# Information Retrieval

PRI 22/23 · Information Processing and Retrieval
M.EIC · Master in Informatics Engineering and Computation

Sérgio Nunes
Dept. Informatics Engineering
FEUP · U.Porto

Based on Chapter 1 from Modern Information Retrieval, Baeza-Yates, R. et al. (2011)
Based on Chapters 1 and 2 from Search Engines in Practice, Croft, B. et al. (2008)

# Project Milestone 1 — Data Preparation

# Project M1: Data Preparation

- M1 presentations are this week.

- Deadline for submitting your materials is the day before the lab class (until 18h).
- Presentation schedules are tight.
  - Be there on time.
  - 5 min presentations + 5 min discussion.
- Don’t forget the individual participation component (10%) — ask questions!

# 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

# Today’s Lecture Outline

- Search systems
- Information retrieval concepts
  - Collections, documents, queries, results
  - Tasks, test collections, measures
  - Users, information needs, evaluation
- Architecture of information retrieval systems
  - Components of an information retrieval system
- History of information retrieval
  - Relation to text and data mining, machine learning, natural language processing

# Search systems

# Search and Information Retrieval

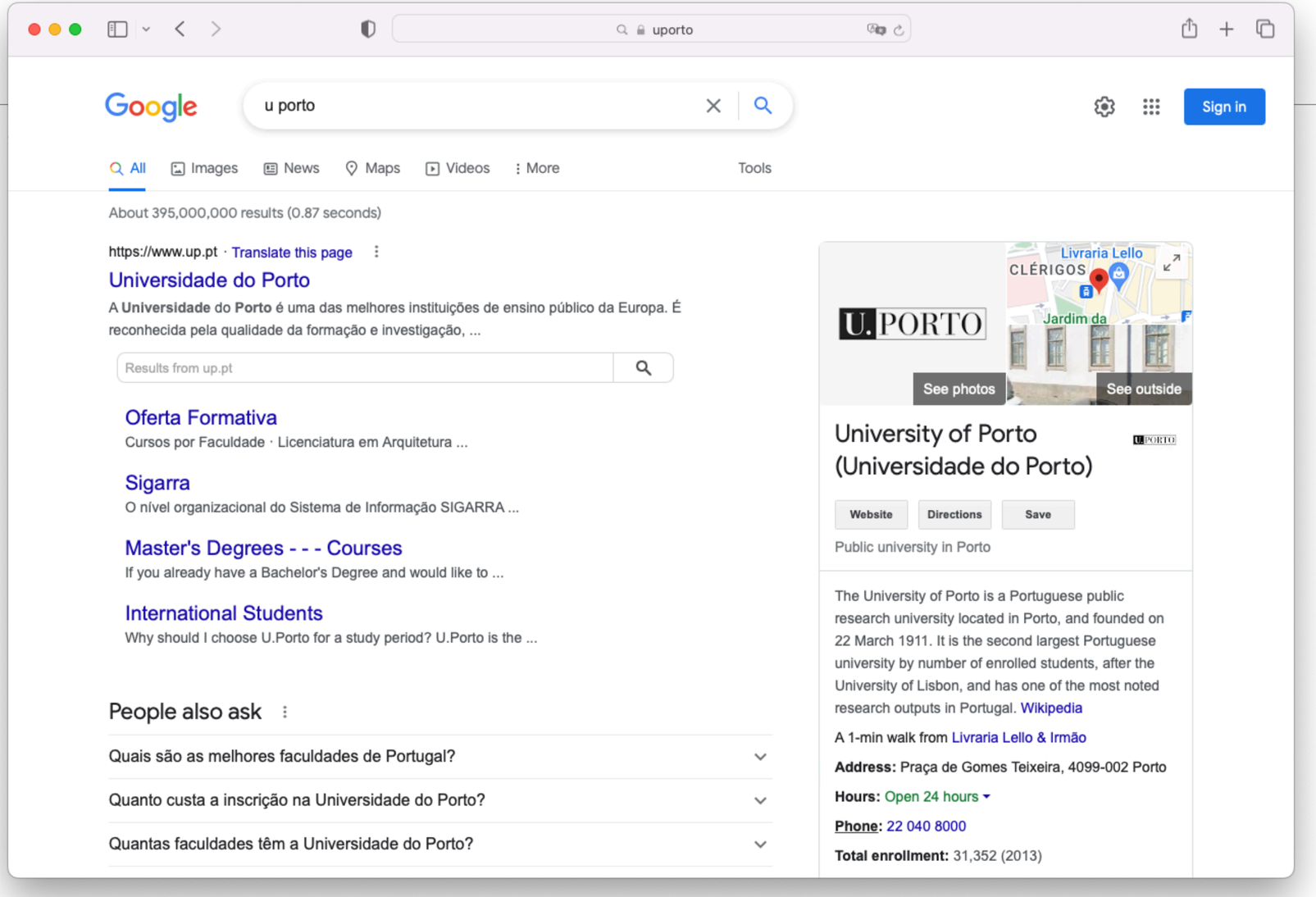
➔ Search is a daily activity for every computer, web, smart phone, … user.

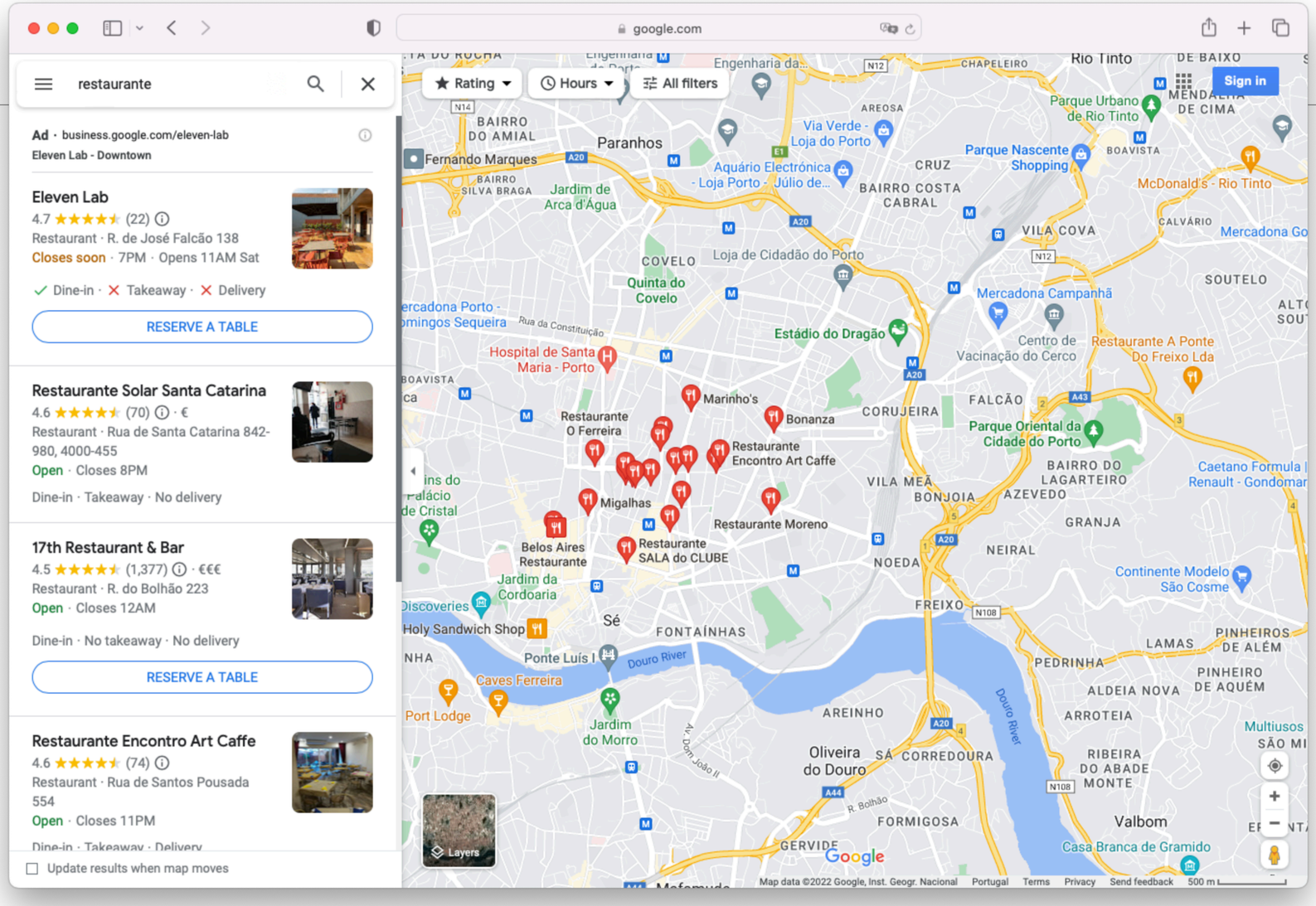
➔ Search exists, *and is expected*, in every application.

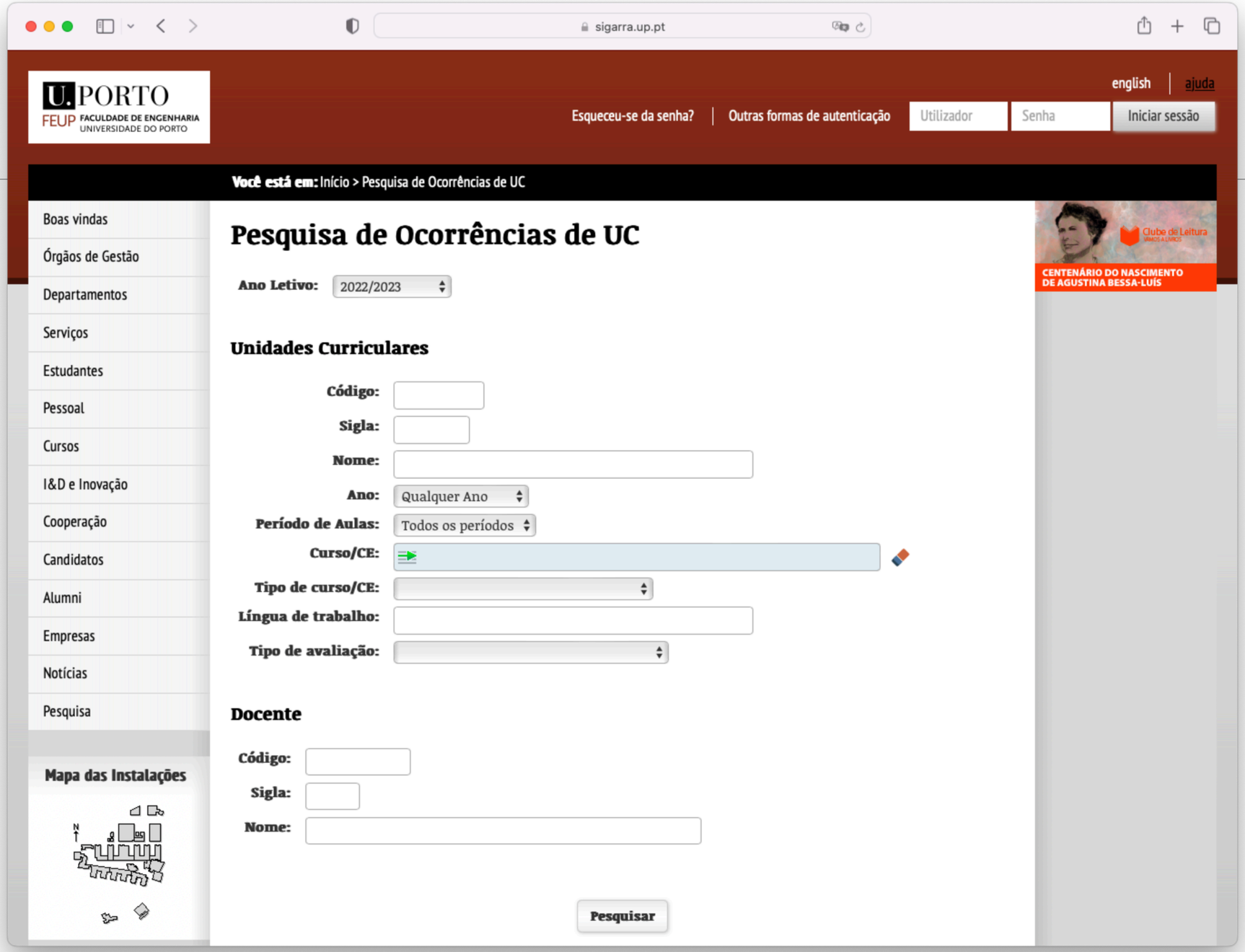
➔ The web has brought the concept of keyword-based search to the center stage.

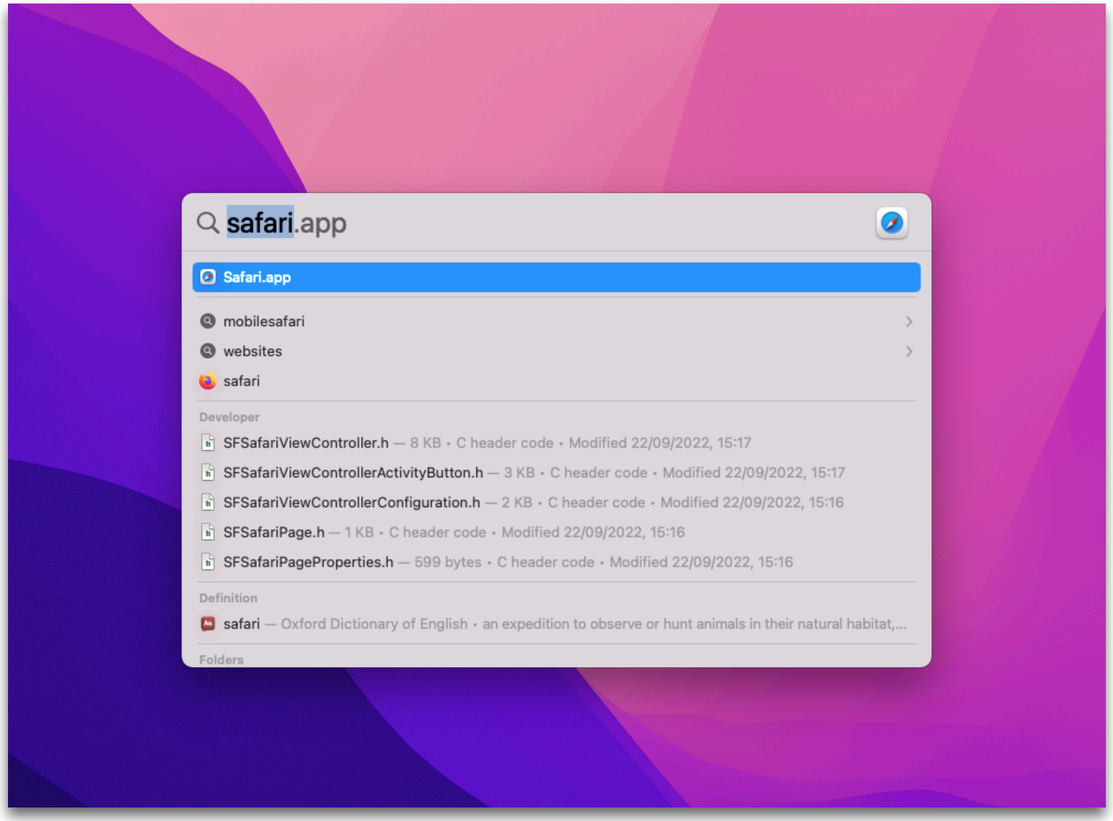
➔ Search contexts: web, site-specific, drive, computer, folders, email, messaging system.

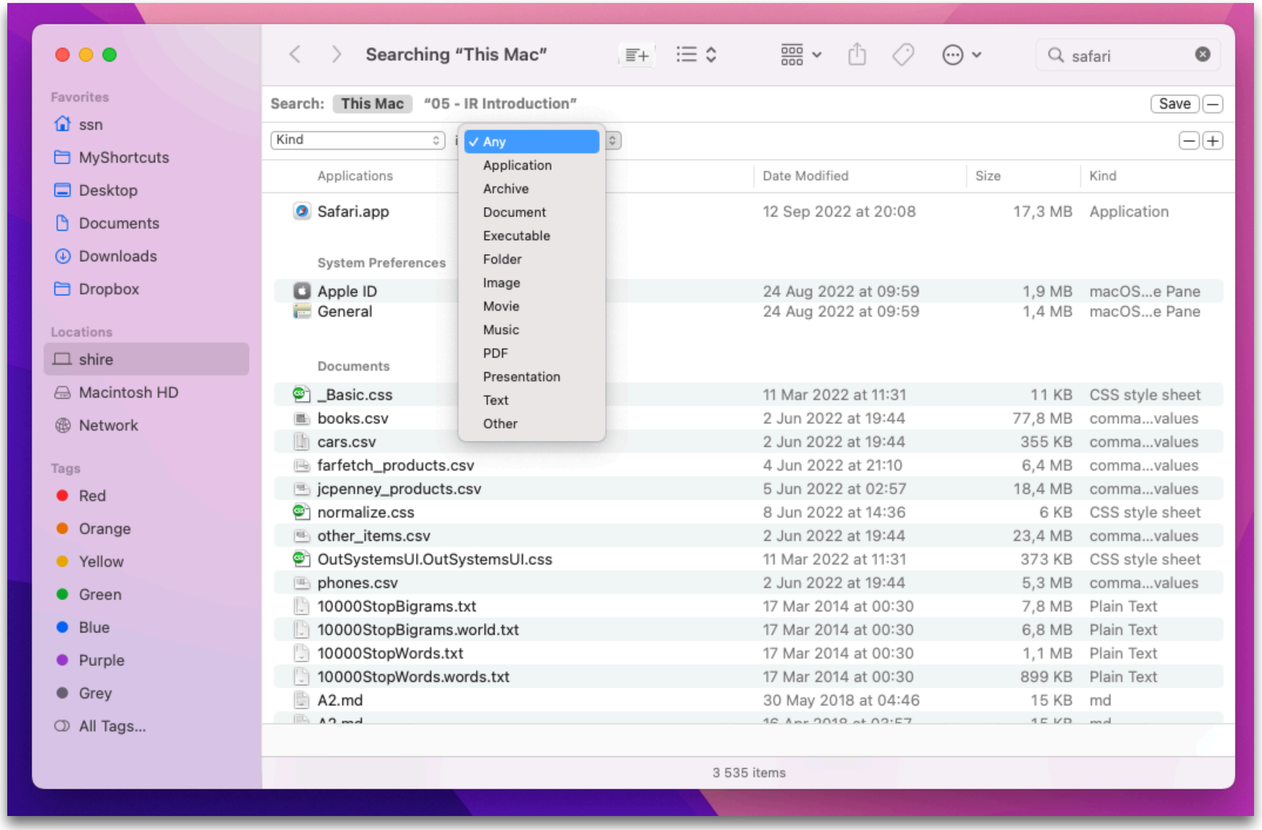
➔ Many different types: web search, maps search, image similarity search, social search, etc.

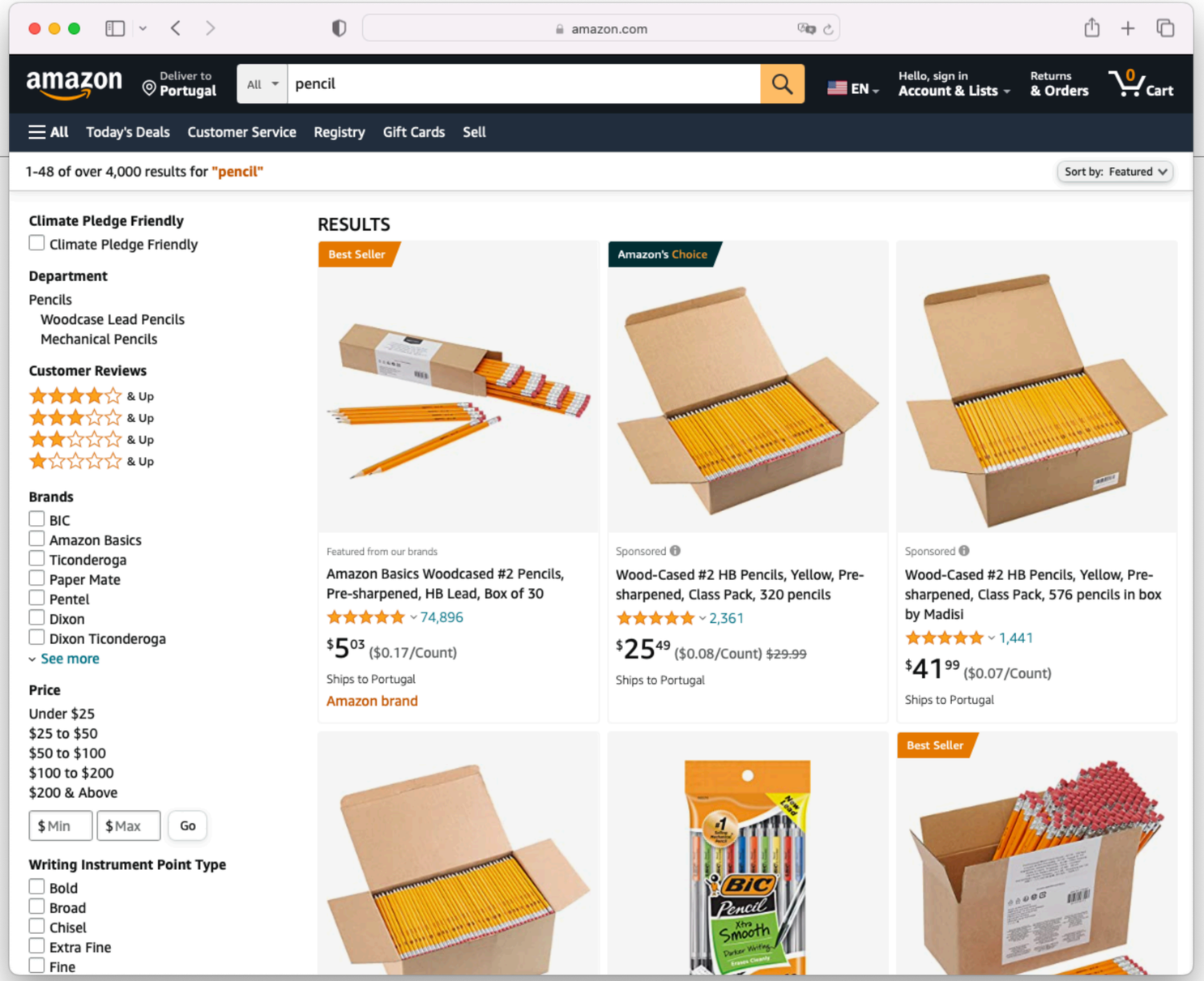

uporto
Google
u porto
Sign in
All
Images
News
Maps
Videos
More
Tools
About 395,000,000 results (0.87 seconds)
https://www.up.pt · Translate this page
Universidade do Porto
A Universidade do Porto é uma das melhores instituições de ensino público da Europa. É reconhecida pela qualidade da formação e investigação, ...
Results from up.pt
Oferta Formativa
Cursos por Faculdade · Licenciatura em Arquitetura ...
Sigarra
O nível organizacional do Sistema de Informação SIGARRA ...
Master's Degrees - - - Courses
If you already have a Bachelor's Degree and would like to ...
International Students
Why should I choose U.Porto for a study period? U.Porto is the ...
People also ask
Quais são as melhores faculdades de Portugal?
Quanto custa a inscrição na Universidade do Porto?
Quantas faculdades têm a Universidade do Porto?
U.PORTO
Livraria Lello
CLÉRIGOS
Jardim da
See photos
See outside
University of Porto (Universidade do Porto)
Website
Directions
Save
Public university in Porto
The University of Porto is a Portuguese public research university located in Porto, and founded on 22 March 1911. It is the second largest Portuguese university by number of enrolled students, after the University of Lisbon, and has one of the most noted research outputs in Portugal. Wikipedia
A 1-min walk from Livraria Lello & Irmão
Address: Praça de Gomes Teixeira, 4099-002 Porto
Hours: Open 24 hours
Phone: 22 040 8000
Total enrollment: 31,352 (2013)


google.com
restaurante
Rating
Hours
All filters
Sign in
Ad · business.google.com/eleven-lab
Eleven Lab - Downtown
Eleven Lab
4.7 (22)
Restaurant · R. de José Falcão 138
Closes soon · 7PM · Opens 11AM Sat
Dine-in · Takeaway · Delivery
RESERVE A TABLE
Restaurante Solar Santa Catarina
4.6 (70) · €
Restaurant · Rua de Santa Catarina 842-980, 4000-455
Open · Closes 8PM
Dine-in · Takeaway · No delivery
17th Restaurant & Bar
4.5 (1,377) · €€€
Restaurant · R. do Bolhão 223
Open · Closes 12AM
Dine-in · No takeaway · No delivery
RESERVE A TABLE
Restaurante Encontro Art Caffe
4.6 (74)
Restaurant · Rua de Santos Pousada 554
Open · Closes 11PM
Dine-in · Takeaway · Delivery
Update results when map moves
Paranhos
Fernando Marques
Jardim de Arca d'Água
Aquário Electrónica - Loja Porto - Júlio de...
Via Verde - Loja do Porto
Parque Urbano de Rio Tinto
Parque Nascente Shopping
McDonald's - Rio Tinto
Loja de Cidadão do Porto
Quinta do Covelo
Mercadona Campanhã
Estádio do Dragão
Hospital de Santa Maria - Porto
Centro de Vacinação do Cerco
Restaurante A Ponte Do Freixo Lda
Marinho's
Bonanza
Restaurante O Ferreira
Restaurante Encontro Art Caffe
Parque Oriental da Cidade do Porto
Migalhas
Restaurante Moreno
Belos Aires Restaurante
Restaurante SALA do CLUBE
Jardim da Cordoaria
Continente Modelo São Cosme
Sé
Holy Sandwich Shop
Ponte Luís I
Douro River
Caves Ferreira
Port Lodge
Jardim do Morro
Oliveira do Douro
Valbom
Casa Branca de Gramido
Layers
Google
Map data ©2022 Google, Inst. Geogr. Nacional
Portugal
Terms
Privacy
Send feedback
500 m

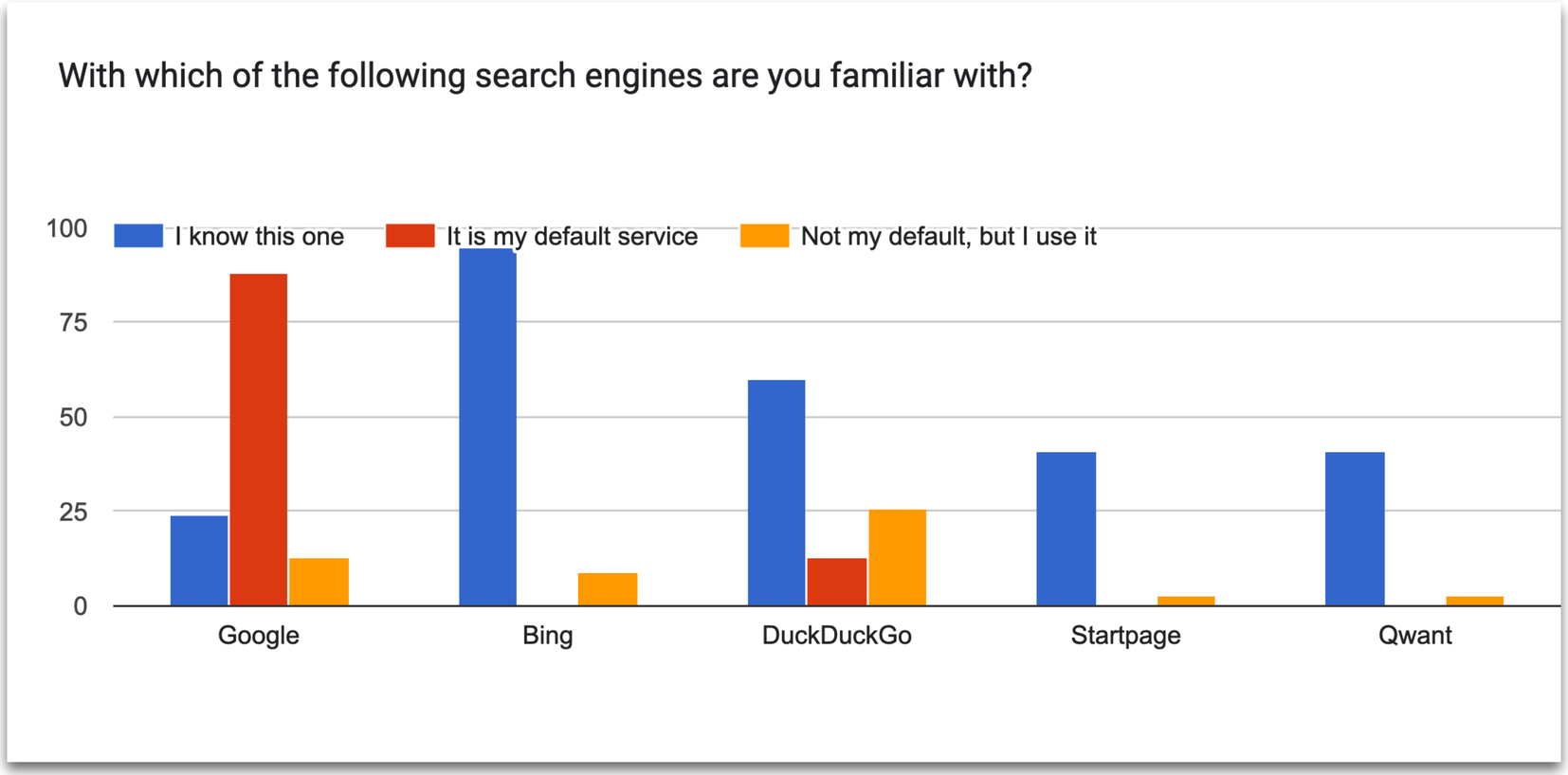

sigarra.up.pt
U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
english | ajuda
Esqueceu-se da senha? | Outras formas de autenticação
Utilizador
Senha
Iniciar sessão
Você está em: Início > Pesquisa de Ocorrências de UC
Boas vindas
Órgãos de Gestão
Departamentos
Serviços
Estudantes
Pessoal
Cursos
I&D e Inovação
Cooperação
Candidatos
Alumni
Empresas
Notícias
Pesquisa
Mapa das Instalações
Pesquisa de Ocorrências de UC
Ano Letivo: 2022/2023
Unidades Curriculares
Código:
Sigla:
Nome:
Ano: Qualquer Ano
Período de Aulas: Todos os períodos
Curso/CE:
Tipo de curso/CE:
Língua de trabalho:
Tipo de avaliação:
Docente
Código:
Sigla:
Nome:
Pesquisar
Clube de Leitura
CENTENÁRIO DO NASCIMENTO DE AGUSTINA BESSA-LUÍS


safari.app
Safari.app
mobilesafari
websites
safari
Developer
SFSafariViewController.h — 8 KB • C header code • Modified 22/09/2022, 15:17
SFSafariViewControllerActivityButton.h — 3 KB • C header code • Modified 22/09/2022, 15:17
SFSafariViewControllerConfiguration.h — 2 KB • C header code • Modified 22/09/2022, 15:16
SFSafariPage.h — 1 KB • C header code • Modified 22/09/2022, 15:16
SFSafariPageProperties.h — 599 bytes • C header code • Modified 22/09/2022, 15:16
Definition
safari — Oxford Dictionary of English • an expedition to observe or hunt animals in their natural habitat,...
Folders


Searching "This Mac"
safari
Favorites
ssn
MyShortcuts
Desktop
Documents
Downloads
Dropbox
Locations
shire
Macintosh HD
Network
Tags
Red
Orange
Yellow
Green
Blue
Purple
Grey
All Tags...
Search: This Mac "05 - IR Introduction"
Save
Kind
Any
Application
Archive
Document
Executable
Folder
Image
Movie
Music
PDF
Presentation
Text
Other
Applications
Date Modified
Size
Kind
Safari.app 12 Sep 2022 at 20:08 17,3 MB Application
System Preferences
Apple ID 24 Aug 2022 at 09:59 1,9 MB macOS...e Pane
General 24 Aug 2022 at 09:59 1,4 MB macOS...e Pane
Documents
_Basic.css 11 Mar 2022 at 11:31 11 KB CSS style sheet
books.csv 2 Jun 2022 at 19:44 77,8 MB comma...values
cars.csv 2 Jun 2022 at 19:44 355 KB comma...values
farfetch_products.csv 4 Jun 2022 at 21:10 6,4 MB comma...values
jcpenney_products.csv 5 Jun 2022 at 02:57 18,4 MB comma...values
normalize.css 8 Jun 2022 at 14:36 6 KB CSS style sheet
other_items.csv 2 Jun 2022 at 19:44 23,4 MB comma...values
OutSystemsUI.OutSystemsUI.css 11 Mar 2022 at 11:31 373 KB CSS style sheet
phones.csv 2 Jun 2022 at 19:44 5,3 MB comma...values
10000StopBigrams.txt 17 Mar 2014 at 00:30 7,8 MB Plain Text
10000StopBigrams.world.txt 17 Mar 2014 at 00:30 6,8 MB Plain Text
10000StopWords.txt 17 Mar 2014 at 00:30 1,1 MB Plain Text
10000StopWords.words.txt 17 Mar 2014 at 00:30 899 KB Plain Text
A2.md 30 May 2018 at 04:46 15 KB md
3 535 items


amazon.com
amazon
Deliver to Portugal
All
pencil
EN
Hello, sign in Account & Lists
Returns & Orders
0 Cart
All
Today's Deals
Customer Service
Registry
Gift Cards
Sell
1-48 of over 4,000 results for "pencil"
Sort by: Featured
Climate Pledge Friendly
Climate Pledge Friendly
Department
Pencils
Woodcase Lead Pencils
Mechanical Pencils
Customer Reviews
& Up
& Up
& Up
& Up
Brands
BIC
Amazon Basics
Ticonderoga
Paper Mate
Pentel
Dixon
Dixon Ticonderoga
See more
Price
Under $25
$25 to $50
$50 to $100
$100 to $200
$200 & Above
$Min
$Max
Go
Writing Instrument Point Type
Bold
Broad
Chisel
Extra Fine
Fine
RESULTS
Best Seller
Featured from our brands
Amazon Basics Woodcased #2 Pencils, Pre-sharpened, HB Lead, Box of 30
74,896
$5 03 ($0.17/Count)
Ships to Portugal
Amazon brand
Amazon's Choice
Sponsored
Wood-Cased #2 HB Pencils, Yellow, Pre-sharpened, Class Pack, 320 pencils
2,361
$25 49 ($0.08/Count) $29.99
Ships to Portugal
Sponsored
Wood-Cased #2 HB Pencils, Yellow, Pre-sharpened, Class Pack, 576 pencils in box by Madisi
1,441
$41 99 ($0.07/Count)
Ships to Portugal
BIC
Pencil
Xtra Smooth
Best Seller

# Student Survey




https://www.youtube.com/watch?v=0eKVizvYSUQ

# Information retrieval concepts

# Information Retrieval

- In computer science, the field that is most involved in search problems and developments is
  - Information Systems > **Information Retrieval**, and includes sub-areas such as:
    - Document representation
    - Information retrieval query processing
    - Users and interactive retrieval
    - Retrieval models and ranking
    - Retrieval tasks and goals
    - Evaluation of retrieval results
    - Search engines architectures and scalability

# Defining Information Retrieval

- “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information”, Salton (1968)
  - General definition that can be applied to many types of search applications.
  - Primary focus since the 50s has been on text and documents.

- “Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”, Manning et al. (2008)

# Documents

- Documents can be web pages, email messages, books, news articles, research papers, computer files (PDF, Word, text), text messages, etc.
- Common properties:
  - Significant text content.
  - Some structure, e.g. title, body, date, tags, author.
- From PRI projects:
  - linux packages, songs lyrics, cooking recipes, MSc dissertations, game plays, etc.

# Different from Databases

- “A database management system is a software system that enables the creation, maintenance, and use of large amounts of data” (Abiteboul et al., 1995)
- Information retrieval and database systems have a lot in common.
- But emphasize different aspects of information management.
  - Databases contains highly structured data, queried using formal query languages, and the results are set-based — results are true or false for a given condition.
  - Information retrieval systems support interactive processes, support natural language queries, and results are rank-based — results are more or less relevant for a given information need. Not just text matching!

# Information Retrieval and Databases

**Table 1.1** Comparison of database systems and information retrieval, based on [40]

| | Database systems | Information retrieval |
|---|---|---|
| Data type | Numbers, short strings | Text |
| Foundation | Algebraic/logic based | Probabilistic/statistics based |
| Search paradigm | Boolean retrieval | Ranked retrieval |
| Queries | Structured query languages | Free text queries |
| Evaluation criteria | Efficiency | Effectiveness (user satisfaction) |
| User | Programmer | Nontechnical person |

Table from Entity-Oriented Search, Balog (2018).

# Documents vs. Records (tuples) Queries

- Example **database queries**
  - All students enrolled in course “ABC” in year 22/23.
  - All bank account transferences between account A and account B, within a given date interval, ordered by transferred amount.

- Example **search engine queries**
  - [ feup enrolled students ABC 22/23 ]
  - [ weather porto site:pt ]

# Information Retrieval Tasks

- Information retrieval applications and research are typically structured around tasks.
- Tasks address specific contexts, and specific information access problems.
- Examples of classical information retrieval tasks:
  - Ad-hoc search, find relevant documents for an arbitrary text query.
  - Filtering, identify relevant user profiles for a new document.
  - Classification, identify relevant labels for documents.
  - Question answering, give specific answer to a question.
- Yearly evaluation initiatives with classical and novel tasks at TREC
  - https://trec.nist.gov/pubs/call2022.html

# Dimensions of Information Retrieval

| Content | Applications | Tasks |
|---|---|---|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | P2P search | |
| | Literature search | |

Image from Search Engines: Information Retrieval in Practice, Croft et al. (2009)

# Central issues in information retrieval

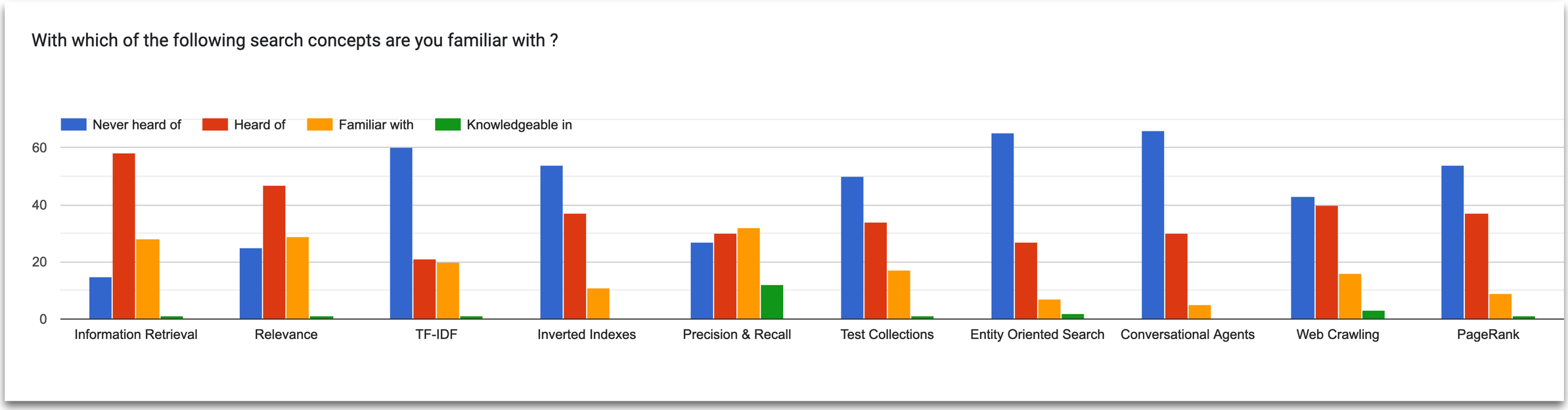
# Relevance

- ➔ The concept of relevance is a key issue in information retrieval.
- ➔ Many factors impact a person's decision of what is relevant,
  e.g. task, context, previous knowledge, etc.
- ➔ Simple definition of relevance,
  - ➔ a relevant document contains the information that a person was looking for when submitting a query to a search engine.

# Evaluation

- ➔ Evaluation setups are based on the use of annotated test collections of documents, queries, and relevance judgements.
- ➔ Evaluation in information retrieval dates back to the 1960s (Cranfield paradigm).
- ➔ IR evaluation methodologies have been adopted by many fields.
  - ➔ Recall and precision are two examples of widely used measures.
- ➔ Information needs drive the search process and result from the underlying task users' have.

# Student Survey - IR Concepts



# Information Retrieval and Search Engines

- ➔ Search engines are practical applications of information retrieval techniques.
- ➔ Existing open-source and commercial solutions that support development and research:
  - ➔ Apache Solr, Elasticsearch, Open Search, Terrier.

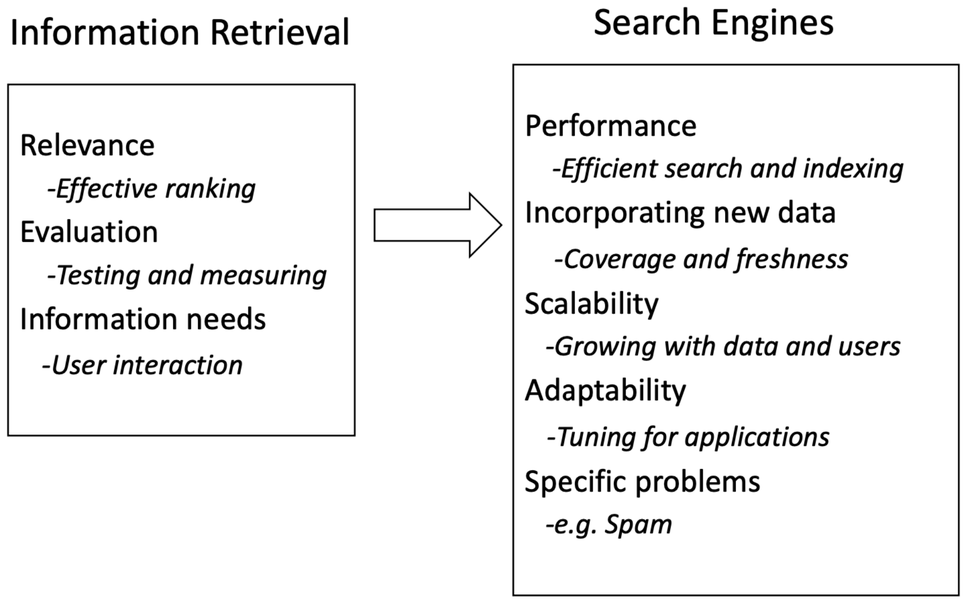
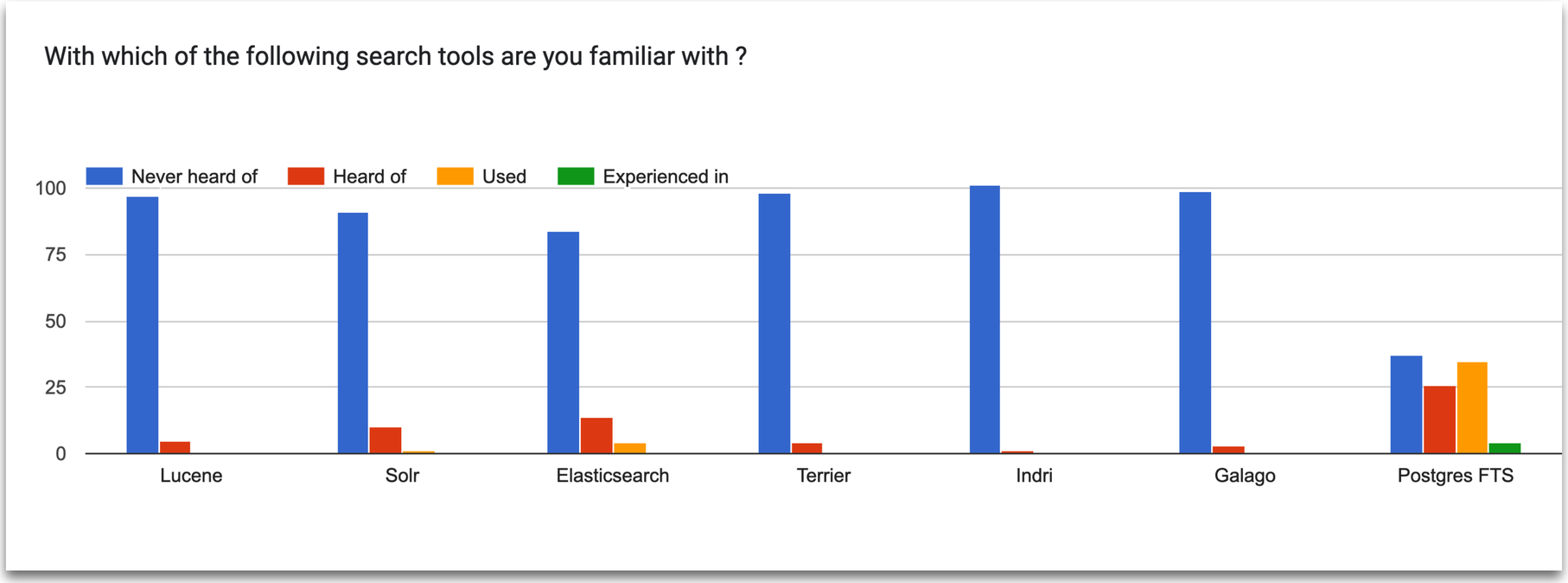


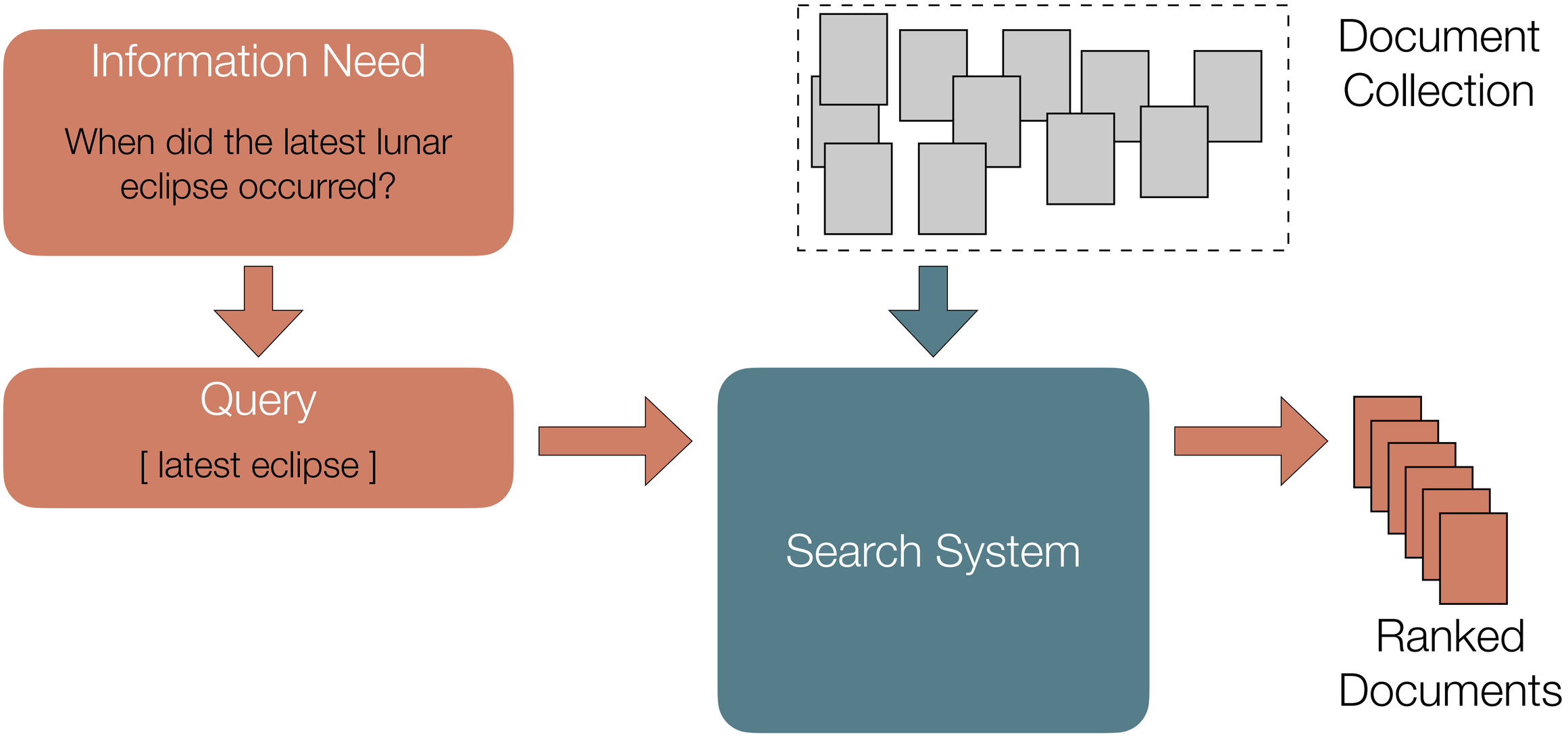

Image from Search Engines: Information Retrieval in Practice, Croft et al. (2009)

# Student Survey - IR Technologies



# Architecture of information retrieval systems

# Classic Information Retrieval



# Architecture of an Information Retrieval System

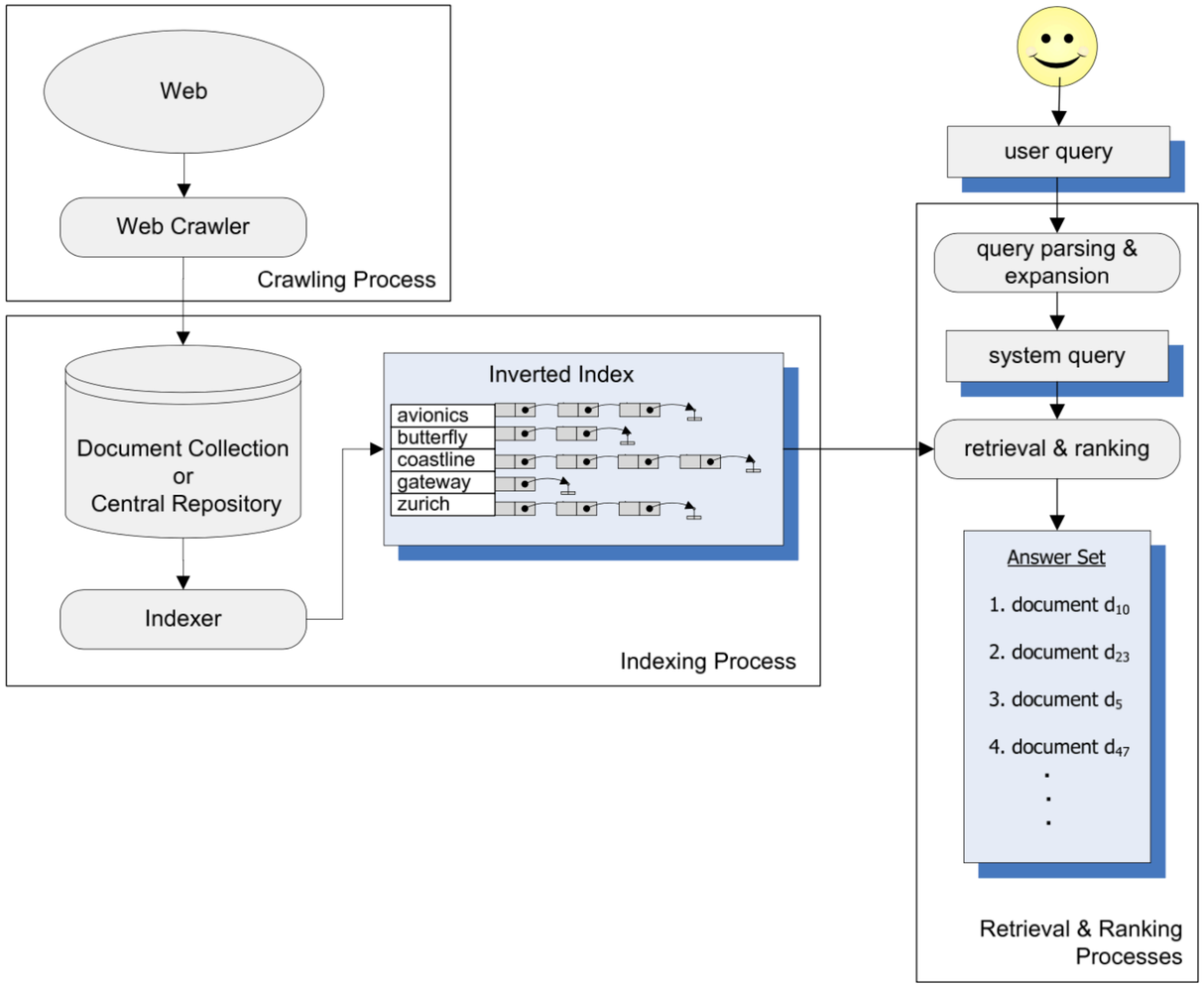



Image from Modern Information Retrieval, Baeza-Yates et al. (2011)

# Indexing Process

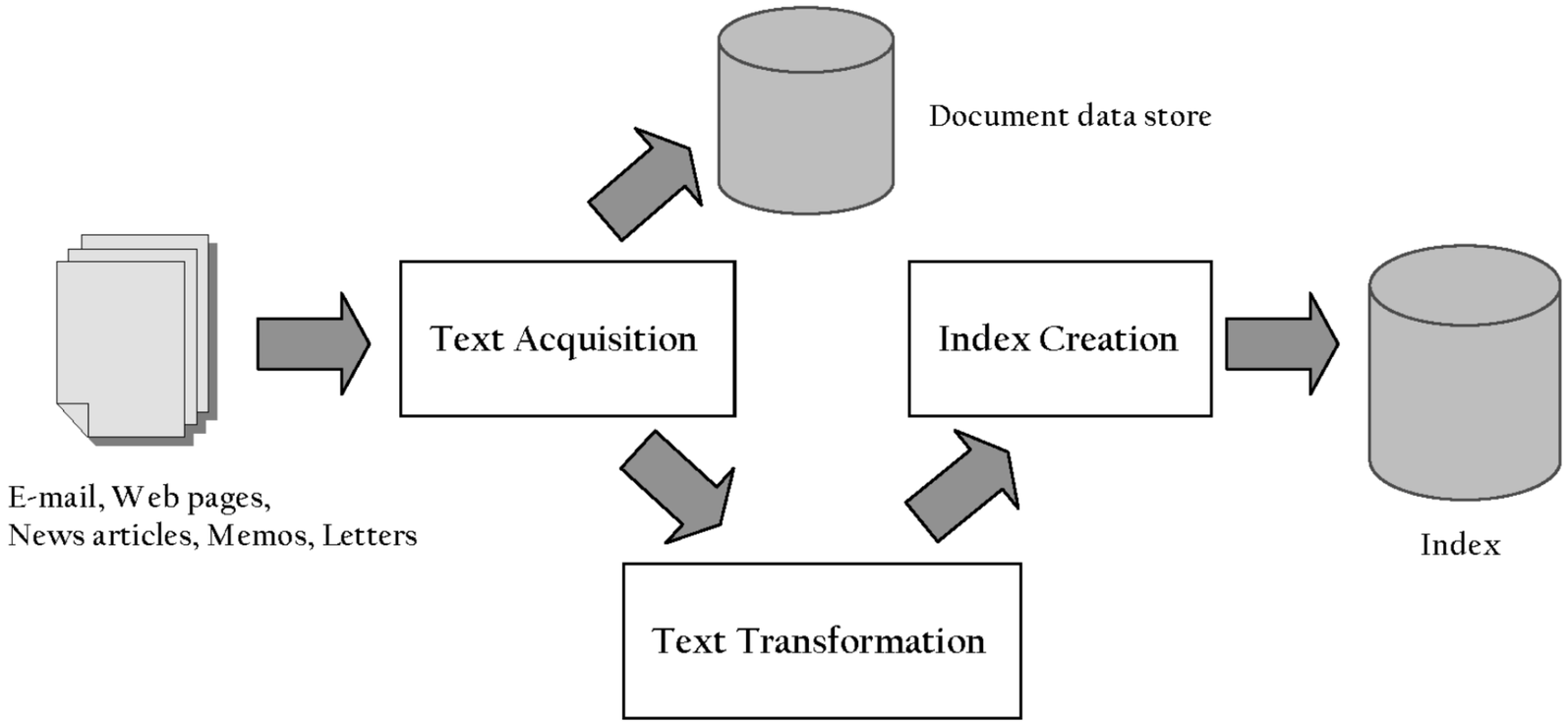



Image from Search Engines: Information Retrieval in Practice, Croft et al. (2009)

# Query Process

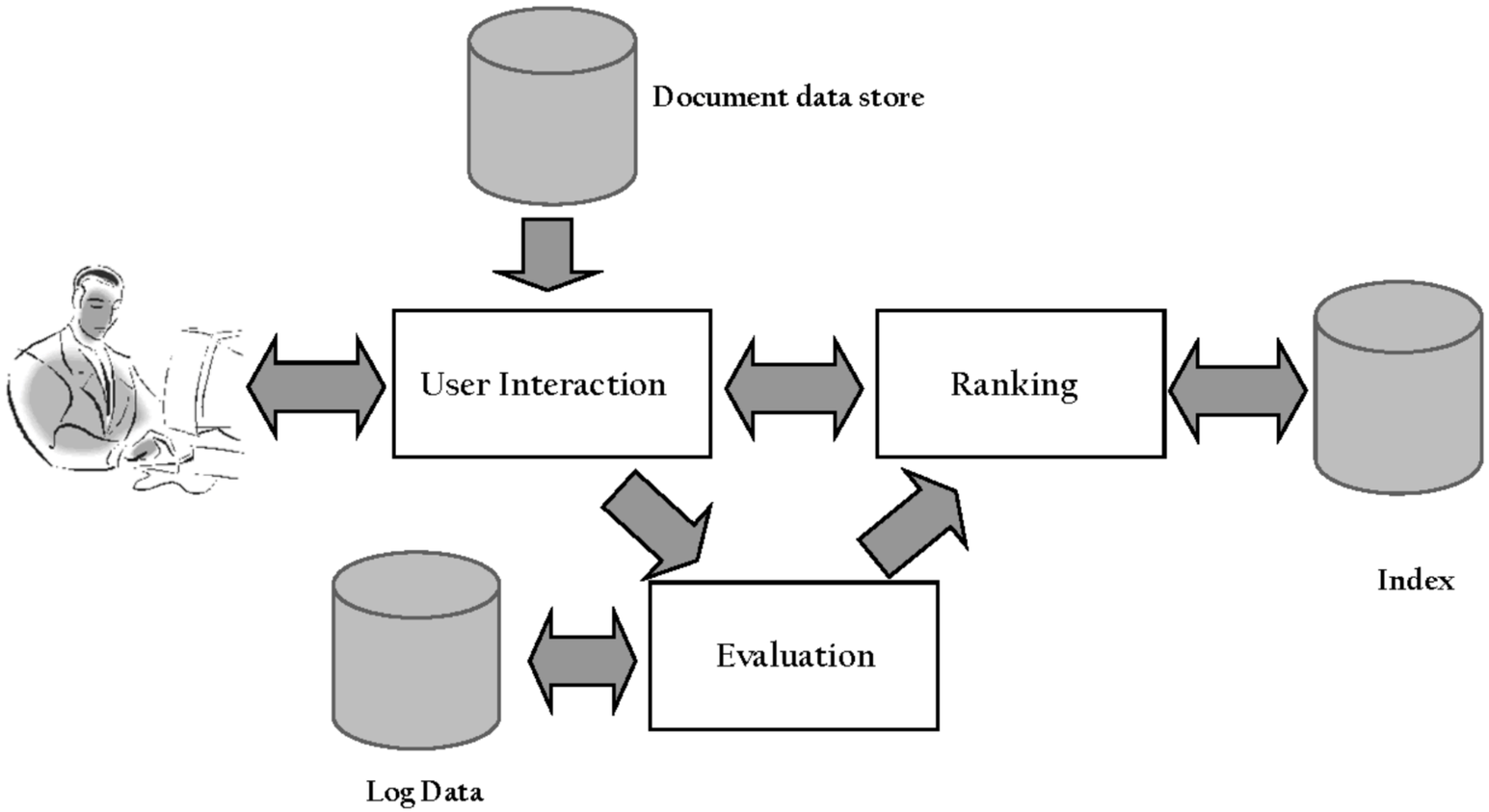



Image from Search Engines: Information Retrieval in Practice, Croft et al. (2009)

# History of information retrieval

# Historic Context

- At the end of World War II, Vannebar Bush, who headed the R&D office during the war, reflected on applications of new technologies in peace times.
- He wrote a landmark essay entitled As We May Think in 1945, where a new device, the Memex, capable of managing and interlinking knowledge is described.
  - https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
- This essay greatly influenced the next generation of innovators.
  - Douglas Engelbart and the development of the computer mouse ("the mother of demos"), 1968.
  - Ted Nelson and the development of the concept of hypertext (Project Xanadu), 1960s.
  - Tim Berners-Lee and the World Wide Web, in 1989.
  - Web search engines, such as Lycos (94), Altavista (95), and Google (98).

# Searching the Web

- ➔ The web represented a unique search challenge.
- ➔ A unique document collection, millions of documents, connected through hyperlinks, and distributed through multiple repositories (required crawling).
- ➔ The size of the collection and the user base, millions of documents and users. Performance and scalability became a key issue.
- ➔ Predicting relevance in such a large and heterogeneous collection is complex. New sources of evidence appeared (e.g. hyperlinks).
- ➔ The web is also a business platform, thus users needs are not limited to text search.
- ➔ Web spam is a real problem on the web, sometimes referred to as “adversarial information retrieval”.

# Information Retrieval Summary

➔ Information Retrieval dates back to 1940s.

➔ Well established field in Computer Science.

➔ Deals with a wide range of information access problems.

➔ Organized in tasks: web search, filtering, question answering, podcast search, conversational, enterprise / intranet search, entity-oriented search

➔ Strong links with information science, NLP, machine learning, databases, etc.

➔ Core concepts: information need, index, query, relevance.

# Course Plan

# Course Plan

- ➔ Information Retrieval Overview (**today**)
- ➔ Evaluation
- ➔ Models and Indexing
- ➔ Retrieval and Web Retrieval
- ➔ Query Processing
- ➔ Entity-Oriented Search
- ➔ Neural IR

# Lab Classes

- Tutorials on Solr and IR Evaluation

- Develop M2: Information Retrieval
  - Index
  - Retrieve
  - Evaluate

# References and Further Reading


- Modern Information Retrieval (2nd Edition) (2011)
  Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Addison-Wesley Professional.
  http://www.mir2ed.org

- Introduction to Information Retrieval (2008)
  Christopher Manning, Prabhakar Raghavan, Hinrich Schütze. Cambridge University Press.
  http://nlp.stanford.edu/IR-book

- Search Engines: Information Retrieval in Practice (2009)
  W. Bruce Croft, Donald Metzler, Trevor Strohman. Pearson.
  http://ciir.cs.umass.edu/downloads/SEIRiP.pdf

- Entity-Oriented Search (2018)
  Krisztian Balog. Springer One.
  https://link.springer.com/book/10.1007/978-3-319-93935-3