CS 6501: Text Mining

Hongning Wang (hw5x@virginia.edu)
Department of Computer Science
University of Virginia

1 Course Overview

Given the dominance of text information over the Internet, mining high-quality information from text becomes increasingly critical. The actionable knowledge extracted from text data facilitates our life in a broad spectrum of areas, including business intelligence, information acquisition, social behavior analysis and decision making. In this course, we will cover important topics in text mining including: basic natural language processing techniques, document representation, text categorization and clustering, document summarization, sentiment analysis, social network and social media analysis, probabilistic topic models and text visualization.

In addition, as we are in the era of Big Data, we will provide you opportunities to gain hands-on experience of handling large-scale data set, i.e., Big Data. Modern data processing architecture, e.g., Apache Hadoop [1], Apache Spark [2] and GraphLab [3], will be incorporated in homework assignments.

2 Prerequisites

It is recommended that you have taken CS 2150 (or equivalent courses in data structure, algorithm) and have a good working familiarity with at least one programming language (Java is recommended, while Python is also ok). Significant programming experience will be helpful as you can focus more on the algorithms being explored rather than the syntax of programming languages. You are expected to independently finish machine problems and collaborate with your team members in the final course project.

Basic mathematics background is also required. Since this is a graduate-level course, you are supposed to know basic concepts of calculus (e.g., derivative and integral), probability (e.g., Bayes's theorem, conditional probability, basic probability distributions), linear algebra (e.g., vector, matrix and inner product) and optimization (e.g., gradient-based methods). Good knowledge in mathematics will help you gain in-depth understanding of the methods discussed in the course and develop your own idea for new solutions.

3 Text Books

There is no official textbook for this course. However, we do recommend the following books for your reference (especially the first one).

[1] `http://hadoop.apache.org`
[2] `http://spark.apache.org`
[3] `http://graphlab.org/projects/index.html`

1. ***Mining Text Data***. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.

2. ***Speech & Language Processing***. Dan Jurafsky and James H Martin, Pearson Education India, 2000.

3. ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

4 Course Content & Schedule

In this course, we will introduce a variety of basic principles, techniques and modern advances in text mining. Topics to be covered include (the schedules are tentative and subject to change, please keep track of it on the course website):

1. Introduction (0.5 week): We will highlight the basic organization and major topics of this course, and go over some logistic issues and course requirements.

2. Natural language processing (2 weeks): We will briefly discuss the basic techniques in natural language processing, including tokenziation, part-of-speech tagging, chunking, syntax parsing and named entity recognition. Public NLP toolkits will be introduced for you to understand and practice with those techniques.

3. Document representation (0.5 week): We will discuss how to represent the unstructured text documents with appropriate format and structure to support later automated text mining algorithms.

4. Text categorization (2 weeks): It refers to the task of assigning a text document to one or more classes or categories. We will discuss several basic supervised text categorization algorithms, including Naive Bayes, k Nearest Neighbor (kNN) and Logistic Regression. (If time allows, we will also cover Support Vector Machines and Decision Trees.)

5. Text clustering (2 weeks): It refers to the task of identifying the clustering structure of a corpus of text documents and assigning documents to the identified cluster(s). We will discuss two typical types of clustering algorithms, i.e., connectivity-based clustering (a.k.a., hierarchical clustering) and centroid-based clustering (e.g., k-means clustering).

6. Topic modeling (2 weeks): Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. We will introduce the general idea of topic modeling, two basic topic models, i.e., Probabilistic Latent Semantic Indexing (pLSI) and Latent Dirichlet Allocation (LDA), and their variants for different application scenarios, including classification, imagine annotation, collaborative filtering, and hierarchical topical structure modeling.

7. Document summarization (1 week): It refers to the process of reducing a text document to a summary that retains the most important points of the original document. Extraction-based summarization methods will be covered.

8. Social media and network analysis (1 week): We will discuss the unique characteristic of social network: inter-connectivity, and introduce Google's winning algorithm PageRank. Based on this, we will discuss social influence analysis and social media analysis.

9. Sentiment analysis (1 week): It refers to the task of extracting subjective information in source materials. We will discuss several interesting problems in sentiment analysis, including sentiment polarity prediction, review mining, and aspect identification,

10. Text visualization (1 week): It refers to the study of (interactive) visual representations of abstract data to reinforce human cognition. We will introduce some mathematical and programming tools to help you visualize a large collection of text documents.

11. Final project presentation (1 week): We will ask you to present your final project in class.

5 Communications

5.1 Lecture Times

We will have our lecture on every Tuesday and Thursday morning from 9:30am to 10:45am, at Rice Hall 340.

5.2 Office Hours

The lecture's office hour will be held on Tuesday and Thursday morning from 11am to 12pm, Rich Hall 408. The TA's office hour will be announced later.

5.3 Course Web Site

The course web site is now under construction, and it will be announced later.

5.4 Piazza

The most important forum for communicating in this class is the course's Piazza. Piazza is like a newsgroup or forum – you are encouraged to use it to ask questions, initiate discussions, express opinions, share resources, and give advice.

We expect that you will be courteous and post only material that is somehow related to the topic of Information Retrieval or course content. The posts will be lightly moderated.

Note that private posts to Piazza can be used for things like conflict requests, or for letting us know that you have that sinking feeling anything you don't really want to share with your classmates.

The Piazza site for this class is under construction and will be announced later.

6 Gradings

The course is a mix of lecture and student presentations. Grading is based on a set of homework assignments (40%), a paper presentation (20%) and a final course project (45%). Since this is a graduate-level course, there is *no* exam and more credits are given towards paper presentation and course project (with 5% extra credit).

6.1 Homework

Homework assignments will be a mix of paperwork and machine problems. Written homework should be finished individually, discussions with peers or instructor is allowed, but copying or any other type of cheating is strictly prohibited. You will be given one week to finish the written homework. Some of the machine problems are designed for teamwork and due day may vary. There will be around four MPs. Everyone will have one chance to ask for extension (extra three days from the deadline). After that, no extension will be granted. And please inform the instructor at least one day prior to the deadline, if you want an extension.

6.2 Paper Presentation

After each lecture, there will be five to seven assigned readings. Everyone is asked to select one paper from the list, and prepare a 20-minutes presentation for the class (including Q&A). One paper can only be presented by one student. Students are required to prepare the slides by themselves (the original authors' slides are not allowed to be used for this presentation). The purpose of this paper presentation is to help students to practice giving talks in front of public at conferences or other situations.

Both the instructor and other students will grade the presentation. The detailed grading criteria are as follows.

Table 1: Evaluation criteria for paper presentation

Aspects	Range
Slides content was clearly visible and self-explainable	[1,10]
Important messages of the paper were properly highlighted	[1,10]
Organization and logic of the presentation were easy to follow	[1,20]
Explained approaches/methods clearly	[1,20]
Presenter did not just read off of the slides	[0,10]
Perfect timing	[0,10]
Responded to audience's questions well	[0,10]
I have learned something from this presentation and would like to read the paper in future	[0,10]

6.3 Course Project

The course project is to give the students hands-on experience on solving some novel text mining problems. The project thus emphasizes either research-oriented problems or "deliverables." It is

preferred that the outcome of your project could be publishable, or tangible, typically some kind of novel research problem or prototype system that can be demonstrated (where bonus points applied). Group work is strongly encouraged, but not required.

More details about the project will be discussed on the course website, including suggested topics and available resources, but it consists of these major parts:

1. Project proposal (20%): State your motivation, research problem, and expected outcome of your course project. Due on the end of 5th week of semester. Discussion with instructor prior to deadline is encouraged.
2. Project presentation (40%): 20 minutes presentation about what you have done for this course project. Format could be tailored according to the nature of the project, e.g., slides presentation and/or system demo.
3. Project report (40%): Detail documentation of your project. Quality requirement is the same as research papers, i.e., in formal written English and rigorous paper format. Due on the last week of course (***before*** project presentation).

6.4 Grade Cutoffs

We will use the standard grade cutoff points:

Table 2: Grade cutoff points

Letter Grade	Point Range
A	[93,105]
A-	[90, 93)
B+	[87, 90)
B	[83, 87)
B-	[80, 83)
C+	[77, 80)
C	[73, 77)
C-	[70, 73)
D+	[67, 70)
D	[63, 67)
D-	[60, 63)
F	[0, 60)

7 Acknowledgements

Thanks to Professor ChengXiang Zhai from University of Illinois at Urbana-Champaign; some teaching materials borrowed from his course site for CS410. And special thanks to Sean Massung from University of Illinois at Urbana-Champaign for his invaluable help in preparing this course.

Thanks to you for reading the entire syllabus. Hopefully it makes your experience a bit easier and less stressful.