CSE 635 Spring 2021

NLP and Text Mining

MW 5:40 – 7:10 pm (Online_
Reg # 23984
Instructor: Rohini K. Srihari
Piazza link: piazza.com/buffalo/spring2021/cse635

Description:

This course will explore various approaches to text, web and social media mining. Since natural language processing (NLP) is the foundation for most text mining solutions, a major focus of the course is on widely used NLP algorithms. This includes topic models, entity tagging, opinion analysis, information extraction, parsing, summarization, machine translation and question answering. We will cover both traditional, feature-based approaches, as well as recent approaches based on neural embeddings and deep learning. Several applications utilizing text mining will be covered including social media mining and recommender systems (algorithms powering Amazon, Facebook and Twitter).

- Recommender Systems

- NLP
  - Language Models, Word Embeddings
  - Deep learning models for NLP
    - Recurrent networks, LSTM
    - Attention models, transformers, contextual embeddings
  - POS tagging, Entity tagging
  - Machine translation
  - Opinion analysis
  - Information Extraction (relationship, event extraction)
  - Summarization
  - Entity resolution, wikification
  - Trending phrases, terminology discovery
  - Semantic Role Labeling, Textual Entailment
  - Chatbots, Discourse

Textbook:

Speech and Language Processing (3rd Edition) Daniel Jurafsky and James Martin, 2020. [SLP]
https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf

Project: Students are expected to work on two programming projects: (i) an individual project involving implementing an NLP algorithm on a standard data set with evaluation, and (ii) a semester long group project involving text/web mining: students will choose between three projects. We will be using common data sets to facilitate evaluation. The project requirements will be discussed in detail during the first week. You will receive guidance regarding data collection, algorithms, evaluation methodology during the semester. Students will be required to present their final group project during the last week of class. Students are also required to write a technical paper describing their project and experiments. You will work in pairs or groups of three for the class project which will satisfy department requirements for the MS project.

Grading: There is no midterm or final for this course. Instead, there will be a weekly or bi-weekly, in-class short quizzes (a few multiple choice questions) based on the previous week's lectures. If you come to class regularly, you should find the quizzes easy. The final grade will be based on all of the above, i.e. (i) quizzes, (ii) individual project, (iii) group project and demonstration, (iii) final paper.

Prerequisites: The required background is a combination of information retrieval (CSE 535), machine learning, and programming expertise.

Piazza: Students should enroll for the piazza site for this course at the link provided. All class related communication will take place through piazza.

Students should read the departmental academic integrity policies. These will be strictly enforced.

Schedule (subject to change**):**

Week of	Topic	Readings
Feb 1	Course Overview Project Recommender Systems Language Models	 Notes [SLP]Ch 3
Feb 8	Naïve Bayes Sentiment Analysis Text Summarization	[SLP] Ch 4
Feb 15	Vector Semantics and Embeddings Neural Language Models	[SLP] Ch 6, 7
Feb 22	POS, Entity tagging Entity Resolution (Wikification)	[SLP] Ch 8 Notes
Mar 1	Deep Learning Architectures for Sequence Processing	Ch 9
Mar 8	Encoder Decoder Models, Machine Translation	[SLP] Ch 11
Mar 15	Attention, Self Attention, Transformers Contextual Embeddings (BERT) Multilingual embeddings	[SLP] Ch 10 Jay Alammar Blog, video
Mar 22	Sentiment, Affect, Coreference	[SLP]20,21
Mar 29	Discourse, Chatbots	[SLP] Ch 22, 24
April 5	Information Extraction: Relationship extraction, event template filling	[SLP] Ch 17
April 12	WordNet, Semantic Role Labeling, Textual Entailment	[SLP] Ch 18,19
April 19	Question Answering	[SLP] Ch 23
April 26	TBD	
May 3	Project Presentations	