Text Analytics

Ion Androutsopoulos, 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 text. Text analytics methods are used, for example, in sentiment analysis and opinion mining, information extraction from documents, search engines and question answering systems. 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 Learning Outcomes

Upon completion of the course, students will be able to:

1. Describe a wide range of possible applications of Text Analytics in Data Science.
2. Describe Text Analytics algorithms that can be used in Data Science applications.
3. Select and implement appropriate Text Analytics algorithms for particular Data Science applications.
4. Evaluate the effectiveness and efficiency of Text Analytics methods and systems.

Requirements and Prerequisites

Basic knowledge of calculus, linear algebra, probability theory. For the programming assignments, programming experience in Python is required.

Required Course Materials

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

- *Speech and Language Processing*, Daniel Jurafsky and James H. Martin. Pearson Education, 2nd edition, 2009, ISBN-13: 978-0135041963. See also the 3rd edition (in preparation): https://web.stanford.edu/~jurafsky/slp3/.
- *Neural Network Methods for Natural Language Processing*, Yoav Goldberg. Morgan & Claypool Publishers, 2017, ISBN-13: 978-1627052986. Available at AUEB's library.
- *Introduction to Natural Language Processing*, Jacob Eisenstein. MIT Press, 2019, ISBN-13: 978-0262042840. See also the on-line draft at: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf.
- *Foundations of Statistical Natural Language Processing*, Christopher D. Manning and Hinrich Schütze. MIT Press, 1999, ISBN-13: 978-0262133609. Available at AUEB's library.

- *An Introduction to Information Retrieval*, Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Cambridge University Press, 2008, ISBN-13: 978-0521865715. Freely available at: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Software/Computing Requirements

An introduction to natural language processing and machine learning libraries (e.g., NLTK, spaCy, scikit-learn, Tensorflow/Keras or PyTorch) will be provided, and students will have the opportunity to use these libraries in the course's assignments. For assignments that require training neural networks, cloud virtual machines with GPUs (e.g., in Google's Colab) can be used.

Grading

In each unit, study exercises are provided (solved and unsolved, some requiring programming), of which 4 or 5 in total are handed in (as assignments). The final grade is the average of the final examination grade (50%) and the grade of the assignments to be submitted (50%), provided that the final examination grade is at least 5/10. Otherwise, the final grade equals the final examination grade.

Participation

In-class contribution is a significant part of our shared learning experience. 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.

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. 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 explicit 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, unless the instructor gives explicit permission. 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.

Unit 1: Introduction, n-gram language models

Introduction, course organization, examples of text analytics applications. *n*-gram language models. Estimating probabilities from corpora. Entropy, cross-entropy, perplexity. Applications in context-aware spelling correction and text generation with beam search decoding.

Units 2 & 3: Text classification with (mostly) linear classifiers

Representing texts as bags of words. Boolean and TF-IDF features. Feature selection and extraction using information gain and SVD. Text classification with *k* nearest neighbors and Naive Bayes. Obtaining word embeddings from PMI scores. Word and text clustering with *k*-means. Linear and logistic regression, stochastic gradient descent. Evaluating classifiers with precision, recall, F1, ROC AUC. Practical advice and diagnostics for text classification with supervised machine learning.

Units 4 & 5: Text classification with Multi-Layer Perceptrons

Perceptrons, training them with SGD, limitations. Multi-Layer Perceptrons (MLPs) and backpropagation. Dropout, batch and layer normalization. MLPs for text classification, regression, window-based sequence labelling (e.g., for POS tagging, named entity recognition). Pre-training word embeddings, Word2Vec. Advice for training large neural networks.

Units 6 & 7: Natural language processing with Recurrent Neural Networks

Recurrent neural networks (RNNs), GRUs/LSTMs. Applications in token classification (e.g., POS tagging, named entity recognition). RNN language models. RNNs with self-attention and

applications in text classification. Bidirectional and stacked RNNs. Obtaining word embeddings from character-based RNNs. Hierarchical RNNs for text classification and token classification. Sequence-to-sequence RNN models with attention, and applications in machine translation.

Units 8 & 9: Natural language processing with Convolutional Neural Networks and Transformers

Quick background on Convolutional neural networks (CNNs) in Computer Vision. Text processing with CNNs. Key-query-value attention, multi-head attention, Transformer encoders and decoders. Pre-trained Transformers and Large Language Models (LLMs), BERT, SMITH, BART, T5, GPT-3, InstructGPT ChatGPT, fine-tuning them, prompting them. Retrieval augmented generation (RAG), LLMs with tools.

Unit 10: Introduction to speech recognition and dialog systems

Introduction to automatic speech recognition (ASR) and systems for spoken and written dialogs. Deep learning encoders of speech segments, wav2vec, HuBERT, encoder-decoder and encoder-only ASR models. Dialog system architectures, intent recognition and dialog tracking using neural models, dialog systems based on pretrained LLMs.