This session includes no mandatory programming, just designing, thinking, and using your imagination. There is no single "good" solution, there can be radically different proposals that are all good. We encourage you to propose projects in which you would be happy to work on for several weeks, or months.

More precisely, in this session:

- We will imagine a system that requires putting together (some of) the topics seen so far.
- We will find out how (if) the specific techniques and tools seen in class can contribute to this system.
- We will reflect about the limitations and additional possibilities of the tool.

1 The setting

Besides general document web search, there are many specialized settings that can benefit from the techniques and tools we have seen so far:

- An investigative journalist could benefit from exploring networks of news. News contain places, people, facts, dates, ... Finding who did business with whom, who appears more often in certain kind of news, where the action seems to happen for certain topics, ... could be of interest.
- Authorities want to discover money laundering or fraud schemes by investigating how money flows. If there is a limit of 100,000 euros for a certain kind of transaction, someone may ask or pay 100 friends to perform a transaction of 1000 euros each. Or, two companies who get a grant ("subvenció") for 100,000 euros each to do some work, actually do work for 50,000 euros each; to justify the rest, they invoice ("facturar") each other as subcontractors for the remaining 50,000 euros, for work that is never done. Networks can be highly complex so that human eyes cannot discover the tricks, but perhaps a machine can.
- In science, we publish papers in journals and conferences. Papers cite other papers. Citations give prestige, which brings professorship positions, good PhD students, and research grants. We propose to work on this setting.

Have a look at these three places where scientific work gets reported:

- Arxiv: `https://arxiv.org/`. Authors can upload papers in pdf, and there is a neatly organized page for each paper with basic info: title, authors, date, abstract, etc.
- Google Scholar: `https://scholar.google.com/`. Enter the name of a researcher (hint: most of your instructors are researchers and have a profile). You see most of the papers s/he has published, a profile with statistics of how many people have cited them, indicators such as h-index, lists of coauthors, papers that cite other papers, etc.
- DBLP, specialized in Computer Science. Google "DBLP researcher-name" and you see a page again with info of her/his publications. Nicely semi-structured format also with authors, titles, sometimes links to sources, bibtex citations, coauthors, etc.

  DBLP also has entries for conferences and journals. For example, searching "DBLP SIGIR" gives the page with the editions of an influential Information Retrieval conference, and searching "DBLP JMLR" gives the page of a top journal for Machine Learning research. From there you can get to lists of published papers and authors. And vice-versa, from an author page you can get to these pages.

As said before, we researchers care a lot about citations to our papers. And we are evaluated for them. Because of this, people try schemes to make their papers look influential even though they are junk. You can cite yourself a lot (self-citations) for no real reason, or you can band with other researchers and agree to cite each other. Some people even create their own conferences and journals where they publish papers that can be written in an afternoon; this should be detectable because nobody other than people in the same circle cites these conferences or journals.

Of course, I would rather be cited once by a Nobel Prize than be cited 1000 times by researchers that nobody cares about.

Hot topics in science appear and disappear all the time. Some topics just go out of fashion, or are abandoned because they don't seem to lead anywhere interesting. While sometimes there is a topic that explodes because one or two papers present spectacular results and everybody starts working on the topic.

Keeping track of the papers relevant to one's research is a real problem for most researchers. There are so many papers written every year, so many conferences and journals to watch, and so many connections within fields, that one discovers papers very relevant too late, or never.

2 What we know that can help

We would like to build a system that helps scientists in their work and in their careers. You now know:

- How to compare text documents with tf-idf and cosine measure.
- How to index large amounts of text, or other information, for efficient search.
- How to arrange large amounts of text documents into topics automatically
- How to crawl the Web or parts of it, and how to scrape content from semi-structured data.
- That Pagerank and topic-sensitive pagerank can approximate the reputation in linked data.
- That the above can be abused with content spam and link spam.
- How to look for duplicate or almost-duplicate items. (well this we will know soon enough!)
- A few techniques for scaling up, distributing indexes and other information, etc.

In the next weeks you will also learn about social network analysis (communities, discovering topics, information propagation) and recommender systems.

Of course there are 100's of useful tools in the Web if you look for them. For example, tools for parsing and extracting text from PDF documents.

3 What to do

The goal of this lab is to conceptualize a system that uses the information in the Web to help researchers with these issues. Implementing the system would be a multi-month, multi-person project. So we focus on the thought experiment.

There is no required programming, although you are welcome to try with parts of the project. For example, scraping DBLP and building a little network of researchers, or scraping arxiv, getting the papers of one particular person and extracting the citations it contains (note that Google Scholar must be doing this somehow).

We have mentioned three sources of information (Arxiv, Google Scholar, DBLP) but feel free to think of using others. For example, authors' homepages.

More precisely, you need to:

- Define the functionalities you would include in the system. Explain what each one would achieve: What the user needs to input, what s/he would get, why that would be interesting, and where to get the information to power it.
- Design the high-level architecture of the system. Draw some kind of block diagram with the different modules and how they connect. Explain the main processes that connect these modules: which are online (interact with the user) and which ones are offline (crawling, indexing...)
- Explain what techniques and technologies you would use for each part.
- Discuss functionalities that you don't know how to implement now, but might learn in the rest of CAI.
- Discuss or speculate freely on any other aspect you wish, including entrepreneurial ones.
- Discuss what limitations and risks you can see in the project. Technical, in data size, in legislation or whatever. Can you index 1000 papers or 10,000,000? Can you extract enough quality information from unstructured data? What volume of machines do you think you would need? What kind of response time would be achievable? How would you measure user satisfaction?

We know that the last part is hard because one does not know until one tries. Do your best. You will have to do this in your professional life all the time. We will value more the arguments you give than the exactness of the answer.

4 Deliverables

Rules: Same rules as in previous labs apply.

To deliver: A report in PDF with, say, 6-8 meaningful pages.

Procedure: Submit your work through the Racó.

Deadline: Work must be delivered within **2 weeks** from the lab. Late deliveries risk being penalized or not accepted at all. If you anticipate problems with the deadline, tell us as soon as possible.