INLS 613: Text Mining

Objective: Gain experience with both the theoretical and practical aspects of text mining. Learn how to build and evaluate computer programs that generate new knowledge from natural language text.

Description: Changes in technology and publishing practices have eased the task of recording and sharing textual information electronically. This increased quantity of information has spurred the development of a new field called text mining. The overarching goal of this new field is to use computers to automatically learn new things from textual data.

The course is divided into three modules: basics, principles, and applications (see details below). The third part of the course will focus on several applications of text mining: methods for automatically organizing textual documents for sense-making and navigation (clustering and classification), methods for detecting opinion and bias, methods for detecting and resolving specific entities in text (information extraction and resolution), and methods for learning new relations between entities (relation extraction). Throughout the course, a strong emphasis will be placed on evaluation. Students will develop a deep understanding of one particular method through a course project.

Prerequisites: There are no prerequisites for this course. We will be using a tool called LightSIDE to train and test machine learned models for different predictive tasks. LightSIDE has a graphical user interface that makes it easy to do this without knowing how to program. That being said, knowing how to program (and manipulate text) may enable you to conduct more interesting experiments as part of your final project.
This course will involve understanding mathematical concepts and procedures. I will cover the basics in order for you to understand these. However, if you strongly dislike math and are unwilling to grapple with and ultimately conquer mathematical concepts and procedures, this may not be a good course for you.

Time & Location: M,W 1:25-2:40pm, Mitchell 205 (In Person).

Instructor: Jaime Arguello (email, web)

Office Hours: By Appointment

Required Textbook: Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition) Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2017. Morgan Kaufman. ISBN 978-0128042915. Available online

Additional Resources: Foundations of Statistical Natural Language Processing. C. Manning and H Schutze. 1999.

Introduction to Information Retrieval. C. Manning, P. Raghavan and H. Schutze. 2008.

Course Policies: Laptops, Attendance, Participation, Collaboration, Plagiarism & Cheating, Late Policy, Use of Generative AI Tools

Grading:
10% Class participation
20% Midterm Exam
30% Homework (10% each)
40% Final project (5% project proposal, 25% project report, 10% project presentation)

Grade Assignments: Undergraduate grading scale: A+ 97-100%, A 94-96%, A- 90-93%, B+ 87-89%, B 84-86, B- 80-83%, C+ 77-79%, C 74-76%, C- 70-73%, D+ 67-69%, D 64-66%, D- 60-63%, F 0-59%

Graduate grading scale: H 95-100%, P 80-94%, L 60-79%, and F 0-59%.

Topics: Subject to change! Readings from the required textbook (Witten, Frank, Hall, and Pal) is marked with a WFHPP bellow.

Lecture	Date	Events	Topic	Reading Due
1	Mon. 8/21		Introduction to Text Mining: The Big Picture	
2	Wed. 8/23		Course Overview: Roadmap and Expectations	WFH Ch. 1, Mitchell '06
3	Mon. 8/28		Predictive Analysis: Concepts, Features, and Instances I	WFH Ch. 2, Dominigos '12
4	Wed. 8/30		Predictive Analysis: Concepts, Features, and Instances II	
5	Mon. 9/4	Labor Day (No Class)		
6	Wed. 9/6	HW1 Out	Text Representation I	
7	Mon. 9/11		Text Representation II	
8	Wed. 9/13		LighSIDE Tutorial	LightSIDE User Manual
9	Mon. 9/18		Machine Learning Algorithms: Naïve Bayes I	WFH Ch. 4.2, Mitchell Sections 1 and 2

Lecture	Date	Events	Topic	Reading Due
10	Wed. 9/20	HW1 Due, HW2 Out	Machine Learning Algorithms: Instance-based Classification I	WFH Ch. 4.7
11	Mon. 9/25	Well-Being Day (No Class)		
12	Wed. 9/27	Project Proposal Due	Machine Learning Algorithms: Instance-based Classification II	
13	Mon. 10/2		Final Project Breakout Group Discussion I	
14	Wed. 10/4	HW2 Due	Machine Learning Algorithms: Linear Classifiers I	WFH 3.2 and 4.6
15	Mon. 10/9		Machine Learning Algorithms: Linear Classifiers II	
16	Wed. 10/11		Predictive Analysis: Experimentation and Evaluation I	WFH Ch. 5
21	Mon. 10/16	Midterm Review	Midterm Review	
22	Wed. 10/18	Midterm	Midterm	
23	Mon. 10/23		Predictive Analysis: Experimentation and Evaluation II	Smucker et al., '07, Cross-Validation, Parameter Tunning and Overfitting
24	Wed. 10/25		Final Project Breakout Group Discussion II	
25	Mon. 10/30	HW3 Out	Predictive Analysis: Experimentation and Evaluation III	
26	Wed. 11/1		Exploratory Analysis: Clustering I	Manning Ch. 16
27	Mon. 11/6		Exploratory Analysis: Clustering II	
28	Wed. 11/8		Sentiment Analysis	Pang and Lee, '08 (skip Section 5 and only skim Section 6), Pang and Lee, '02
29	Mon. 11/13	HW3 Due	Discourse Analysis	Arguello '15
30	Wed. 11/15	TREC Conference (No Class)		
31	Mon. 11/20		Detecting Viewpoint	Weibe '10
32	Wed. 11/22	Thanksgiving (No Class)		
33	Mon. 11/27		Text-based Forecasting	Lerman et al., '08
34	Wed. 11/29		Final Project Presentations I	
35	Mon. 12/4		Final Project Presentations II	
36	Wed. 12/6		Final Project Presentations III	
37	TBD	Final Project Due		