CSE 635 Spring 2022

NLP and Text Mining

Wed 12 – 2:30 pm (Online)
Reg # 24771

Instructor: Rohini K. Srihari
Piazza link: http://piazza.com/buffalo/spring2022/cse635

Description:

This course covers a comprehensive set of topics in natural language processing (NLP). The course begins with some information retrieval (IR) based approaches to NLP tasks such as question answering, recommender systems and text summarization. This is followed by topics relating to early stages in the NLP pipeline including language models, POS analysis and entity recognition. The next section is a sequence of topics related to deep learning in NLP, including neural embeddings, encoder-decoder models, and transfer learning using pretrained contextual models. These topics are presented in the context of NLP tasks such as machine translation and sentiment analysis. We will cover knowledge-based and traditional feature-based approaches in addition to deep learning approaches to gain a more intuitive understanding. The latter part of the course covers advanced NLP topics such as discourse, chatbots, semantic role labeling along with information extraction. Several text mining applications utilizing NLP will be discussed including social media mining and recommender systems (algorithms powering Amazon, Facebook, and Twitter). Each session will have a lecture component followed by a code demonstration session for hands-on learning.

Textbook:

Speech and Language Processing (3rd Edition) Daniel Jurafsky and James Martin, 2021. [SLP]
https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.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 four topics. 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 quiz (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 as follows.
Quizzes: 30%
Individual Project: 20%
Group Project: 50%

Prerequisites: The required background is a combination of information retrieval (CSE 535), machine learning (CSE 574), 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.

Team Projects:
We are planning on the following team topics this semester. This will be discussed more in class at our first meeting.

- Socialbots (Persona)
- Predicting social unrest (Event extraction)
- Social media mining (for health)
- Disinformation (stance detection)

Schedule (subject to change**):**

Lecture Date	Topic	Recitation	Readings
Feb 2	Course Overview Question Answering Recommender Systems	Semester-long project release	Notes [SLP]Ch 23
Feb 9	Text Summarization Language Models Naïve Bayes Sentiment Analysis	PageRank for summarization, n-gram language model, Naive-Bayes classification	Notes [SLP] Ch 3, 4
Feb 16	Vector Semantics and Embeddings Neural Language Models	Word2Vec, gensim	[SLP] Ch 6, 7
Feb 23	POS, Entity tagging Constituency Grammars Entity Resolution (Wikification)	POS, NER basics using Spacy, StanzaNLP, etc.	[SLP] Ch 8, 12 Notes
Mar 2	Deep Learning Architectures for Sequence Processing ** Individual Project assigned **	PyTorch tutorial	Ch 9
Mar 9	Encoder-Decoder Models Machine Translation	Seq2Seq tutorial	[SLP] Ch 10
Mar 16	Transformers Attention Contextual Embeddings (BERT) Transfer Learning	BERT based sequence classification using 'transformers' library	[SLP] Ch 11 Jay Alammar Blog, video
Mar 23	*** **Spring Break** ***		
Mar 30	Sentiment, Affect Discourse Models	Targeted-Aspect Based Sentiment Analysis	[SLP]20, 22
April 6	Chatbots & Dialogue Systems	BlenderBot demo	[SLP] Ch 24
April 13	Information Extraction: Relationship/Event extraction, Co-reference	Spacy & AllenAI coreference	[SLP] Ch 17, 21
April 20	Dependency Parsing, WordNet	Spacy, Stanford Core NLP	[SLP] Ch 14, 18
April 27	Semantic Role Labeling, Textual Entailment	AllenAI entailment using MNLI & SNLI	[SLP] Ch 19 Notes
May 4	Bias in NLP	TBD	Notes
May 11	Project Presentations		