MASTER OF SCIENCE IN ENGINEERING

Module Description

Analysis of Text Data

General Information

Number of ECTS Credits

3

Module code

TSM_AnTeDe

Responsible of module

Andrei Popescu-Belis, HES-SO

Language

Explanations regarding the language definitions for each location:

- Instruction is given in the language defined below for each location/each time the module is held.
- Documentation is available in the languages defined below. Where documents are in several languages, the percentage distribution is shown (100% = all the documentation).
- The examination is available 100% in the languages shown for each location/each time it is held.

	Bern	Lausanne		Lugano	Zurich		
Instruction	□ E 100%		☑ F	□ E 100%	☑ E 100%		□ D 100%
Documentation	□ E 100%	☑ E 100%	□ F	□ E 100%	☑ E 100%	□ E %	□ D %
Examination	□ E 100%	☑ E 100%	☑ F 100%	□ E 100%	☑ E 100%	□ E 100%	□ D 100%

Module category

□ FTP Fundamental theoretical principles

☑ TSM Technical/scientific specialization module

□ CM Context module

Lessons

2 lecture periods and 1 tutorial period per week (during 14 weeks)

Entry level competencies

Prerequisites, previous knowledge

- Mathematics: basic linear algebra (e.g. matrix multiplications), probability theory (e.g. Bayes theorem)
- Statistics: basic descriptive statistics (e.g., mean, variance, hypothesis testing)
- Programming: good command of a structured programming language (e.g., Python, C++, Java, etc.)
- Machine learning: experimental framework, simple classifiers (e.g. decision trees, Naive Bayes, SVMs)

Brief course description of module objectives and content

This module introduces the main methods of text analysis using natural language processing (NLP) techniques, from a computer / data science perspective. The methods are introduced in relation to concrete applications, in order to extract meaningful, structured knowledge in several dimensions from large amounts of unstructured texts. The knowledge and applications are complementary to those of information retrieval, with several commonalities (e.g. document representation), and advanced IR topics will be included as well.

This module is divided into three parts, each of them starting with the description of one or more text analysis problems. Then, the main methods needed to address them are defined, emphasizing their generality and reusability. Finally, for each part, the methods are instantiated and combined to enable concrete applications.

The three parts are organized by increased sophistication of the analysis of language in texts:

- Text analysis using bags-of-words (i.e. texts are considered as sets of independent words)
- Text analysis using sequences of words
- Text analysis using sentence structure (i.e. considering also the dependencies between words)

Aims, content, methods

Learning objectives and acquired competencies

- The students are able to categorize a text analysis problem and relate the type of analysis that is required and the features to be extracted to a range of known problems.
- The students are able to identify text processing methods to leverage for solving a new problem.
- The students are aware of text processing tools and can adapt off-the-shelf systems to their needs.
- The students understand the role of data and evaluation metrics. Given a text analysis problem they are able to design comparative experiments to identify the most promising solution.

Contents of module with emphasis on teaching content

Introduction [5%]**:** importance of text analysis; layers of language analysis; basic text processing tools and notions of statistics; basic notions of information retrieval; data sources; evaluation methods; overview of the course.

Part A. Text analysis using bags-of-words *[40%]*

Motivating examples: text classification and sentiment analysis, need for word representations accounting for meaning and similarity, distributional semantics.

Methods for learning low-rank word representations from data with illustration of resulting vectors: topic models from LSA to LDA; word embeddings; word sense disambiguation (statistical vs. knowledge-based).

Apply low-rank word representations to text classification, sentiment analysis, information retrieval and content-based text recommendation using bag-of-words models.

Part B. Text analysis using sequences of words *[20%]*

Motivating examples: predict the next word in a sequence, POS tagging, named entity detection.

Methods and their applications: collocation extraction with mutual information, POS tagging with HMMs, NE detection with CRFs, language modeling with n-grams and neural networks.

Part C. Text analysis using sentence structure *[20%]*

Motivating example: natural language inference (reasoning over sentences).

Methods: parsing, semantic role labeling, named entity linking, relationship and fact extraction, neural network models of sentence structure (e.g. CNNs or HANs).

Applications: solving logical entailment with deep neural networks, revisiting sentiment analysis with DNNs, question answering system; automatic information extraction from texts (entities, relationships, facts, events) and linking with ontologies (e.g. DBpedia).

Part D. Special chapters *[15%]*

Perspectives on other text analysis tasks, on multilingual issues, question answering and dialogue, information retrieval and recommendation.

Teaching and learning methods

Classroom teaching; programming exercises

Literature

Foundations of Statistical Natural Language Processing, Christopher Manning & Hinrich Schütze, MIT Press, 1999.

Speech and Language Processing, 2nd edition, Daniel Jurafsky and James H. Martin, Prentice-Hall, 2008.

Introduction to Information Retrieval, Christopher Manning, Prabhakar Raghavan and Hinrich Schütze, 2008.

Natural Language Processing with Python, Steven Bird, Ewan Klein and Edward Loper, O'Reilly, 2009.

Neural Network Methods for Natural Language Processing, Yoav Goldberg, Morgan & Claypool, 2017.

Supplemental material (articles) will be indicated for each lesson.

Assessment

Certification requirements for final examinations (conditions for attestation)

75% of homework passed.

Basic principle for exams:
All the standard final exams for modules are written exams.
The repetition exams can be either written or oral.

Standard final exam for a module and written repetition exam

Kind of Exam	written
Duration of exam	120 minutes
Permissible aids	□ no aids
	☑ permissible aids:
	□ Electronical aids: *no such aids allowed*
	☑ Hardcopy form:1 A4 page (front and back) of handwritten notes
	□ _____

Special case: Repetition exam as an oral exam

If an oral exam is set (only possible for ≤ 4 students), the following applies:

Kind of Exam	oral
Duration of exam	30 minutes
Permissible aids	no aids