PRI Project

PRI 23/24 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Project Overview

Group Project

- Design and implementation of an information processing and retrieval system
- The project is developed in groups of 4 students and starts with the selection of a topic and the relevant data sources
- The project is organized in deliveries and partial presentations, which correspond to the project development phases
    - Milestone 1: Information Processing (week of Oct 9th)
    - Milestone 2: Information Retrieval (week of Nov 13th)
    - Milestone 3: Search System (week of Dec 11th)

Milestones

- Each project delivery (milestone) has a corresponding presentation and discussion
- Electronic submissions of the project deliverables are accepted up to 18:00 on the day before the in-class presentation
- Reports are written as short scientific papers, using a two-column format (4 pages max in each delivery). Each report is a self-contained work-in-progress and is based on the previous deliveries
- In the weeks assigned to project presentations, the practical class will be organized in a workshop format, with project presentations and discussions according to an established schedule
- The final project evaluation corresponds to a weighted average of the milestones evaluations.

Milestones Workshops

- The lab class is organized as a workshop for presentation and discussion.
- Each group has a 10 minute slot: 5 min for presentation, plus 5 min for questions.
- Questions are part of the 10% individual assessment for this milestone.
  - *The grade for deliveries has an individual component, that is positive if the student contributes to the workshop with questions or comments, and negative if the student is not present or shows unprofessional behavior.*
- Planned schedule and additional notes available in Moodle.

M1: Information Processing

- The first milestone is achieved with the preparation and characterization of the datasets selected for the project

- Work on these tasks depends on the nature, volume, organization and accessibility of the selected datasets. As a result of this milestone, a well-documented and reproducible pipeline of data processing is expected

  - search repositories for datasets
  - select convenient data subsets
  - assess the authority of the data source and data quality
  - perform exploratory data analysis
  - prepare and document a data processing pipeline
  - characterize the datasets, identifying and describing some of their properties
  - identify the conceptual model for the data domain
  - define and characterize the documents in the final collection
  - identify and characterize follow-up information needs for the project (**important**)

M2: Information Retrieval

- The second milestone is achieved with the implementation and use of an information retrieval tool on the project datasets and its exploration with free-text queries

- This task makes use of state-of-the-art retrieval tools and involves the view of the datasets as collections of documents, the identification of a document model for indexing, and the design of queries to be executed on the indexed information

    - choose the information retrieval tool (Solr, Elasticsearch, ...)
    - analyze the documents and identify their indexable components
    - use the selected tool to build the indexes
    - use the selected tool to configure and execute the queries
    - demonstrate the indexing and retrieval processes
    - implement and evaluate two distinct retrieval setups
    - manually evaluate the returned results
    - evaluate the results obtained for the defined information needs

M3: Search System

- The third milestone is achieved with the development of the final version of the search system
- This version is an improvement over the previous milestone, making use of features and techniques with the goal of improving the quality of the search results
- For this milestone, each group is expected to explore innovative approaches and ideas, and will heavily depend on the context and data of each group
- Additionally, an extended evaluation of the results and a comparison with the previous version of the search system is also expected
- Examples of topics to explore include: introduce semantic search using embeddings, incorporate new information retrieval algorithms; expand the information available for each document by adding and linking new datasets; work on user interfaces by developing a frontend for the search system

Project Themes

- Project topics are “free”, but cannot be repeated in the same class
- Need to be approved by the end of the **first practical class**
- Data source(s) must be of unstructured nature and rich in textual data
- Consider your personal interests and motivations
- Avoid too common topics (e.g. recipes, books).

- Many possibilities: education, sports, law, government, media …

Working in Groups

- The project is developed in groups of four students.
- Obtaining approval in the project **requires the participation of each student in all phases of the project**, namely in the selection of the data sources, in the selection of technologies, in identifying and characterizing the problem, in designing and implementing the solution, in writing the reports, and in the project presentations.
- The individual final grade of the project can vary from element to element of the same group, by plus or minus 3 values, based on the opinion of the teachers and in the self-assessment and hetero-assessment to be carried out internally in each group.

Project Report

LaTeX

- LaTeX should be used to prepare the final manuscript.
- LaTeX is a document processing and preparation system.
- Focus is on the content, separating writing from presentation and styling.
- Commonly used in academia to prepare scientific manuscripts and publications.
- Helpful for preparing tables in LaTeX — www.tablesgenerator.com
- You can either use a local installation or a web-based collaborative application such as Overleaf (overleaf.com).
- Use the two-column ACM template (see project rules).
- Each milestone report is limited to 4 pages, excluding references.

Report Structure

- Final report will be organized in three parts, one for each milestone.
- The **Abstract**, **Introduction**, **Conclusions** and **References** are expected to be reviewed/extended in each iteration.

Writing Style

- Scientific writing is expected.
- Scientific writing is technical writing for communication in science.
- Factual, concise, evidence-based.
- Avoid creative, verbose prose.
- Attention to expressions with rigorous meanings in CS, Math or Engineering, e.g. "significant" (statistics), "exponential growth" (math, CS).
- Do not use "decorative" figures.
- Abstracts are not an introduction to the document, they should act as 'mini-documents' with all the main elements: why, what, how, results.

Reports: Writing Style

- ➔ Be gender neutral: “he”, “his” are not neutral, “he/she” is cumbersome; the plural is well accepted, even if it does not agree: “The user is planning their trip”.
- ➔ “This figure” versus “Figure 1”: the first is a common name (lower case) the second the name of the figure (upper case).
- ➔ Some expressions have rigorous meanings in Math, CS or Engineering and we should avoid using them informally; examples: Significant (statistics); Exponential growth (math, CS).
- ➔ Mentions to the project, the course, etc: leave them out of the document. They may appear as a note (footnote on the first page, final note on the report) but not on the text; this must focus on the problem, solutions, considerations on these.
- ➔ When making comparisons (“a better solution”, “an improved method”) make sure we know what you are comparing with.
- ➔ Beware of false friends: some words have very different meaning from their similar forms in Portuguese, e.g. notorious.

Reports: Writing Style

- ➔ The abstract is not an introduction, it is a single paragraph summary of full content. It includes the motivation, the problem, the solution, the results, and findings.
- ➔ Do not use a cover page.
- ➔ Tables, figures and code listings must be presented and referenced in the text.
- ➔ Scientific writing is: factual, concise, precise.
- ➔ Pay attention to opening and closing quotes in LaTeX, e.g. use “quoted text”, and not ” quoted text”.
- ➔ Always use vector formats (e.g. PDF) for images to include (unless impossible, e.g. external source, digitization).

Milestone 1: Information Processing

Milestone 1 - Data Preparation

- The first milestone is achieved with the preparation and characterization of the datasets selected for the project.
- Work on these tasks depends on the nature, volume, organization and accessibility of the selected datasets.
- Typical actions (depending on the data source):
    - search repositories for datasets
    - select convenient data subsets
    - assess the authority of the data source and data quality
    - perform exploratory data analysis
    - prepare and document a data processing pipeline
    - characterize the datasets, identifying and describing some of their properties
    - identify the conceptual model for the data domain
    - define and characterize the documents in the final collection
    - identify follow-up information needs in the data domain

Project Themes

- Discussion on project themes
- Case studies
    - Electoral programs: break in proposals / sections / paragraphs
    - Sports: search for "entity x" is not sufficient (trivial case)
- Common: combine different data sources
- Different starting points, e.g. already prepared datasets to raw unstructured data
    - If using already prepared datasets, two or more datasets need to be linked (combined).
- Important: value added by the data processing pipeline must be clear
- Prepare data for the next task - search system.

Data Processing

- Wikipedia and Wikidata are key resources in the data ecosystem.

- Wikipedia is often used as a source of text about entities.
  - An API is available - https://en.wikipedia.org/w/api.php
  - Organized in language-specific projects, e.g. https://pt.wikipedia.org
  - DBpedia produces structured snapshots of Wikipedia subsets - https://www.dbpedia.org

- Wikidata is used as a hub to link different collections, e.g. exploring references as entry points to distinct collections / domains.
  - Luís Vaz de Camões — https://www.wikidata.org/wiki/Q590
  - Portugal — https://www.wikidata.org/wiki/Q45

Domain Data Model

- The data preparation report includes a presentation of the domain data model, i.e. a representation of the domain data concepts and their interrelationships.
- Conceptual modeling is the process of discovering the entity types that represent the things and concepts, their relationships, pertinent to the system.
- It establishes a vocabulary and a shared vision for the system.
- Use a UML class diagram to document the domain data model.

Data Processing Pipeline

- The data preparation report includes a detailed presentation of the data processing pipeline.
- Document the pipeline describing the main decisions and key processes.
- We interested on the physical perspective, i.e. focus on implementation details - including data formats, libraries, storage solutions, software tools, etc.
- Include a data flow diagram (DFD). Key elements are:
    - External entities, which are sources or destinations of data.
    - Processes, that represent data processing operations (automatic or manual).
    - Data flows, that describe how and what data moves around.
    - Data stores, representing data storage structures (files, databases, etc).

Documents and Information Needs

- As a result of Milestone 1, you should be able to clearly answer these questions:

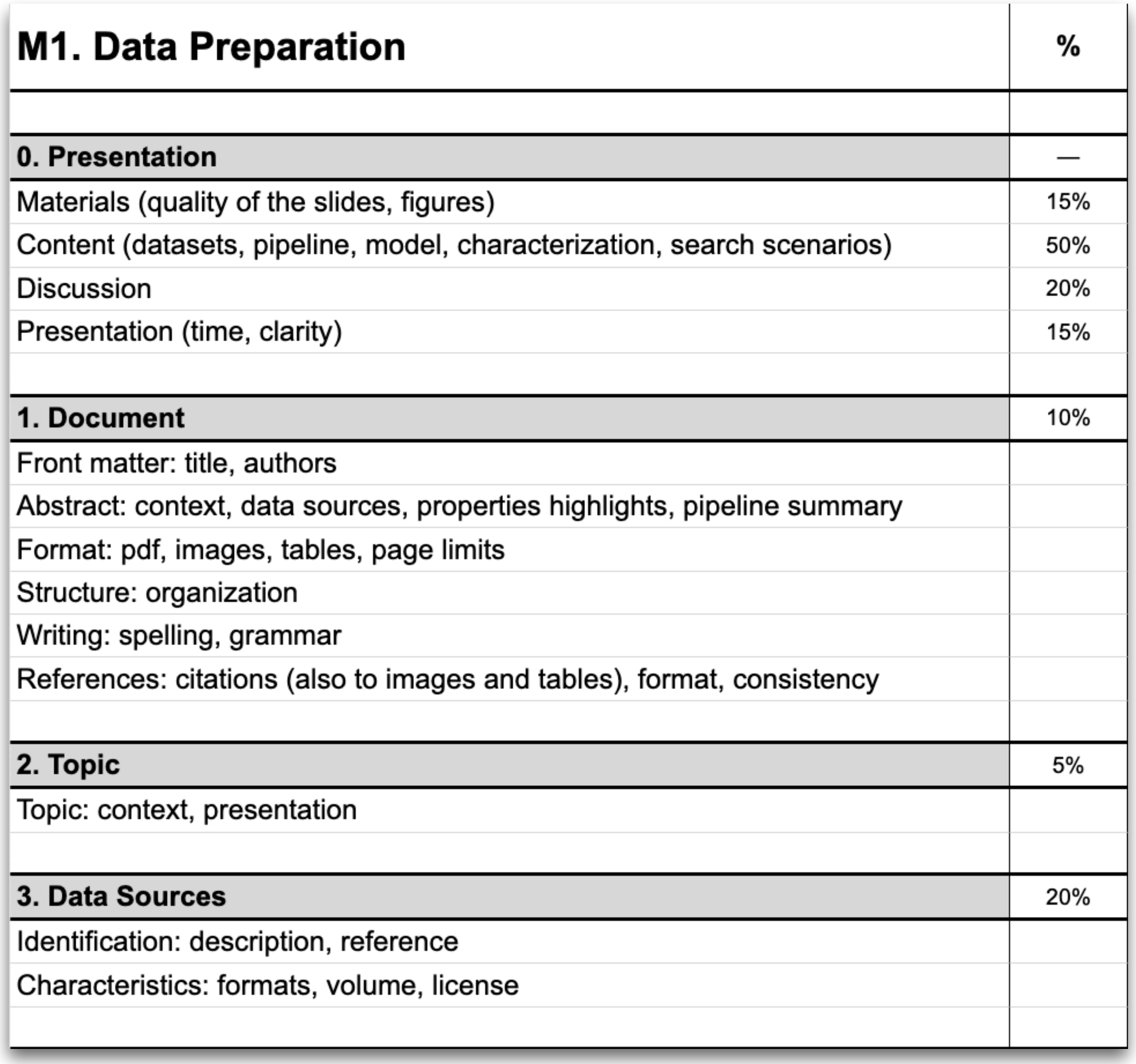
- What is a document in my final collection?

- What are examples of information needs in this context?
  - What are relevant documents?
  - What are non-relevant documents?

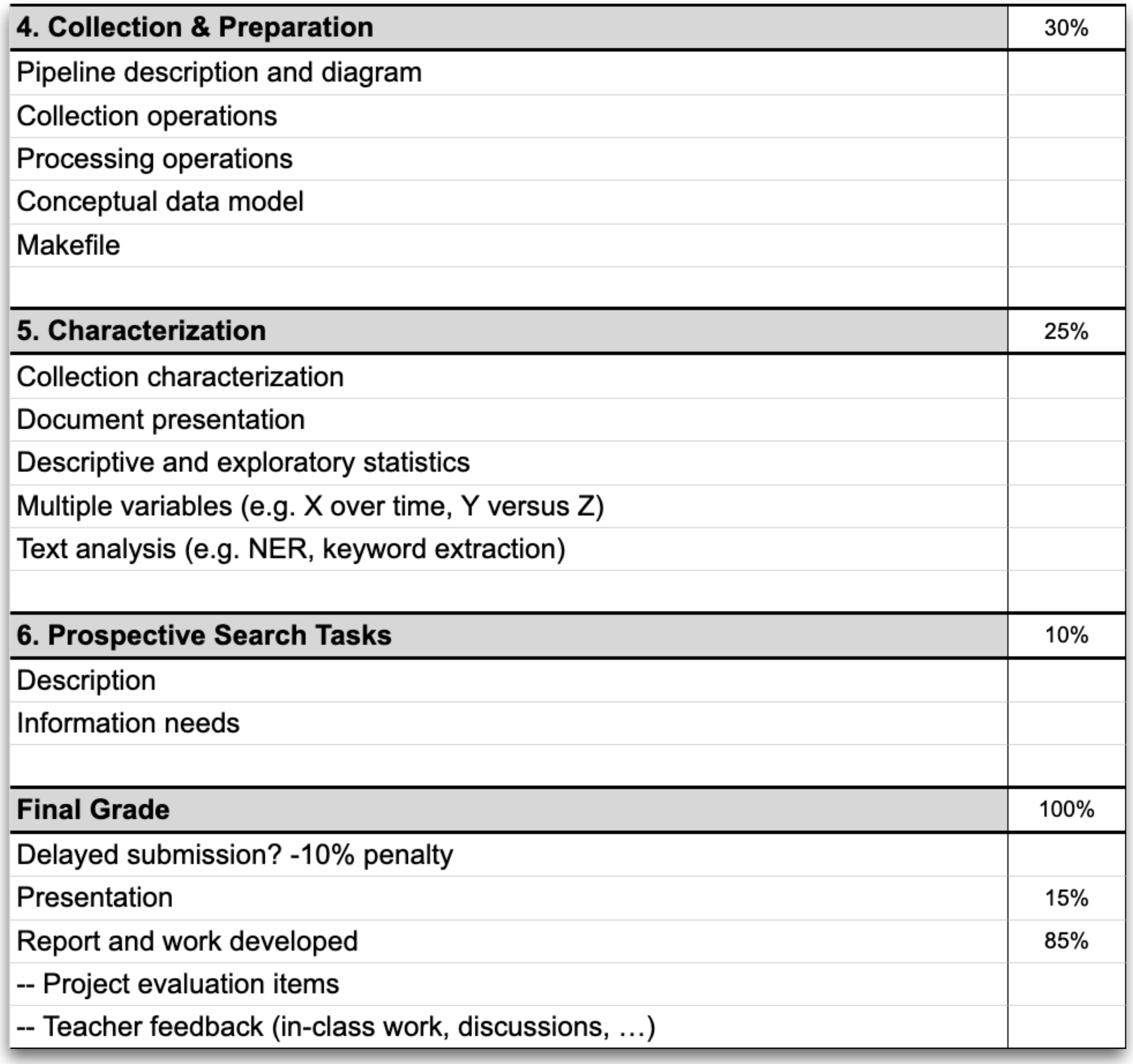
Information Needs

- Prospective search tasks should be described considering "information needs" and the document types to be returned by the search system to be developed.
- Information needs represents the underlying motivation of a user to obtain information to satisfy a need, i.e. use a search system.
- Example of an information need:
  - Information on whether drinking red wine is more effective at reducing your risk of heart attacks than drinking white wine.
- This might be translated into a query such as:
  - [wine red white heart attack effective]
- A search system is successful if retrieved documents address the stated information need, not because they just contain all the words in the query.

M1 Checklist



M1. Data Preparation	**%**
0. Presentation	—
Materials (quality of the slides, figures)	15%
Content (datasets, pipeline, model, characterization, search scenarios)	50%
Discussion	20%
Presentation (time, clarity)	15%
1. Document	10%
Front matter: title, authors	
Abstract: context, data sources, properties highlights, pipeline summary	
Format: pdf, images, tables, page limits	
Structure: organization	
Writing: spelling, grammar	
References: citations (also to images and tables), format, consistency	
2. Topic	5%
Topic: context, presentation	
3. Data Sources	20%
Identification: description, reference	
Characteristics: formats, volume, license	



4. Collection & Preparation	30%
Pipeline description and diagram	
Collection operations	
Processing operations	
Conceptual data model	
Makefile	
5. Characterization	25%
Collection characterization	
Document presentation	
Descriptive and exploratory statistics	
Multiple variables (e.g. X over time, Y versus Z)	
Text analysis (e.g. NER, keyword extraction)	
6. Prospective Search Tasks	10%
Description	
Information needs	
Final Grade	100%
Delayed submission? -10% penalty	
Presentation	15%
Report and work developed	85%
-- Project evaluation items	
-- Teacher feedback (in-class work, discussions, …)	

Milestone 2: Information Retrieval

Project Overview

- M1:
  - **from…** selected data sources
  - **to…** an organized (and studied) collection of documents
  - **plus…** and a set of prospective search tasks

- M2:
  - **from…** a collection of documents
  - **to…** the first version of a search system

Information Needs

- ➔ An information need (PT: *necessidade de informação*) is the main motivation for a person using a search engine.

- ➔ There are many types of information needs, categorized considering dimensions such as:
  - ➔ the number of relevant results expected;
  - ➔ the type of information needed;
  - ➔ the tasks that led to that information need.

- ➔ In some cases it might be difficult for people to define exactly what their information need is, because that information is a gap in their knowledge.

- ➔ **Results are evaluated against the information need, not the query.**

Search Queries

- ➔ A **search query** corresponds to the user's materialization of an information need for a specific search engine.
- ➔ Queries represent very different information needs and may require different search techniques and ranking algorithms to produce best rankings.
- ➔ A query may be a poor representation of the information need.
  - ➔ This can happen if the user finds it difficult to express the information need; has a limited knowledge of the problem's context and vocabulary; lacks expertise in the search engine's features.

Examples

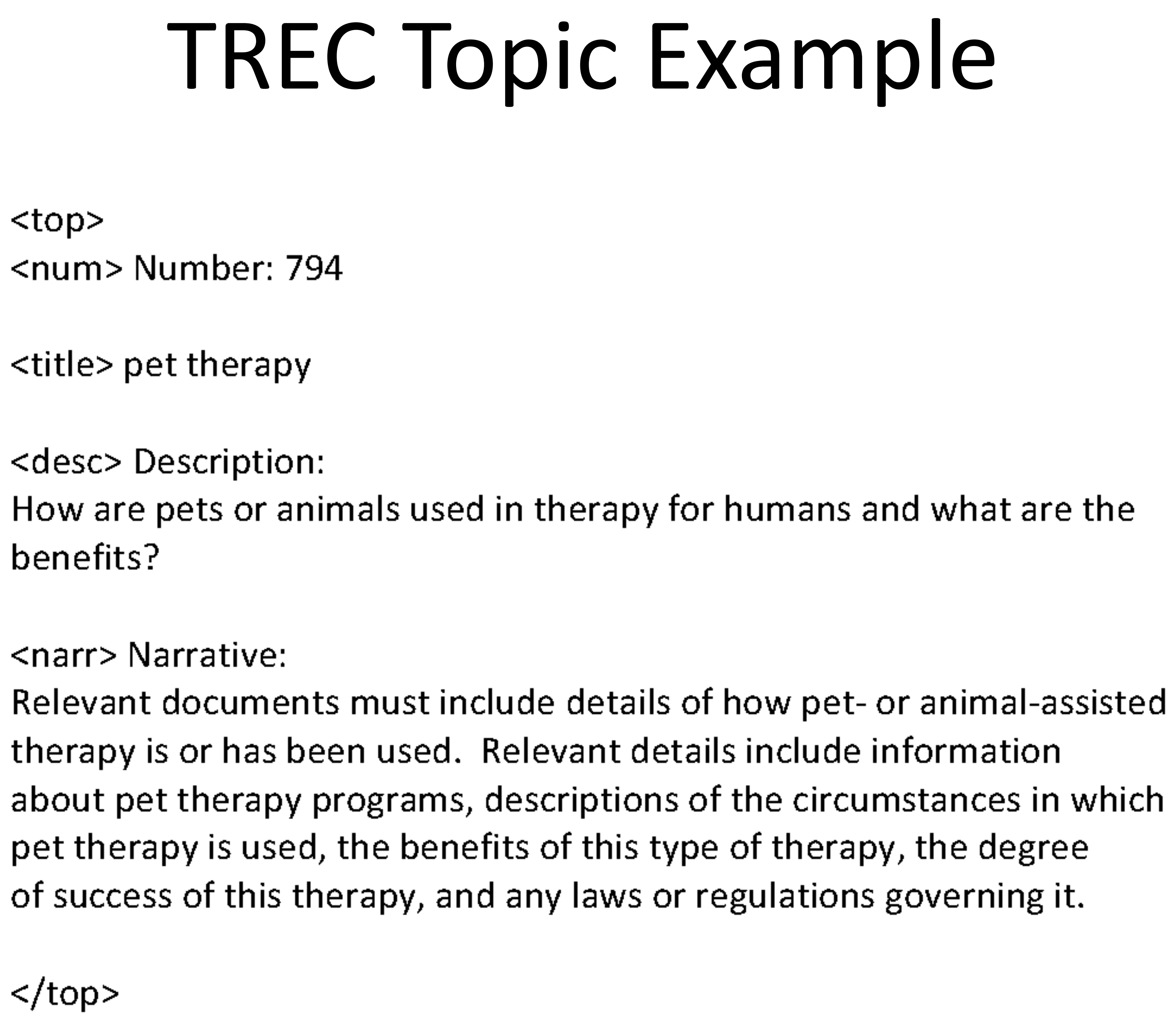
- IN: What is the address of restaurant X?
  Q: [restaurant x address]

- IN: What is the web site of the University of Porto?
  Q: [university of porto]

- IN: How to solve a "permission denied" error in Y?
  Q: ["permission denied" y]

- Tip for selecting good keywords: use words that you expect to find in a relevant result; not necessarily the words from the information need.

Complete Example



TREC Topic Example

```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Evaluating Results

- ➔ Complete manual assessment of a collection can only be done for tiny collections.
- ➔ This is unfeasible for most search problems or collections where the number of returned documents can easily be on the thousands.
- ➔ In these cases (also yours in the PRI project), only a subset of the returned documents can be evaluated for each query.
- ➔ The most standard approach is pooling, where relevance is assessed over a subset of the collection that is formed from the top k documents returned by the different IR systems.
- ➔ In TREC, between 50 and 200 top results are evaluated.
- ➔ The collection of all evaluated results are used to evaluate and compare the systems.

Additional Notes on Evaluation

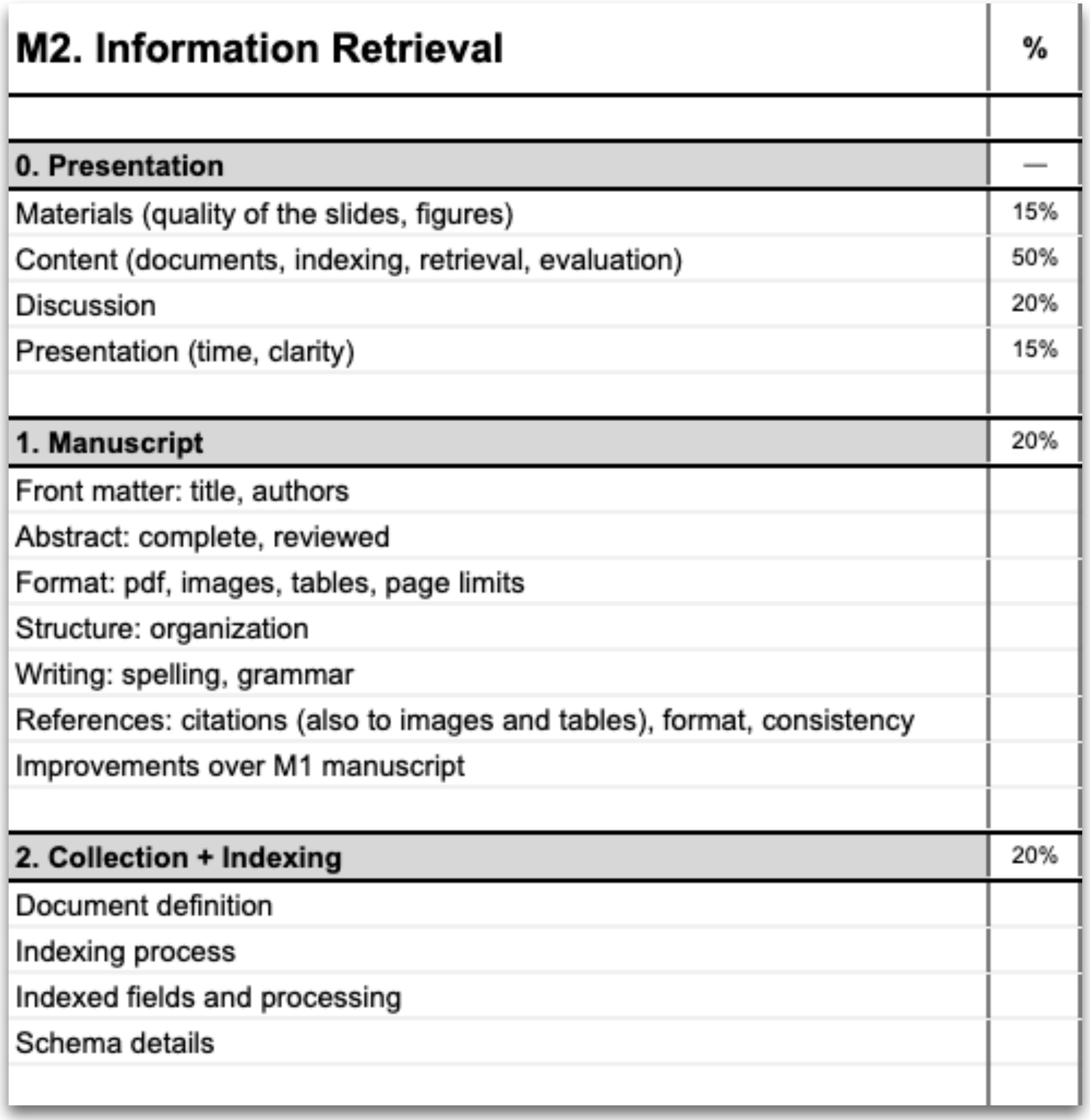
- ➔ *Only apples can be compared to apples.*

- ➔ You can make changes to the indexing process or the retrieval process.
  - ➔ But use the same approach for all queries.
- ➔ If you make optimizations per query, these are not comparable.
- ➔ Example: you cannot compare results where different optimization strategies where used; you can only compare identical systems.

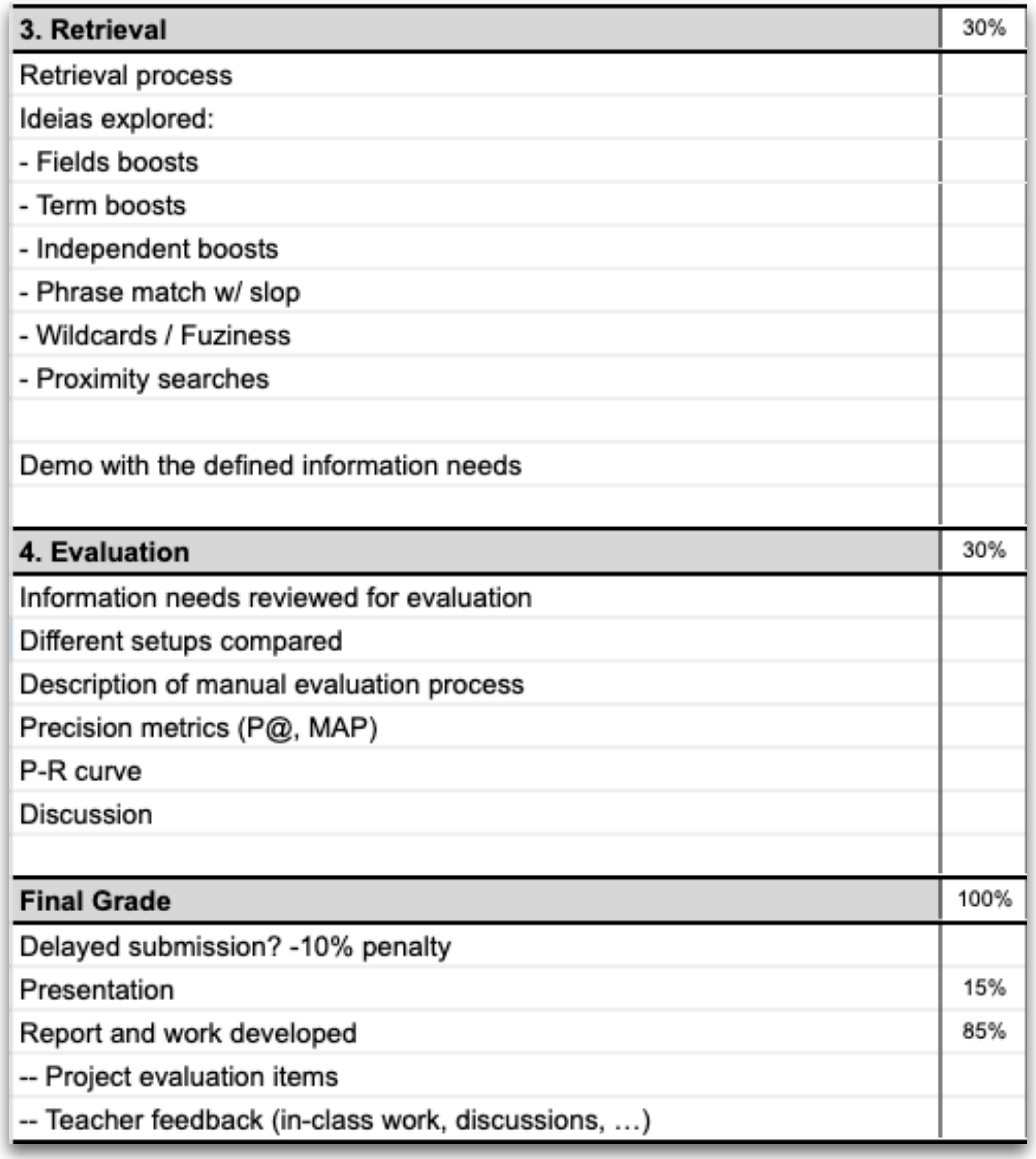
Milestone #2

- The second milestone is achieved with the implementation and use of an information retrieval tool on the project datasets and its exploration with free-text queries.
  - choose the information retrieval tool (**Solr**, Elasticsearch, Lucene, Terrier, …);
  - analyze the documents and identify their indexable components;
  - use the selected tool to build the indexes;
  - use the selected tool to configure and execute the queries;
  - demonstrate the indexing and retrieval processes;
  - manually evaluate the returned results;
  - evaluate the results obtained for the defined information needs.

M2 Checklist



M2. Information Retrieval	%
0. Presentation	—
Materials (quality of the slides, figures)	15%
Content (documents, indexing, retrieval, evaluation)	50%
Discussion	20%
Presentation (time, clarity)	15%
1. Manuscript	20%
Front matter: title, authors	
Abstract: complete, reviewed	
Format: pdf, images, tables, page limits	
Structure: organization	
Writing: spelling, grammar	
References: citations (also to images and tables), format, consistency	
Improvements over M1 manuscript	
2. Collection + Indexing	20%
Document definition	
Indexing process	
Indexed fields and processing	
Schema details	



3. Retrieval	30%
Retrieval process	
Ideias explored:	
- Fields boosts	
- Term boosts	
- Independent boosts	
- Phrase match w/ slop	
- Wildcards / Fuziness	
- Proximity searches	
Demo with the defined information needs	
4. Evaluation	30%
Information needs reviewed for evaluation	
Different setups compared	
Description of manual evaluation process	
Precision metrics (P@, MAP)	
P-R curve	
Discussion	
Final Grade	100%
Delayed submission? -10% penalty	
Presentation	15%
Report and work developed	85%
-- Project evaluation items	
-- Teacher feedback (in-class work, discussions, …)	

Milestone 3: Search System

Milestone #3

- ➔ The 3rd milestone is achieved with the development of the final version of the search system.
- ➔ This version is an **improvement over the previous milestone**, making use of features and techniques with the goal of improving the quality of the search results.
- ➔ For this milestone, each group is expected to explore innovative approaches and ideas, and will heavily depend on the context and data of each group.
- ➔ An **extended evaluation of the results** and **a comparison with the previous version** of the search system is expected.
- ➔ As examples of topics to explore, groups **may**:
    - ➔ Incorporate new information retrieval algorithms;
    - ➔ Expand the information available for each document by adding and linking new datasets;
    - ➔ Work on user interfaces by developing a frontend for the search system (**not sufficient**);

Examples

- Semantic search.
- Combine multiple indexes, e.g. search over multiple entity types — persons, organizations, news, and combine the results.
- Expand the information indexed by integrating new information (external), e.g. linking to new data sources.
- Expand the information indexed by further processing the existing information (internal), e.g. entity extraction using NER, other information extraction techniques.
- Propose new ranking signals using the existing information, e.g. PageRank signal based on citation data.
- Query processing improvements, e.g. pseudo-relevance feedback, supporting cross-language retrieval, incorporating global thesaurus, incorporating synonyms (Wikidata is a good source for expansion).
- Work on user interaction features must be measured using these metrics (*harder*).
- …
- Improvements must be measured using standards metrics such as P@, AvP, MAP, PR curves.

Improvements over Milestone 2

- In some cases, fixing the problems (or missing parts) identified in Milestone 2 is necessary before developing a new version of the system.

- Prepare your ideas in advance of this week's lab classes.

Demo Video

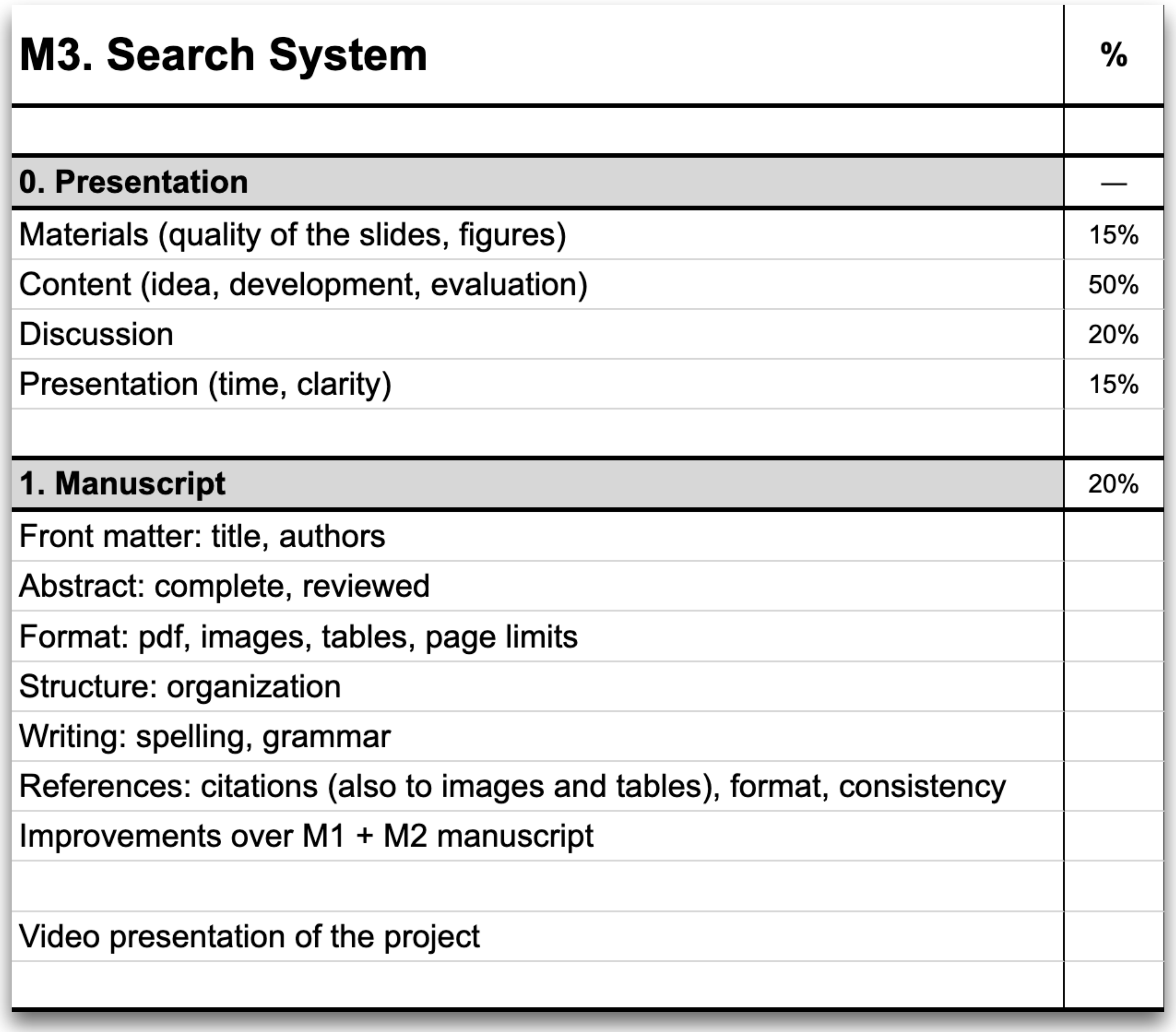
- M3 submission requires the submission of a short video showcasing the main features of the developed search system.

- ~2 minute video showing interactions with the system, highlighting the main features.

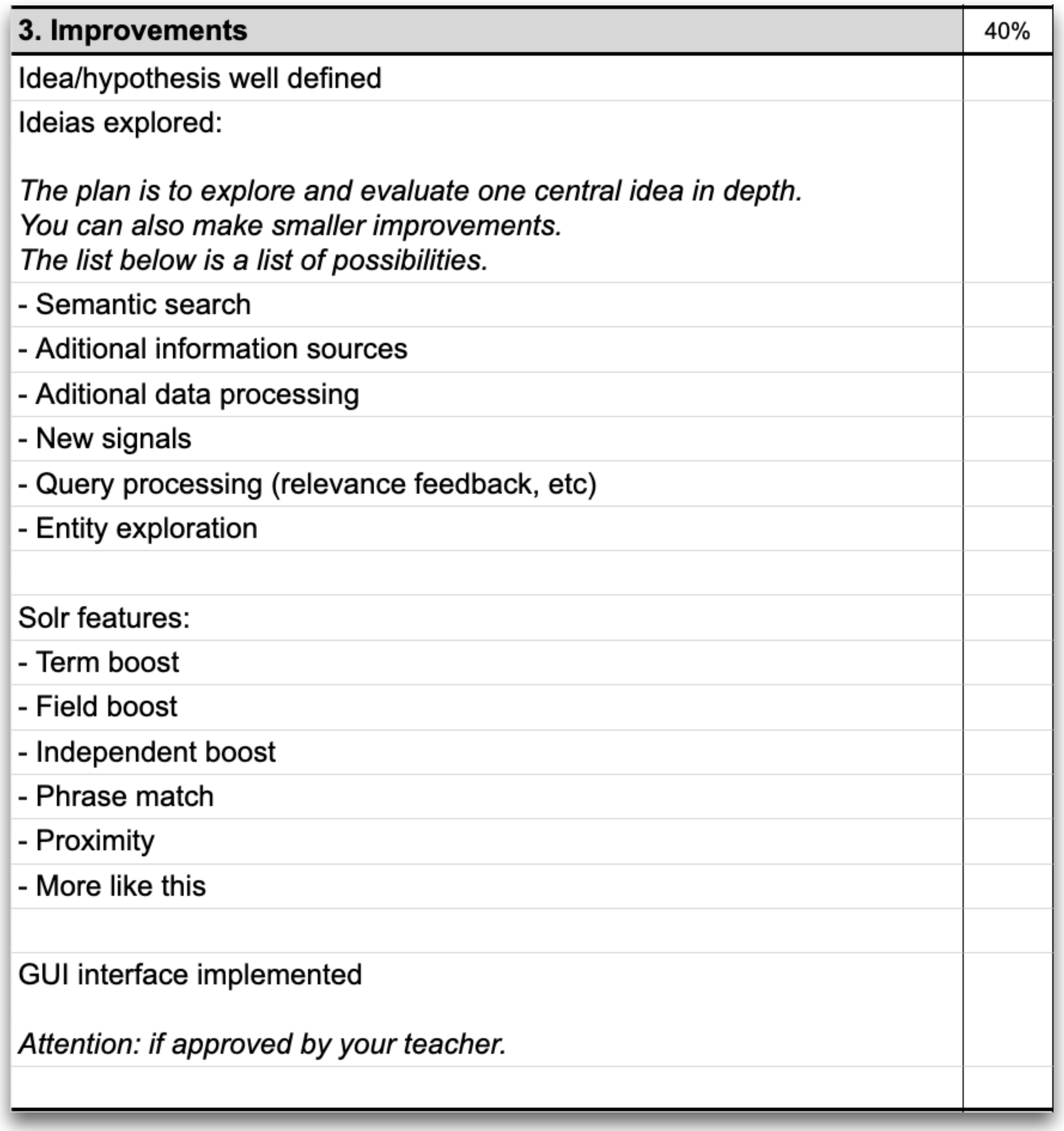
- Including subtitles (plus optional audio) describing what is being presented.

- Goal: keep archive of PRI project; showcase for future students.

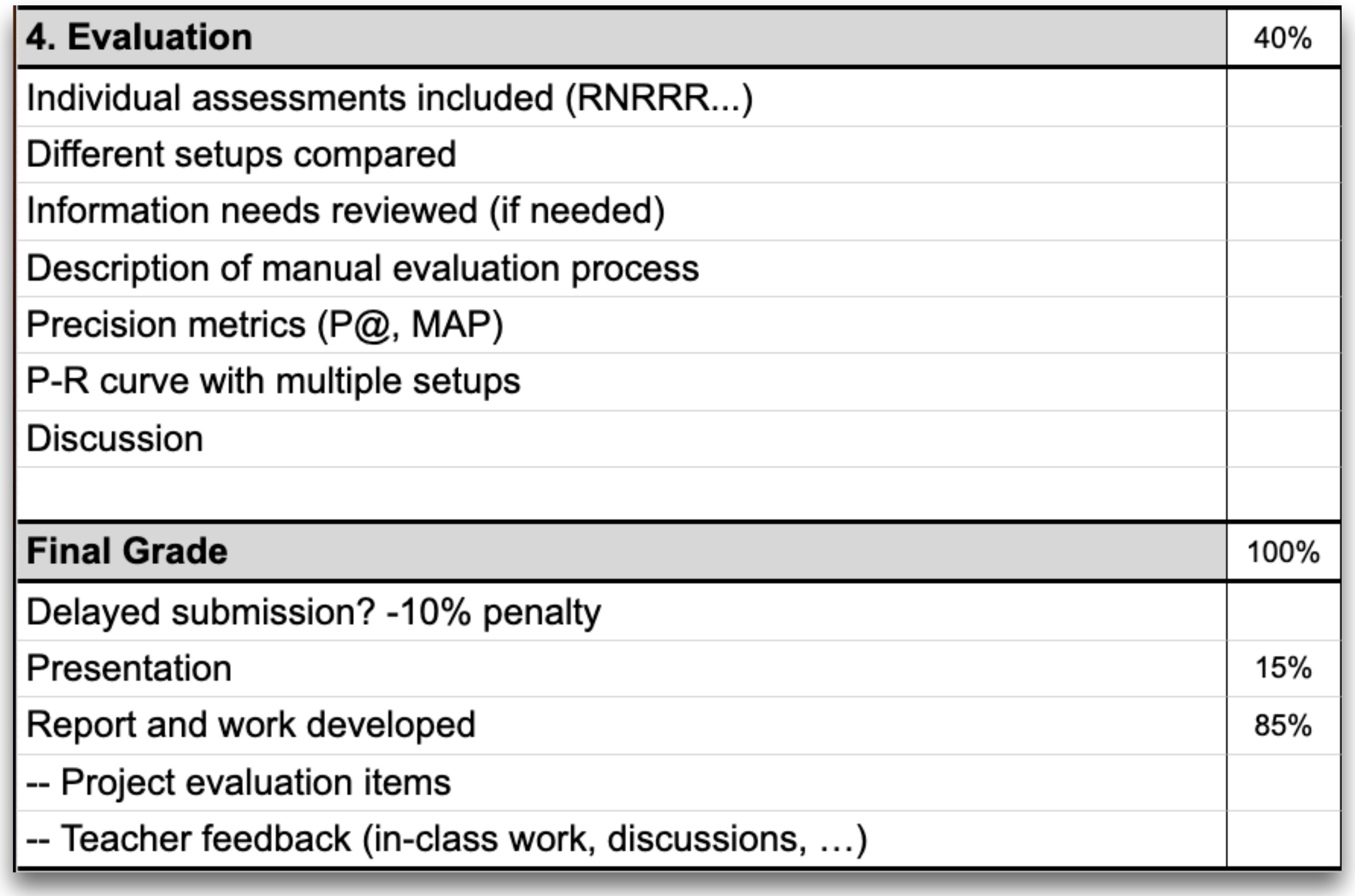
M3 Checklist



M3. Search System	**%**
0. Presentation	—
Materials (quality of the slides, figures)	15%
Content (idea, development, evaluation)	50%
Discussion	20%
Presentation (time, clarity)	15%
1. Manuscript	20%
Front matter: title, authors	
Abstract: complete, reviewed	
Format: pdf, images, tables, page limits	
Structure: organization	
Writing: spelling, grammar	
References: citations (also to images and tables), format, consistency	
Improvements over M1 + M2 manuscript	
Video presentation of the project	



3. Improvements	40%
Idea/hypothesis well defined	
Ideias explored: *The plan is to explore and evaluate one central idea in depth.* *You can also make smaller improvements.* *The list below is a list of possibilities.*	
- Semantic search	
- Aditional information sources	
- Aditional data processing	
- New signals	
- Query processing (relevance feedback, etc)	
- Entity exploration	
Solr features:	
- Term boost	
- Field boost	
- Independent boost	
- Phrase match	
- Proximity	
- More like this	
GUI interface implemented *Attention: if approved by your teacher.*	



4. Evaluation	40%
Individual assessments included (RNRRR...)	
Different setups compared	
Information needs reviewed (if needed)	
Description of manual evaluation process	
Precision metrics (P@, MAP)	
P-R curve with multiple setups	
Discussion	
Final Grade	100%
Delayed submission? -10% penalty	
Presentation	15%
Report and work developed	85%
-- Project evaluation items	
-- Teacher feedback (in-class work, discussions, …)	

Project Examples

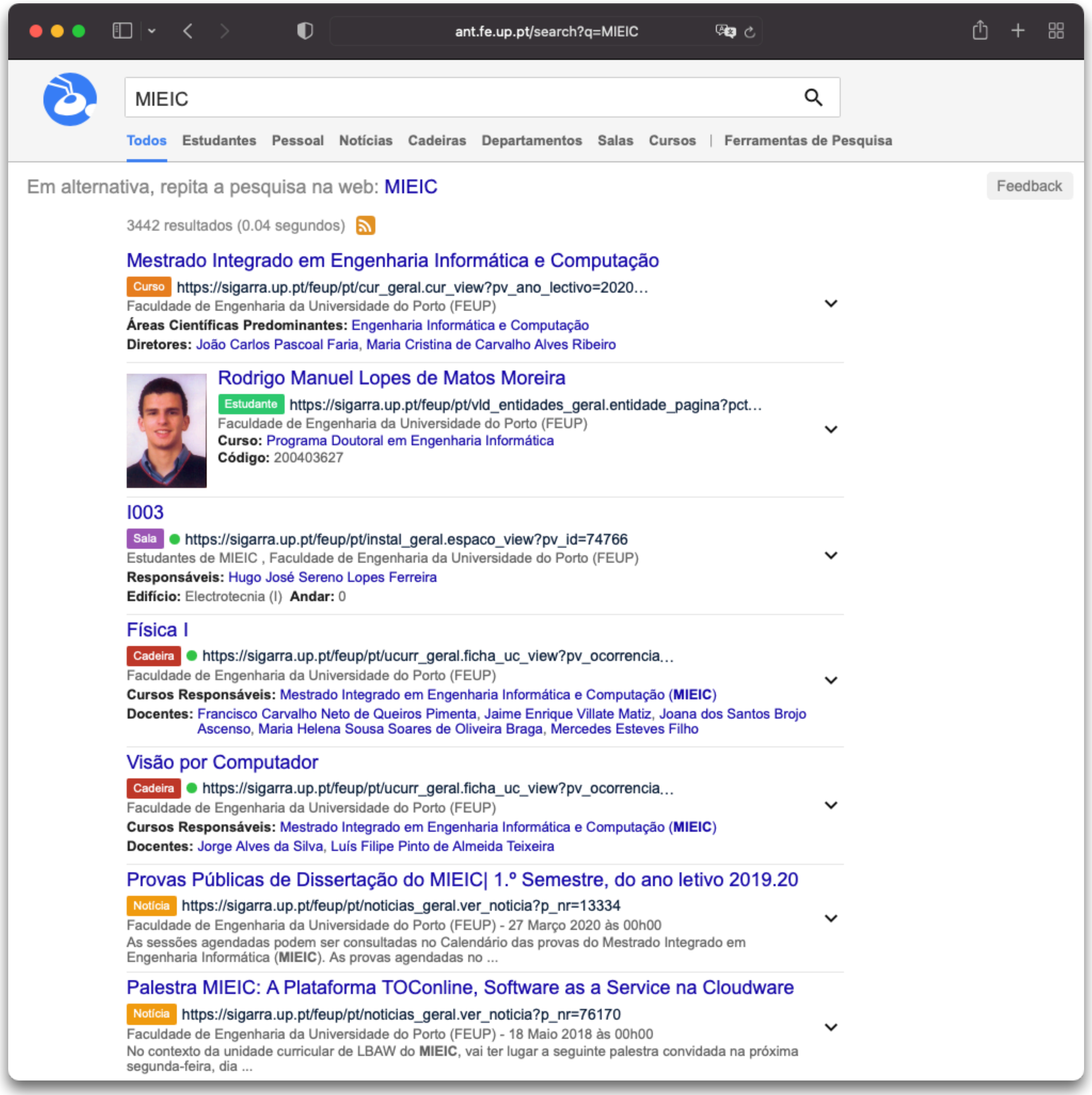
Topic Examples (1)

- Search over MIEIC / M.EIC dissertations:
  - Data sources: U.Porto Open Repository
  - (Fallback: scientific articles)
  - Collection: web scraping
  - Processing: PDF parsing
- Search over political parties web pages throughout history:
  - Data source: arquivo.pt API
  - Collection: JSON API requests
  - Processing: parsing HTML

Topic Examples (2)

- Search over movies subtitles:
  - Data sources: existing datasets + Wikipedia / Wikidata
  - Document: individual sentence; other ?
  - Document enrichment with: movies datasets; wikipedia pages, etc.
    - Include speaker / actor, time, movie, director. Many opportunities for filtering.
- Search for European universities:
  - Data sources: existing datasets + Wikipedia / Wikidata
  - Document: university info including city info
  - Document enrichment with: schools and courses?

ANT



Em alternativa, repita a pesquisa na web: MIEIC

3442 resultados (0.04 segundos)

Mestrado Integrado em Engenharia Informática e Computação
https://sigarra.up.pt/feup/pt/cur_geral.cur_view?pv_ano_lectivo=2020…
Faculdade de Engenharia da Universidade do Porto (FEUP)
Áreas Científicas Predominantes: Engenharia Informática e Computação
Diretores: João Carlos Pascoal Faria, Maria Cristina de Carvalho Alves Ribeiro

Rodrigo Manuel Lopes de Matos Moreira
https://sigarra.up.pt/feup/pt/vld_entidades_geral.entidade_pagina?pct…
Faculdade de Engenharia da Universidade do Porto (FEUP)
Curso: Programa Doutoral em Engenharia Informática
Código: 200403627

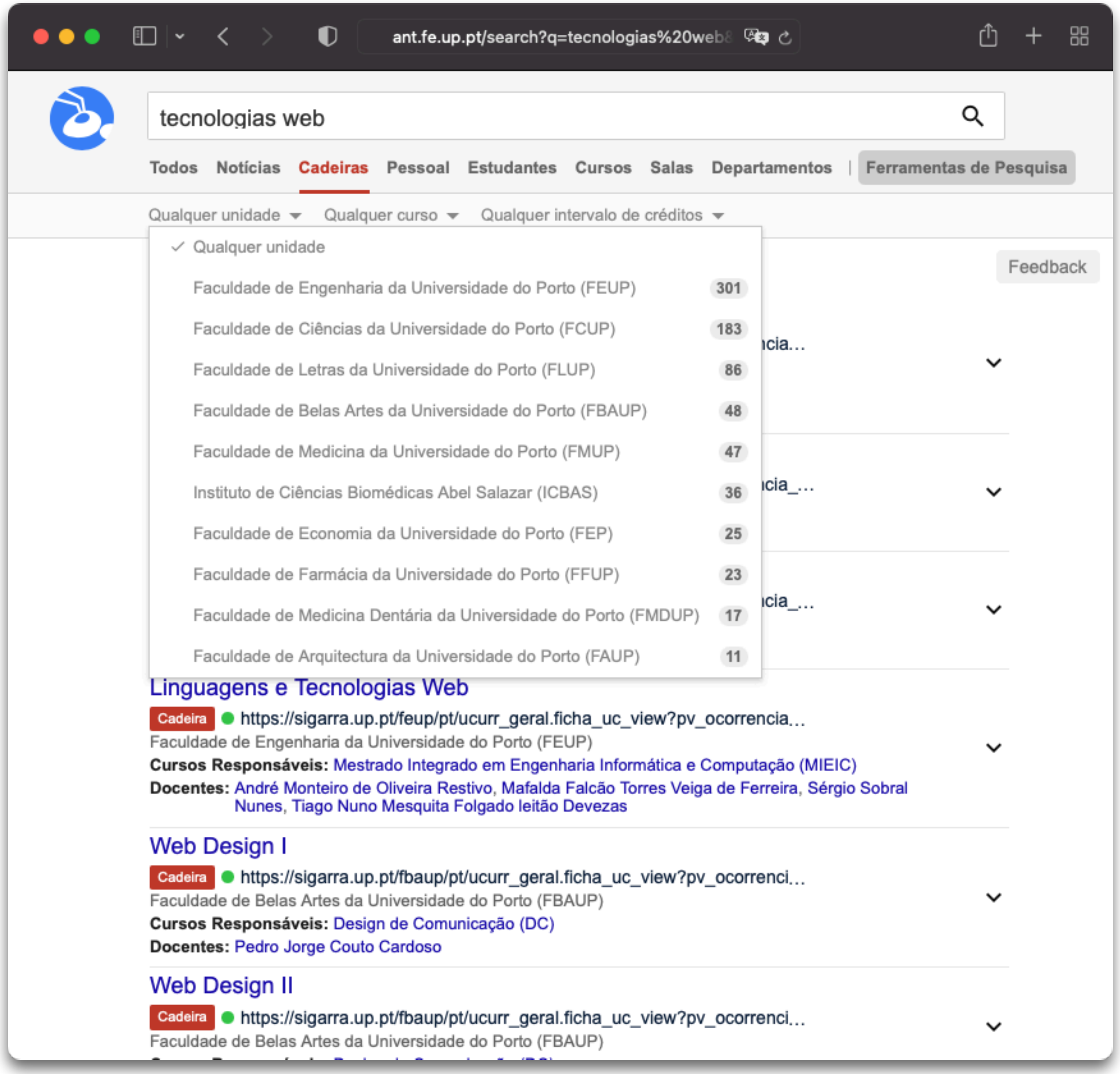
I003
https://sigarra.up.pt/feup/pt/instal_geral.espaco_view?pv_id=74766
Estudantes de MIEIC , Faculdade de Engenharia da Universidade do Porto (FEUP)
Responsáveis: Hugo José Sereno Lopes Ferreira
Edifício: Electrotecnia (I) **Andar:** 0

Física I
https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia…
Faculdade de Engenharia da Universidade do Porto (FEUP)
Cursos Responsáveis: Mestrado Integrado em Engenharia Informática e Computação (**MIEIC**)
Docentes: Francisco Carvalho Neto de Queiros Pimenta, Jaime Enrique Villate Matiz, Joana dos Santos Brojo Ascenso, Maria Helena Sousa Soares de Oliveira Braga, Mercedes Esteves Filho

Visão por Computador
https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia…
Faculdade de Engenharia da Universidade do Porto (FEUP)
Cursos Responsáveis: Mestrado Integrado em Engenharia Informática e Computação (**MIEIC**)
Docentes: Jorge Alves da Silva, Luís Filipe Pinto de Almeida Teixeira

Provas Públicas de Dissertação do MIEIC| 1.º Semestre, do ano letivo 2019.20
https://sigarra.up.pt/feup/pt/noticias_geral.ver_noticia?p_nr=13334
Faculdade de Engenharia da Universidade do Porto (FEUP) - 27 Março 2020 às 00h00
As sessões agendadas podem ser consultadas no Calendário das provas do Mestrado Integrado em Engenharia Informática (**MIEIC**). As provas agendadas no ...

Palestra MIEIC: A Plataforma TOConline, Software as a Service na Cloudware
https://sigarra.up.pt/feup/pt/noticias_geral.ver_noticia?p_nr=76170
Faculdade de Engenharia da Universidade do Porto (FEUP) - 18 Maio 2018 às 00h00
No contexto da unidade curricular de LBAW do **MIEIC**, vai ter lugar a seguinte palestra convidada na próxima segunda-feira, dia ...



✓ Qualquer unidade	
Faculdade de Engenharia da Universidade do Porto (FEUP)	301
Faculdade de Ciências da Universidade do Porto (FCUP)	183
Faculdade de Letras da Universidade do Porto (FLUP)	86
Faculdade de Belas Artes da Universidade do Porto (FBAUP)	48
Faculdade de Medicina da Universidade do Porto (FMUP)	47
Instituto de Ciências Biomédicas Abel Salazar (ICBAS)	36
Faculdade de Economia da Universidade do Porto (FEP)	25
Faculdade de Farmácia da Universidade do Porto (FFUP)	23
Faculdade de Medicina Dentária da Universidade do Porto (FMDUP)	17
Faculdade de Arquitectura da Universidade do Porto (FAUP)	11

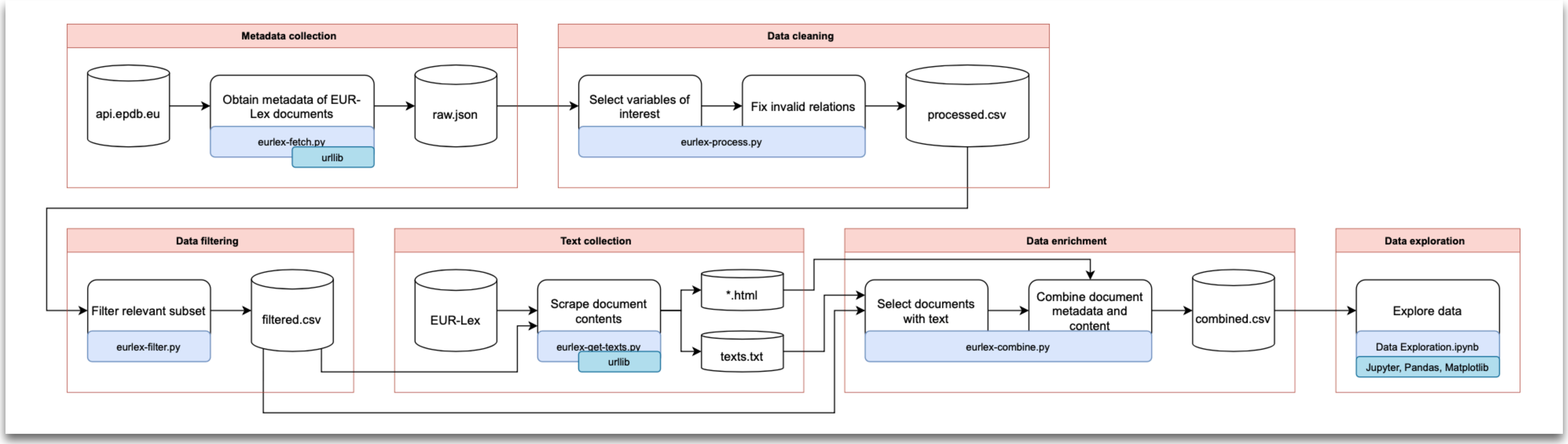
Linguagens e Tecnologias Web
https://sigarra.up.pt/feup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia…
Faculdade de Engenharia da Universidade do Porto (FEUP)
Cursos Responsáveis: Mestrado Integrado em Engenharia Informática e Computação (MIEIC)
Docentes: André Monteiro de Oliveira Restivo, Mafalda Falcão Torres Veiga de Ferreira, Sérgio Sobral Nunes, Tiago Nuno Mesquita Folgado leitão Devezas

Web Design I
https://sigarra.up.pt/fbaup/pt/ucurr_geral.ficha_uc_view?pv_ocorrenci…
Faculdade de Belas Artes da Universidade do Porto (FBAUP)
Cursos Responsáveis: Design de Comunicação (DC)
Docentes: Pedro Jorge Couto Cardoso

Web Design II
https://sigarra.up.pt/fbaup/pt/ucurr_geral.ficha_uc_view?pv_ocorrenci…
Faculdade de Belas Artes da Universidade do Porto (FBAUP)

Example

- *EUR-Lex is an European legislation database that offers access to European Union (EU) law, case-law by the Court of Justice of the EU and other public EU documents. This project aims to retrieve, process and prepare the data from this database, in order to create an information retrieval system of EU legislation. ~100 000 documents*
- *Example Information Needs: (1) Legislation related to Portugal; (2) Most recent regulation on Agriculture*



Example

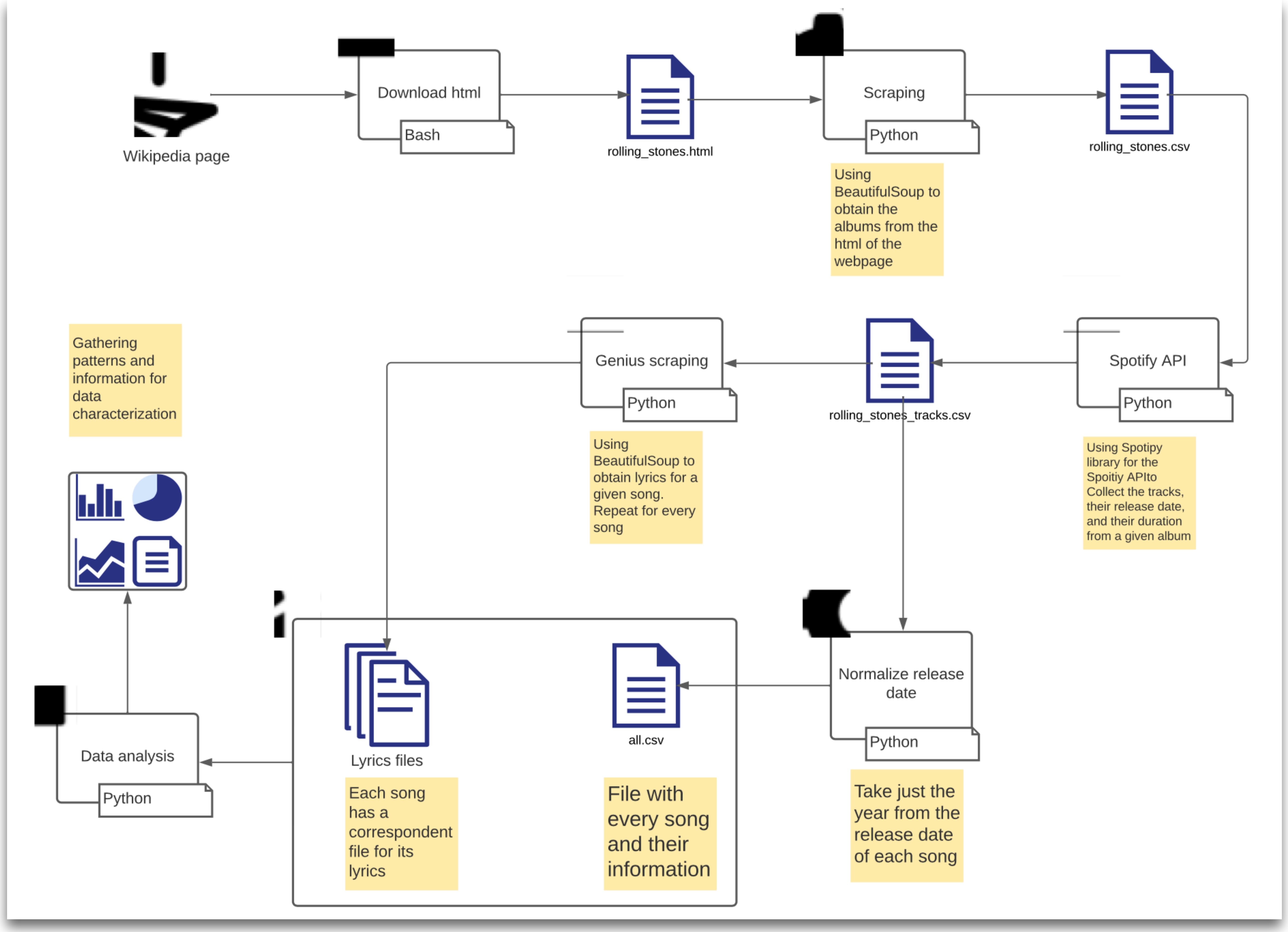
- *An Information Retrieval system about the 500 best music albums ever produced according to Rolling Stone magazine. Thanks to Wikipedia, Spotify and Genius it was possible to assemble the underlying data set with relevant attributes to identify and search for a song considering album and song information, metadata, and also lyrics.*

- *Three different data sources.*

- *Example Information Needs:*

  - *Looking for a song with a sentence like "I like her"*

  - *Songs from Rolling Stones about love.*



Using BeautifulSoup to obtain the albums from the html of the webpage

Using Spotipy library for the Spotiy APIto Collect the tracks, their release date, and their duration from a given album

Using BeautifulSoup to obtain lyrics for a given song. Repeat for every song

Gathering patterns and information for data characterization

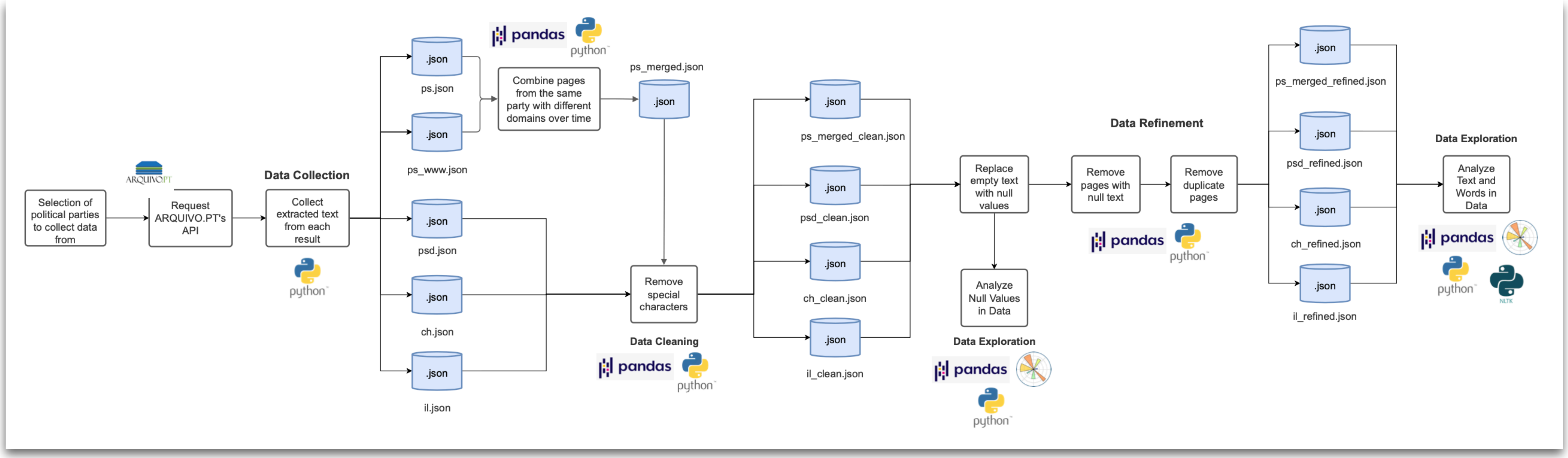
Each song has a correspondent file for its lyrics

File with every song and their information

Take just the year from the release date of each song

Example

- *Search in past political parties web pages.*
- *Example information needs:*
  - *Who was the candidate for party Bloco de Esquerda to the Presidential Elections in 2016?*
  - *What is the political position of parties regarding “global warming”?*



Selection of political parties to collect data from

ARQUIVO.PT

Request ARQUIVO.PT's API

Data Collection

Collect extracted text from each result

python

.json ps.json

.json ps_www.json

.json psd.json

.json ch.json

.json il.json

pandas python

Combine pages from the same party with different domains over time

ps_merged.json .json

Remove special characters

Data Cleaning

pandas python

.json ps_merged_clean.json

.json psd_clean.json

.json ch_clean.json

.json il_clean.json

Replace empty text with null values

Analyze Null Values in Data

Data Exploration

pandas python

Data Refinement

Remove pages with null text

Remove duplicate pages

pandas python

.json ps_merged_refined.json

.json psd_refined.json

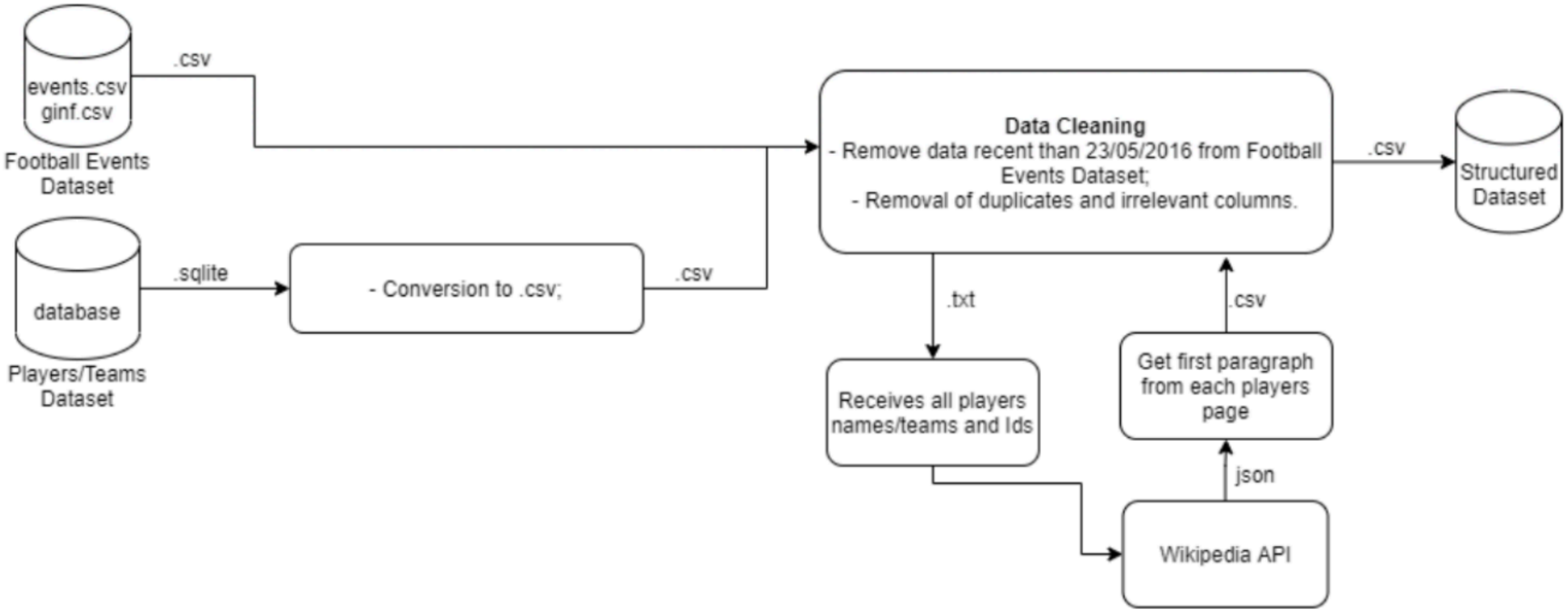
.json ch_refined.json

.json il_refined.json

Data Exploration

Analyze Text and Words in Data

pandas python NLTK



events.csv
ginf.csv

Football Events Dataset

.csv

database

Players/Teams Dataset

.sqlite

- Conversion to .csv;

.csv

Data Cleaning
- Remove data recent than 23/05/2016 from Football Events Dataset;
- Removal of duplicates and irrelevant columns.

.csv

Structured Dataset

.txt

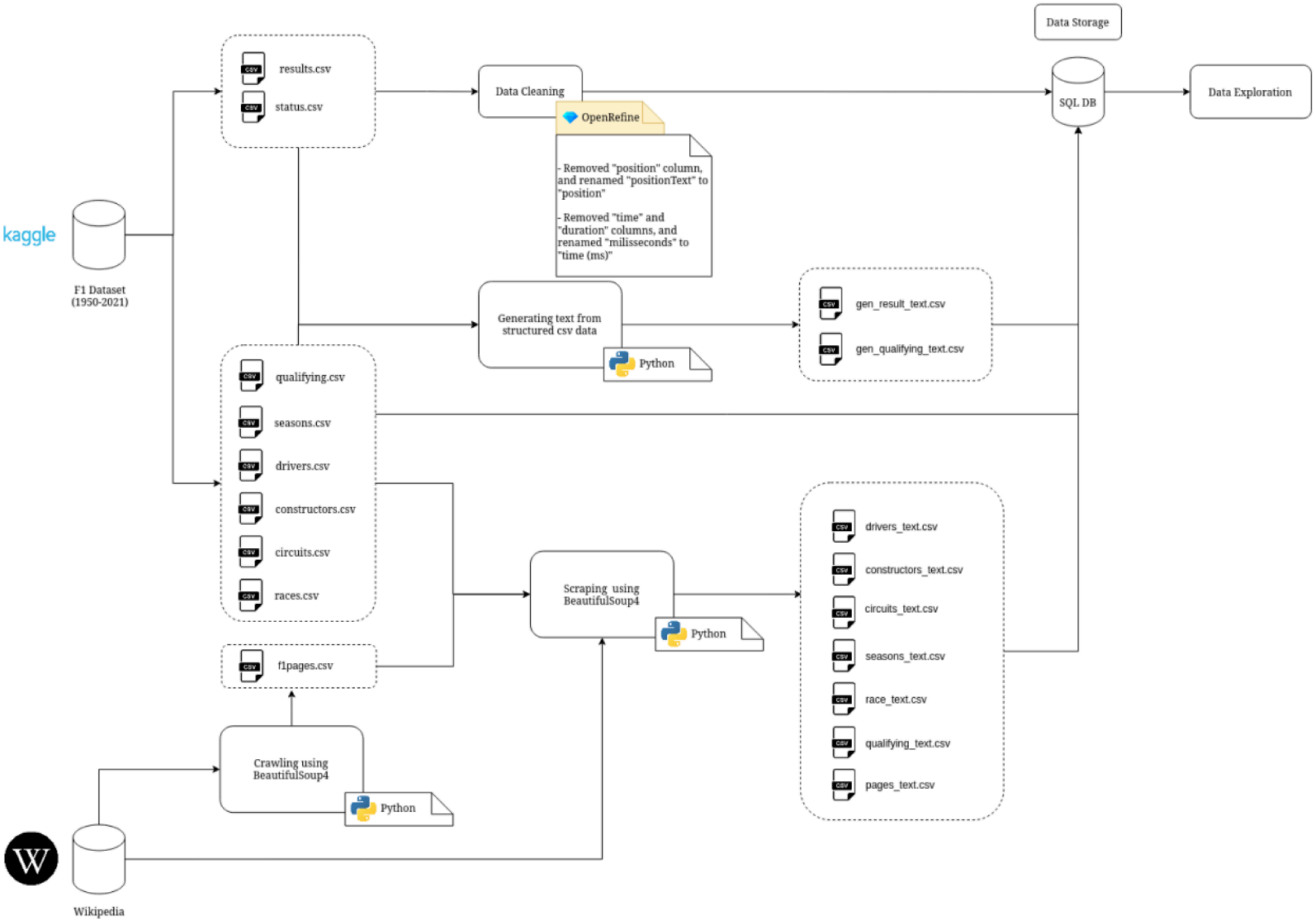
Receives all players names/teams and Ids

Wikipedia API

json

Get first paragraph from each players page

.csv



kaggle

F1 Dataset
(1950-2021)

results.csv

status.csv

Data Cleaning

OpenRefine

- Removed "position" column, and renamed "positionText" to "position"

- Removed "time" and "duration" columns, and renamed "milisseconds" to "time (ms)"

Data Storage

SQL DB

Data Exploration

Generating text from structured csv data

Python

gen_result_text.csv

gen_qualifying_text.csv

qualifying.csv

seasons.csv

drivers.csv

constructors.csv

circuits.csv

races.csv

f1pages.csv

Scraping using BeautifulSoup4

Python

drivers_text.csv

constructors_text.csv

circuits_text.csv

seasons_text.csv

race_text.csv

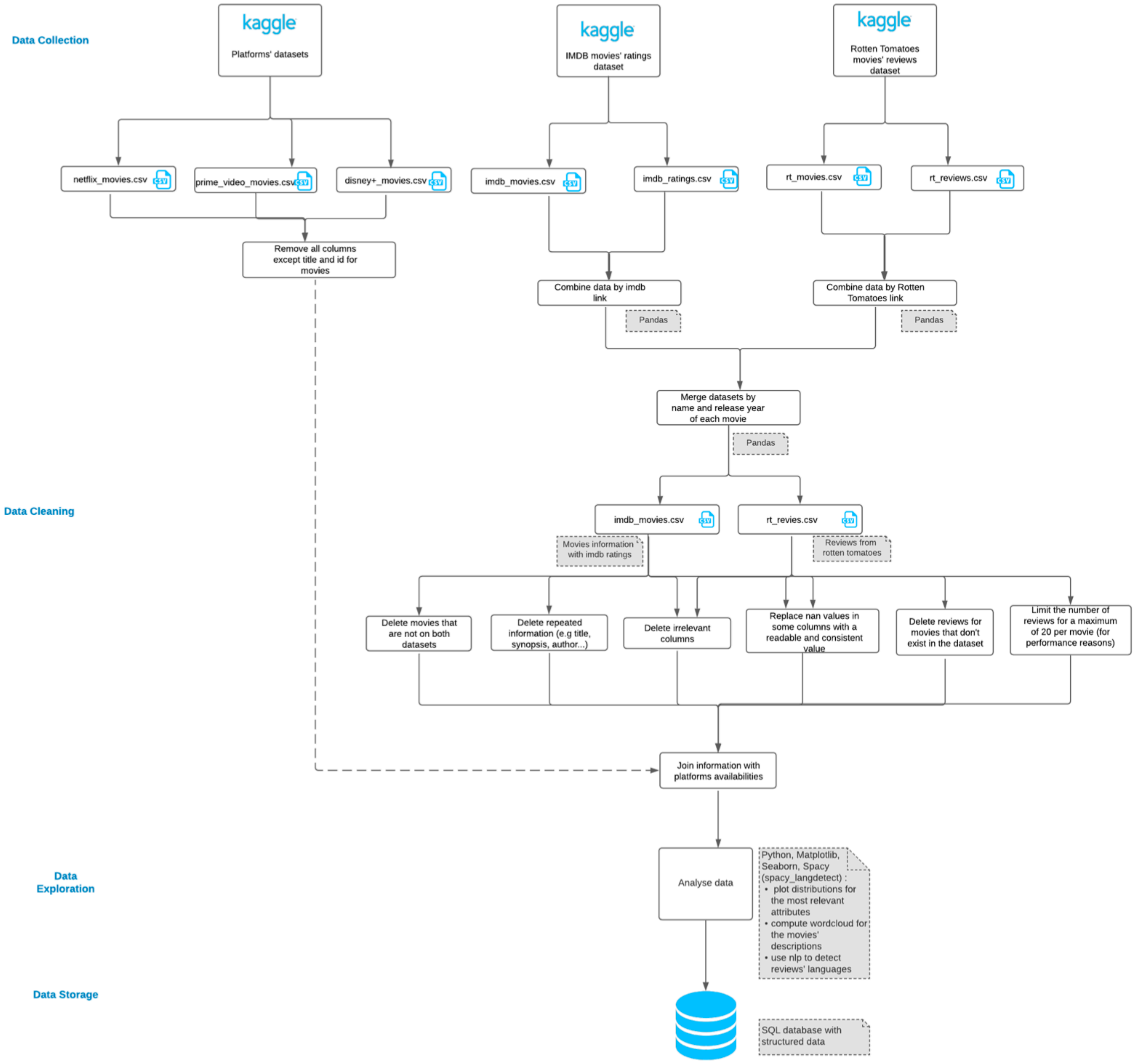
qualifying_text.csv

pages_text.csv

Crawling using BeautifulSoup4

Python

Wikipedia



Data Collection

kaggle
Platforms' datasets

kaggle
IMDB movies' ratings dataset

kaggle
Rotten Tomatoes movies' reviews dataset

netflix_movies.csv

prime_video_movies.csv

disney+_movies.csv

imdb_movies.csv

imdb_ratings.csv

rt_movies.csv

rt_reviews.csv

Remove all columns except title and id for movies

Combine data by imdb link

Pandas

Combine data by Rotten Tomatoes link

Pandas

Merge datasets by name and release year of each movie

Pandas

Data Cleaning

imdb_movies.csv

Movies information with imdb ratings

rt_revies.csv

Reviews from rotten tomatoes

Delete movies that are not on both datasets

Delete repeated information (e.g title, synopsis, author...)

Delete irrelevant columns

Replace nan values in some columns with a readable and consistent value

Delete reviews for movies that don't exist in the dataset

Limit the number of reviews for a maximum of 20 per movie (for performance reasons)

Join information with platforms availabilities

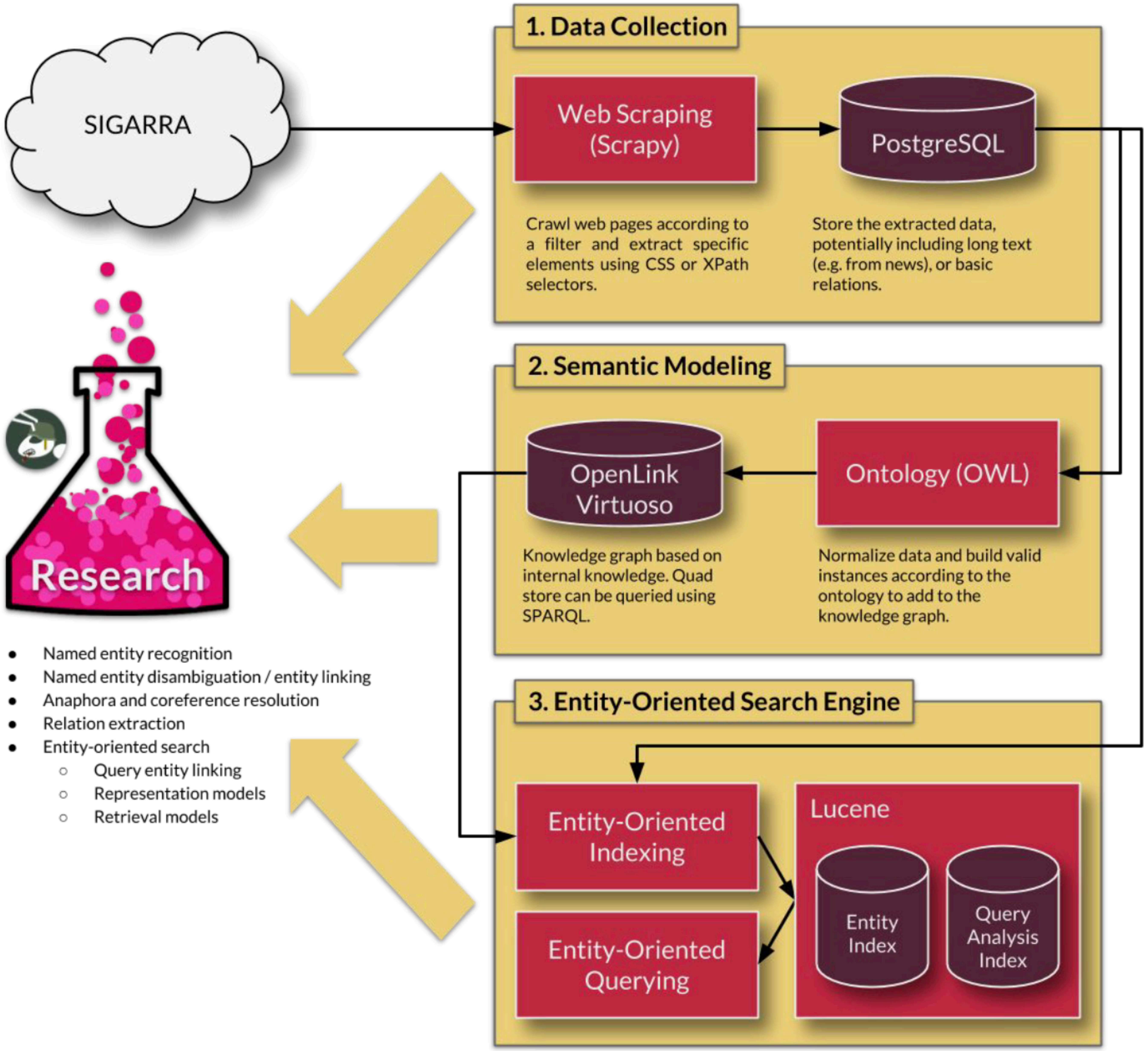
Data Exploration

Analyse data

Python, Matplotlib, Seaborn, Spacy (spacy_langdetect) :
- plot distributions for the most relevant attributes
- compute wordcloud for the movies' descriptions
- use nlp to detect reviews' languages

Data Storage

SQL database with structured data



SIGARRA

1. Data Collection

Web Scraping (Scrapy)

Crawl web pages according to a filter and extract specific elements using CSS or XPath selectors.

PostgreSQL

Store the extracted data, potentially including long text (e.g. from news), or basic relations.

2. Semantic Modeling

OpenLink Virtuoso

Knowledge graph based on internal knowledge. Quad store can be queried using SPARQL.

Ontology (OWL)

Normalize data and build valid instances according to the ontology to add to the knowledge graph.

3. Entity-Oriented Search Engine

Entity-Oriented Indexing

Entity-Oriented Querying

Lucene

Entity Index

Query Analysis Index

Research

- Named entity recognition
- Named entity disambiguation / entity linking
- Anaphora and coreference resolution
- Relation extraction
- Entity-oriented search
  - Query entity linking
  - Representation models
  - Retrieval models