CSC 143U/291U
Introduction to Information Retrieval
Computer Science Department

Course Number: CSC 143U/291U
Course Title: Introduction to Information Retrieval
Semester hours: 3 s.h.
Instructor: Simona Doboli
Office hours: M: 10 - 11:00 and 4:30 to 6:00 and Tu: 5:00 – 5:30.
Office: Adams 101
Email: Simona.Doboli@hofstra.edu

Course Description: The course covers techniques used in text retrieval and text analysis applications such as search engines, text categorization and clustering, topic extraction, summarization, sentiment analysis. Topics include: natural language processing techniques for extracting relevant terms out of text data, vector space and probabilistic methods for computing similarity between documents, document ranking, clustering and classification methods for text analysis. ***Prerequisites:*** CSC 17, basic knowledge of linear algebra (matrix and vector computations).

Course learning outcomes:

1. Understand the internal workings of a search engine.
2. Apply basic natural language processing techniques to preprocess text data.
3. Apply vector space methods to retrieve and rank a large number of text sources.
4. Apply probabilistic methods to retrieve and rank a large number of text sources.
5. Evaluate, analyze and compare the performance of different methods.
6. Apply machine learning techniques to text categorization and clustering.
7. Apply text retrieval and analysis methods to summarize a large text.
8. Apply machine learning techniques to extract and analyze the topic structure.

Required Text/Readings

Text Data Management and Analysis, ChengXiang Zhai, Sean Massung, http://www.morganclaypoolpublishers.com/catalog_Orig/product_info.php?cPath=24&products_id=954

Recommended

Introduction to information retrieval, C. Manning, P. Raghavan, H, Schutze, https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf , http://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html

Course Outline:

1. Introduction to Text Information Systems
2. Basics of Natural Language Processing.
3. Text retrieval methods: tf-idf, vector space, probabilistic ranking, language models.

4. Text analysis methods: categorization, clustering, summarization, topic modeling, sentiment/opinion analysis.

Grading Criteria

Assignments (~ 5)	30 %
Mid-term Project	30%
Final Project	40%

Weekly schedule

Weeks	Topic	Assignment
1	Introduction	
2	Basic NLP	
3	Intro to NLTK	Assignment 1
4	Vector space models	
5-6	Intro to probabilities, probabilistic models, language models,	Assignment 2
7	Feedback methods, Anatomy of a search engine	Midterm project
8	Text clustering	Assignment 3
9	Text categorization	
10	Text summarization	Assignment 4
11	Topic modeling	
12	Sentiment/opinion analysis	Assignment 5
13-14	Final project	
Final Exam week		Project report and presentations due

Attendance Policy

Attendance is required for each seminar and is part of the grade. If absent to more than 3 weeks, an incomplete grade will be given.

Other requirements:

Incompletes will be solved by the faculty responsible for the seminar and require completion of the missing work, or reduced grade.

Disabilities Policy

If you believe you need accommodations for a disability, please contact Services for Students with Disabilities (SSD). In accordance with Section 504 of the Rehabilitation Act of 1973 and the Americans with Disabilities Act of 1990, qualified individuals with disabilities will not be discriminated against in any programs, or services available at Hofstra University. Individuals with disabilities are entitled to accommodations designed to facilitate full access to all programs

and services. SSD is responsible for coordinating disability-related accommodations and will provide students with documented disabilities accommodation letters, as appropriate. Since accommodations may require early planning and are not retroactive, please contact SSD as soon as possible. All students are responsible for providing accommodation letters to each instructor and for discussing with him or her the specific accommodations needed and how they can be best implemented in each course.

For more information on services provided by the university and for submission of documentation, please contact the Services for Students with Disabilities, 212 Memorial Hall, 516-463-7075.

Academic Honesty

Plagiarism is a serious ethical and professional infraction. Hofstra's policy on academic honesty reads: "The academic community assumes that work of any kind [...] is done, entirely, and without assistance, by and only for the individual(s) whose name(s) it bears." Please refer to the "Procedure for Handling Violations of Academic Honesty by Undergraduate Students at Hofstra University" to be found at http://www.hofstra.edu/PDF/Senate_FPS_11.pdf , for details about what constitutes plagiarism, and Hofstra's procedures for handling violations.

NOTICE ON CAMPUS SEXUAL ASSAULT AND DISCRIMINATORY HARASSMENT

Hofstra prohibits sexual and other discriminatory harassment, stalking, domestic and dating violence, sexual assault and other sexual misconduct (collectively, "Gender Based Offenses"). If you or someone you know believes they have been subjected to any of these Gender Based Offenses, help is available. To make a report, or for more information about Hofstra's Student Policy Prohibiting Discriminatory Harassment, Relationship Violence, and Sexual Misconduct (available at http://hofstra.edu/sexualmisconduct), please contact the Title IX Coordinator at (516) 463-5841 or TitleIXCoordinator@hofstra.edu, or Public Safety at (516) 463-6606. Confidential resources and support are also available from clinicians in Student Counseling Services (516-463-6791), medical professionals at the Health and Wellness Center (516-463-6745), and clergy in the Interfaith Center.