Text mining

from basics to deep learning tools

ECAS – SFdS course

11-15 October 2021 - La Villa Clythia - CAES du CNRS, Fréjus, France

Main topic

Text data are pervasive and can be leveraged to help solving a wide range of problems. This new source of information coupled with recent advances in text mining have incontestably impacted the industry and academic research. While classical approaches yield reasonable performances on diverse text mining tasks, they make restrictive assumptions incompatible with some properties of natural language. In the last decade, these assumptions have been partly relaxed thanks to important breakthroughs in representation learning and deep learning, enhancing the performance for several tasks.

The school is a first introduction to text mining aimed at a broad audience of practitioners. We'll present the classical way of pre-processing, encoding and leveraging text data. Then, we'll introduce recent techniques to learn more meaningful text representations and ways to deal with them using deep neural networks. We'll stress out the importance of using modern approaches to represent the text through case studies with industrial applications.

Some experience in programming with Python is a plus but not a prerequisite. No prior knowledge of any deep learning framework is required.

Detailed program

Course 1 - Introduction

This lecture will serve as a global introduction for the school. First, we will present archetypal applications of text classification at different levels of granularity (*i.e.* document-level classification, sentence-level classification and word-level classification). Second, we will cover the basics of language, both in terms of linguistics (*e.g.* semantics, syntax, complexity) and in terms of statistics (*e.g.* word occurrence frequency, co-occurrence frequency). Last, we will discuss two different ways of encoding text before processing it (*i.e.* bag-of-words encoding versus document/sentence/word embedding).

Course 2 - Basics of text classification

This class will focus on explaining how to adapt classical techniques to learn from text. More specifically, we'll see how to actually encode text under the bag-of-word assumption, how to improve this encoding using weighting schemes and/or dimension reduction techniques and how to adapt classical linear models to solve classification tasks in this particular space.

The audience will have the opportunity to experiment during this class through working on a real-word case study, namely predicting archetypes (informative, practical, entertaining and inspirational) in online content, with the help of the lecturer.

Course 3 - Advanced text classification (I)

Distributed feature vectors are an important breakthrough in the modelling of natural language by means of neural probabilistic models both for academic research as well as practical applications. A suitable embedding of the words on a vocabulary is the key element to enhance the performance of classical classification task. The words are represented by dense vectors in such a way that similar meanings are coded by nearby vectors. We show in this lecture the foundations of these approaches with special attention on word2vec and Glove algorithms. Two case studies are proposed to detect relationships between geological identities, and to predict the load of an electrical system by means of textual data. Finally, we highlight some interesting geometrical properties of the word-vectors obtained by these embedding techniques.

Course 4 - Advanced text classification (II)

The large success on natural language processing using deep neural networks has boosted the research in this direction. While words are satisfactory embedded using distributed feature vectors, longer structures as phrases, paragraphs or whole documents need a special treatment. For instance, words in a sentence seems to be connected by some kind of dependence that need to be explicated. Some neural network architectures (e.g. convolutional or recurrent networks) can learn this dependence to some extent. An intrinsically different approach is to weight explicitly the dependence between words by what is known as an attention mechanism. In this lecture we present some of the recent works on this area to give a landscape of the most up to date challenges of text mining.

Lecturers

Jairo Cugliari is Associate Professor on Statistics at University of Lyon. His research is oriented to industrial data science problems involving complex data such as functions, texts, multicriteria or time series.

http://eric.univ-lyon2.fr/jcugliari/

Adrien Guille is Associate Professor of Computer Science at the University of Lyon. His research interest lies in developing machine learning methods to deal with textual corpora, with an emphasis on structured corpora (i.e. networks of documents).

http://mediamining.univ-lyon2.fr/people/guille/