposes, as long as credit is given.

Data Storage

Azure SQL Database

- **Language:** SQL Server
- **Tools:** Azure Data Studio

Data Enrichment

- UPDATE statements to add the overviews scraped from Wikipedia to the database.

Data Cleaning

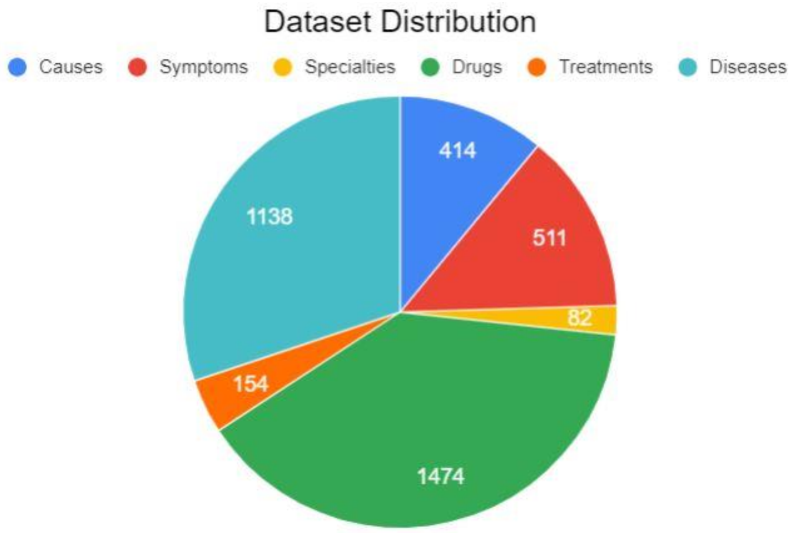
On the structural data:

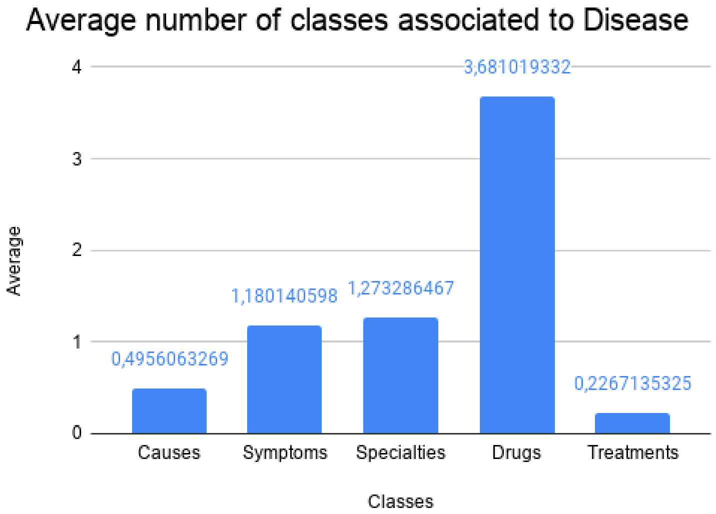
- Remove diseases that had less than 2 connections to other classes.
- Remove all symptoms, treatments, drugs, causes and specialties that did not have a connection to a disease.

On the textual data:

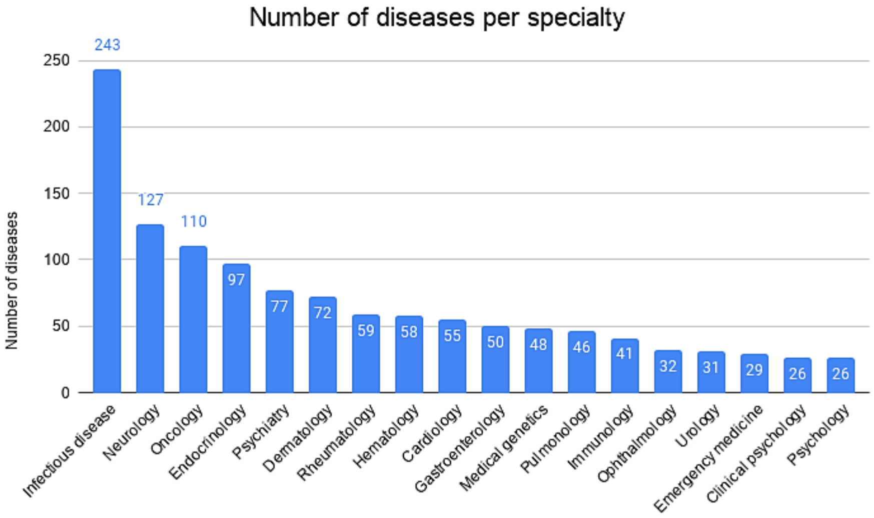
- Extract text using **BeautifulSoup**.
- Remove special characters.
- Remove references.

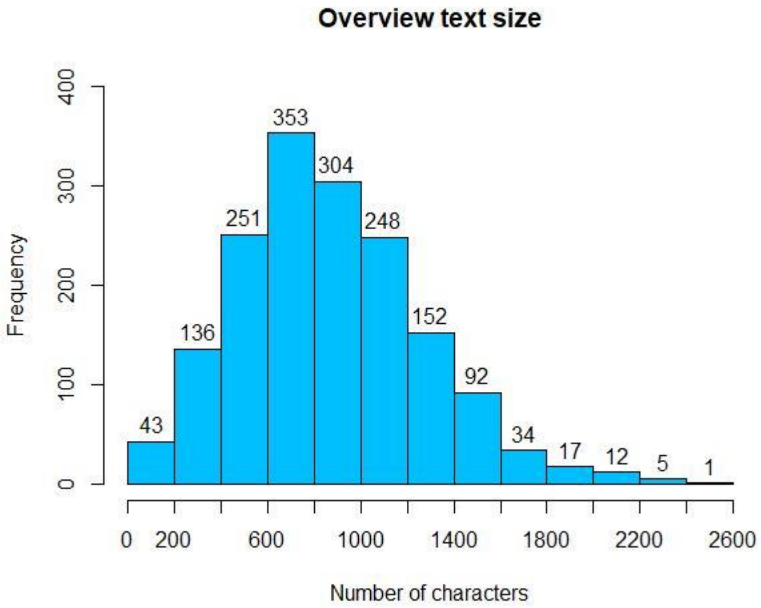
Data Characterization





Data Characterization





System Results

Disease:

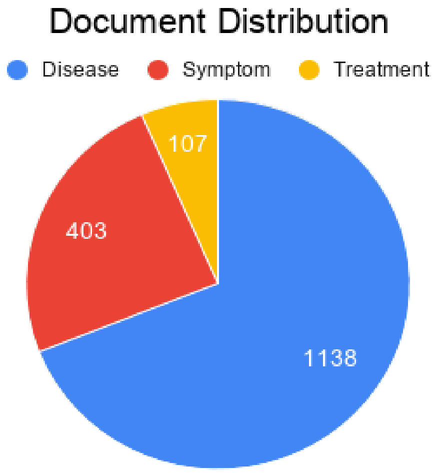
- Overview.
- List of symptoms, treatments, drugs, causes and health specialties.

Treatment:

- Overview.
- List of diseases.

Symptom:

- Overview.
- List of diseases.



Retrieval Tasks

- Retrieve disease, treatment or symptom based on its information (name or word in overview)
- Retrieve disease by symptom.
- Retrieve disease by health specialty.
- Retrieve treatment by disease.
- Retrieve symptom by disease.

CS:GO Professional Matches and News

Dataset Preparation and Characterisation

Data collection and preparation

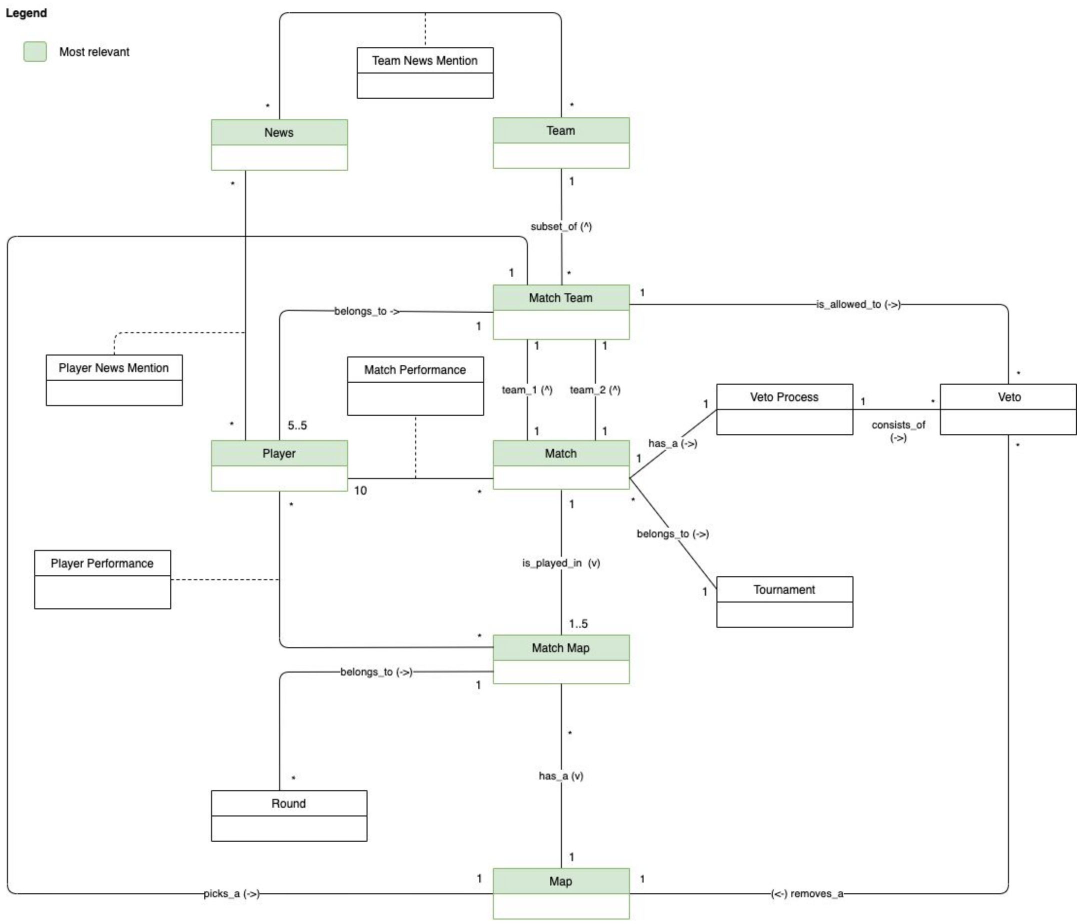
Professional Matches

- **Source:** HLTV.org
- **External pre-processing:** Web scraping and upload of dataset to Kaggle
- **Collection method:** Download of Kaggle dataset
- **Preparation steps:**
  - Removal of invalid map values
  - Date formatting
  - Removal of unnecessary columns
- **Tools used:** OpenRefine

News

- **Source:** HLTV.org
- **Collection method:** Web scraping
  - Scraping limited to news from 2018 and 2019
- **Preparation steps:**
  - Date formatting
  - Entity extraction
- **Tools used:** Scrapy, spaCy

Conceptual model *



* Simplified

Data characterization

Professional Matches

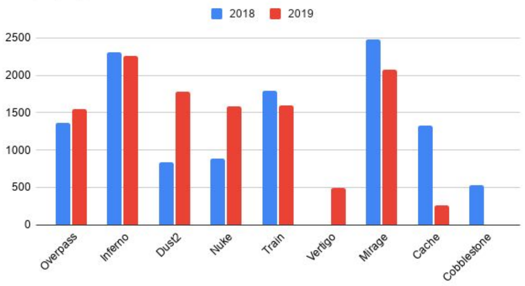
- Dataset comprised of **4 CSV files**: players, results, economy, picks.
- **Players (~383,000 rows):** performance of a players in a given **map** (within a match)**;**
- **Results (~45,700 rows): rounds** played in a given **map** (within a **match**) and the respective outcome;
- **Economy (~43,000 rows):** money earned by the teams in all **rounds** of a **map** (within a **match**)**;**
- **Picks (~16,000 rows): maps** picked and removed by **teams** in a given **match.**

News

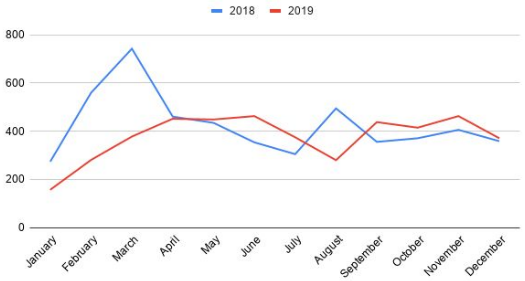
- Dataset comprised of **2 CSV files**: news and news_tag;
- **News (~6,300 rows): news** collected from HLTV.org, spanning 2018 and 2019;
- **News Tag (~280,000 rows)**: (~2700 unique) **entities** extracted from an article and their respective **type** (**player** or **team**)

Data characterization

Professional Matches

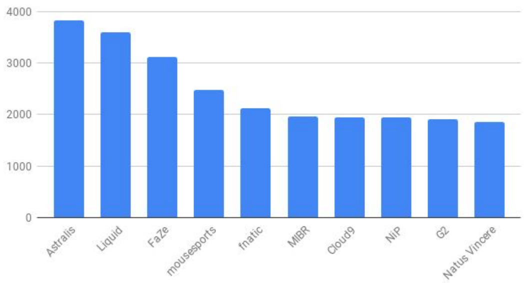


Maps played in 2018/2019

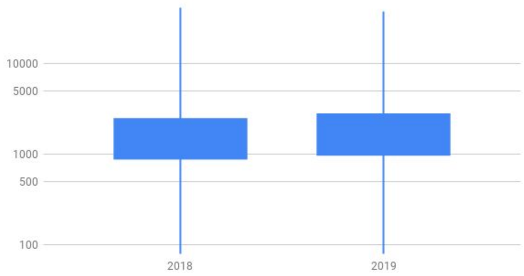


Matches in 2018/2019

News



Top 10 Most Mentioned (Relevant) Entities in 2018-2019



News length distribution

Retrieval tasks

Search for	Order by	Restrict on
Players	- Most/Least {Kills, Assists, Deaths, HS, Flash Assists} - Best/Worst {KAST, KD, ADR, FKDIFF, Rating}	- Map - Team Against - Date Interval - Side (T/CT) - Team - Nationality
Teams	- More/Less {Wins, Games Played, Round Win %, Force Buy %, Upset Potential, Pick Win Rate}	- Map - Team Against - Date Interval - Side (T/CT)
Matches	- Date	- Map - Teams - Date Interval - Event
News	- Date	- Date Interval

Luís Silva (up201503730)
Mariana Costa (up201604414)
Pedro Fernandes (up201603846)
(Group 7)

BILLBOARD 200: POPULAR ALBUMS AND ARTISTS

DATASET PREPARATION

Grupo 8
João Miguel (up201604241@fe.up.pt)
José Azevedo (up201506448@fe.up.pt)
Ricardo Ferreira (up200305418@fe.up.pt)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DATASET CHARACTERIZATION

Acoustic and meta features of albums and songs on the Billboard 200

- SQL Database Format
- Two Tables (Albums and acoustic features)
- Free for use and download

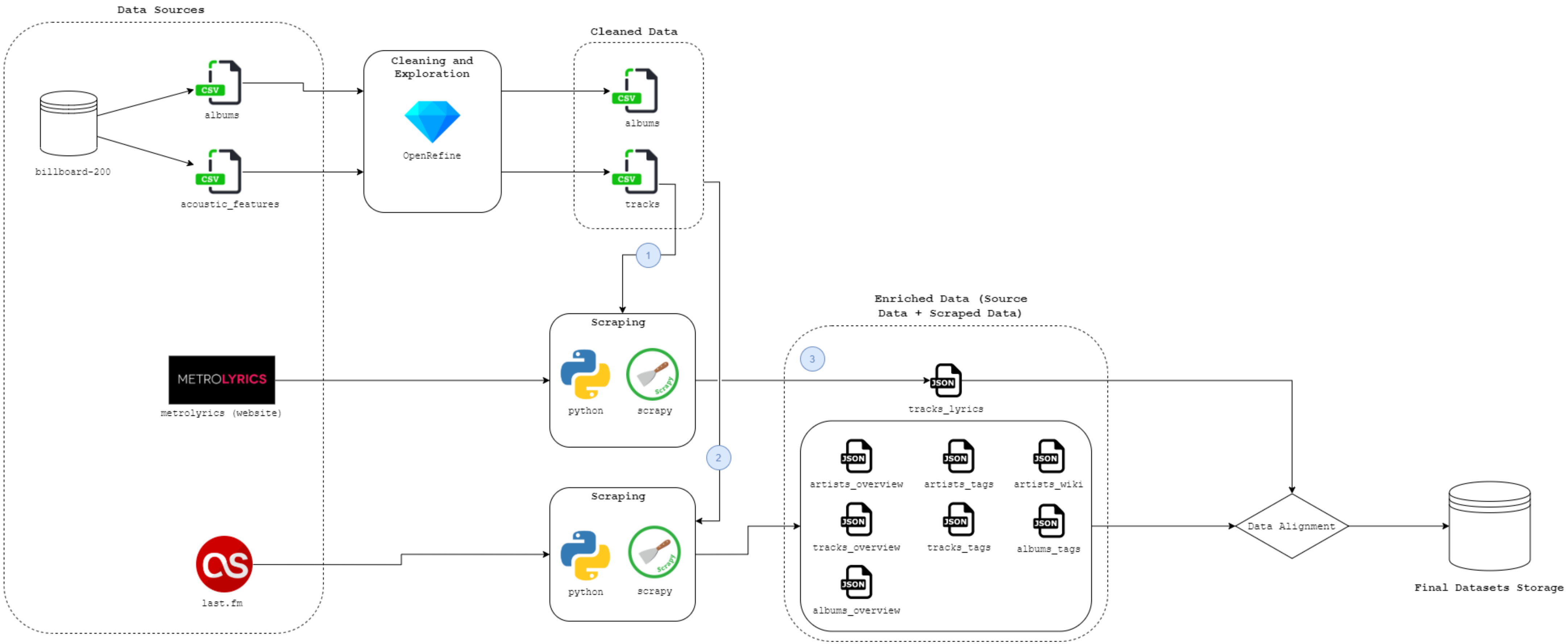
Last FM

- Website with a huge list of artists, songs and albums
- Easy to build urls
- Crawlers allowed

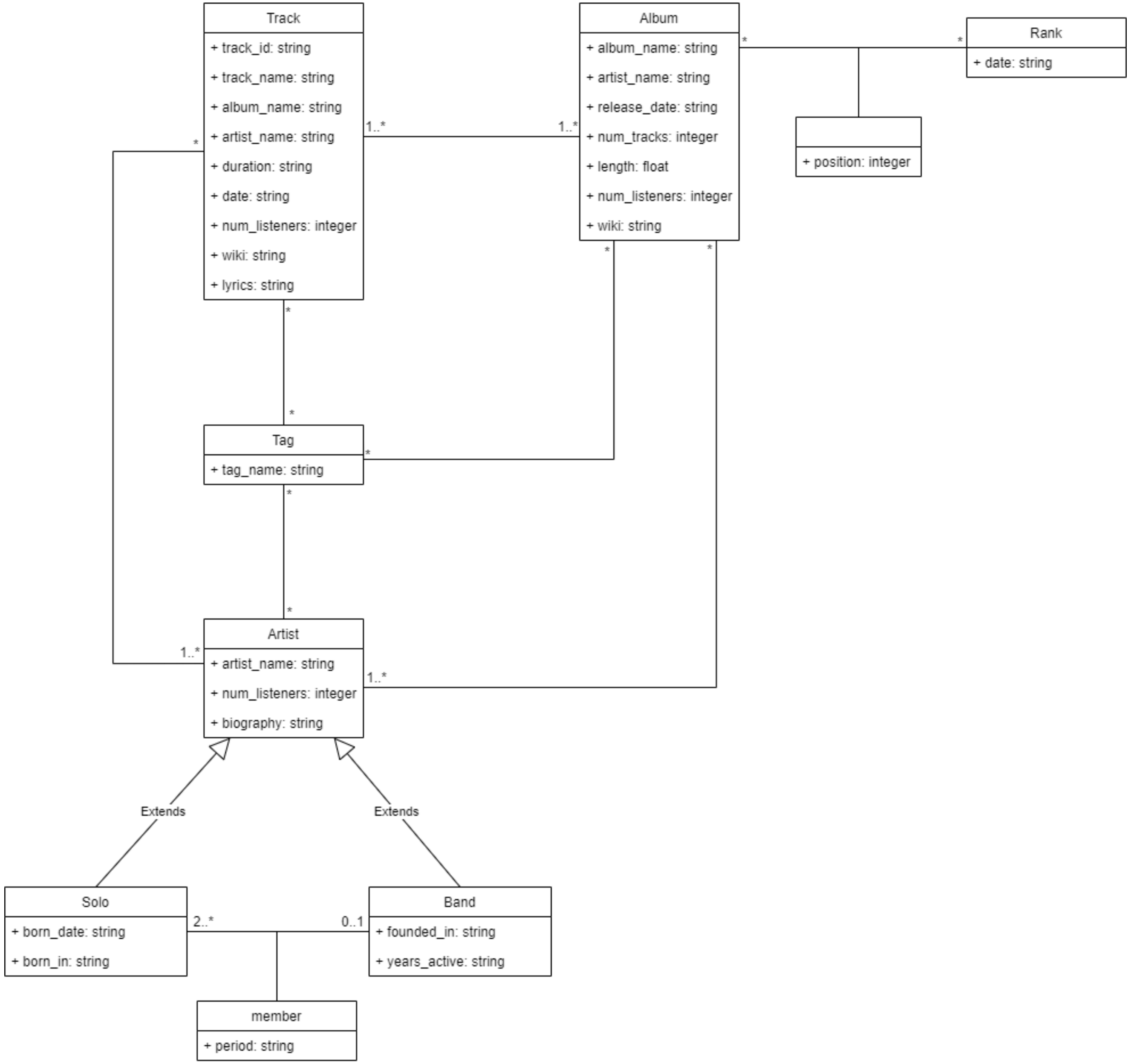
Metro Lyrics

- Website with songs lyrics
- Easy to build urls
- Crawlers allowed

DATA PIPELINE



CONCEPTUAL MODEL



Track
+ track_id: string
+ track_name: string
+ album_name: string
+ artist_name: string
+ duration: string
+ date: string
+ num_listeners: integer
+ wiki: string
+ lyrics: string

Album
+ album_name: string
+ artist_name: string
+ release_date: string
+ num_tracks: integer
+ length: float
+ num_listeners: integer
+ wiki: string

Rank
+ date: string

+ position: integer

Tag
+ tag_name: string

Artist
+ artist_name: string
+ num_listeners: integer
+ biography: string

Solo
+ born_date: string
+ born_in: string

Band
+ founded_in: string
+ years_active: string

member
+ period: string

SEARCH AND RETURNED DOCUMENTS

Returned documents

- Albums
- Artists
- Musics
- Rank

Possible search tasks:

- **Rank by date (year, month, day)**
  - Will return: Albums, Artists, Rank
- **Artists (Band or Solo)**
  - Will return: Artist, Albums
- **Location**
  - Will return: Artists
- **Album**
  - Will return: Album, Artist, Musics, Best Rank
- **Release Date (year, month, day)**
  - Will return: Albums, Artists
- **Musical Genre**
  - Will return: Albums, Artists, Musics
- **Musics (By name or words/sentences from the lyrics)**
  - Will return: Musics

DATASET PREPARATION

ANIMATION IN JAPAN

DAPI 2020/2021

INTRODUCTION

- Japanese-style animated film
- Popular form of entertainment for all kinds of audience.
- Originated from novels or vídeo games adaptations

DATASET

- Fans gather in plataforms to talk about animes
- Information regarding animes and animes reviews are collected and can be accessed.
- Gathering of data separately
- All users can have na overview of anime rating

INFORMATION RETRIEVAL

kaggle
MyAnimeList

+

IMDb

Used to get general information about animes and their rating in MyAnimeList.

Used to get rating information from IMDb.

DATA PREPARATION

Cleaning and organizing the dataset