Information Retrieval and Web Search (IRWS)

Department of Computational and Data Sciences

January 19, 2021

Course Description:

Information Retrieval forms the foundation of the modern search engines, and IR (popular acronym for Information Retrieval) is often called as the *science behind search*. Although IR systems are mostly associated with Web search engines (e.g., Bing, Google, Yandex etc.), there are significant applications of IR in digital library search, patent search, and automatic question-answering, to name a few. Likewise, IR models (the underlying algorithm behind retrieval systems) are adopted to solve a wide range of problems, such as organizing documents into an ontology, recommending news stories to users, detecting spam, and efficiently address information need to the users. This course will provide an overview of the theory, implementation, and evaluation of IR techniques. In particular, we will explore how search engines work, how they "interpret" human language, what different users expect from them, how they are evaluated, why they sometimes fail, and how they might be improved in the future. For hands-on experience, we will use *PyLucene*[1], a robust, industry standard search engine with a Python wrapper.

Prerequisite(s):

- Basic concepts of Computer Science and Data Structures (CS3101, CS3201).
- Basic probability (conditional probability, Bayes theorem etc.).
- Programming knowledge for practicals (Programming in Python: knowledge of packages, modules, functions etc.).

Textbooks:

- *Introduction to Information Retrieval*
  C. D. Manning, P. Raghavan and H. Schutze
  ISBN: 978-0-521-86571-5
  `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`
- *Information Retrieval: Implementing and Evaluating Search Engines*
  S. Buttcher, C. L. A. Clarke, G. Cormack.
  ISBN: 978-0-262-02651-2
  `http://www.ir.uwaterloo.ca/book/`

[1] `https://lucene.apache.org/pylucene/`

Course Outline (tentative) and Syllabus:

MODULE 1: INTRODUCTION, MODELS, EVALUATION

Lecture 1 - 2: Introduction:

- Brief history and evolution of Information Retrieval (IR).
- Overview of applications.

Lecture 3 - 5: Basic idea of IR:

- Document representation.
- Controlled vocabulary.
- Free text representation.
- Inverted index.
- Term-document incidence matrix.
- Text processing: stopword removal, stemming, lemmatization.
- Statistical properties of text:
  - Zipf's law.
  - Heaps's law.
  - Term co-occurrence.
- Keyword search.
- Text search (retrieval).

Lecture 6 - 8: Index structures:

- Index creation.
- Index compression.
- Query execution.

Lecture 9 - 11: Lucene@work - 1:

- Introduction to Lucene and PyLucene.
- Different components of Lucene.
- Indexing using PyLucene.

Lecture 12 - 15: Retrieval Models:

- Boolean retrieval.
- Ranked retrieval.
- Vector space model.
- Term weighting.
- Axiomatic justification of term weighting (brief introduction).

Lecture 16 - 17: Probabilistic model for IR:

- BM25.
- Other variants of TF-IDF models.

Lecture 18 - 20: Language modeling for IR:

- Query likelihood.
- KL divergence.
- Smoothing (Dirichlet, Jelinek-Mercer).

Lecture 21 - 22: Lucene@work - 2:

- Searching with PyLucene.

Lecture 23 - 25: Evaluation:

- Confusion matrix.
- Recall, Precision, MAP, NDCG, MRR, other commonly used metrics.
- Evaluation forums and their roles.
- Statistical testing of a pair of evaluation.

Lecture 26: Lucene@work - 3:

- How to Implement a new retrieval model.

Lecture 27 - 29:

- Relevance feedback: Rocchio's method, variations.
- Query expansion in the vector space model.
- Pseudo relevance feedback.

Lecture 30 - 32:

- Relevance based language model and variation.

Lecture 33 - 34:

- Web Search.
- Web document preprocessing: parsing, segmentation, deduplication, shingling.
- Crawling, focused crawling, meta-crawlers.
- Link analysis: hubs and authorities, Google PageRank.

Lecture 35 - 36:

- Discussion on different corpora, forums (TREC, CLEF, NTCIR, FIRE) and their tasks.
- Presentation on some notable recent works.