EPL660: Information Retrieval and Web Search Engines
FINAL PROJECT

Written **research projects** must be done **individually or in groups of no more than two**. **Implementation projects** can be done in groups of **up to 3**.

Each group or individual must submit your preference projects proposal (e.g. W2, I2, W3) to be approved (email to instructors: P. Antoniou and cc-ed G. Pallis) **no later than March 11**.

Written Projects

Written projects involve doing an in-depth study, survey, or evaluation of one or more topics related to information retrieval and Web search engines. The project can take a form of a research paper examining the use of a specific technique or model in various IR systems, or it can be a detailed case study involving two or more existing IR systems. In either case, the paper should contain a summary and a technical evaluation of the state-of-the-art related to the particular topic studied. If the paper involves a case study, then a thorough comparative evaluation with other similar systems must be provided. A research paper should present a new idea or provide a detailed survey of methods to solve a specific IR-related problem. The approach presented should be, at least in part, a novel and original contribution, and should ideally be evaluated experimentally. A research paper could be good start for a Masters or Ph.D. research project. The maximum length for the written projects is 20 single-spaced pages (12 point font), including figures and references. The evaluation of the papers will be based on clarity, thoroughness, and soundness of ideas and concepts presented, as well as the overall organization of the paper.

Note: Written project should not simply be a summary of some of the material covered directly in the lectures, but rather should go beyond this material in one or more specific areas related to that material.

W1. A comparative study of implementation techniques for scalable information retrieval on large-scale search engines or Web-based information systems (such as Google, Facebook, etc.). This study must include an analysis of challenges in managing and leveraging large data repositories and various proposed and implemented solutions (such as Big Table, Map Reduce, and other approaches based on "cloud computing"). The study can also focus on implementation platforms that enable scalable retrieval (e.g., Hadoop).

W2. Deep Learning for Information Retrieval. Recent years have observed a significant progress in information retrieval and natural language processing with deep learning technologies successfully applied into almost all of their major problems. The key to the success of deep learning is its capability of accurately learning distributed representations (vector representations) of natural language expressions such as sentences, and effectively utilizing the representations in various tasks. This study aims at summarizing and introducing the recent results of research on deep learning for information retrieval, in order to stimulate and foster more significant research and development work on the topic in the future.

W3. Study of the use of social network analysis (SNA) and its use in information retrieval. This study should include a detailed summary of various techniques from SNA and their use in providing relevant information to users in online social network and/or traditional search engines. This project should also include related works in social network aware search, and the related subtasks of information retrieval, aggregation, and ranking in a social context.
Implementation Projects

I1. Implement an advanced search engine for Research Projects at CS department:

- Your system should be able to crawl the research projects that have been published at CS UCY web site as well as fetch related projects by searching in other portals (e.g., EU Commission portal, NDF portal).
- Provide advanced search capabilities (e.g., search by name, principal investigator, topic, time).
- The results will be clustered by their description in topics.
- Provide a dynamic summary of the results and visualization statistics.

I2. Design a search agent for CS Web site:

- Your system should be able to search CS Web site for documents relating to a user query.
- An index of the extracted information can be available locally and updated on a regular basis.
- The query interface (and language) can be restricted (based on characteristics of the CS Web site) to make the search and matching process easier and more efficient.

I3. Propose a project:

Email a paragraph describing your proposed project to the instructors of the course. Include as much detail as possible. The instructors will reply with any concerns about the content or scope of the project. If you are proposing a project with a partner, one partner should email the description and the other partner should email a confirmation of his/her involvement.

Guidelines

Written Projects

Written projects will be evaluated based on thoroughness, soundness, clarity and organization. The overall structure of the paper is up to you, but you must have the following sections in addition to the main body of the paper:

- **Abstract:** This is a short synopsis of the main points of the paper. This should be 200-300 words, and should appear along with the title and your name, ID number, and email, on the first page. The rest of the paper should start on page 2.
- **Conclusion:** Summarize your conclusions and findings. Keep this to 300-400 words.
- **References:** This is a list of references that you have used in doing your research and throughout your paper. The references should be numbered and the number for the reference should appear in the appropriate places in the text of the paper where the reference was used (it is not enough to list a bibliography at the end of the paper without actually using any references within the body of the paper). You can look at any of the papers included in the optional reading for acceptable uses of references. URL references should only be used for referring to specific system Web sites and not as a way to reference published work.

Your final paper should be sent by email either PDF or Postscript formats. Submissions in other formats will not be accepted.

Project presentation (15 min per team + 5 min Q/A)

Implementation Projects

You will need to electronically submit a compressed file by email containing your project distribution files and documentation. Your project documentation should contain the following components:

- A detailed description of your system (including specific techniques and algorithms you used), and the interaction between the components (make references to code segments, modules, methods, functions, etc. as necessary). **If you used any outside**

sources in your implementation, please clearly indicate which sources, and how and where they were used.

- A complete sample run of your program with description, illustrating how your system works, along with any intermediate input or output used for the sample run.

Your project distribution files should contain the following:

- Complete source code (**be sure that your source code is fully documented and easy to read**).
- Binary files (e.g., executables, DLLs, Class files) or other components necessary to run your program.
- Readme file containing instructions on how to compile, install, and/or run your program.
- If your application is CGI-based or otherwise has a server component, please provide a URL for a demo version of your system.

Project presentation (15 min per team + 5 min Q/A)