

# Advanced Information Retrieval - 2022

## Tentative Course Plan

Term: Jan-May 2022

Welcome to the Advanced Information Retrieval course. Information Retrieval has matured on several aspects in the last decade. We have moved from keyword-based search to semantic search. Several concepts such as knowledgebases and entities have been widely accepted in practice. Ontologies, RDF Triplets and SPARQL have found its way into the industry. In this course, we survey these ideas as we delve into recent trends of information retrieval. This is a four-credit course. You **must have** completed Information Retrieval course with a passing grade to register for this course.

## Key Learning Objectives

The key objectives of this course are as follows:

- To transition from keyword-based search to semantic search
- To take a deeper dive into query understanding, retrieval models and ranking
- To understand the recent trends in IR practice and research

## Course Plan

Couse is divided into two parts. Firstly, there is a semester long project starting from day 1. We use the project to drive the course. Entire class teams up to do a single project. We may have more than one team if the class size grows beyond five. We meet once a week to discuss topics relevant to the project and to take stock of the progress. Secondly, we cover the following topics through a mix of instructor-led teaching (20%) and student presentations (80%). The topics and the sources are listed below. We meet once every week to discuss these topics.

1. Information Retrieval
  - a. Science of Information [IT]
  - b. Search Techniques [Chapter 4 of BBH]
  - c. Survey of Advanced Search Techniques [Chapter 5 of BBH]
2. Query Understanding
  - a. Query understanding challenges - Misspellings, Query Segmentation, Need for Query Expansion, Need for Query Reformulation [CPS]
  - b. Query annotation with semantic information [SQA]
  - c. Generic Intent Representation [AH]
  - d. Semantic Query Understanding [SQU, TQI]
3. Review of Retrieval Models
  - a. Boolean, VSM, vs. Probabilistic [CPS]
  - b. Pivoted Length Normalization
  - c. Learning to Rank

4. Semantic Search
  - a. Limitations of Keyword Search [SWP]
  - b. Semantic Search [SWP]
  - c. Embeddings - Word2Vec [MIK]
  - d. Ontology [OWL], RDF Triplets, SPARQL [RDF, SPQL]
  - e. Constructing knowledgebases and Knowledge Graphs [KB]
  - f. Entity Retrieval [RT]
5. Tools
  - a. Lucene
  - b. Solr
  - c. Stanford NER

## Evaluation

- Project: 40% (10% for progress till mid-term; 30% for the rest;)
- (Reading) Assignments & student Presentations: 3 \* 10% = 30%
  - Individual. Lead a discussion in the class on the topic for 30 minutes. Submit a one-page summary of the topic.
- Final Exam: 30%

## References

- [IT] Chapters 1, 2 (Sections 2.1 - 2.5) of A First Course in Information Theory
- [AH] Generalized Syntactic and Semantic Models of Query Reformulation
- [SQU] Semantic Query Understanding
- [SQA] Intent-Aware Semantic Query Annotation
- [TQI] Understanding Temporal Query Intent
- [CPS] An Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze.
- [MIK] Efficient Estimation of Word Representations in Vector Space
- [OWL] Sections 2, 3, 4 and 5 of OWL Primer at w3.org.  
<https://www.w3.org/2012/pdf/REC-owl2-primer-20121211.pdf>
- [RDF] Chapter 1 of RDF Primer <https://www.w3.org/TR/rdf11-concepts/>
- [SPQL] Chapters 1 and 2 of SPARQL Primer <https://www.w3.org/TR/rdf-sparql-query/>.
- [KB] <http://pages.cs.wisc.edu/~anhai/papers/kcs-sigmod13.pdf>
- [RT] Chapter 4 of Rhode's thesis  
[https://research.utwente.nl/files/6040940/thesis\\_H\\_Rode.pdf](https://research.utwente.nl/files/6040940/thesis_H_Rode.pdf)
- [SWP] A Semantic Web Primer, Grigoris Antoniou and Frank van Harmelen
- [SWR] The Semantic Web Revisited  
[https://eprints.soton.ac.uk/262614/1/Semantic\\_Web\\_Revisited.pdf](https://eprints.soton.ac.uk/262614/1/Semantic_Web_Revisited.pdf)
- [CGP]  
<https://github.com/vvtesh/vvtesh.github.io/blob/v1/teaching/advir2021/VLDBTemplate.zip>