Lab Project

Do it yourself

Task 1:

- Build a search engine for a specific domain.

Requirements / Results / Contributions:

- A collection of documents for the domain. At least 1000 documents.
- A search engine implementation for the document collection
  - A copy of the code repository; open source license of your choosing
  - Data and resources required
  - Sufficient documentation to compile and run it
- A short report (∼10 pages)
  - Motivation for the domain; what makes the domain/documents special?
  - Description of the search engine's architecture and retrieval model.
  - Evaluation of the search engine.
- A final presentation

Lab Project

Do it yourself

Task 2:

- Reimplement and reproduce a research paper.

Requirements / Results / Contributions:

- A selected paper from information retrieval or a related subject area.
- Implementation of its main algorithmic contribution and experiments
  - A copy of the code repository; open source license of your choosing
  - Data and resources required
  - Sufficient documentation to compile and run it
- A short report (∼10 pages)
  - Explanation of the paper and its motivation
  - Assessment of the paper in terms of reproducibility
  - Report on experiments carried out and how they match the original
- A final presentation

Lab Project

Organization

Group work:

- Work in groups of 2–6 people.
- Keep track of your work: code repository (e.g., Git).
  (the entire revision history must be handed in!)
- Code sharing allowed, but must be approved.
- Give small lightning talks about your progress in the lab class.

Tips:

- Get a vertical prototype going ASAP.
- Use APIs and libraries, but understand them.
- Share ideas and know-how among groups: irlecture.slack.com
- Search for solutions to specific problems: Google, StackOverflow.

Lab Project

Datasets / Papers

Search for a dataset. Items must include a plain text field.

Possible sources:

- Our datasets
- Google Dataset search
- Kaggle datasets
- Amazon public datasets,
- Data dumps from Wikipedia, StackOverflow, Reddit, etc.

What are your ideas?

Regarding paper reproducibility, see: CENTRE

Paper choice needs to be discussed on an individual basis.

Topic and team settlement in next week's lab class.