TC3 Information Retrieval

Master 1 AI, Upsay
T4, March to April

Kim Gerdes
Lisn

Me

- mathematician by trade (maîtrise and DEA de Paris 11 / Polytechnique)
  - Arithmetic geometry
- then formal and computational syntax, PhD in NLP
- corpus linguistics, treebanks, Natural Language Processing, typology
- work on text generation: variations of word order, NLG systems, idea generation (patent text generation, massive defensive publishing)

  CTO of Qatent
- Habilitation on epistemology of NLP

- speaking too fast, especially if I like the subject. let me know!

The class on Information Retrieval

- textual information retrieval and text mining
- continuation of Hands-On NLP
- work on speeches, then technical texts
  - indexing
  - NER
  - spacy/prodigy
- evaluation:
  - notebooks
  - project:
    - exploratory: taxonomy development, including definitions???

Today

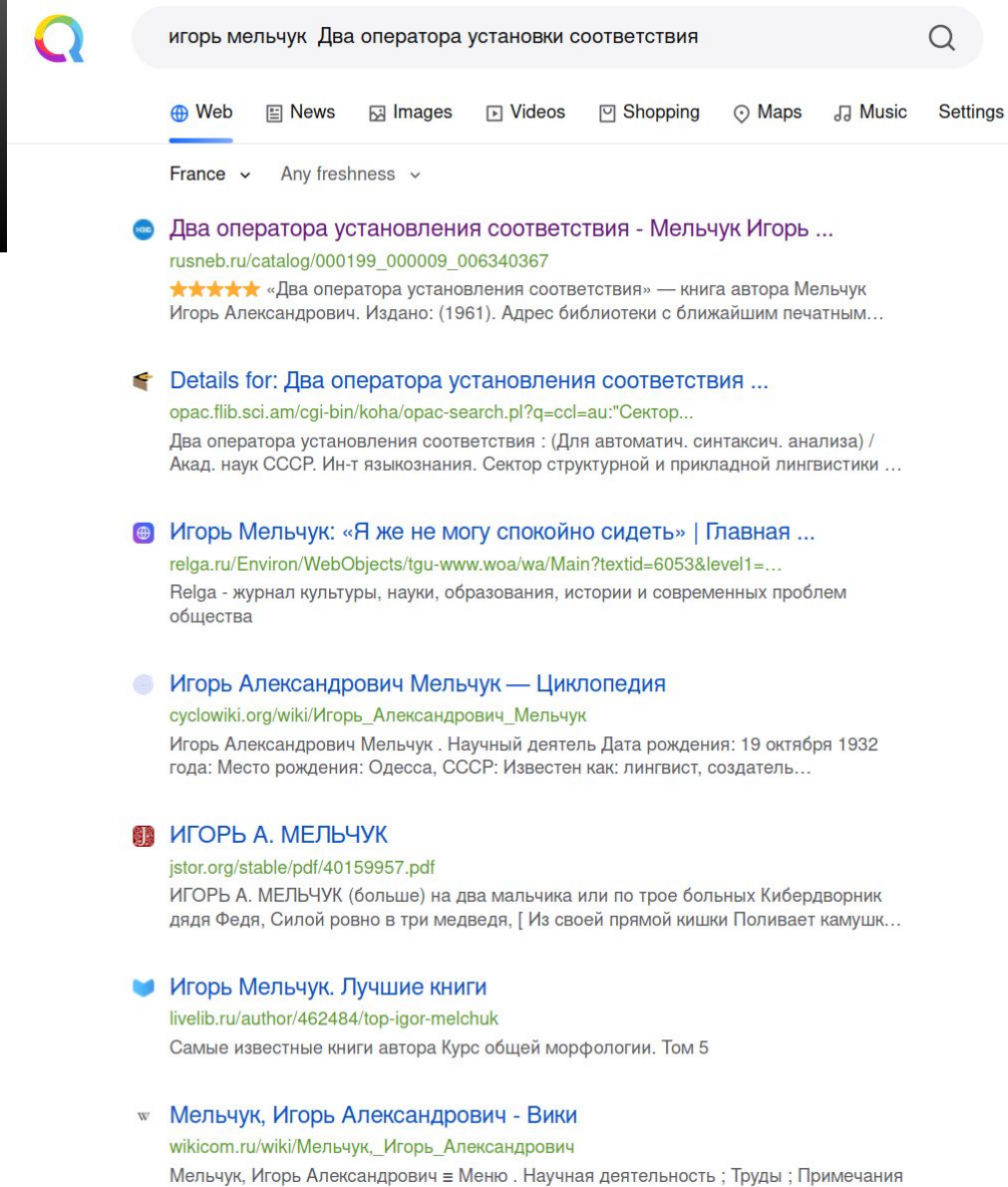
- introduction
- terms and domains
- tutorial:
  - gentle introduction on new textual dataset
    - if you haven’t followed Hand-On NLP, spend some time on the notebook this week. it should be very easy!
  - indexes and counting

How to find a book here?

How to find a book here?

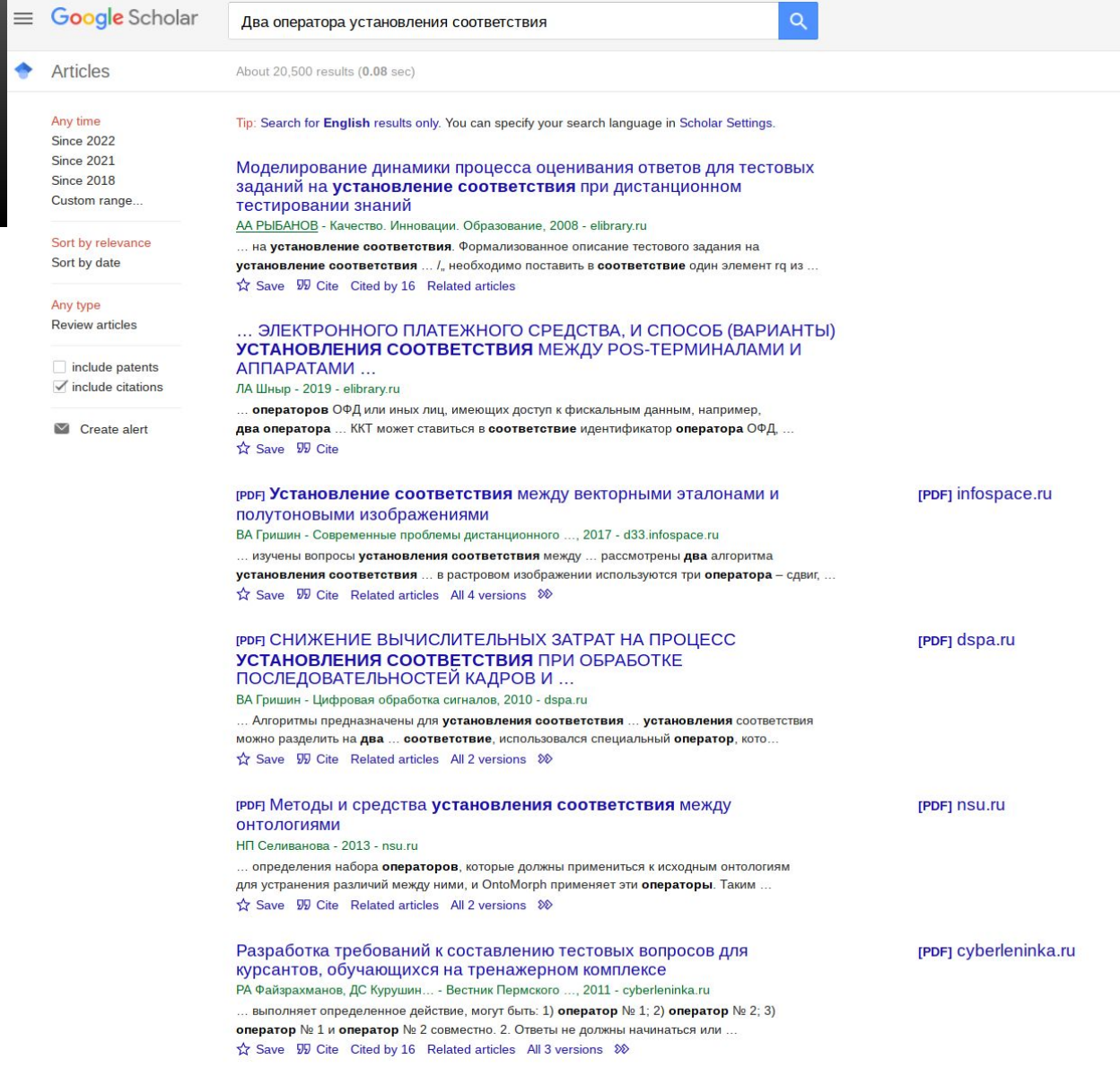
Today

Search engines:



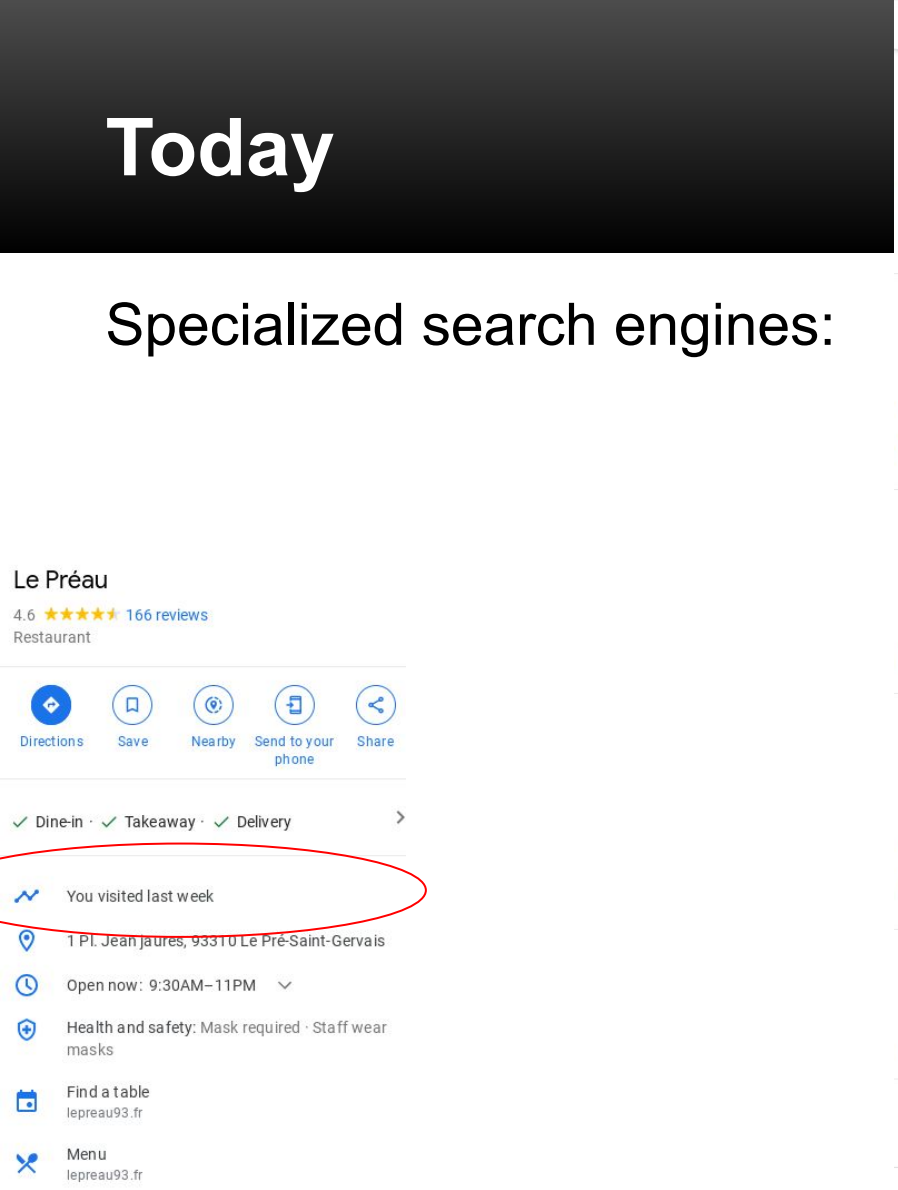
Today

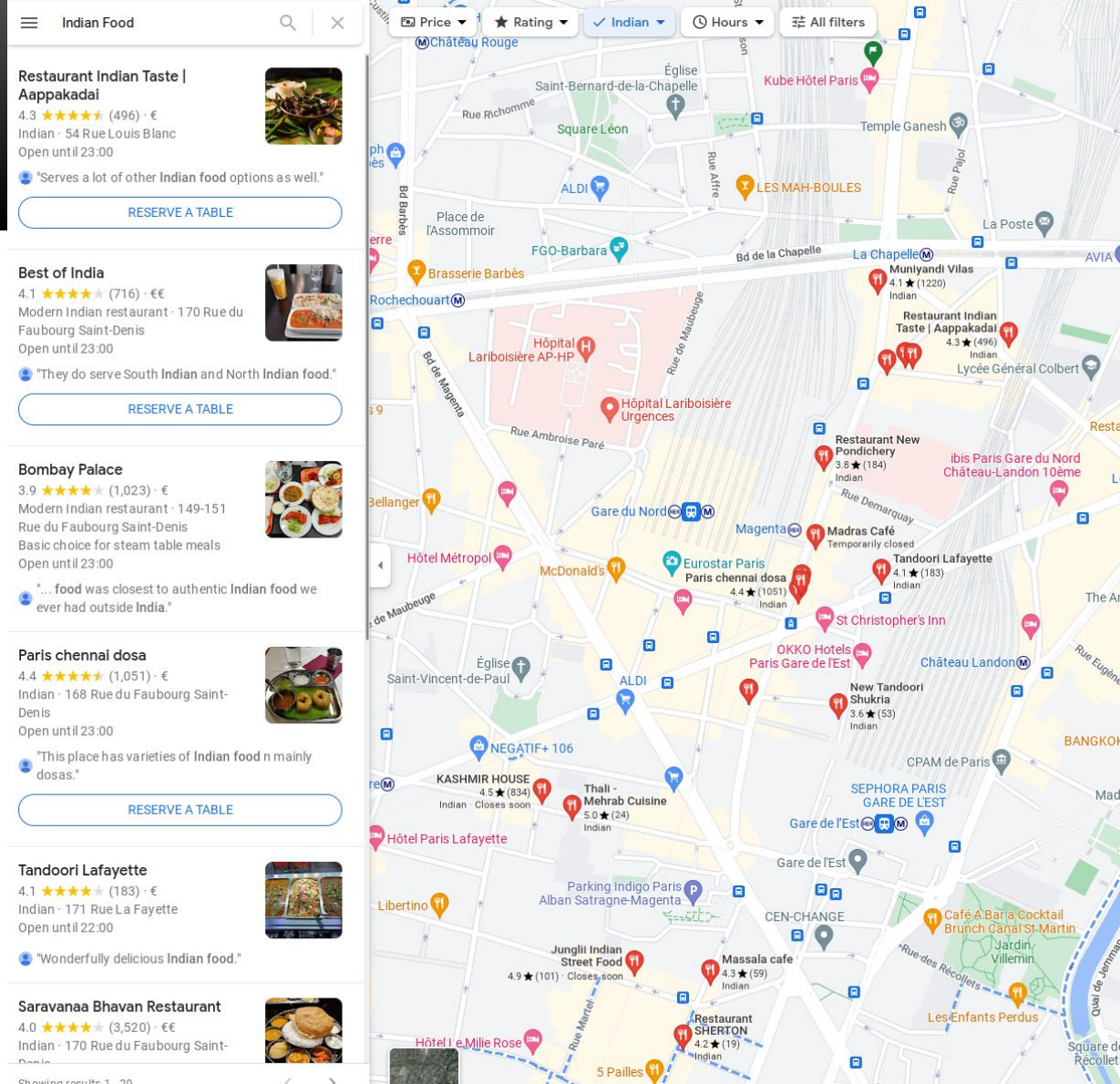
Specialized search engines:



Today

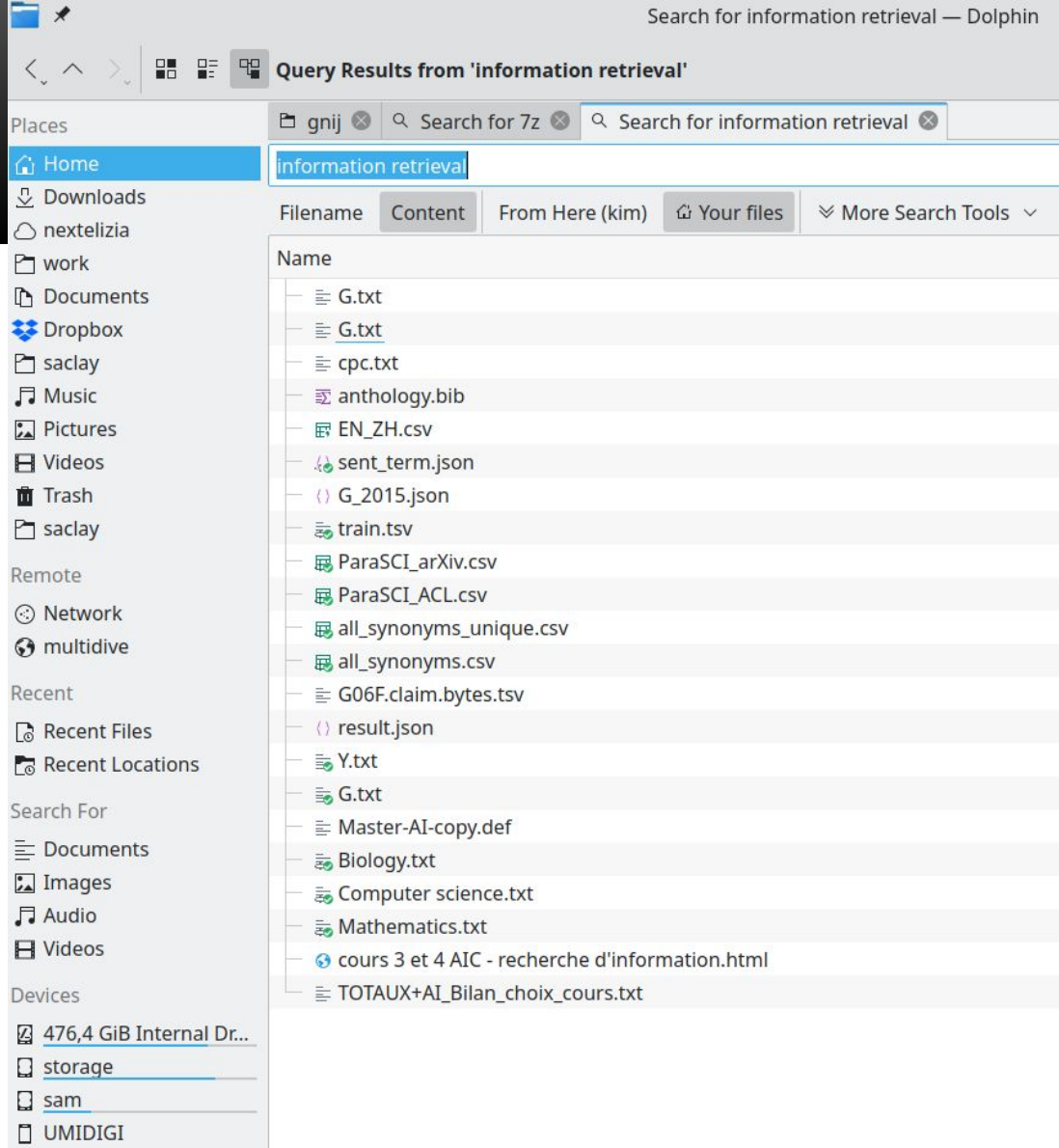
Specialized search engines:





Today

Local search engines:



beyond search engines:

- digital library search
- E-mail search, Searching your desktop and laptop computers
- Corporate knowledge bases, local business search, expert search
- Legal information retrieval, patent search
- news search
- image and video search
- (micro-)blog search
- product search, federated search
- social search, community Q&A, question-answering
- recommender systems
- opinion mining → see Hands-On NLP

Information Retrieval?

query → document

- answer is a set of documents

fill out a form, a table

- What is the capital of France? Paris

often two types of participants: query system builder + users

- ‘advanced’ search: quotes?
  - who knows more?

- Intelligent interpretation of information need

Chronology of word use:

- Google ngram: https://books.google.com/ngrams

Information Retrieval?

query → document

- answer is a set of documents

fill out a form, a table

- What is the capital of France? Paris

often two types of participants: query system builder + users

- ‘advanced’ search: quotes?
  - who knows more?
    -dash ~tilde site: link:
    2019..2021 allintitle: filetype: ….
- Intelligent interpretation of information need

Chronology of word use:

- Google ngram: https://books.google.com/ngrams

Information Retrieval?

- **The task of information retrieval is to retrieve relevant documents in response to a query.**
- 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).
  –from IIR book.
  - Introduction to Information Retrieval, by C. Manning, P.Raghavan, and H. Schutze. Cambridge University Press
  - book website https://nlp.stanford.edu/IR-book/

Other definitions

- Jaime Arguello
  - Information retrieval (IR) is the science and practice of designing, developing, and evaluating systems that match information seekers with the information they seek.
- Gerard Salton, 1968:
  - Information retrieval is a field concerned with the structure, analysis, organization, storage, and retrieval of information.

Information Retrieval?

- **The task of information retrieval is to retrieve relevant documents in response to a query.**

Text mining?

- **The task of text mining is to discover or derive new information from text, finding patterns across datasets, and separating signal from noise.**
- **possibly: text mining techniques help to improve index systems**

Text mining = ML?

- **ML may be used in information retrieval, but it's not always (think full text search, no predictions or pattern discovery).**
- **Text mining contains tasks of storage of big data, index structures, parallelization, …**

Information Retrieval = Text Mining = ML?

- **information retrieval: find data (texts, but also pictures, sound files, videos)**
  - **here: only texts**
- **machine learning: generalize examples to new data**
- **text mining: find new properties of existing data**

Why should you care?

- You use information retrieval techniques every day
- You have too much data
  - On your hard drive
  - Accessible online

→ You need tools to access the data

  - Actively
    - Search queries
  - Passively
    - Spam protection: classification task
  - Indirectly
    - You depend on others that use information retrieval tools:
      - search in (medical) research results, jurisprudence, patents

→ You need to understand how the system works

  - because you rely on them for understanding the world

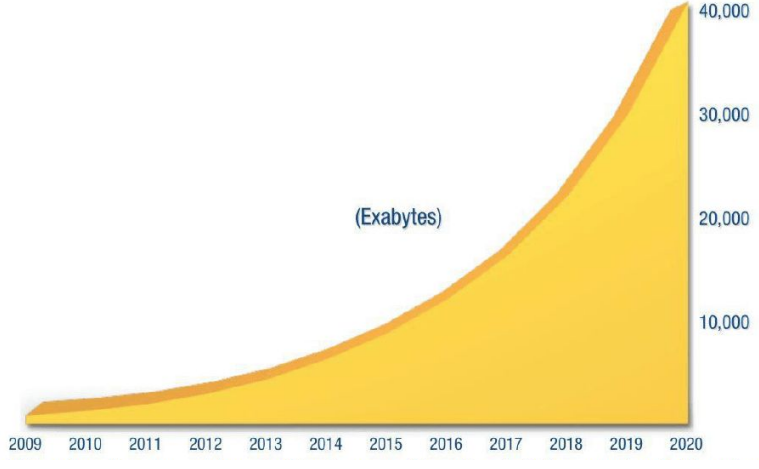
- How does the black-box work?

Why should you care?

- People are generating 2.5 quintillion bytes of data each day.
- Nearly 90% of all data has been created in the last two years.
- Each person on earth generates an average of about 1.7 MB of data per second.
- Moore's law

https://kommandotech.com/statistics/big-data-sta
https://curriculum.code.org/csp-19/unit4/1/

- singularity?



The search task

- Given a query and a corpus, find relevant items

- query: user’s expression of their information need
- corpus: a repository of retrievable items
- relevance: satisfaction of the user’s information need

Why IR is hard

Information retrieval is an uncertain process

- Query
  - users don’t know what they want
  - users don’t know how to convey what they want
  - computers can’t elicit information like a librarian
  - computers can’t understand natural language text
  - A query is an impoverished description of the user’s information need
  - Highly ambiguous to anyone other than the user
- Relevance
  - the search engine can only guess what is relevant
  - the search engine can only guess if a user is satisfied
  - over time, we can only guess how users adjust their short- and long-term behavior

Retrieval Model

measures of an information retrieval system:

- speed
- size
- price
- user happiness:
  - Speed of response
  - Size of index
  - UI
  - price: free?
  - Most important: relevance
- A formal method that predicts the degree of relevance of a document to a query

User?

1. searcher
2. advertiser
3. buyer
4. seller
5. ?

Retrieval Model

- measures of relevance:
  - precision? relevant/retrieved
  - recall? retrieved/relevant
  - F-score
  - accuracy? relevant/non-relevant? useless.

(Manning et al's class on IR)

Retrieval Model

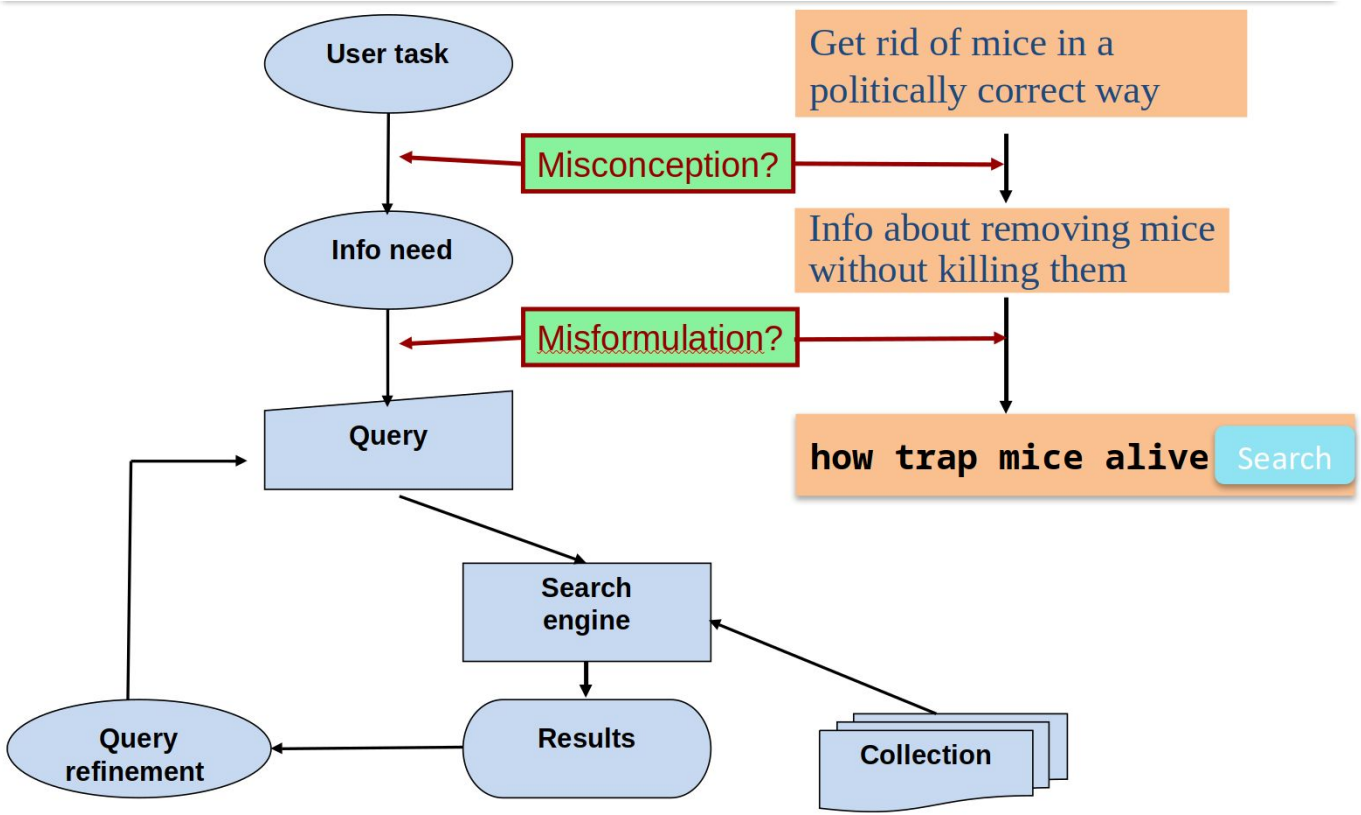
- Relevance to what?
  - First take: relevance to the query
  - “Relevance to the query” is very problematic.
- Information need i:
  - “I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.”
- This is an information need, not a query.
  - Query q: [red wine white wine heart attack]

- Consider document d′:
  - At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.

- d′ is an excellent match for query q ...
- d′ is not relevant to the information need i.

(Manning et al’s class on IR)

classic search model



from: Manning

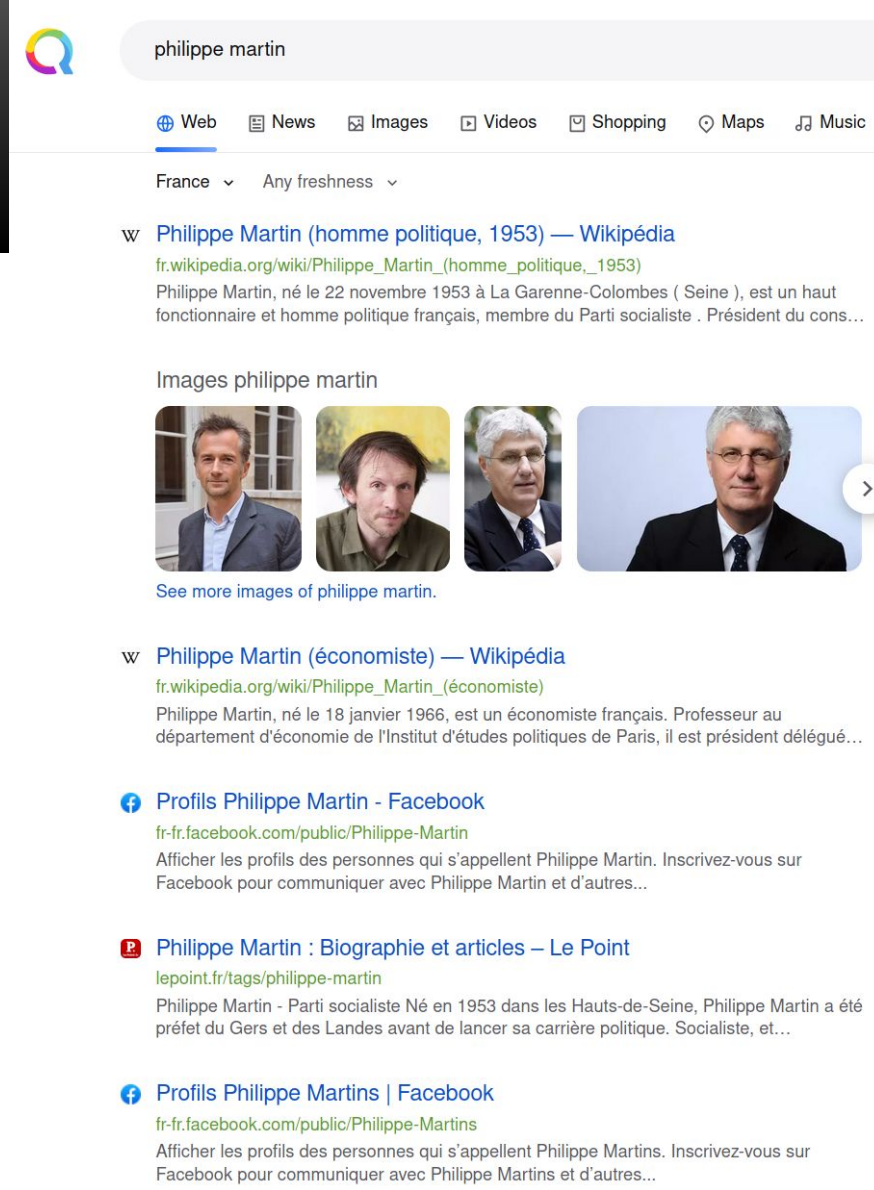
Who?

James Smith?
Zhang Wei - 张玮?

entity linking!

Given a knowledge base, chose the entity corresponding to the mention (referent)

- Named Entity Recognition
  - typing?
- Named Entity Disambiguation
- Named Entity Linking



Entity linking

hard task
https://services.d4science.org/web/tagme/demo

Developed by Paolo Ferragina and Ugo Scaiella at A³ Lab Dipartimento di Informatica, University of Pisa.

"The Australian Government is sanctioning 10 people of strategic interest to Russia for their role in encouraging hostility towards Ukraine and promoting pro-Kremlin propaganda to legitimise Russia's invasion," reads a new statement from Australian Foreign Minister Marise Payne. "This includes driving and disseminating false narratives about the 'de-Nazification' of Ukraine, making erroneous allegations of genocide against ethnic Russians in eastern Ukraine, and promoting the recognition of the so-called Donetsk People's Republic and Luhansk People's Republic as independent." A report by the Australian Associated Press and the Daily Mail says that the men targeted with these new sanctions are "journalists, authors or Putin's press officers." This move follows earlier waves of sanctions directed at Russian government, military and financial institutions, as well as economic sanctions on the Luhansk and Donetsk regions in Ukraine.

Entity linking

https://www.dbpedia-spotlight.org/

- confidence 0.5:

"The Australian Government is sanctioning 10 people of strategic interest to Russia for their role in encouraging hostility towards Ukraine and promoting pro-Kremlin propaganda to legitimise Russia's invasion," reads a new statement from Australian Foreign Minister Marise Payne. "This includes driving and disseminating false narratives about the 'de-Nazification' of Ukraine, making erroneous allegations of genocide against ethnic Russians in eastern Ukraine, and promoting the recognition of the so-called Donetsk People's Republic and Luhansk People's Republic as independent."
A report by the Australian Associated Press and the Daily Mail says that the men targeted with these new sanctions are "journalists, authors or Putin's press officers." This move follows earlier waves of sanctions directed at Russian government, military and financial institutions, as well as economic sanctions on the Luhansk and Donetsk regions in Ukraine.

- confidence 0.2: overlaps!!

"The AustralianAustralianAustralian Government is sanctioning 10 people of strategic interest to Russia for their role in encouraging hostility towards Ukraine and promoting pro-Kremlin propaganda to legitimise Russia's invasion," reads a new statement from Australian Foreign Minister Marise Payne. "This includes driving and disseminating false narratives about the 'de-Nazification' of Ukraine, making erroneous allegations of genocide against ethnic Russians in eastern Ukraine, and promoting the recognition of the so-called Donetsk People's Republic and Luhansk People's Republic as independent."
A report by the Australian Associated Press and the Daily Mail says that the men targeted with these new sanctions are "journalists, authors or Putin's press officers." This move follows earlier waves of sanctions directed at Russian government, military and financial institutions, as well as economic sanctions on the Luhansk and Donetsk regions in Ukraine.

Topics

Australia Person Strategy Attention Russian Empire Hostility Ukraine Promotion (marketing) Propaganda 2003 invasion of Iraq Press release Ministry of Foreign Affairs (Russia) Marise Payne Social influence False flag Narrative Denazification Ukraine Genocide Russians Russians Eastern Ukraine Diplomatic recognition Donetsk People's Republic Luhansk People's Republic Independence 9/11 Commission Report Australian Associated Press Daily Mail International sanctions during the Ukrainian crisis Journalist Author Vladimir Putin Vladimir Putin News media Russians Government of Russia Military Financial institution Oil well International sanctions during the Ukrainian crisis Luhansk Donetsk Oblast Ukraine

NER

Named-entity recognition:

- independent field of research
- ~ entity linking
  - https://en.wikipedia.org/wiki/Entity_linking
  - entity-linking: non-named entities
  - NER:
    - https://en.wikipedia.org/wiki/Named-entity_recognition
    - typing: person vs. organization etc.
    - un-linked terms: “*Peter* came to see me”
- ~ coreference resolution
  - https://en.wikipedia.org/wiki/Coreference
  - “*Peter* came to see me. *He* brought a cake.”
  - He knocked the vase against the sink. *It* broke.
- ~ Vinograd schemes
  - https://en.wikipedia.org/wiki/Winograd_schema_challenge
  - world knowledge

Hearst Patterns (1992)

- Pairs of terms that are in hyponym-hypernym relationships tend to occur in certain lexico-syntactic patterns:
  - The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.
    (Hearst, 1992)
- What are the hyponym and hypernym in this passage?
  - bonus: meronyms?
  - project!
  - recent Facebook AI article https://arxiv.org/pdf/1806.03191.pdf
  - synonym harvesting (towards datascience article)
- taxonomy vs ontology vs linked database

Hearst's Original Patterns

NP such as {NP}* {and|or} NP
such NP as {NP ,}* {or|and} NP
NP {, NP}* {,} or other NP
NP {, NP}* {,} and other NP
NP {,} including {NP, }* {or|and} NP
NP {,} especially {NP ,}* {or|and} NP

- needs: some chunking, maybe syntax
- Exercise: label each NP as indicating hyponym or hypernym

Normalization / Normalisation

semantic units:

USA = U.S.A. = U.S. = America?
fr: US = ricain = gringo ?
normalize = normalization?

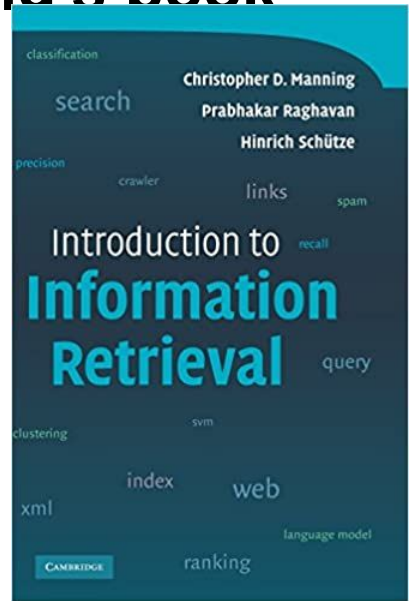
1. spell correction in documents
2. spell correction in queries

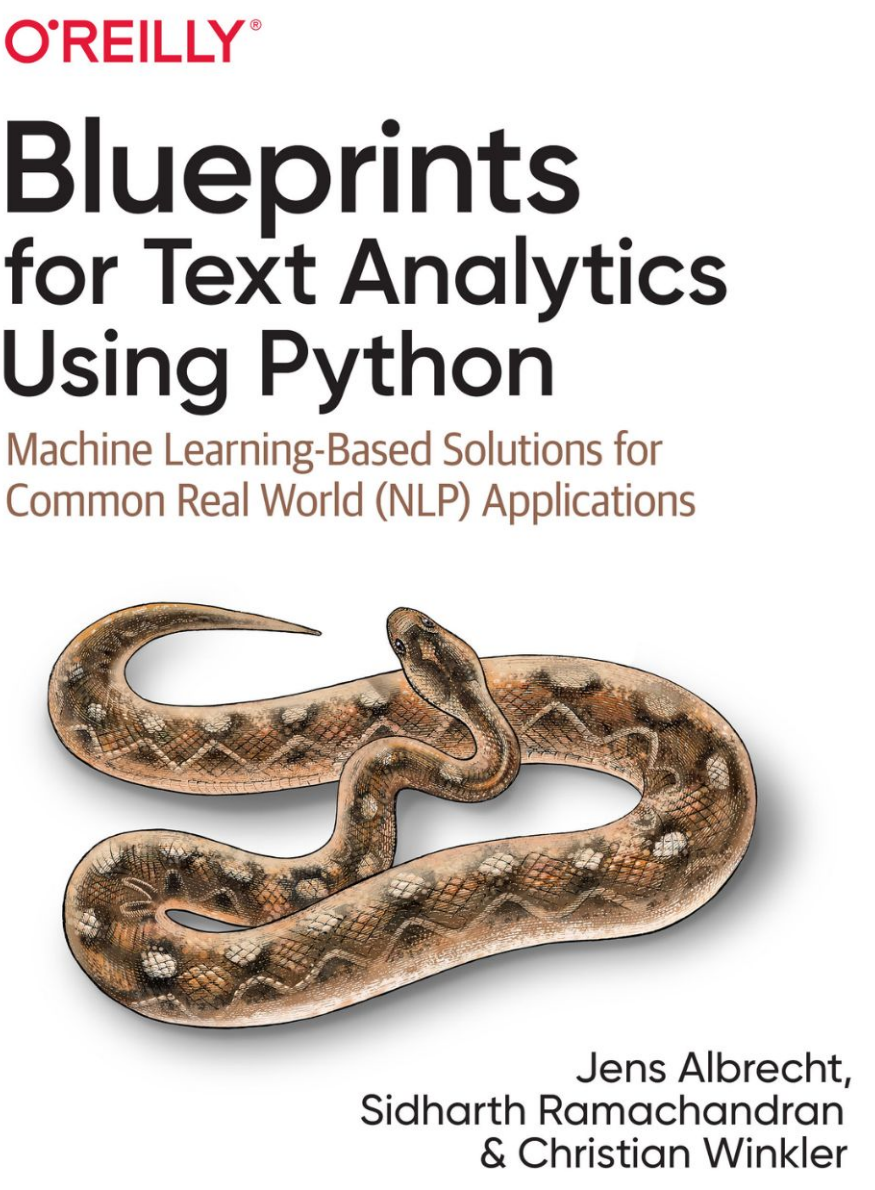
→ unique entry in the dictionary, queries and documents point to the dictionary.

- big subject: Levenshtein editing distance
  - weighted? OCR, typing, transcription
  - context?
  - word embedding! multi-words?

References

- First lessons based on this book: [Code on Github]
- Some more theoretical notions come from Manning's book also on Youtube
- Many online resources





Collaboration policies

- Homework: individual
- Final project: Groups of 2 or 3 students
  - Share the work evenly!
- Get inspired around the Web, take what you want, but cite it!!!
- Copy-paste of a few lines of code must be cited
  - “solution adapted from stackoverflow.com/xxxx”
- Copy-paste of a few words of explanations must be cited
  - “The tf–idf is the product of two statistics, term frequency and inverse document frequency.” (Wikipedia page on Tf-idf)
- It’s good to show that you know your way around. No shame!
- No quote → 0 points

Next steps

- retrieval techniques and evaluation
- grouping similar terms
- NER
- Hearst patterns

How will the field of Information Retrieval change with the arrival of chatbots such as chatGPT?

Chatbots like ChatGPT are built on top of language models that are trained on large datasets of text, and they can use this knowledge to understand user queries and generate responses. In the field of information retrieval, chatbots can provide a more natural and conversational interface for users to interact with information systems.

One potential impact of chatbots on information retrieval is that they could make it easier for users to find relevant information. With traditional search engines, users typically have to type in a few keywords and then sift through a list of search results to find what they're looking for. Chatbots, on the other hand, can engage in a natural language conversation with users and help them narrow down their search query to find exactly what they need.

Another potential impact of chatbots on information retrieval is that they could help automate some of the more tedious tasks involved in information retrieval. For example, a chatbot could be trained to search through large volumes of data and present the most relevant results to users, without requiring users to manually sift through the data themselves.

Overall, the arrival of chatbots like ChatGPT is likely to lead to new advances in the field of information retrieval, as researchers and developers explore new ways to leverage these powerful language models to help users find and interact with information more effectively.

The future of Information Retrieval

- Is the basis of IR wrong?
  - Do we actually want to find the most relevant document?
- What's the problem with chatGPT compared to traditional search engines?

The future of Information Retrieval

- Is the basis of IR wrong?
  - Do we actually want to find the most relevant document?
- What's the problem with chatGPT compared to traditional search engines?
  - No link to the source (SEO? Bard?)
  - combination of information from different sources (good or bad?)
  - No actual real-world understanding (and Google?)

What new discoveries from the James Space Webb Telescope can I tell my 9 year old about?

Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

- **In 2023, The JWST discovered a number of galaxies nicknamed "green peas."** They were given this name because they are small and round, like peas.
- **The telescope captured images of galaxies that are over 13 billion years old.** This means that the light from these galaxies has been traveling for over 13 billion years to reach us.
- **JWST took the very first pictures of a planet outside of our own solar system.** These distant worlds are called "exoplanets". Exo means "from outside".

These discoveries can spark a child's imagination about the infinite wonders of the universe

Check it

Information Retrieval in the age of Transformers

- IR as a subfield of NLG?
- Problem with project
  - last year’s project on line
  - programming projects in the age of Copilot
  - needs lots of data + GPU to do good stuff
    - ideas:
      - transforming documents in question and answer pairs
      - training T5 on it.
      - definitions, relations, facts

let’s break and move over to discord

- grab the notebook on https://ecampus.paris-saclay.fr/course/view.php?id=28064#section-8
  - start the download right away…
- and go to https://discord.gg/PKYTuvTF to the lounge
- today: VERY easy notebook (easy for students of Hands-On NLP)
  - try to look up stuff that isn’t that clear
- brace for next time ;)

  (BIG files, bring space on your hard drive)

Merci de votre

- attention
- considération
- intérêt
- écoute
- présence
- curiosité
- question

!