Course Outline

The course consists of **6 weekly modules**. Please note: There are no required readings for this course. All readings listed below are optional and are primarily from the textbook

C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016.

Week 1

Key Concepts:

- Part of Speech tagging, syntactic analysis, semantic analysis, and ambiguity
- “Bag of words” representation
- Push, pull, querying, browsing
- Probability ranking principle
- Relevance
- Vector space model
- Dot product
- Bit vector representation

Recommended Readings:

- N. J. Belkin and W. B. Croft. 1992. *Information filtering and information retrieval: Two sides of the same coin?* Commun. ACM 35, 12 (Dec. 1992), 29-38.
- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapters 1 - 6**.

Week 2

Key Concepts:

- Term frequency (TF)
- Document frequency (DF) and inverse document frequency (IDF)
- TF transformation
- Pivoted length normalization
- BM25
- Inverted index and postings
- Binary coding, unary coding, gamma-coding, and d-gap
- Zipf’s law

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 6 - Section 6.3, and Chapter 8**.
- Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*, Second Edition. Morgan Kaufmann, 1999.

Week 3

Key Concepts:

- Cranfield evaluation methodology
- Precision and recall
- Average precision, mean average precision (MAP), and geometric mean average precision (gMAP)
- Reciprocal rank and mean reciprocal rank
- F-measure
- Normalized Discounted Cumulative Gain (nDCG)
- Statistical significance test

Recommended Readings:

- Mark Sanderson. *Test collection based evaluation of information retrieval systems.* Foundations and Trends in Information Retrieval 4, 4 (2010), 247-375.
- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 9**

Week 4

Key Concepts:

- p(R=1|q,d), query likelihood, and p(q|d)
- Statistical and unigram language models
- Maximum likelihood estimate
- Background, collection, and document language models
- Smoothing of unigram language models
- Relation between query likelihood and TF-IDF weighting
- Linear interpolation (i.e., Jelinek-Mercer) smoothing
- Dirichlet Prior smoothing

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapter 6 - Section 6.4**

Week 5

Key Concepts:

- Relevance feedback
- Pseudo-relevance feedback
- Implicit feedback
- Rocchio feedback
- Kullback-Leiber divergence (KL-divergence) retrieval function
- Mixture language model
- Scalability and efficiency
- Spams
- Crawler, focused crawling, and incremental crawling
- Google File System (GFS)
- MapReduce
- Link analysis and anchor text
- PageRank and HITS

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapters 7 & 10**

Week 6

Key Concepts:

- Learning to rank, features, and logistic regression
- Content-based filtering
- Collaborative filtering
- Beta-Gamma threshold learning
- Linear utility
- User profile
- Exploration-exploitation tradeoff
- Memory-based collaborative filtering
- Cold start

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*, ACM Book Series, Morgan & Claypool Publishers, 2016. **Chapters 10 - Section 10.4, Chapter 11**.

Elements of This Course

The course is comprised of the following elements:

Lecture videos. Each week your instructor will teach you the concepts you need to know through a collection of short video lectures. You may either stream these videos for playback within the browser by clicking on their titles, or you can download each video for later offline playback by clicking the download icon.

The videos in each week is usually total 1.5 to 2 hours, but you generally need to spend at least the same amount of time digesting the content in the video. The actual amount of time needed to digest the content would naturally vary according to your background.

Quizzes. Each week will include one for-credit quiz. Your cumulative score will be used when calculating your final score in the class. There is no time limit on how long you take to complete each quiz, and you will be allowed 3 attempts at the quiz every 8 hours. The deadline for all quizzes is the last day of the course.

The quiz in each week should take less than 1 hour to finish, assuming you have mastered the materials to be tested in the quiz, and you should make sure that you have mastered the materials before you attempt to work on the quizzes. You should use the practices to help you understand and master the materials.

Programming assignments. The programming assignments for this course are optional, but they provide an opportunity for you to practice your programming skills and apply what you've learned in the course.

The programming assignment is optional and you can budget about 2 hours each week to work on it if you plan to finish it; you may need to budget more time for this if you are not familiar with C++ programming.

Information about Lectures