CS657A: Information Retrieval Introduction

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs657/

2nd semester, 2021-22

Tue 1030-1145, Thu 1200-1315

Rules

- No prerequisites except general aptitude (and, of course, basics of computer science)
- Email `arnabb@cse.iitk.ac.in` to set up appointment
- Put "CS657A" in the subject so that automatic filter catches it
- Participate
  - Attend classes
  - Clear doubts
  - Answer questions
- Do assignments individually unless asked otherwise
- No extension of deadlines for degradation of health of
  - Your computer
  - Your family members, friends, etc.
- If you are sick, follow IITK procedure
  - Produce a sick certificate, etc.

Grading Policy

- Exams: 15-25%
- Project: 30-45%
  - Form your own idea
  - Discuss with instructor and get it approved
  - Just implementation or survey will not be enough
  - Back it up with analysis
- Assignments: 20-30%
  - Possibly 2
- Paper presentation and discussion: 10%
  - Read abstracts and shortlist
  - Needs instructor's approval
- Changes may be done with mutual consent
  - Will be discussed and fixed in the class

Course Material

- Slides
- Books
  - "Introduction to Information Retrieval" by C. D. Manning, P. Raghavan, H. Schütze, Cambridge University Press.
  - "Modern Information Retrieval: The Concepts and Technology behind Search" by R. Baeza-Yates, B. Ribeiro-Neto, Addison-Wesley.
  - "Information Retrieval: Algorithms and Heuristics" by D. A. Grossman, O. Frieder, Springer.
- Journals/conferences/forums for project and paper ideas, etc.
  - SIGIR, CIKM, WWW, ECIR, SPIRE, etc.
  - FIRE, TREC, CLEF, etc.
  - TOIS, IP&M, JASIST, IR, TALLIP, etc.

Course Contents

1. What is information retrieval (IR)?
   - Retrieval of information (mostly text) efficiently from a large collection of objects (mostly documents)
   - Motivation
2. Basic document retrieval
   - Inverted index
   - Querying using inverted index
3. Performance evaluation
   - Precision, recall, F-score, etc.
4. Tokenization
   - Word segmentation
   - Stopwords
   - Stemming
5. Document scoring
   - Zone scoring
   - Term Frequency
   - Inverse Document Frequency
   - Tf-idf

Course Contents (contd.)

6. Document as a vector
   - Vector model
   - Document similarity
   - Document vector models
     - LDA
     - GLoVe
     - Word2Vec
7. Scalability
   - Skip list
   - Champion list
   - Tiered index
8. IR as a system
   - Indri – `http://sourceforge.net/projects/lemur/`
   - Terrier – `http://terrier.org/`
   - Lucene – `http://lucene.apache.org/`
9. IR for non-documents
   - Images
   - Graphs
   - Audio