03-60-538: Information Retrieval Systems (2018 Fall)

The instructor is Professor Jianguo Lu from School of Computer Science, University of Windsor.

- Email: jlu@uwindsor.ca
- Phone: 519 253 3000 ext 3786
- Course web site: Blackboard system and `http://cs.uwindsor.ca/~jlu/538`
- Instructor's web site: `http://cs.uwindsor.ca/~jlu`
- Office: 5111 Lambton Tower
- Office hours: Tuesday and Thursday 11:00-12:00.

1 Course overview

"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)." **?**, Chapter 1. The most common application of information retrieval is web search engines. This course will cover the key issues in search engine construction. The tentative topics covered in this course include major components of search engines:

- Text operations before indexing, such as stop word removal, stemming;
- Indexing, constructing an inverted index of word to document pointers;
- Searching, retrieving documents that contain a given query token from the inverted index; Use Lucene to construct a practical and large search engine.
- Various retrieval models, including the classic boolean model and vector space model. TF-IDF and their variants;
- Statistic properties of text, power laws, Zipf's law, Heaps' law.
- Language models, unigram model and bigram model.
- Ranking, scoring retrieved documents according to relevance or importance metrics. PageRank algorithm. Markov chain.
- Evaluation criteria, precision and recall, F1.
- Document classification. Naive Bayes classification (multinomial and Bernoulli) , Feature selection. Mutual information. Feature transformation.

- Neural Network based text processing, Vector representation of words and documents. Latent semantic indexing. Singular value decomposition, word2vec, and doc2vec.
- Document clustering, K-means, HAC.
- Crawling, collecting web documents. Crawling tools. Near duplicate detection. MinHash algorithm.

2 Learning outcomes

After the course you will be able to

- Construct an industrial size search engine;
- Understand text statistics and language models, the power law in natural language, Zipf's law, Heaps' law;
- Understand some fast algorithms in search engine constructing, indexing, vector space and TF-IDF information retrieval model;
- Understand the link/graph analysis, in particular PageRank algorithm;
- Classify and cluster documents using various machine learning algorithms;
- Apply deep learning in text processing;
- Crawl web documents.

3 Grading scheme

The grading will be based on mainly on exam and project. The weight of the final exam is 40%, the project is 50%, class participation and presentation is 10%. The project is about constructing a real search engine for academic papers. It is divided into two components. You need to accomplish each component in time.

4 Text books

I will mainly follow the IIR book listed below. This is an excellent book and also available online. Some algorithms for large data processing are described in more detail in the MMD book, which is also available online. The third book (LA) is a practical introduction to Lucene search engine.

IIR Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze,ŁIntroduction to Information RetrievalŁ, Cambridge University Press. 2008.

MMD Anand Rajaraman and Jeff Ullman, Mining of massive datasets , 2013.

LA Michael McCandless, Erik Hatcher, and Otis Gospodnetic, Lucene in Action, Second Edition. 2010.

5 SET

Student Evaluation of Teaching forms will be administered in the first class of week 12, the last week.