asandeep@unc.edu

[INLS 613: Text Mining]

Course Objective

This course will allow the student to develop a general understanding of knowledge discovery and gain a specific understanding of text mining. Students will become familiar with both the theoretical and practical aspects of text mining and develop a proficiency with data modeling.

Description

Production and consumption of information have drastically changed. Today, we inherently expect websites and apps to understand what we need and provide the appropriate service. At the same time, there is a growing call to provide better transparency on what such services provide. This course will provide you a **theoretical** and **practical** knowledge on how to tackle such problems. You will learn how to parse through text data to gain insights, and also learn why certain algorithms work the way they do.

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).

[Assignments]

In this course, you will work on three homework assignments (based on lectures), one midterm and one final project.

This course is generally taught over three months with a major emphasis on a course project. I don't expect you to do a three-month project in one month while also paying attention to the course. But I do want you to have practical experience. For that purpose, I have divided the final course project into 5 small sub-goals, each of which would be graded. Don't get intimidated by the 5 sub-goals. The point of splitting them into smaller goals is to make the course more enjoyable! At the end of certain lectures (See schedule) I will allocate time to discuss and monitor your progress. Given that the course will introduce different toolkits, ML methods, and evaluation techniques, the subgoals will also be positioned in a manner where you will be able to apply your knowledge from the lectures. These are the sub-goals.

Sub-goal 1: Identification of the problem. The goal here is to narrow your focus on the kind of problem you are interested in. For example, are you interested in analyzing restaurant reviews, tweets, search queries, hashtags on Instagram, etc?

Sub-goal 2: Narrow down of the problem statement. At this point I expect you to further narrow down your problem statement. For example, are you interested in the ratings of restaurants or the sentiment or tweets? If so why and what is interesting about this problem?

Sub-goal 3: Methods. What methods are suitable for

your problem statement. If this is a prediction problem, how do we want to approach it? If it is an exploratory problem, how do we do that? What toolkit do we use?

Sub-goal 4: Evaluation. Now that we know what methods we want to use, how do we evaluate our results?

Sub-goal 5: You present in class!

Sub-goal 6: Final Project due!

Please check the instructions on the **deliverables** here.

[Schedule]

#Lecture	Date	Topic	Events	Readings
1	5/17	Introduction and course outline		WFHP Ch. 1, Mitchell '06, Hearst '99
2	5/21	Predictive analysis	Sub-goal 1.	WFHP Ch. 2, Dominigos '12
3	5/22	Text Representation	HW1 out	WFHP Ch. 4.2, Mitchell Sections 1 and 2

#Lecture	Date	Topic	Events	Readings
4	5/24	Machine Learning: Linear Classifiers and Naïve Bayes		Re-read the readings from the last class
5	5/28	Memorial day		
6	5/29	Machine Learning: Instance based classification	HW1 Due.	
7	5/31	Machine Learning: Linear Classifiers + Review and HW discussion.	Sub goal 1-2 due.	Andrew Ng's notes
8	6/4	Mid-Term		
9	6/5	Machine Learning: Linear Classifiers (Continued) + ML-Toolkits tutorial(Weka; LightSIDE) + Mid Term Review.	HW2 Out.	(Read again) Andrew Ng's notes

#Lecture	Date	Topic	Events	Readings
10	6/7	Predictive analysis: Experimentation and evaluation - Part I	HW1 grades. Sub-goal 3.	WFHP Ch. 5
11	6/11	Predictive analysis: Experimentation and evaluation - Part II	HW2 due.	Smucker et al. '07, Cross-Validation, Parameter tuning and overfiting
12	6/12	Machine Learning: Clustering	Sub-goal 4. HW3 out.	Manning Ch.16
13	6/14	Applied Machine Learning : Sentiment analysis, View Points, Text-based forecasting.		
14	6/18	Class Presentation	Sub-goal 5; HW 3 due.	

#Lecture	Date	Topic	Events	Readings
15	6/20	Final Project submission	Sub-goal 6.	

Summer - I (2018)