COURSE SYLLABUS
IST 736 Text Mining

Updated: 09/25/2023

Instructor: Bei Yu
Email: byu@syr.edu
Time: Wednesdays 9:30-12:15pm
Location: Hinds 243A (Ice Box II)
Office hour: Tuesdays 1-2pm (Hinds 320)
Requirement: bring your own laptop and headphone

Course Description: This course takes a historical perspective to introduce four generations of text mining techniques: rule- and knowledge-based methods, traditional machine learning, transformers, and prompt engineering for large language models. This course also teaches text mining as a research method for solving data science problems in various domains, such as social media analysis and science literature mining. Case studies are used to teach the design and evaluation of text mining solutions.

Prerequisite:

- Basic Python Programming skills. This course uses Python. If you have basic programming knowledge but are new to Python, you can use the free online book "Python for Everybody" (chapters 1-12) or other materials to get familiar with the Python syntax, especially the data structures like list, dict, and hash, and basic controls like if-condition and loop. Regular expressions can be very helpful in this course.
- A laptop with relevant software packages installed (see below details for the software list), since IceBox II is a regular classroom, not a computer lab.
- A pair of earphones. Sample code demos will be pre-recorded; students will watch the videos in class individually for best learning outcome.

Audience: Graduate Students

Credits: 3

Learning Objectives:

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

- Describe basic concepts and methods in text mining, for example text representation, text classification and clustering, and topic modeling;
- Use the text mining concepts and methods to model real-world problems into text mining tasks, develop technical solutions, and evaluate the effectiveness of the solutions.
- Communicate text mining process, result, and major findings to various audience including both experts and laypersons.

Required Texts / Supplies:

None. The instructor will provide slides, tutorials, readings, sample data, and sample scripts.

Texts / Supplies – Optional:

The instructor consulted the following books to design this course. You are encouraged to explore these books.

- Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. " O'Reilly Media, Inc.".

- Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining*. New York: Springer.

- 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 and computing environment:

This course uses popular open source toolkits for text mining. Please make sure to install these software on your laptop as we will use them in class. You can use Google Colab for GPU access.

The easiest way to install both sci-kit learn and nltk is to install the Anaconda package.

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

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

- Sci-kit Learn http://scikit-learn.org/stable/
- NLTK http://www.nltk.org/

Course Requirements and Expectations:

- **Communications**

This course will use the SU BlackBoard System as the main communication platform in

and out of class time. Students are required to check their BlackBoard accounts on a regular basis. Important announcements will be posted to the Announcements board. Failure to read the class announcements will not be considered a suitable excuse for not being informed. The BlackBoard can be accessed at http://blackboard.syr.edu.

- **Tips for success in this course**

  - **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**: 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.

Grading:

- **In-class exercise (25%):** Students are required to 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 on BlackBoard within a week without grade penalty. In-class exercise grades will be calculated at the end of the semester using the formula x/y*25, denoting y as the total number of discussions/exercises, and x as the actual number the student participated.

- **Assignments (40%):** Assignments must be written in academic writing format and submitted electronically to the BlackBoard. All assignments should be submitted in Microsoft Word files named as "**HW_Num_Lastname_Firstname.doc(x)**", e.g. "HW_1_Smith_John.doc". **Do not convert Word Files to PDFs – it is difficult to comment on PDF files**. Grades and comments for the assignments will be made available in the BlackBoard.

- **Final Project (35%):** Students will work on course projects in the second half of the semester. Students can choose to work individually or in groups. Group size is up to three students. The instructor provides three rounds of feedback. These check points include **project proposal presentation (5 points)**, **project progress presentation (5 points)**, and **project result presentation (5 points)**. **Final project report (20 points)** is due one week after project result presentation. Students can use this week for final revisions based on feedback from instructor and peer students.

COURSE SYLLABUS
IST 736 Text Mining

Grading Table:

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.

Sample Schedule

Date	**Week**	**Topic**	Item due
08/30	1	Introduction	
09/06	2	1st generation text classifier – pattern matching with rules and lexica	
09/13	3	Case study – sentiment lexica	
09/20	4	2nd generation text classifier – fully supervised learning, e.g. naïve Bayes and SVMs	HW1
09/27	5	Evaluating and interpreting linear classifiers	
10/04	6	3rd generation text classifier – fine-tuned language models, e.g. BERT	HW2
10/11	7	Sentence embedding for text clustering	
10/18	8	4th generation text classifier – large language models (LLMs), e.g. GPT4	HW3
10/25	9	Case study – prompt engineering	
11/01	10	Quality of training labels	HW4
11/08	11	Project proposal presentation	Checkpoint #1
11/15	12	Ethics in NLP applications	
11/22		Thanksgiving break	
11/29	13	Project progress presentation	Checkpoint #2
12/06	14	Project result presentation	Checkpoint #3
12/13		One week after final presentation	Final project report

Course Policy on Use of Generative AI

In this class students are welcome to use generative AI tools to assist their learning and writing. The tools for programming help (e.g. Copilot) and writing help (e.g. chatGPT) may be the most relevant to this class. Students should follow these two guidelines when using generative AI tools in this class.

#1 Critical Evaluation – when using generative AI tools, students should exercise their

critical thinking skills to verify answers provided by these tools, evaluate the benefits and risks of these tools for their academic learning, and acknowledge that these tools may pose risks such as inaccurate answers, biases in answers, or fake information in answers.

#2 Declare Use – if generative AI tools were used in submitted assignments or project reports, make sure to add a paragraph to declare what tools were used, how they were used, and your critical evaluation on the pros and cons of using these tools for this task. By sharing your experience with the class, we help each other to pursue best practice for using generative AI for this class.

Academic Work

The professor may share sample homework assignment reports and project reports with the class for peer learning purpose. By default, student names will be shown to recognize your excellent work. If you would like to be anonymous instead, please notify the professor.

Other Policies

Please refer to the appendix for Syracuse University Student Policies and Services.