Information Retrieval & Text Mining [MS Data Science]

By: **Dr. Saeed Ul Hassan**

TAs:
Hadia Irshad [MSDS19016@itu.edu.pk]
Muhammad Sohaib Khalid [MSDS19096@itu.edu.pk]

Course Description

Recent years have seen a dramatic growth of natural language text data, including web pages, news articles, scientific literature, emails, enterprise documents, and social media such as blog articles, forum posts, product reviews, and tweets. Text data are unique in that they are usually generated directly by humans rather than a computer system or sensors, and are thus especially valuable for discovering knowledge about people's opinions and preferences, in addition to many other kinds of knowledge that we encode in text.

This course will cover technologies, which play an important role in any data mining applications involving text data for two reasons. First, while the raw data may be large for any particular problem, it is often a relatively small subset of the data that are relevant, and a search engine is an essential tool for quickly discovering a small subset of relevant text data in a large text collection. Second, search engines are needed to help analysts interpret any patterns discovered in the data by allowing them to examine the relevant original text data to make sense of any discovered pattern. You will learn the basic concepts, principles, and the major techniques in text retrieval, which is the underlying science of search engines.

Course Goals and Objectives

By the end the course, students will be able to do the following:

- Explain many basic concepts and multiple major algorithms in text retrieval and search engines.
- Explain how search engines and recommender systems work and how to quantitatively evaluate a search engine.
- Create a test collection, run text retrieval experiments, and experiment with ideas for improving a search engine.

Course Book

- 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.

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**.

Course Outline

Phase 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

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**.

Phase 2:
Key Concepts:
- Term frequency (TF)
- Document frequency (DF) and inverse document frequency (IDF)
- TF transformation
- BM25
- Inverted index and postings
- Binary coding, unary coding, gamma-coding, and d-gap

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.

Phase 3:
Key Concepts:
- 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**

Mid Term Exam

Phase 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 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**

Phase 5:

Key Concepts:

- Relevance feedback
- Pseudo-relevance feedback
- Implicit feedback
- Rocchio feedback
- 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**

Phase 6

Key Concepts:

- Learning to rank, features, and logistic regression
- Content-based filtering
- Collaborative filtering
- Beta-Gamma threshold learning
- User profile
- Exploration-exploitation tradeoff

Term paper submission

Final Exam