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 will be also fine). 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.

Solid mathematics background is also required. Since this is a graduate-level course, you are supposed to know important concepts in calculus (e.g., derivative and integral, and how to compute them), probability (e.g., Bayes's theorem, conditional probability, basic probability distributions, moments), linear algebra (e.g., vector, matrix and inner product, and matrix operations) and optimization (e.g., gradient-based methods and optimality conditions). 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 two books. Suggested readings from the first two books will

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

be provided in the corresponding lecture slides.

1. ***Speech & Language Processing*** - *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd edition).* Dan Jurafsky and James H Martin, Pearson Education India, 2000.

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

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

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. Document representation (2 weeks): We will discuss how to represent the unstructured text documents with appropriate format and structure to support later automated text mining algorithms.

3. 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.

4. Text categorization (2.5 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 (2.5 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 (1.5 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. Text Mining Applications (2 weeks):

(a) Document summarization: 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.

(b) Sentiment analysis: 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,

(c) Text visualization: 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.

8. Final project presentation (1.5 week): We will ask you to present your final project in class.

5 Communications

5.1 Lecture Hours

We will have our lectures on every Monday and Wednesday afternoon from 5:00pm to 6:15pm, at Olsson Hall 009.

5.2 TAs

Lin Gong (lg5bt@virginia.edu) and Qingyun Wu (qw2ky@virginia.edu).

5.3 Office Hours

The instructor's office hour will be held on Monday and Wednesday afternoon from 4:00pm to 5:00pm, at Rice Hall 408. And the TAs' office hour will be held on Tuesday and Thursday afternoon from 2:00pm to 3:00pm, at Rice Hall 414.

5.4 Course Web Site

To be announced later.

5.5 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 text mining 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 sign-up link to our Piazza page is at: `piazza.com/virginia/spring2016/cs6501`

6 Gradings

The course is a mix of lectures and student presentations. Grading is based on a set of homework assignments (40%), a set of in-class quizzes (15%), a paper presentation (15%) and a final course project (35%).

6.1 Homework

Homework assignments will consist of four machine problems. Homework should be finished individually (unless it is a group assignment): 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 homework. Some of the machine problems are designed for teamwork and the due date may vary. 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 TA at least *one day prior to the deadline*, if you want an extension. Any late submission without granted extension request will receive a ***15% late penalty*** within the first week of due date; after that it will be a ***30% late penalty***.

6.2 Quizzes

We will have five in-class quizzes to cover all the important concepts we have learned in class. The quizzes will be in the form of True/False questions, multiple choice questions and short answer questions. All the quizzes are **closed** book and **closed** notes; no electronic aids or cheat sheets are allowed. And each quiz is 15 minutes long in time, and the detailed date of the quiz will be announced one week ahead.

6.3 Paper Presentation

After each lecture, there will be five to seven assigned readings. Students are required to form a team (of 2-3 students), select one paper from the list, and prepare a 15-minutes presentation for the class (including Q&A). One paper can only be presented by one team of students. Note, this team should be the same as your course project team, and it is preferred for you to present a paper that is directly related to your project topic. 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, and better prepare the course project.

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

Paper presentation sign-up is due in *the end of 4th week*. Presenters must present the selected/assigned paper on the scheduled date. No extension will be given due to the tight schedule of this course.

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	[1,15]
Explained approaches/methods clearly	[1,15]
All students in the team well understood the paper	[0,10]
Presenter(s) 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]

6.4 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, and each team can have no more than three students.

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).

Project proposal is due in *the end of 5th week*. The detailed grading criteria for project presentation and report are as follows. No extension will be given for any part of this course project due to the tight schedule of our class.

6.5 Grade Cutoffs

We will use the standard grade cutoff points defined in Table 4. We will strictly follow this cut-off table; no curves will be applied onto the final grades.

Table 2: Evaluation criteria for project presentation

Aspects	Range
Slides content was clearly visible and self-explainable.	[0,10]
Presenters were confident about their work and clearly explained it to me.	[0,10]
Background and research question were clearly highlighted, and the logic and argument were reasonable.	[0,10]
There was sufficient discussion of state-of-the-art and why do we need this new method.	[0,10]
Description of the proposed method is clear, comprehensive, coherent and consistent with the claim in the introduction.	[0,15]
Thorough experimentations that proved all necessary components in the proposed method and detailed analysis of the experimental results.	[0,15]
The presenters well managed their time during presentation.	[0,10]
The presenters did a good job in answering the questions.	[0,10]
I like this work!	[0,10]

Table 3: Evaluation criteria for project report

Aspects	Range
Strictly follow the provided template.	[0,10]
Background and research question were clearly stated in the introduction and the logic and argument were reasonable.	[0,10]
Contribution of the work was properly articulated in the introduction.	[0,10]
Sufficient discussion of state-of-the-art in related work section.	[0,10]
Description of the proposed method was clear, comprehensive, coherent and consistent with the claim in the introduction.	[0,20]
Precise description of experiment design and experimental data set.	[0,10]
Thorough experimentations that proved all necessary components in the proposed method and detailed analysis of the experimental results.	[0,20]
Summarization of the work, reasonable discussion of limitation of the proposed solution and future work.	[0,10]

The policy stated in this syllabus will be strictly followed in order to ensure fairness among all the students. Any request violating our policy will not be considered.

Table 4: Grade cutoff points (all in real numbers)

Letter Grade	Point Range
A+	[97,105]
A	[93,97)
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.