Module Description, available in: EN, FR

Analysis of Text Data

General Information

Number of ECTS Credits

3

Module code

TSM_AnTeDe

Valid for academic year

2023-24

Last modification

2021-12-15

Coordinator of the module

Andrei Popescu-Belis (HES-SO, andrei.popescu-belis@heig-vd.ch)

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.

	Lausanne			Lugano	Zurich		
Instruction		**X** F 100%			**X** E 100%		
Documentation			**X** E 100%		**X** E 100%		
Examination		**X** F 100%	**X** E 100%		**X** E 100%		

Module Category

TSM Technical scientific module

Lessons

2 lecture periods and 1 tutorial period per week

Entry level competences

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 Python or another programming language (C++, Java, etc.)
- Machine learning: experimental framework (incl. data partitioning), simple classifiers (e.g. decision trees, Naive Bayes, SVMs), fundamentals of neural networks

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 main 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 a range of text processing tools and libraries 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 deep learning; basic notions of information retrieval; data sources; evaluation methods; overview of the course.

Part A. Text analysis using bags-of-words [35%]
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 using LSA ; word embeddings using feed-forward neural networks.

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 [25%]
Motivating examples: predict the next word in a sequence, POS tagging, named entity detection, contextual word embeddings.

Methods: Hidden Markov models (HMMs), Conditional Random Fields (CRFs), n-grams, seq2seq neural networks, attention-only sequence-to-sequence models (Transformers).

Applications: POS tagging, NE recognition, language modeling, machine translation.

Part C. Text analysis using sentence structure [25%]
Motivating example: natural language inference (reasoning over sentences).

Methods: parsing, semantic role labeling, named entity linking, relationship and fact extraction, neural network models of dialogue.

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).

Part D. Special chapters [10%]
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

Speech and Language Processing, Daniel Jurafsky and James H. Martin, 2nd edition, Prentice-Hall, 2008 / 3rd edition draft, online, 2021.

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

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

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

Assessment

Certification requirements

Module uses certification requirements

Certification requirements for final examinations (conditions for attestation)

75% of homework passed; homework will also contribute towards 20% of the final grade.

Basic principle for exams

As a rule, all the standard final exams for modules and also all resit exams are to be in written form

Standard final exam for a module and written resit exam

Kind of exam

written

Duration of exam

120 minutes

Permissible aids

Aids permitted as specified below:

Permissible electronic aids

Non-programmable pocket calculator.

Other permissible aids

Two A4 sheets (front and back) of personal notes (4 pages).

Special case: Resit exam as oral exam

Kind of exam

oral

Duration of exam

30 minutes

Permissible aids

No aids permitted

Description du module, disponible en: EN, FR

Analyse des Données Textuelles

Informations générales

Nombre de crédits ECTS

3

Code du module

TSM_AnTeDe

Valable pour l'année académique

2023-24

Dernière modification

2021-12-15

Coordinateur/coordinatrice du module

Andrei Popescu-Belis (HES-SO, andrei.popescu-belis@heig-vd.ch)

Explication des définitions de langue par lieu :

- Les cours se dérouleront dans la langue définie ci-dessous par lieu/exécution.
- Les documents sont disponibles dans les langues définies ci-dessous. Pour le multilinguisme, voir la répartition en pourcentage (100% = documents complets)
- L'examen est disponible à 100% dans chaque langue sélectionnée pour chaque lieu/exécution.

	Lausanne			Lugano	Zurich		
Leçons		**X** F 100%			**X** E 100%		
Documentation			**X** E 100%		**X** E 100%		
Examen		**X** F 100%	**X** E 100%		**X** E 100%		

Catégorie de module

TSM approfondissement technico-scientifique

Leçons

2 leçons et 1 leçon de pratique par semaine

Compétences préalables

Connaissances préalables, compétences initiales

- Mathématiques: algèbre linéaire de base (p.ex. multiplication de matrices), notions de probabilités (p.ex. formule de Bayes)
- Statistiques: statistiques descriptives de base (p.ex. moyenne, variance, test d'hypothèse)
- Programmation: maitrise de Python ou d'un autres langage de programmation (C++, Java, etc.)
- Apprentissage automatique (machine learning) : principes des expérimentations (incluant la partition des données), classifieurs élémentaires (p.ex. arbres de décision, classifieur bayésien naïf, machines à vecteur support), fondements des réseaux de neurones

Brève description du contenu et des objectifs

Ce module présente les principales méthodes d'analyse des données textuelles, utilisant le traitement automatique des langues (TAL), dans la perspective de la science des données (data science). Les méthodes sont présentées en relation à des applications concrètes, pour extraire des connaissances sur plusieurs plans, à partir de grandes quantités de textes non-structurés. Ces connaissances et applications sont complémentaires à celles internenant dans le domaine de la recherche d'information (RI), avec toutefois plusieurs points communs (p.ex. la représentation des documents) ; des notions avancées de RI seront également présentées.

Ce module est divisé en trois parties, chacune commençant par la présentation d'un ou plusieurs problèmes d'analyse des données textuelles. Puis, les principales méthodes requises pour résoudre ces problèmes sont définies, en mettant l'accent sur leur généralité et leur réutilisabilité. Enfin, pour chaque partie, les méthodes sont mise en œuvre et combinées en vue d'applications concrètes.

Les trois parties principales sont organisées par ordre croissant de la complexité des analyses textuelles utilisées :

- Analyse de textes utilisant des « sacs de mots » (les textes sont considérés comme des ensembles de mots indépendants)
- Analyse de textes utilisant les séquences (ordonnées) de mots
- Analyse de textes utilisant la structure des propositions (i.e. les relations entre mots)

Objectifs, contenus, méthodes

Objectifs d'apprentissage, compétences à acquérir

- Les étudiants sont capables de classer un problème d'analyse de textes, d'identifier les analyses nécessaires et les traits à extraire, et de les relier à la gamme d'applications déjà étudiées.
- Les étudiants sont capables de choisir les méthodes de traitement automatique des langues à utiliser pour résoudre un problème nouveau.
- Les étudiants connaissent une gamme d'outils et de librairies de TAL et peuvent adapter des systèmes génériques existants à leurs propres besoins.
- Les étudiants comprennent le rôle des données et des métriques d'évaluation. Étant donné un problème d'analyse de textes, les étudiants sont capables de concevoir des expériences comparatives pour identifier la solution la plus prometteuse.

Contenu des modules avec pondération du contenu des cours

Introduction [5%]: importance de l'analyse des données textuelles ; niveaux d'analyse des langues ; outils fondamentaux d'analyse des textes ; bases du *deep learning* et de la recherche d'information ; sources de données ; méthodes d'évaluation ; vue d'ensemble du cours.

Partie A. Analyse de textes comme ensemble de mots [35%]
Motivation (exemples): classification de textes, analyse des sentiments ; nécessité de représenter les mots en tenant compte de leurs sens et leur similarité ; sémantique distributionnelle.

Méthodes: apprentissage de représentations de mots en dimensions réduites, illustration des vecteurs résultants : modèles de *topics* avec LSA ; plongements de mots (embeddings) utilisant des réseaux de neurones *feed-forward*.

Application des représentations en dimensions réduites à la classification de textes, à l'analyse des sentiments, la recherche d'information, et la recommandation de textes basée sur le contenu (modèles « sacs de mots »).

Partie B. Analyse de textes utilisant les séquences de mots [25%]
Motivation (exemples): prédire le mot suivant dans une phrase, étiquetage morphosyntaxique, reconnaissance d'entités nommées, plongements contextuels de mots.

Méthodes : modèles de Markov cachés (HMM), champs aléatoires conditionnels (CRFs), n-grammes, réseaux neuronaux seq2seq, modèles de séquence purement attentionnels (Transformeurs).

Applications : étiquetage morphosyntaxique, reconnaissance d'entités nommées, modèles de langue, traduction automatique.

Partie C. Analyse de textes utilisant les structures des propositions [25%]
Motivation (exemples): capacité à faire des inférences à partir de phrases.

Méthodes: analyse syntaxique, étiquetage des rôles sémantiques, liage des entités nommées, extraction de faits et de relations, modèles neuronaux pour le dialogue.

Applications: identification de l'implication logique ou analyse des sentiments avec des réseaux de neurones ; systèmes de question-réponse ; extraction d'information textuelle (entités, relations, faits, événements).

Partie D. Morceaux choisis [10%]
Perspectives sur les autres tâches d'analyse de textes, le cas des données multililingues, le dialogue humain-machine, la recherche et la recommandation d'information.

Méthodes d'enseignement et d'apprentissage

Enseignement magistral, exercices utilisant la programmation

Bibliographie

Speech and Language Processing, Daniel Jurafsky and James H. Martin, 2e édition, Prentice-Hall, 2008 / 3e édition en ligne, 2021.

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

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

Le matériel supplémentaire (articles) sera indiqué pour chaque cours.

Evaluation

Conditions d'admission

Le module utilise les conditions d'admission

Conditions d'admission à l'examen de fin de module (exigences du certificat)

75% des devoirs à la maison validés ; les notes de ces devoirs contribueront à 20% de la note finale.

Principe pour les examens

En règle générale, tous les examens de fin de module réguliers et les examens de rattrapage sont organisés sous la forme écrite

Examen de fin de module régulier et examen écrit de répétition

Type de l'examen

écrit

Durée de l'examen

120 minutes

Aides autorisées

Les aides suivantes sont autorisées:

Aides électroniques autorisées

Calculatrice de poche non-programmable.

Autres aides autorisées

Notes personnelles sur deux feuilles A4 recto-verso (4 pages).

Cas spécial: examen de répétition oral

Type de l'examen

oral

Durée de l'examen

30 minutes

Aides autorisées

Sans aides