COURSE SYLLABUS
IST 736 Text Mining

Course Lead: Professor Bei Yu (byu@syr.edu)
Section Instructor: See 2SU LMS Section
Note: Students should contact section instructor for questions

Prerequisite: IST 687: Introduction to Data Science. The purpose of setting up this prerequisite is to ensure that incoming students have basic programming and algorithmic thinking skills. Exceptions may be given to students who have acquired equivalent skills.

Audience: Graduate students.

Description: Introduces concepts and methods for knowledge discovery from large amounts of text data and the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis.

Additional Course Description: The main goal of this course is to increase student awareness of the power of large amounts of text data and computational methods to find patterns in large text corpora. This course is designed as a general introductory-level course for all students who are interested in text mining. Programming skill is preferred but not required in this class. This course will introduce the concepts and methods of text mining technologies rooted from machine learning, natural language processing, and statistics. This course will also showcase the applications of text mining technologies in (1) information organization and access, (2) business intelligence, (3) social behavior analysis, and (4) digital humanities.

What is the difference between IST 565 Data Mining and IST 736 Text Mining?

A number of students have asked the question, “What is the main difference between the two courses: data mining and text mining?” The two classes share the theoretical foundation in machine learning. Therefore, the fundamental concepts in machine learning, such as classification and clustering, are covered in both classes. However, these two classes differ in the following aspects:

- **Content wise:** The data mining class focuses on structured data, meaning the data sets we play in the class are usually in .csv format. Text mining focuses on unstructured text data, which come in words. How to convert text to numbers that still bear the meaning of text is an important topic in text mining. In text mining we will have to deal with some problems that do not exist in mining structured data, such as the subjectivity in annotations, for example, how to determine if a tweet is positive, negative, or neutral. Different people might give different assessments.
- **Technology wise:** The data mining class uses Weka and R. The text mining class uses a Python-based command line tool call scikit-learn.
- **Challenge wise:** Text mining requires even more critical thinking as it is research oriented. The students are usually a mix of doctoral and masters students from multiple disciplines, such as iSchool, linguistics, communications, business, political science, etc.

What is the difference between IST 736 Text Mining and IST 664 Natural Language Processing?

NLP and TM do share some foundations. NLP focuses on deep analysis of language, such as POS tagging, sentence structure parsing, named-entity recognition, semantic role labeling, etc. Deep analysis can be highly informative and also time consuming. In contrast, text mining focuses on using machine learning and large amounts of data for fast predictions, usually on shallow features like just words. Depending on real-world applications, sometimes we need deep analysis if time is not a concern, sometimes we do fast and shallow analysis, and sometimes combine both. IST 664 is not a prerequisite for IST 736.

Credits: 3

Learning Objectives:

After taking this course, the students will be able to:

- describe basic concepts and methods in text mining, for example document representation, information extraction, text classification and clustering, and topic modeling;
- use benchmark corpora, commercial and open-source text analysis and visualization tools to explore interesting patterns;
- understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems; and
- choose appropriate technologies for specific text analysis tasks and evaluate the benefit and challenges of the chosen technical solution.

Bibliography/Texts/Supplies—Required:

- Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining*. New York: Springer. ISBN: 978-1849962254
- The instructor will also provide slides, tutorials, readings, sample data, and sample scripts.

Bibliography/Texts/Supplies—Additional:

The following books are not required. But, the instructor also consulted the following books to design this course. You are encouraged to explore these books.

- Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*, Chapters 6 and 13–18, Cambridge University Press. Available online at: http://nlp.stanford.edu/IR-book/
- Mitchell, T. (1990). *Machine learning*. McGraw-Hill.
- Severance, C. (2016). *Python for everybody: Exploring data in Python 3*. Online book: https://www.py4e.com/

Software:

All of the following tools are widely used open source toolkits for machine learning and text mining. Although they will be installed on lab computers, students should also install these packages on their own computers to be able to use them conveniently during the semester and in the future.

The easiest way to install both scikit-learn and NLTK is to install the Anaconda package. On iSchool lab machines, students can install Anaconda in their H drive.

http://docs.continuum.io/anaconda/pkg-docs.html

This package is large, containing 224 useful packages for all kinds of data analysis, including both scikit-learn and NLTK. The scikit-learn and NLTK websites provide comprehensive documentations and tutorials.

scikit-learn: http://scikit-learn.org/stable/

NLTK: http://www.nltk.org/ Weka
Weka: http://www.cs.waikato.ac.nz/ml/weka/

Mallet: http://mallet.cs.umass.edu/

Requirements:

- **Communication**: This course will use the 2SU LMS as the main communication platform in and out of class time. Students are required to check their accounts on a regular basis. Important announcements will be posted to the Announcement wall. Failure to read the class announcements will not be considered a suitable excuse for not being informed.
- **Class participation (15%):** Class participation is important for this class. Students are required to attend classes and actively participate in class discussions and exercises. If a student missed a class for legitimate reasons, such as health problems and job interviews, the student should make up the exercises within a week without grade penalty.
  - Participation grade will be calculated at the end of the semester using the formula x/y*15, denoting y as the total number of discussions/exercises, and x as the actual number the student participated in.
  - For each week, the exercise questions in the asynchronous content need to be answered 24 hours before that week's live session, so that the instructor can review them and address any issues during the live session. For example, all exercises for Week 1 should be finished 24 hours before Live Session 1.
- **Assignments (60%):** Assignments must be professionally prepared and submitted electronically. All assignments should be submitted in Word files named as "HW_Num_Lastname_Firstname.doc(x)," e.g., "**HW_1_Smith_John.doc.**" **No PDF files please**. To ensure fast return, all assignments should be submitted on time.
  - Each assignment accounts for 100 points. At the end of the semester, the total points from all assignments will be rescaled to 60% of the final score.
  - Each assignment is **due** 48 hours before the next week's live session, so that the instructor can grade and return them to students before they work on the next assignment. For example, HW1 is due 48 hours before Live Session 2.
- **Final project (25%):** Students will work on projects starting from Week 9. Checkpoints include project idea presentation (Week 9), project clinic (Week 10), and final project report (one week after Week 10 live session).

Students can use this week for final revisions based on feedback from professor and peer students.

Grading:

Grade	Points	Grade	Points	Grade	Points	Grade	Points
		B+	87–89	C+	77–79	F	0–69
A	93–100	B	83–86	C	73–76		
A–	90–92	B–	80–82	C–	70–72		

Grades of D and D– may not be assigned to graduate students.

Tips for success in this class: curiosity, critical thinking, math, and programming.

- Curiosity: Curious about language and meaning, pay attention to the data details. Don't treat a data set as a blackbox. Don't treat an algorithm as a blackbox. Try see through them.
- Critical thinking: Text mining is essentially research. You will learn and practice methods to discover patterns, and also evaluate whether and why the discovered patterns are true and useful.
- Math: You will need some math knowledge, such as algebra and probability, to understand how the data mining algorithms work.
- Programming: Although GUI tools like Weka would allow users with no programming skills to play with data sets, data sets are rarely immediately ready for analysis in these tools. The results from off-the-shelf tools may need additional transformation to see patterns. Python programming skills would help you pre- and post-processing text data. Programming would also help you gain more convenient control over algorithm tuning in your scripts.

Academic Integrity Policy:

Syracuse University's Academic Integrity Policy reflects the high value that we, as a university community, place on honesty in academic work. The policy defines our expectations for academic honesty and holds students accountable for the integrity of all work they submit. Students should understand that it is their responsibility to learn about course-specific expectations, as well as about university-wide academic integrity expectations. The policy governs appropriate citation and use of sources, the integrity of work submitted in exams and assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The policy also prohibits students from submitting the same work in more than one class without receiving written authorization in advance from both instructors. Under the policy, students found in violation are subject to grade sanctions determined by the course instructor and non-grade sanctions determined by the School or College where the course is offered as described in the Violation and Sanction Classification Rubric. SU students are required to read an online summary of the University's academic integrity expectations and provide an electronic signature agreeing to abide by them twice a year during pre-term check-in on MySlice. For more information about the policy, see http://academicintegrity.syr.edu.

Disability-Related Accommodations:

Syracuse University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. If you believe that you need accommodations for a disability, please contact the Office of Disability Services (ODS), disabilityservices.syr.edu, located at 804 University Avenue, room 309, or call 315.443.4498 for an appointment to discuss your needs and the process for requesting accommodations. ODS is responsible for coordinating disability-related accommodations and will issue “Accommodation Authorization Letters” to students as appropriate. Since accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible. Our goal at the iSchool is to create learning environments that are useable, equitable, inclusive and welcoming. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please meet with me to discuss additional strategies beyond official accommodations that may be helpful to your success.

Religious Observances Notification and Policy:

SU’s religious observances policy, found at supolicies.syr.edu/emp_ben/religious_observance.htm , recognizes the diversity of faiths represented in the campus community and protects the rights of students, faculty, and staff to observe religious holy days according to their tradition. Under the policy, students should have an opportunity to make up any examination, study, or work requirements that may be missed due to a religious observance provided they notify their instructors no later than the end of the second week of classes through an online notification form in MySlice listed under **Student Services/Enrollment/My Religious Observances/Add a Notification**.

Student Academic Work Policy:

Student work prepared for University courses in any media may be used for educational purposes, if the course syllabus makes clear that such use may occur. You grant permission to have your work used in this manner by registering for, and by continuing to be enrolled in, courses where such use of student work is announced in the course syllabus.

Sample Schedule

Week	Topic	Readings before class	Items due
1	Introduction to Text Mining	Ch. 1	
2	Document Vectorization	Ch. 2	HW1
3	Corpus Analysis	Ch. 2	HW2
4	Naïve Bayes for Text Categorization	Ch. 3.1–3.4	HW3
5	Text Classifier Evaluation and Human Annotation	Ch. 3.5	HW4
6	scikit-learn		HW5
7	SVMs for Text Categorization	Ch. 3.4	HW6
8	Document Clustering and Topic Modeling	Ch. 5	HW7
9	Literature Review on Text Mining Applications/Project Idea Presentation		HW8 Student presentation
10	New Topics in Text Mining/Project Clinic and Presentation		Student presentation
11			Final project report