Text Engineering and Analytics

Ion Androutsopoulos, Associate Professor, AUEB (http://www.aueb.gr/users/ion)

Overview

The course is concerned with algorithms, models, and systems that can be used to process and extract information from natural language texts. A brief introduction to speech processing will also be given. Text and speech analytics methods are used, for example, in sentiment analysis and opinion mining, search engines, question answering, and call centers. They are particularly important in corporate information systems, where knowledge is often expressed in natural language (e.g., minutes, reports, regulations, contracts, product descriptions, manuals, patents). Companies also interact with their customers mostly in natural language (e.g., via e-mail, call centers, web pages describing products, blogs and social media).

Key Outcomes

By the end of the course, the students will be able to design and implement applied text analytics systems and couple them to speech recognition engines. They will also have the necessary background to pursue a research path in human language technology.

Requirements and Prerequisites

Basic knowledge of calculus, linear algebra, probability theory. For the programming assignments, programming experience is required (e.g., in Java, C, C++, Python).

Required Course Materials

There is no required textbook. Extensive notes in the form of slides are provided.

Books

The course is mainly based on the book *Speech and Language Processing*, by D. Jurafsky and J.H. Martin, 2nd edition, Pearson, 2009, which can be found at AUEB's library.

Software/Computing Requirements

The students will be allowed to implement the programming assignments of the course in any language, though Python is recommended. Interested students will be given the option (and some support) to use existing natural language processing and machine learning libraries (e.g., NLTK, scikit-learn, Keras). No specialized hardware is required.

Grading

In each unit, study exercises are provided (solved and unsolved, some requiring programming), of which one or two per unit are handed in (as assignments). Students are graded for their participation in class (10%), assignments (45%), and their performance at the final exam (45%).

Participation

In-class contribution is a significant part of your grade and an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. You can excel in this area if you come to class on time and contribute to the course by:

- Providing strong evidence of having thought through the material.
- Advancing the discussion by contributing insightful comments and questions.
- Listening attentively in class.
- Demonstrating interest in your peers' comments, questions, and presentations.
- Giving constructive feedback to your peers when appropriate.

Please arrive to class on time and stay to the end of the class period. Chronically arriving late or leaving class early is unprofessional and disruptive to the entire class. Repeated tardiness will have an impact on your grade.

Turn off all electronic devices prior to the start of class. Cell phones, tablets, and other electronic devices are a distraction to everyone. If the course requires you to use a laptop or other device in class, you will be informed to do so.

Late Assignments

Late assignments will either not be accepted or will incur a grade penalty unless due to documented serious illness or family emergency. Exceptions to this policy for reasons of civic obligations will only be made available when the assignment cannot reasonably be completed prior to the due date, you make suitable arrangements, and give notice for late submission in advance.

Attendance Requirements

Class attendance is essential to succeed in this course and is part of your grade. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence.

Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission.

Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one's own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Please report observed violations of this policy. Any violations will incur a fail grade at the course and reporting to the senate for further disciplinary action.

Course Syllabus

The course comprises ten units of three hours each.

Units 1 and 2: Intro, language models, spelling correction, text normalization

Introduction, course organization, examples of text and speech analytics applications. *n*-gram language models. Estimating probabilities from corpora. Entropy, cross-entropy, perplexity. Edit distance. Applications in context-sensitive spelling correction and text normalization.

Units 3 and 4: Text classification and clustering

Representing texts as bags of words (or bags of *n*-grams). Boolean and TF-IDF features. Feature selection and extraction using information gain and Principal Components Analysis (PCA). Text classification with *k* nearest neighbors and Naive Bayes. Semi-supervised classification with Expectation Maximization. Word and text clustering with *k*-means. Word embeddings with Pointwise Mutual Information (PMI) scores. Linear and logistic regression, (stochastic) gradient descent (SGD). Lexicon based features, constructing and using sentiment lexica. Support Vector Machines (SVMs) and kernels. Practical advice and diagnostics for text classification with machine learning.

Unit 5: Sequence labeling for texts

Hidden Markov Models (HMMs), Viterbi decoding, unsupervised and semi-supervised learning with Baum-Welch. Maximum Entropy Markov Models (MEMMs) and Conditional Random Fields (CRFs). Applications in part-of-speech tagging, chunking, named entity recognition (NER).

Unit 6: Neural Networks for natural language processing

Natural and artificial neural networks. Perceptrons, training perceptrons with stochastic gradient descent. Multi-layer perceptrons (MLPs), backpropagation. MLPs for text classification, regression, windows-based sequence labeling (e.g., POS tagging, NER). Recurrent neural networks (RNNs) for language modeling. Obtaining dense word embeddings via backpropagation. GRU and LSTM cells, attention, bidirectional RNNs, stacked RNNs. RNNs for sentence classification, sequence labeling, machine translation. Convolutional neural networks (CNNs), sentence classification with CNNs.

Unit 7: Parsing

Phrase-structure grammars, Chomsky's grammar hierarchy, relation to automata. Chart parsing with Earley's parser. Chomsky normal form, CKY parsing. Augmented context-free grammars, DCG grammars.

Probabilistic context-free grammars. Dependency trees, transition-based and graph-based dependency parsers. Neural network approaches to dependency parsing.

Unit 8: Semantics and discourse analysis

Semantic parsing with DCG grammars. Lambda calculus basics. Lexical semantics, WordNet. Thematic roles, FrameNet. Word-sense disambiguation. Obtaining dense word embeddings via Word2Vec, GloVe. Textual entailment recognition with neural networks. Cohesion, discourse segmentation, TextTiling. Coherence, discourse relations, Rhetorical Structure Theory. Types of referring expressions, referring expression resolution. Centering Theory.

Unit 9: Speech recognition and spoken dialogues

Speech sampling, Fourier transform, MFCC features. Using HMMs, *n*-gram language models, and neural networks in speech recognition. Spoken dialogue systems. Dialogue management. Chatbots.

Unit 10: Text analytics systems

Relation extraction from texts, information extraction systems; examples from news feeds. Question answering and text summarization; examples from the biomedical domain. Sentiment analysis and opinion mining; examples from social media and customer product reviews.