PRI Presentation, 23/24 Edition

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

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Today's Plan

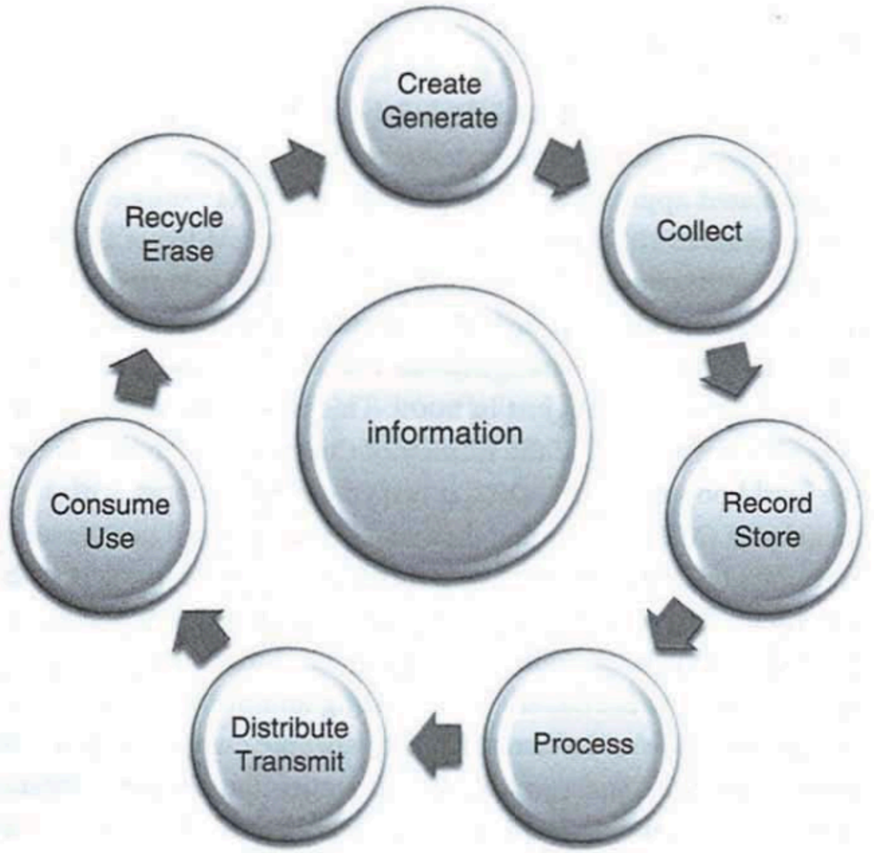
- Information processing and retrieval: context and motivation
- Course presentation: topics, materials, classes, evaluation
- Course projects: groups, themes, rules
- Q&A

Context and Motivation

Information Society

- Information communication technologies are ubiquitous in modern societies.
- An ever-growing number of activities depend on the ability to extract value from information.
- Human progress and welfare is largely dependent on an efficient management of the life cycle of information.
- New professional profiles: data engineer, data architect, data analyst, data scientist.
- This course is an introduction to information processing and to information retrieval.

Information Life Cycle



1. A typical information life cycle

Image from Information - A Very Short Introduction. Luciano Floridi, Oxford University Press, 2010

The Unreasonable Effectiveness of Data

- In 2009, Google researchers published a paper highlighting the virtues of data in successfully tacking complex language-related problems.

- *The Unreasonable Effectiveness of Data (2009) [link]*
  Alon Halevy, Perter Norvig, Fernando Pereira

  - *"[I]nvariably, simple models and a lot of data trump more elaborate models based on less data"*

  - *"simple n-gram models or linear classifiers based on millions of specific features perform better than elaborate models that try to discover general rules"*

An Era of Information Abundance

- **Information consumes attention** — leading to attention scarcity
  - *"In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."*
    Herbert A. Simon (1971!)

- **Information is complemented by analysis** — information growth results in growth in information analysis
  - *"The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."*
    Hal Varian (2009)

Information Processing and Retrieval Tasks

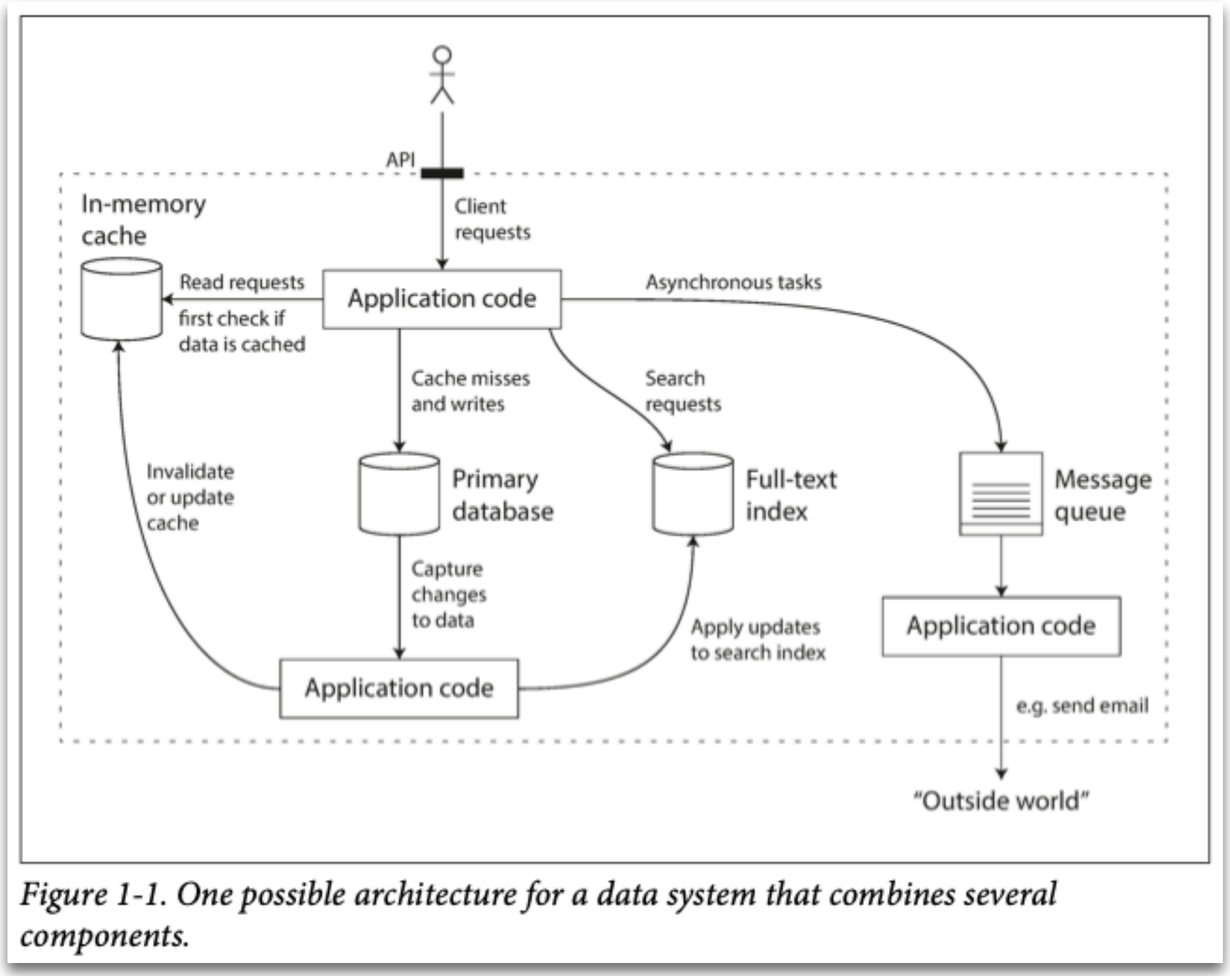
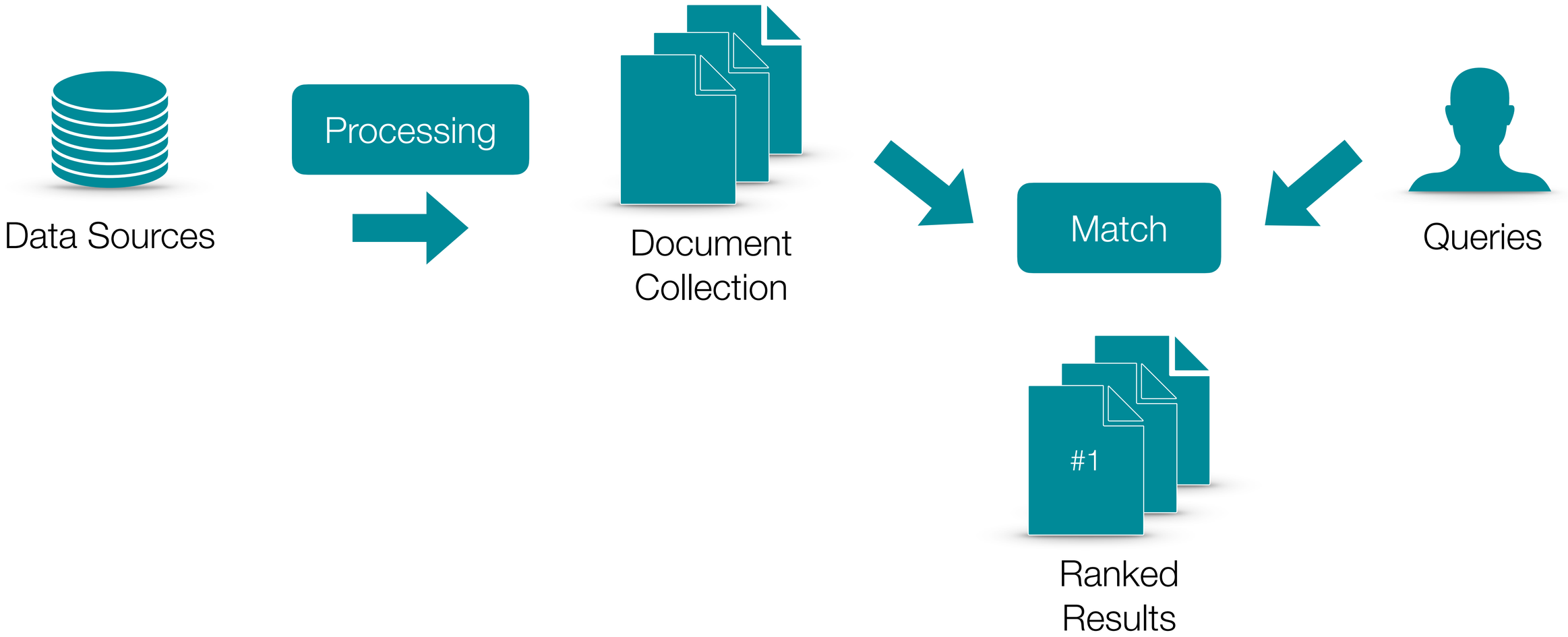


Figure 1-1. One possible architecture for a data system that combines several components.

- Classic workflow and typical tasks.
- Data Ingestion
  - Collect data
  - Describe data
  - Move data
- Data Transformation
  - Data modeling
  - Data migration
  - Pipeline orchestration
- Data Optimization
  - Selection, export, assessment

Image from Designing Data-Intensive Applications O'Reilly, 2017

Information Processing and Retrieval Pipeline



Information Processing

- This “world of data” is supported by numerous and heterogeneous components.
- Even a single system typically involves many complex and interlinked subsystems.
- We will only “scratch the surface on this area” and focus on information processing tasks related to search systems.
- Other related topics not covered: analytics, recommendation, reliability, etc.

Information Retrieval

- Information Retrieval if the field of Computer Science focused on
  - finding information that is **relevant**
  - to a user's **information need**.
- Managing and accessing information is a classical problem in computer science.
- Vannevar Bush authored "As We May Think" in 1945 (!) [link]
- Tim Berners-Lee proposed the Word Wide Web in 1989 to address information management challenges at CERN. [link]

Apple’s Knowledge Navigator (1987)

https://en.wikipedia.org/wiki/Knowledge_Navigator

Information Access

- Why not use pattern matching and structured languages (e.g. SQL)?
- Information Retrieval distinguishes itself from Data Retrieval in two central aspects:
  - Not limited to exact matching (e.g. multiple words, synonyms)
  - Is ordered by relevance, i.e. importance to the user's query (a central concept in IR).
- The web brought information retrieval to the center stage,
  but search systems can be found everywhere.
  - Web search, vertical search, enterprise search, desktop search, etc.

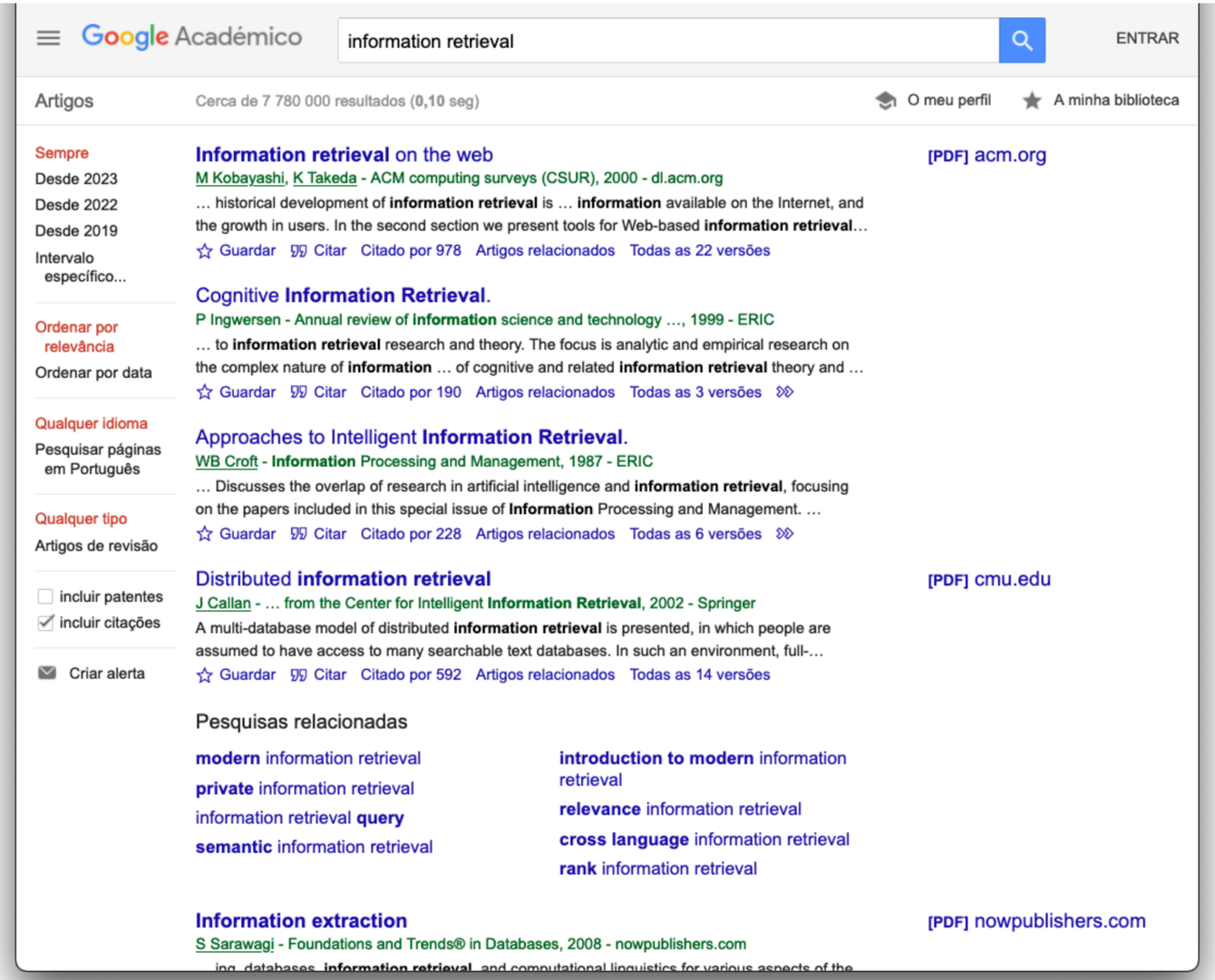
Ad Hoc Search

- “Ad hoc” search is the most common type of search task.
- “Ad hoc” refers to an independent search episode to retrieve information for an isolated information need. Contrast with a search.
- A common Google search is typically an “ad hoc search” task.

- Web search systems have significantly evolved since the 90s.
- See: A study on the evolution of user interfaces (MIEIC MSc, 2021)

Vertical Search

- Vertical search focuses on domain-specific information such as news, travel, music, sports, academic papers, etc.
- Usually very specific information needs, e.g. paper on a given topic published by researchers from a specific institution; flights between two destinations on a specific date.
- Results not limited to “documents”, e.g. flight information; a game result and list of participants.



Enterprise Search

- Enterprise Search is designed to allow search across an enterprise's content.
- High number and diversity of information sources, e.g. human-resources, document repository, projects, forums, etc.
- These are typically internal systems, thus outside the scope of general public search engines.

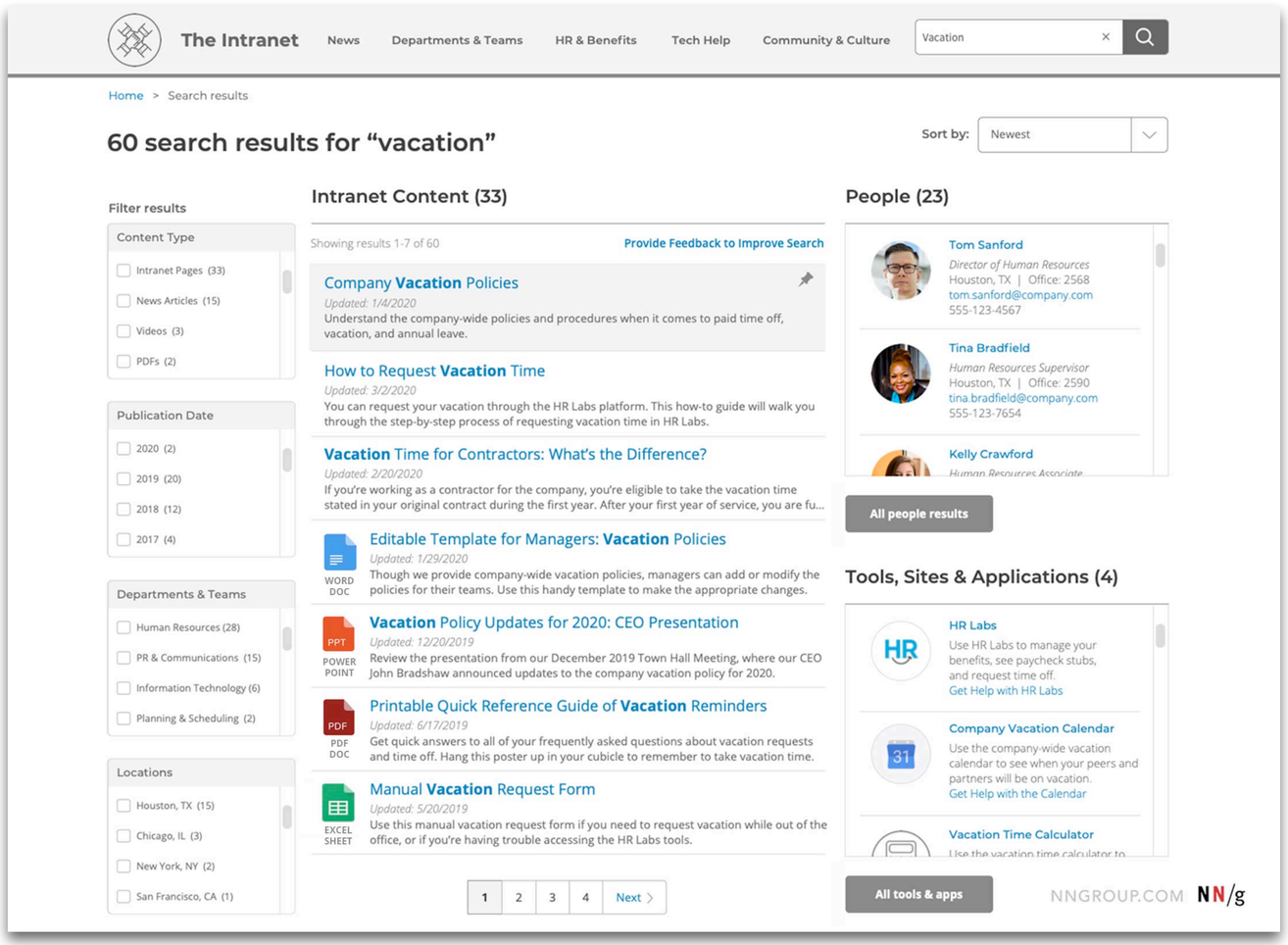
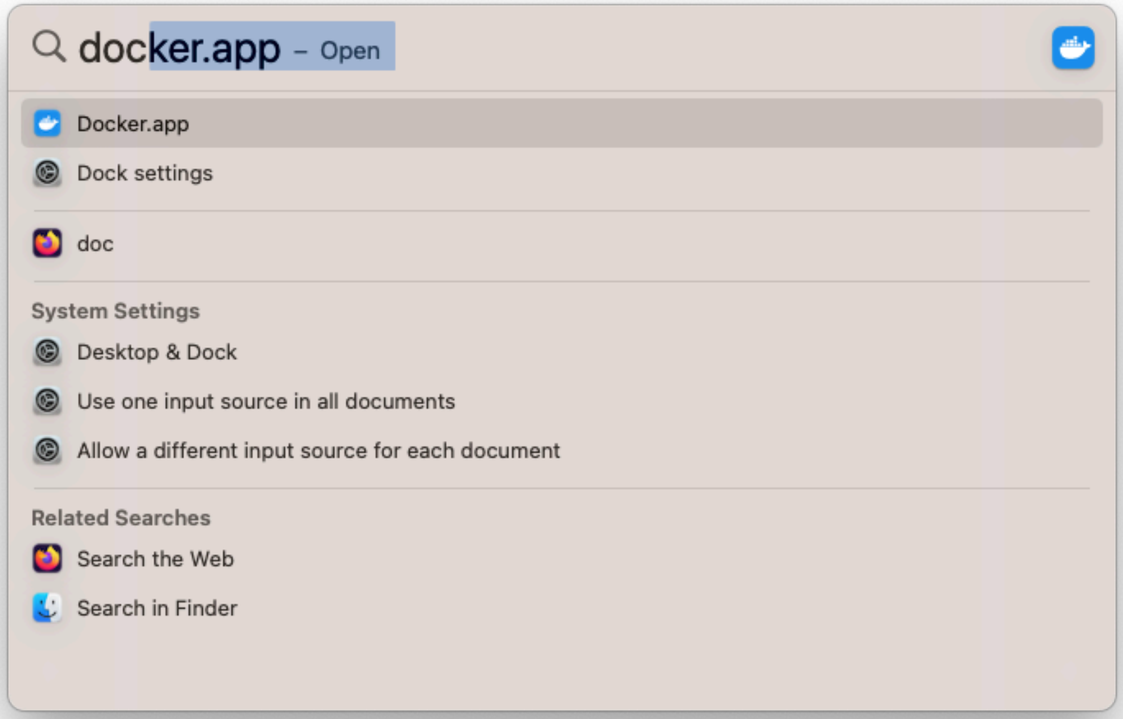


Image from Intranet-Search Essentials, NN Group (2022)

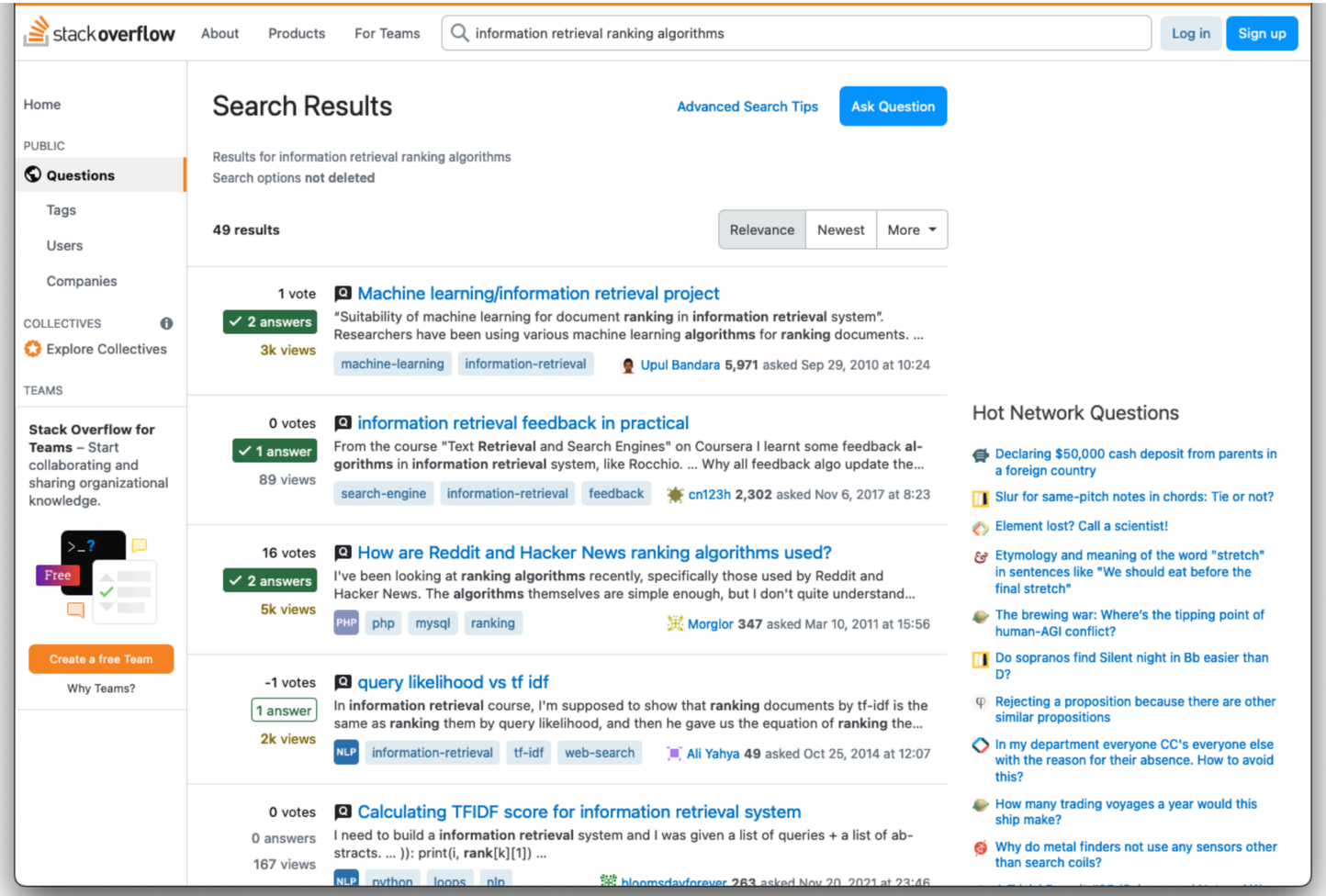
Desktop Search

- Desktop Search is a single-user system focused on searching the contents stored in a computer.
- The collection of documents includes files (text, multimedia), folders, settings, contacts, etc.
- Very different across operating systems.



macOS search feature.

App Specific Search



Course Presentation

PRI Team, 21/22 Edition

Sérgio Nunes
(regente)

João Damas

Sara Fernandes

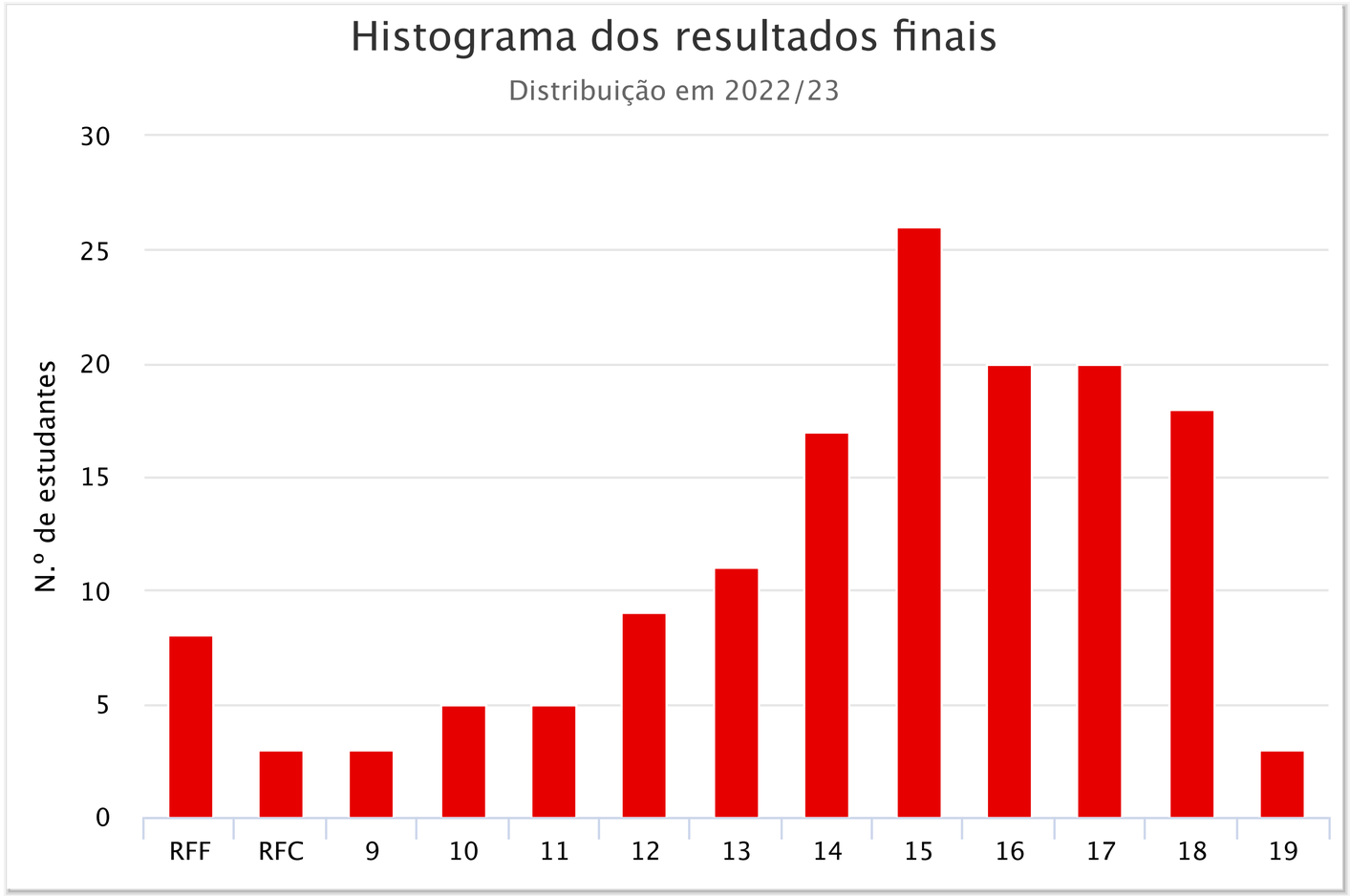
Daniel Garrido

PRI @ M.EIC

- M.EIC 1st-year course from the area of Information Systems
- 8 classes / ~180 students.

- Lectures will be on Mondays, 18h00, in person.
- Practical classes will be in person, throughout the week.

Previous Year (22/23 edition)



Course Objectives

- Prepare students to know, understand, design and develop solutions for information processing and retrieval
  - Make students aware of the challenges associated with building information search systems
  - Familiarize students with the main concepts and techniques associated with information processing and retrieval
  - Enable students to design, implement and evaluate information search systems on document collections

Learning Outcomes

- Identify and describe the main tasks associated with information processing and retrieval
- Describe the architecture and functioning of an information search system
- Describe the tasks associated with the processing phases of a collection (offline) and interrogation processing (online)
- Distinguish the different information retrieval models, identifying their principles, models for document representation, and similarity measures
- Describe and implement different techniques for indexing information
- Describe and implement different techniques for retrieving and ordering results

Information Processing Topics

- Data sources, data provenance and datasets
- Data acquisition and data exploration
- Data pipelines
- Data processing and extraction
- Data characterization

Information Retrieval Topics

- Field of Information Retrieval: history, basic concepts and tools
- Architecture of IR Systems: indexing and retrieval processes
- IR Models: ranking, boolean model, vector space model
- Evaluation in IR: methods and metrics
- Web IR: link analysis, classic algorithms
- Also: Neural IR, Entity-Oriented Search, User Interfaces for Search, Applications

Classes

- Lectures, 2h
  - Topic presentation and discussion.
- Practical, 2h
  - Brief presentation of working guides and tutorials
  - Status report with each group
  - 3 milestone presentations on selected dates

Evaluation

- Final grade =
  - 60% Group Project +
  - 40% Exam
- Group Project =
  - 20% Data Processing (M1) +
  - 40% Information Retrieval (M2) +
  - 40% Search System (M3)

- The final grade of the project can vary between members of the same group, by plus or minus 3 values, based on the opinion of the teachers and in the self- and hetero- assessment to be carried out internally in each group.
- Minimum grade of 40% (8) required (**but not sufficient!**) in the exam and in each milestone.
- Minimum grade of 50% (10) required in the final individual project evaluation.

Semester Plan

Week	Lecture (2h)	Lab (2h)	Activities / Deliveries
1 (11 Set)	PRI Presentation; Projects and Themes.	*No lab classes.*	
2 (18 Set)	Datasets (1): Sources; Collection; Storage.	Student Groups Setup; Themes Exploration.Approve Datasets (deadline).	
3 (25 Set)	Datasets (2): Preparation; Pipelines.	Data Acquisition and Exploration.	
4 (2 Out)	Datasets (3): Processing; Characterization; Extraction.	Data Processing; Characterization.	
5 (9 Out)	Information Retrieval (1): Basics; Tools; Relevance.	M1 Presentations.	**M1: Data Processing**
6 (16 Out)	Information Retrieval (2): Solr; Evaluation.	Solr Practice.	
7 (23 Out)	Information Retrieval (3): Models; Indexing.	Information Indexing and Retrieval.	
— (30 Out)	*FEUP Week*		
8 (6 Nov)	Information Retrieval (4): Retrieval; Web.	Evaluation.	
9 (13 Nov)	Information Retrieval (5): Query Processing.	M2 Presentations.	**M2: Information Retrieval**
10 (20 Nov)	Information Retrieval (6): Neural IR.	Project Development.	
11 (27 Nov)	Information Retrieval (7): Entity-Oriented Search.	Project Development.	
12 (4 Dez)	Information Retrieval (8): Search User Interfaces.	Project Development.	
13 (11 Dez)	IR Applications or Invited Talk / Projects	M3 Presentations	**M3: Search System**

Main Bibliography

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.
  Introduction to Information Retrieval. Cambridge University Press, 2008
  https://nlp.stanford.edu/IR-book/information-retrieval-book.html

- W. Bruce Croft, Donald Metzler, Trevor Strohman.
  Search Engines: Information Retrieval in Practice. Pearson Education, 2015
  https://ciir.cs.umass.edu/irbook/

- Kristian Balog. Entity-Oriented Search. Springer, 2018
  http://link.springer.com/978-3-319-93935-3

- Martin Kleppmann. Designing Data-Intensive Applications. O'Reilly, 2017

- Jeroen Janssens. Data Science at the Command Line, 2nd Edition, O'Reilly, 2021
  https://www.datascienceatthecommandline.com/2e/

ChatGPT Survey

- New technology on the block.
- Understanding and being proficient in these tools is an emerging skill.
- But we also need to have clear expectations to avoid misunderstandings.

- Short survey to understand how it is being used and your perspective on it.

ChatGPT Policy

- Use of ChatGPT (or other generative AI) technologies
  - is permitted in specific contexts
  - and with clear acknowledgements.
- Possible uses in the context of PRI:
  - Brainstorm project topic ideas
  - Improve text (not write text from scratch)
  - Improve or debug code.
  - Other? *When in doubt ask in Moodle.*
- In the end, what you submit must be of your authorship.
- If you use these tools, add a annex to your report describing in detail how you have used them, namely the services and prompts used. This annex does not count towards the page limit.

Group Project

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.

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).

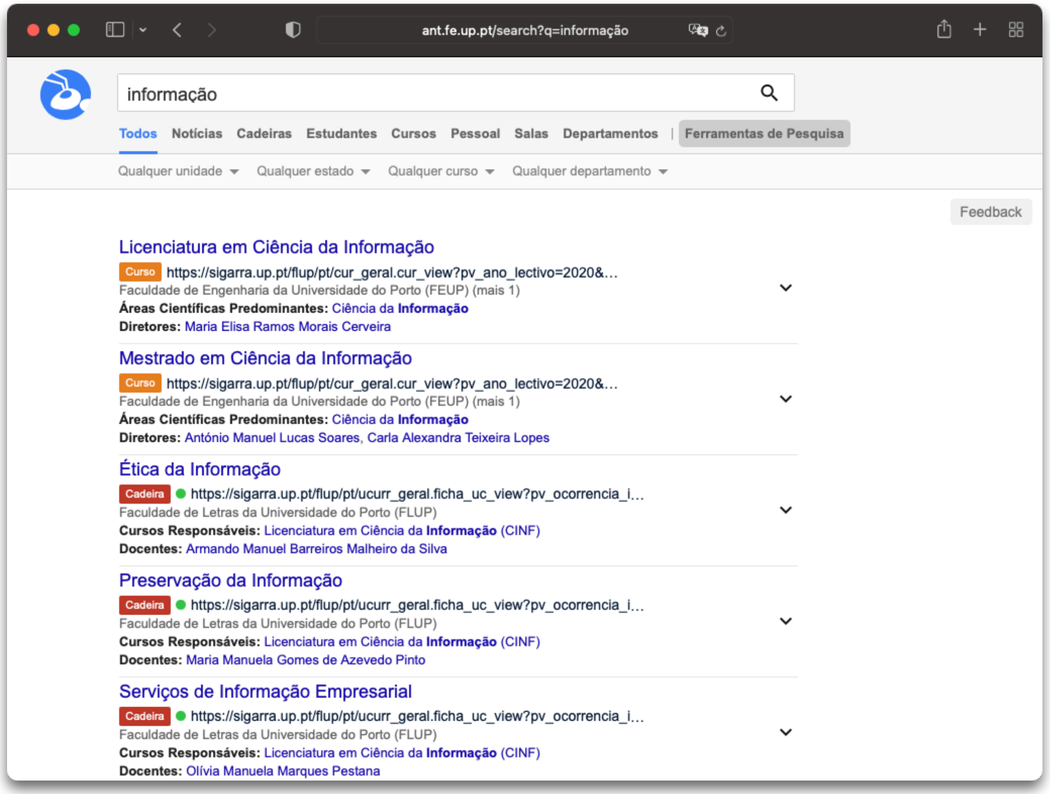
- Many possibilities: education, sports, law, government, media ...

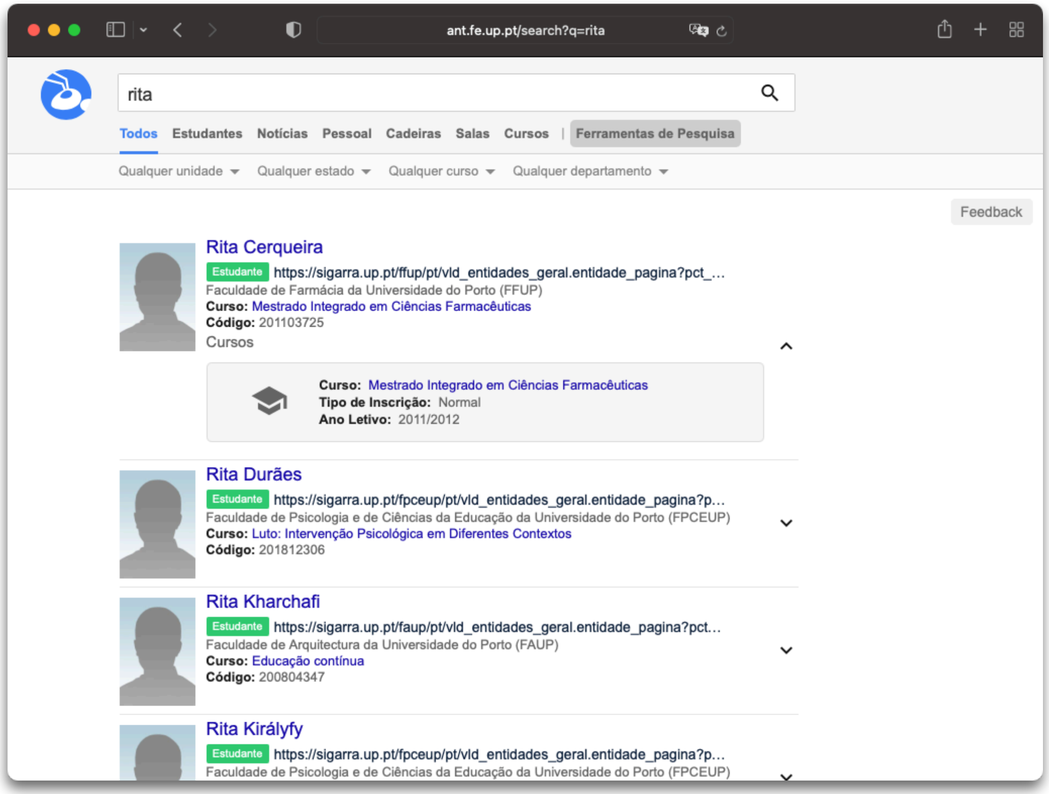
Working in Groups

- 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.

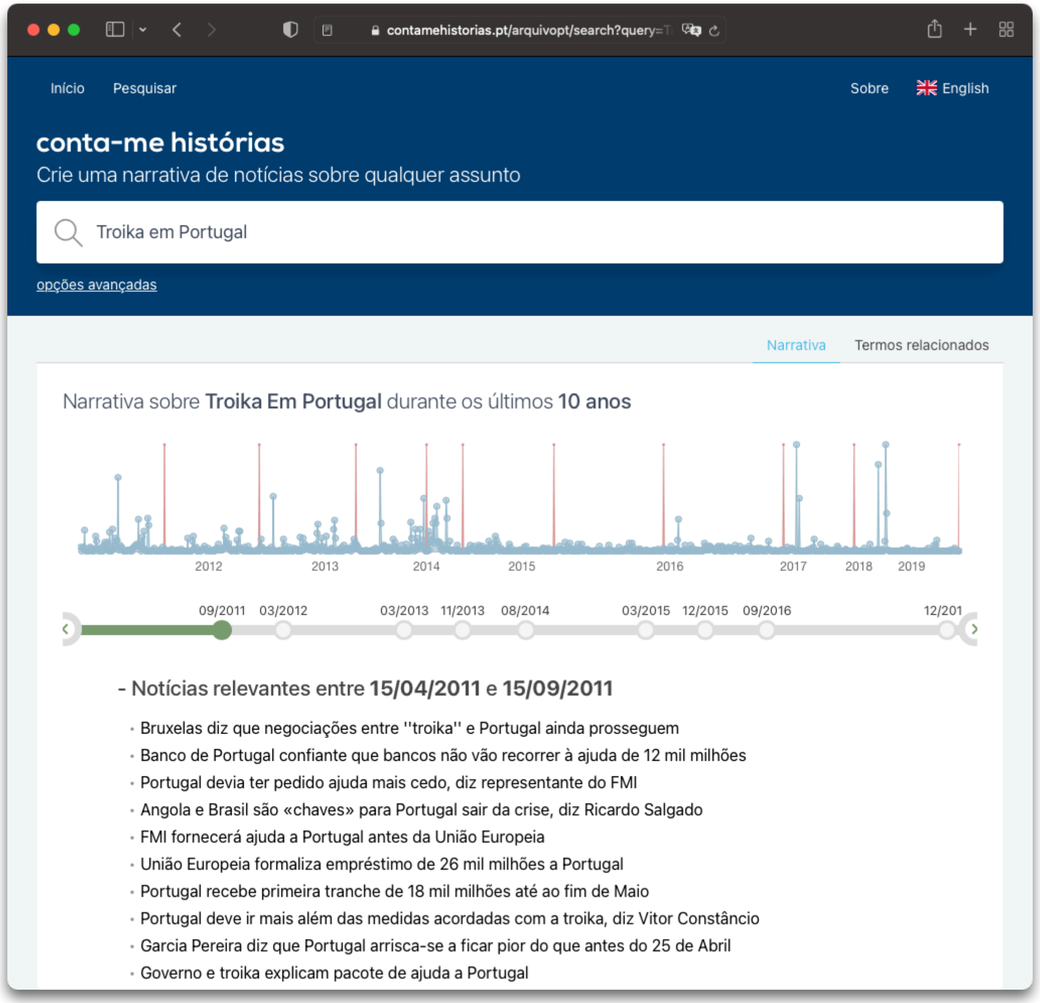
Project Examples

ANT


ant.fe.up.pt/search?q=informação
informação
Todos Notícias Cadeiras Estudantes Cursos Pessoal Salas Departamentos | Ferramentas de Pesquisa
Qualquer unidade Qualquer estado Qualquer curso Qualquer departamento
Feedback
Licenciatura em Ciência da Informação
Curso https://sigarra.up.pt/flup/pt/cur_geral.cur_view?pv_ano_lectivo=2020&…
Faculdade de Engenharia da Universidade do Porto (FEUP) (mais 1)
Áreas Científicas Predominantes: Ciência da Informação
Diretores: Maria Elisa Ramos Morais Cerveira
Mestrado em Ciência da Informação
Curso https://sigarra.up.pt/flup/pt/cur_geral.cur_view?pv_ano_lectivo=2020&…
Faculdade de Engenharia da Universidade do Porto (FEUP) (mais 1)
Áreas Científicas Predominantes: Ciência da Informação
Diretores: António Manuel Lucas Soares, Carla Alexandra Teixeira Lopes
Ética da Informação
Cadeira https://sigarra.up.pt/flup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_i…
Faculdade de Letras da Universidade do Porto (FLUP)
Cursos Responsáveis: Licenciatura em Ciência da Informação (CINF)
Docentes: Armando Manuel Barreiros Malheiro da Silva
Preservação da Informação
Cadeira https://sigarra.up.pt/flup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_i…
Faculdade de Letras da Universidade do Porto (FLUP)
Cursos Responsáveis: Licenciatura em Ciência da Informação (CINF)
Docentes: Maria Manuela Gomes de Azevedo Pinto
Serviços de Informação Empresarial
Cadeira https://sigarra.up.pt/flup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_i…
Faculdade de Letras da Universidade do Porto (FLUP)
Cursos Responsáveis: Licenciatura em Ciência da Informação (CINF)
Docentes: Olívia Manuela Marques Pestana


ant.fe.up.pt/search?q=rita
rita
Todos Estudantes Notícias Pessoal Cadeiras Salas Cursos | Ferramentas de Pesquisa
Qualquer unidade Qualquer estado Qualquer curso Qualquer departamento
Feedback
Rita Cerqueira
Estudante https://sigarra.up.pt/ffup/pt/vld_entidades_geral.entidade_pagina?pct_…
Faculdade de Farmácia da Universidade do Porto (FFUP)
Curso: Mestrado Integrado em Ciências Farmacêuticas
Código: 201103725
Cursos
Curso: Mestrado Integrado em Ciências Farmacêuticas
Tipo de Inscrição: Normal
Ano Letivo: 2011/2012
Rita Durães
Estudante https://sigarra.up.pt/fpceup/pt/vld_entidades_geral.entidade_pagina?p…
Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP)
Curso: Luto: Intervenção Psicológica em Diferentes Contextos
Código: 201812306
Rita Kharchafi
Estudante https://sigarra.up.pt/faup/pt/vld_entidades_geral.entidade_pagina?pct…
Faculdade de Arquitectura da Universidade do Porto (FAUP)
Curso: Educação contínua
Código: 200804347
Rita Királyfy
Estudante https://sigarra.up.pt/fpceup/pt/vld_entidades_geral.entidade_pagina?p…
Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP)

Conta-me Histórias



Topics from Previous Editions

- Electoral manifests
- Political parties webpages
- Portuguese laws
- Portuguese judicial decisions
- MIEIC / M.EIC MSc dissertations
- Linux packages
- …

- *A lot of books, musics and recipes, that we will try to avoid.*

Prémio Arquivo.pt

- *O objetivo é galardoar trabalhos inovadores que utilizem informação histórica preservada da Web.*
- *Estes trabalhos deverão utilizar os serviços de pesquisa e acesso disponibilizados publicamente pelo Arquivo.pt e demonstrar claramente a utilidade do serviço.*
- *Não há restrição de tema. Porém, é necessário ter o Arquivo.pt como fonte primordial de informação.*
- *O concurso está aberto a todos os interessados, a título individual ou em grupo.*

- https://arquivo.pt/premios

3rd Prize in Prémio Arquivo.pt 2023

- A project for the 2022/23 edition of PRI won the 3rd prize in Prémio Arquivo.pt 2023.
- More information at: Conheça os vencedores do Prémio Arquivo.pt 2023!

Data Sources

- A diversity of data sources exist.
- Things to keep in mind when choosing data sources:
  - Direct access, i.e. not dependent on third party actions
  - Volume, ideally thousands
  - Rich in textual data, i.e. long texts, not just labels or titles
  - If using an already prepared dataset, you need to combine two data sources
- Moodle > Data Sources

Resources

Materials

- The course's web page in Moodle is the starting point
- For each lecture and lab class an information page is available
- Moodle is also used for:
  - Group registration
  - Announcements and discussion
  - Submission of milestone materials (report and presentation)
- Slack:
  - Last minute warnings (rare), and class and in-group communication

PRI Tutorials

- You have access to a collection of tutorials to introduce concepts and tools, and guide your explorations during the semester.
- https://git.fe.up.pt/pri/tutorials
- Start this week with the first tutorial — Command Line Practice

- *A work in progress. Your feedback is welcomed.*

Next steps

- Answer 'PRI Survey' (if you haven't done so)
- Read the project rules
- Prepare for the first practical class:
  - organize groups before class (4 students) — register in Moodle (you can change later)
  - explore data sources and contexts to identify datasets
- Explore the first PRI tutorial — Command Line

- First delivery in four weeks (October 9th) - Milestone 1: Data Processing.

Questions or comments?