Text Analytics, MSIA - 414

Course Syllabus

Text Analytics
MSIA 414
Friday, 1 pm – 4 pm
Autumn 2023
9/19/2023-12/02/2023
Location: North Garage Krebs Room 1440

Yuri Balasanov, Ph.D.
yuri.balasanov@northwestern.edu

COURSE DESCRIPTION

The course explores a breadth of Natural Language Processing (NLP) applications with a focus on contemporary, state-of-the-art systems, often based on deep learning techniques. Topics include word embeddings and common deep learning NLP architectures; approaches to a variety of NLP tasks such as text classification, named entity recognition, machine translation, information retrieval, etc. The course also includes the necessary background in linguistics, discusses potential biases and harms of language models, and data sets necessary for training the models.
The course has been substantially updated to include the latest applications of Generative AI.

BOOK

Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third Edition draft. Daniel Jurafsky, James H. Martin, © 2023 (Recommended).

SOFTWARE AND HARDWARE

Students will use Python and related packages (https://www.python.org/)
It is recommended that students have their laptops with Python installed during the class. We will use them for data assignments in class.

LEARNING OBJECTIVES

After completing this course, students should be able to:

- Develop familiarity with a variety of NLP applications and state-of-the-art solutions.
- Develop skills to understand non-trivial scientific NLP publications and NLP / ML libraries and framework for continuous and independent learning.
- Develop NLP / ML engineering skills and familiarity with common industry NLP tasks.
- Develop skills to creatively approach business problems and create practical NLP solutions.

ATTENDANCE

Attendance of all class sessions is mandatory. To attend the class remotely on Zoom the students must notify the instructor in advance.

Students can miss not more than two sessions out of ten; every missed lecture or seminar must be arranged with the instructor in advance.

LATE WORK

All assignments must be submitted on the due date before 11:59 pm. If you turn in an assignment late, 10% will be deducted from the total score for each day after the deadline. Assignments turned in more than one week late will not receive credit. In the case of unexpected events, you must contact the instructor before the assignment due date to receive a grace period. Students can only receive up to two grace periods in the course.

ACADEMIC HONESTY & PLAGIARISM

It is contrary to justice, academic integrity, and to the spirit of intellectual inquiry to submit another's statements or ideas of work as one's own. To do so is plagiarism or cheating, offenses punishable under the University's disciplinary system. Because these offenses undercut the distinctive moral and intellectual character of the University, we take them very seriously.

Proper acknowledgment of another's ideas, whether by direct quotation or paraphrase, is expected. In particular, if any written or electronic source is consulted and material is used from that source, directly or indirectly, the source should be identified by author, title, and page number, or by website and date accessed. Any doubts about what constitutes "use" should be addressed to the instructor.

At any time during or after the course students are encouraged to help developing this course by providing their feedback to the instructor in any form, as long as it is constructive, respectful and in compliance with the ethical norms of the University.

GRADE COMPOSITION

Final grades will be created based on:

- Homework assignments (70%)
- Class participation (20%)
- Group presentation (10%)

COURSE SCHEDULE

Important Note: Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via email and in-class announcement.

SESSION 1

Introduction to NLP
Regular expressions

SESSION 2: Text normalization and word embeddings

Words and tokens
Morphological analysis
Language models
Distributional semantics
Lexical semantics
Vector semantics, embeddings
Introduction to NLTK
Potential harms from Language models

SESSION 3: Naive Bayes and logistic classifiers for bag-of-words models

Naive Bayes as a generative model
Naive Bayes as a graph model
Naive Bayes as a language model
Bag-of-words Bernoulli and multinomial models
Laplace smoothing
Generative v.s. discriminative classifiers
Logistic regression for text classification
Words polarity analysis
Potential harms of text classification models

SESSION 4: Main architectures of neural networks in NLP

Motivation for models with memory
Recurrent neuron, memory cell
Motivation for LSTM
LSTM structure
Encoder-decoder architectures, autoencoder

SESSION 5: Attention-based models

Transfer learning generation of models
Self-attention layer
Transformer architecture
BERT, GPT, T5
Applications to summarization, classification, question answering

SESSION 6: Searching for patterns and information extraction

Searching for patterns
Part of Speech tagging
Named Entity tagging
Conditional random fields
Information extraction: relation, event, temporal
Semantic role labeling
Introduction to SpaCy

SESSION 7: Machine translation

Linguistic typology
Machine translation using encoder-decoder architecture based on RNN and Attention
Greedy decoding
Beam search
MT corpora
Back translation
Bias and ethical issues of MT

SESSION 8: Chatbots and dialogue systems

Properties of human conversation
Rule-based, corpus-based and hybrid chatbots
Task-based dialogue
ChatGPT, other LLMs and applications of Generative AI

SESSION 9: Speech recognition

Speech chain
Waveform in time and frequency domains
Feature extraction
Recognition of spoken words
Generating waveforms of sounds

SESSION 10: Applications of Generative AI

Student presentations