KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

CS 4422 — Information Retrieval

Syllabus

Fall 2021 · Kennesaw State University

General Information

Lectures	**[CS4422]** MWF 1:25 PM - 2:15 PM
Classroom	Atrium Building (**J251**), Marietta campus
Course Webpage	`https://jiho.us/teaching/ir/21f`
Instructor	Jiho Noh (jnoh3@kennesaw.edu)
Office Hours	M 2:30 PM - 4:30 PM or by appointment at J341 Atrium Building

Course Description

Information retrieval (IR) methods are an indispensable tool in the current landscape of exponentially increasing text-based information, especially on the Web. Conventionally, IR tasks involves fetching and ranking a set of documents from a large corpus in terms of relevance to a user's information need. IR has been one of the most important problems in the domain of natural language processing (NLP). The application of IR techniques has been evolved with the advances of machine learning (deep learning) and data science.

Prerequisites

CS 3305 and (CS 3410 or CSE 3153)
Students need to be familiar with basic algebra, calculus, probability and statistics.

Proficiency in Python
Students should be familiar with programming in Python which is one of the most popular language choice for data science. All the assignments and lecture demos will be in Python.

Textbook

Textbooks are available online. It is recommended to read the textbook sections specified in the schedule table and the additional reading materials.

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines Information Retrieval in Practice, 2015

Course Schedule and Lecture Topics

In this course, we will cover basic and advanced techniques for building text-based information systems, including the following topics:

Week of	Topics
Aug. 16th	IR introduction - Search engines, Document retrieval pipeline, challenges in IR Crawling basics - HTTP Request, Web exploration strategies - Politeness policy, Duplicate detection
Aug. 23rd	Text preprocessing - Zipf's law, Heap's Law - Regular expression, Parsing structured documents - Tokenization / Lemmatization / Stemming
Aug. 30th	Indexing - Inverted index, Vocabulary - Index construction (Sort-based / BSBI / SPIMI)
Sep. 6th	Set-based retrieval models - Boolean retrieval Algebraic retrieval models - Document-Term matrix, Term weighting using TF.IDF - Vector space models (VSM) and cosine similarity - Latent semantic Analysis (LSA)
Sep. 13th	Probabilistic retrieval models - Bayes' rule, Binary independence model (BIM) - Language models, Query-likelihood model, Okapi BM25, Smoothing techniques
Sep. 20th	Evaluation metrics for IR - confusion matrix, precision, recall, F-score - precision@K, AP, R-prec, nDCG, ROC, Precision-recall curves Test collections - Cranfield paradigm, TREC
Sep. 27th	Link graph - Webgraph, Markov Chains, PageRank, HITS, SALSA
Oct. 4th	Relevance feedback and query transformation - spell checking and suggestions - query expansion, controlled vocabulary, relevance feedback
Oct. 11th	Text classification - Naive Bayes classifier, kNN, Rocchio, Decision tree Named entity recognition (NER)
Oct. 18th	Document clustering - topic modeling, LDA, PCA
Oct. 25th	Recommender Systems - collaborative / content-based filtering
Nov. 1st	Learning to rank - SVM, Boosting techniques - RankSVM, RankBoost, LambdaMART, NDCG Boost
Nov. 15th	NN for IR - Basics of neural networks - Distributed representations for words - semantic matching, query transformation - document summarization, named entity recognition
Nov. 22th	Fall break
Nov. 29th	Literature review

Course Outcomes

Students who complete this course successfully will be able to

1. L01 Describe the Web crawling system architecture and write/run a crawling program that scrapes webpages under the specified domain. (PA1)
2. L02 Explain the motivations of different retrieval models, which include Okapi BM25 and Query-likelihood model. (PS1, PS2)
3. L03 Write a program that parses textual data and creates an index using Elasticsearch. (PA2)
4. L04 Analyze performance characteristics of the IR systems in terms of the commonly used evaluation metrics. (PS3)
5. L05 Perform machine learning algorithms with textual data, such as the classification and clustering tasks. (PA3, PS4, PS5)
6. L06 Apply the trending deep learning methods for the IR tasks, such as conversational information retrieval, natural language understanding/generation. (Test1, PS6)

(* PS: Problem Set, PA: Programming Assignment, Test: Final Exam)

Attendance

Students are expected to attend class *regularly*, although attendance is not required. In the event that a student must miss a class, the student is responsible for finding out what assignments were made, what due dates were announced, what handouts were given, and what material was covered. All the assignments and problem sets will be posted on the course website promptly. Any student who violates the rules for civil behavior in class will be told to leave the class.

Course Communication

Students are encouraged to use only their official KSU email account (or via D2L) since emails from other accounts may not successfully reach the instructor. Please, include the course number in the email subject. The emails sent via D2L will be responded within max 2 days. If you need immediate response, please use the instructor's email address directly.

Evaluation Criteria

	CS 4422
programming assignments	40%
problem sets	40%
final exam	20%

Programming Assignments

There will be *three programming assignments*. All the codes and documentation should be zipped and submitted before the specified due date. You may be asked to demo the code and results during office hours for particular assignments. More details will be provided during the classes.

Problem Sets

There will be *six problem sets*. Each set will cover approximately two weeks of topics. You may collaborate on the problems, but your write-up must be your independent work. Transcribed solutions are unacceptable. The lowest of the six score will be dropped.

Late Acceptance

Due dates for the programming assignments and problem sets will be specified on the instructions. Late submissions will be accepted up to 24 hours after the due date with a 20% credit penalty. Any assignment turned in more than 24 hours late will not be accepted.

Final Exam

There will be only one exam. The final exam contains job interview-style questions pertaining to the IR and NLP topics. The class, as a group, will prepare a pool of possible questions. The instructor will choose a few from the pool at random for the final exam.

Grading System

We will use the conventional letter grading scale as below (understanding your GPA):

Letter grade	Percent grade	4.0 scale
A (Excellent)	87–100	4.00
B (Good)	77–86	3.00
C (Satisfactory)	67–76	2.00
D (Passing, but less than satisfactory)	57–66	1.00
F (Failing)	87–100	0.00

KSU Academic Integrity

Important! *Do not, under any circumstances, copy another person's code or answers for assignments.* Incorporating someone else's code into your program in any form is a violation of academic regulations. In our class, all assignments will be reviewed for plagiarism using automatic or manual means throughout the semester. Any suspected cases of Academic Integrity violations will be sent to the Department of Student Conduct and Academic Integrity (SCAI).

Every KSU student is responsible for upholding the provisions of the Student Code of Conduct, as published in the Undergraduate and Graduate Catalogs. Section 5c of the Student Code of Conduct addresses the university's policy on academic honesty, including provisions regarding plagiarism and cheating, unauthorized access to university materials, misrepresentation/falsification of university records or academic work, malicious removal, retention, or destruction of library materials, malicious/intentional misuse of computer facilities and/or services, and misuse of student identification cards. Incidents of alleged academic misconduct will be handled through the established procedures of the SCAI, which includes either an "informal" resolution by a faculty member, resulting in a grade adjustment, or a formal hearing procedure, which may subject a student to the Code of Conduct's minimum one semester suspension requirement.

Disability Accomodation

Any student with a disability requiring accommodation in this course are encouraged to contact the Student Disability Services (SDS) during the first week of class.
website: `https://sds.kennesaw.edu/`, phone: (470) 578-2666

Course Withdrawal

The last day to withdraw without academic penalty is **October 21st, 2021**. Please find the pertaining information from 2021-22 Graduate / Undergraduate Catalog.

COVID-19 Information

Course Delivery

KSU may shift the method of course delivery at any time during the semester in compliance with University System of Georgia health and safety guidelines. In this case, alternate teaching modalities that may be adopted include hyflex, hybrid, synchronous online, or asynchronous online instruction.

COVID-19 Illness

If you are feeling ill, please stay home and contact your health professional. In addition, please email your instructor to say you are missing class due to illness. Signs of COVID-19 illness include, but are not limited to, the following:

- Cough
- Fever of 100.4 or higher
- Runny nose or new sinus congestion
- Shortness of breath or difficulty breathing
- Chills
- Sore Throat
- New loss of taste and/or smell

COVID-19 vaccines are a critical tool in "Protecting the Nest." If you have not already, you are strongly encouraged to get vaccinated immediatly to advance the health and safety of our campus community. As an enrolled KSU student, you are eligible to receive the vaccine on campus. Please call (470) 578-6644 to schedule your vaccination appointment or you may walk into one of our student health clinics.

For more information regarding COVID-19 (including testing, vaccines, extended illness procedures and accommodations), see KSU's official Covid-19 website

Face Coverings

Based on guidance from the University System of Georgia (USG), all vaccinated and unvaccinated individuals are encouraged to wear a face covering while inside campus facilities. Unvaccinated individuals are also strongly encouraged to continue to socially distance while inside campus facilities, when possible.

General Campus Policies

- Confidentiality and Privacy Statement (FERPA)
- Ethics Statement
- Sexual Misconduct Policy

Additional Student Resources

- CCSE student resources
- KSU Service Desk: email studenthelpdesk@kennesaw.edu
- Academic Advising
- Department of Career Planning & Development
- Counseling and Psychological Services
- Center for Health Promotion and Wellness
- Student Health Services