CSCI 599: Special Topic
Spring 2012 TOPIC: Applications of Natural Language Processing: Information Retrieval
(3 UNITS)

Applications of Natural Language Processing covers various applications of human language technology in greater depth than in CSCI 544 or CSCI 562. There are three complementary linked courses: the first covers Machine Translation; the second covers Information Retrieval, and the third covers Information Extraction and Text Summarization. These are taught successively, once per year.

This syllabus describes Information Retrieval.

Instructors: Anton Leuski (leuski@ict.usc.edu), Don Metzler (metzler@isi.edu). Office hours immediately follow each lecture.

Prerequisites: Permission of instructor. Students should have familiarity with natural language processing and be comfortable with medium-sized programming projects.

Goals: Information Retrieval (IR) is the science of searching for and making sense of information from large collections of text. It synthesizes topics from computer science, mathematics, linguistics, and psychology. As the amount of digital content continues to grow, there is an increased need to be able to effectively and efficiently search, organize, and understand it. While Web search engines (e.g., Bing, Google, and Yandex) are the best-known IR applications, there are many other areas to which IR can be applied. These include targeted advertising, recommender systems (e.g., Amazon user suggestions), cross-lingual search, and spam filtering. Two important areas where IR saw significant growth in recent years are e-discovery and medical search. The former deals with automatic organization and sense-making of legal documents. The latter operates in the domain of medical articles and focuses on the development of techniques for automatic knowledge extraction. Constantly emerging application areas means that there are many employment opportunities in the field of IR. The job market is poised to grow as more and more companies and organizations accumulate large collections of digital content.

Textbook: W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice.* 2009.

Other recommended reading:

- S. Buettcher, C. L. A. Clarke, G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*. 2010.
- C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval.* 2008.
- C. J. van Rijsbergen, *Information Retrieval.* 1979.
- I. H. Witten, A. Moffat, T. C. Bell, *Managing Gigabytes.* 1999.
- A. Moffat, J. Zobel, D. Hawking, *Recommended Reading for IR Research Students.* 2004.

Requirements

- 3 programming/homework assignments: 35%
- Midterm exam: 20%
- Final exam: 20%
- Final project: 25%

Course outline (29 lectures)

1. Course introduction, administrativia, What is IR? Problems in IR.
2. Architecture of a search engine. Building blocks. [CMS 2.]
3. Crawls and feeds. Documents. Conversion. Storage. Duplicates. [CMS 3.]
4. Text processing. Text statistics. Term weighting. Parsing. [CMS 4.]
5. Tokenization. Stopping. Stemming. Phrases. N-grams. Markup. Links. IE. Internationalization. [CMS 4.]
6. Indexes. Inverted index. Counts, positions, fields, scores, ordering. Compression. [CMS 4 & 5]
7. Query processing. Synonymy, Polysemy and Relevance Feedback. [CMS 5 & 6.]
8. Clustering and classification. [CMS 9.]
9. User Interfaces. User models. Snippets. [CMS 6.]
10. Retrieval models. Boolean. Vector space. [CMS 7.]
11. Retrieval models. Probabilistic, inference networks, and logic models. [CMS 7.]
12. Retrieval models. Language models. Relevance models. [CMS 7.]
13. Evaluation. Effectiveness and efficiency. Logging. Metrics. [CMS 8.]
14. Midterm exam
15. Filtering. Topic detection and tracking.
16. Image and video retrieval.
17. Learning to Rank
18. Web search. Local search.
19. Advertising.
20. Distributed IR. MapReduce. P2P.
21. Cross-language retrieval
22. Question answering
23. Social Search [CMS 10.]
24. Semi-structured data. Term dependency. [CMS 11.]
25. E-discovery. Legal search.
26. NLP approaches. [CMS 11.]
27. The future of IR
28. Final project presentations
29. Final project presentations

Course policies

Students are expected to submit only their own work for homework assignments. They may discuss the assignments with one another but may not collaborate with or copy from one another. University policies on academic integrity will be closely observed. All assignments and the project will be due at the beginning of class on the due date. Late assignments will be accepted with a 7% penalty for each day after the due date, up to a week after the due date. No exceptions can be made except for a grave reason.

Statement for Students with Disabilities

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to the instructor as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776.

Statement on Academic Integrity

USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor,

and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. Scampus, the Student Guidebook, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov/. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/.