Caltech | Center for Technology & Management Education

California Institute of Technology

Natural Language Processing

2022–2023

Syllabus

==========

Course Page: https://ctme.caltech.edu/nlp-open

Course Description

Natural Languages have evolved from thousands of years of human existence as they pass from generation to generation. The grammar of any natural language is complex and different from other languages. Moreover, it is evolutionary. This makes Natural Language Processing (NLP) a complex challenge.

The main goal of NLP is to understand the meaning of text. Only when computers understand the real meaning of the text, can they take decisive action which must be the intended action. Sentiment analysis of text is one of the important applications of NLP. The use case of sentiment analysis is for the purpose of analyzing customer feedback and tweets. The translation of text between languages is another significant NLP application. Search engines (Google) use NLP to understand the searcher's intent and provide relevant content.

There are currently two different approaches to NLP. The first one is the analysis of words, sentences, and the semantics of text. There are various software packages that can provide these capabilities. These software packages are Natural Language Tool Kit (NLTK), TextBlob, and spaCy.

The other approach to NLP is using the Machine/Deep Learning strategy to analyze the text. Neural Network models are used to train a model by feeding it a lot of text data. Google Cloud Platform (GCP) provides a Machine Learning (ML) API (Application Programming Interface) for the analysis of Natural Languages and provides translation service between languages.

This course explores both approaches to NLP.

For the first approach, the fundamental mathematical analysis of NLP will be covered. Students will write Python code to access NLTK, TextBlob and spaCy software packages. Text will be broken down into tokens. Tokenization process will be covered using Regular Expressions, NLTK and TextBlob software. Text tokens get converted into vectors. The vectorization process will be covered which includes count vectorizer, cosine similarity computation and TF-IDF (Term Frequency Inverse Document Frequency). The semantics of text will be analyzed using Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

For the second approach, students will explore the Machine/Deep Learning models for NLP. Naïve Bayes machine learning model will be used for document classification. Deep learning tools will be used to generate Word Embeddings like Word2Vec. In the end transformers will be covered that includes GPT, BERT for semantic analysis of text.

Course Outline & Schedule

Software Based NLP:

- Analysis of text to understand the meaning of the text.
- Software: Python + TextBlob + Natural Language Tool Kit (NLTK) + spaCy + Pattern
- Analysis of Words + Sentences + Semantics + Polarity + Subjectivity
- Inflection: Pluralization + Singularization
- Normalization: Stemming + Lemmatization
- Semantics using nGrams
- Entity Recognition: spaCy
- Similarity Detection: spaCy
- Language Detection + Translation

Tokenization

- Tokenization Using Regular Expressions
- Tokenization Using NLTK + TextBlob

Vectorization

- Count Vectorizer: Bag of Words (BOW)
- Cosine Similarity
- Zipf's Law
- TF-IDF (Term Frequency + Inverse Document Frequency)

Semantic Analysis

- Eigen Vectors + Values
- Singular Value Decomposition
- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)

Machine Learning Models for NLP

- Machine Learning models for text processing
- Naïve Bayes model for text classification
- Laplace Smoothing

Deep Learning Models for NLP – Word Embeddings

- Deep Learning Models for Text Processing
- Word Embedding: Word2Vec: Skipgram + CBOW

Language Models

- Recurrent Neural Networks (RNN) + LSTM
- Transformers
- GPT 1/2/3
- BERT: Bidirectional Encoder Representations Transformers

Schedule

L#	Subject
1	**Introduction to NLP + Tools + Regular Exp + Tokenization**
	Tools for NLP
	Regular Expressions
	Tokenization Using Regular Exp + NLTK
2	**Text Analysis + Vectorization**
	Text Analysis Using TextBlob
	Similarity: Cosine Similarity + Distance
	Bag of Words
	Zipf's Law
	TF-IDF: Term Freq + Inverse Document Freq

L#	Subject
3	**Semantic Analysis**
	Eigen Vectors + Values
	Singular Value Decomposition
	Latent Semantic Analysis (LSA)
	Latent Dirichlet Allocation (LDA)
4	**NLP Using Machine Learning**
	Naïve Bayes Theory
	Laplace Smoothing
	Naïve Bayes Models for Document Classification

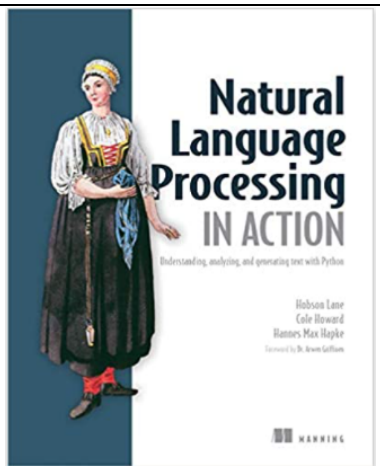
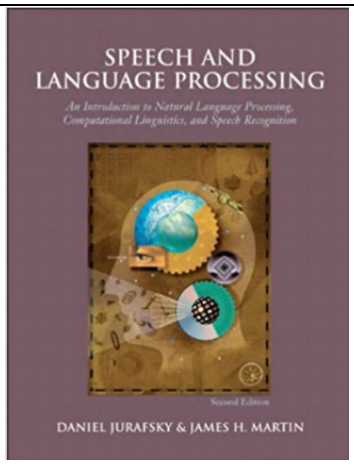
L#	Subject
5	**NLP Using Deep Learning – Word Embeddings**
	Neural Networks Math
	Gradient Descent Optimization + Back Propagation
	Word Embedding: Word2Vec: Skipgram + CBOW
	Generating Word2Vec + Glove
6	**Language Models**
	RNN + LSTM
	Transformers
	GPT-1 + GPT-2 + GPT-3
	BERT: Bidirectional Encoder Representations From Transformers

Learning Objectives

At the end of this course, students will be able to:

- Search text data using Regular Expressions.
- Tokenize text data using NLTK and TextBlob libraries.
- Vectorize text by creating Bag-of-Words (BOW) vectors using Count Vectorizer.
- Vectorize words using Term Frequency + Inverse Document Frequency (TF-IDF).
- Understand Zipf's Law.
- Perform Latent Semantic Analysis (LSA) using Singular Valued Decomposition (SVD).
- Build Naïve Bayes Machine Learning Model for text classification.
- Build Deep Learning Models for creation of Word2Vec vectors.
- Analyze BERT Language Model for NLP Applications.

Textbooks:

1	Natural Language Processing in Action By • Hobson Lane • Cole Howard • Hannes Max Hapke Publisher: Manning Purchase of the book is optional. All class material will be posted on LMS.	
2	Speech and Language Processing Second Edition By • Daniel Jurafsky • James Matin Publisher: Pearson Prentice Hall Purchase of the book is optional. All class material will be posted on LMS.	

Instructor Information

Dr. Ash Pahwa
Office Phone: (949) 378-1229
Email: ash@ashpahwa.com
Website: www.AshPahwa.com

Ash Pahwa, Ph.D., is an educator, author, entrepreneur, and technology visionary with three decades of industry and academic experience. He has founded several successful technology companies during his career, the most recent being A+ Web Services.

Dr. Pahwa earned his doctorate in Computer Science from the Illinois Institute of Technology in Chicago. He is listed in *Who's Who in the Frontiers of Science and Technology*. He is also a Google Certified Analytics Consultant. His expertise includes Deep Learning, Machine Learning, Data Science, Digital Image Processing, Database Management, Digital Video, and Data Storage Technologies.

In Industry, Dr. Pahwa has worked for General Electric, AT&T Bell Laboratories, Xerox Corporation, and Oracle. He founded CD-Gen, Inc. and DV Studio Technologies, LLC., which introduced successful products for CD-Recording (CDR) and MPEG encoding. His book, *CD-Recordable Bible* was published in English, Japanese, and German.

In Academia, Dr. Pahwa teaches courses at California Institute of Technology (Pasadena) and the University of California system. Since 2008, he taught many courses at UC Irvine, UCLA, and UC San Diego.