Module Description, available in: EN, FR

Advanced Natural Language Processing

General Information

Number of ECTS Credits

3

Module code

TSM_AdvNLP

Valid for academic year

2024-25

Last modification

2023-09-22

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** E 100%		**X** E 100%		

Module Category

TSM Technical scientific module

Profile Recommendations

This module is recommended for the following profiles:

	Lausanne	Zurich
Profile	DS	DS

Lessons

2 lecture periods and 1 tutorial period per week

Entry level competences

Prerequisites, previous knowledge

- Mathematics: basic linear algebra, probability theory (e.g. Bayes theorem), descriptive statistics and hypothesis testing.
- Machine learning and deep learning (e.g., classifiers, neural networks), basic notions of natural language processing and information retrieval (e.g., preprocessing and manipulating text data, tokenization, tagging, TF-IDF, query-based text retrieval).
- Programming for data science: good command of Python, ability to handle the entire data science pipeline (data acquisition and analysis, design and training of ML models, evaluation and interpretation of results).

Brief course description of module objectives and content

This module enables students to understand the main theoretical concepts that are relevant to text and speech processing, and to design applications which, one the one hand, find, classify or extract information from text or speech, and on the other hand, generate text or speech to summarize or translate language data, or in response to user instructions. The module briefly reviews fundamentals of natural language processing from a data science perspective, with emphasis on methods that support recent approaches based on deep learning models. The module emphasizes the origins and rationale of foundation models, which can be fine-tuned, instructed, or given adequate prompts to achieve a wide range of tasks, thus paving the way towards generative artificial intelligence. The module also provides practical knowledge regarding multi-task models for spoken or written input, multilingual models, and interactive systems, as well as practical skills through hands-on exercises using open-source libraries and models, focusing on the rapid prototyping of solutions for a range of typical problems.

The module is divided into four parts. The first part reviews the main concepts of language analysis and then focuses on the representation of words and the uses of bags-of-words, from the vector space model to non-contextual word embeddings with neural networks; applications include document retrieval and text similarity. In the second part, deep learning models for sequences of words are discussed in depth, preceded by a review of statistical sequence models, with application, e.g., to part-of-speech tagging and named entity recognition. The module presents a paradigm based on foundation models with Transformers – encoders, decoders, or both – which can be fine-tuned to various tasks or used for zero-shot learning. The third part surveys neural models for speech processing and synthesis, along with typical tasks, data and evaluation methods. Finally, the module presents methods that enable natural interaction with generative AI systems, including instruction tuning and reinforcement learning from human feedback, along with spoken and written chatbots, concluding with a discussions of the limitations and risks of such systems.

Aims, content, methods

Learning objectives and competencies to be acquired

- The students are able to frame a problem in the domain of text and speech processing and generation. They can relate a new problem to a range of known problems and adapt solutions to their needs.
- The students are able to specify the characteristics of the data and features needed to train and test models, along with the suitable evaluation metrics. Given a language processing problem, they are able to design comparative experiments to identify the most promising solution.
- The students are able to select, among statistical and neural models, the most effective ones for a given task in language or speech processing and generation. Moreover, they know how to select, between existing libraries and pretrained models, the most suitable ones for a given task. The students are aware of the capabilities of foundation models, and know how to adapt them to specific task, through additional layers, fine-tuning, or prompt engineering.

Module content with weighting of different components

Part I: Words [ca. 20%]

1. Brief review of basic notions of natural language processing: properties of language, speech, and text; subword tokenization, including BPE and SentencePiece; main processing stages, tasks, evaluation metrics, and applications.

2. Text classification and sentiment analysis based on statistical learning with a bag-of-words representation; evaluation metrics for these tasks.

3. Word vectors and their uses: (a) high-dimensional vectors, the VSM model, and application to document retrieval; (b) low-dimensional vectors, non-contextual word embeddings, LSA, word2vec, FastText, and applications to text similarity.

Part II: Word Sequences [ca. 35%]

4. Statistical modeling of word sequences for word-level, span-level and sentence-level tasks; application to part-of-speech (POS) tagging, named entity recognition (NER), and parsing; evaluation methods for these tasks.

5. Language modeling, from n-grams to neural networks; sequence-to-sequence models using deep neural networks, RNNs, Transformers; application to machine translation and text summarization; evaluation methods for these tasks.

6. Foundation models: encoders, decoders, and encoder-decoder models; pre-training tasks; adaptation of models to other tasks using additional layers; fine-tuning pre-trained models; few-shot learning in large language models.

Part III : Speech [ca. 20%]

7. Representation and processing of speech with neural networks; statistical models vs. neural architectures based on RNNs and Transformers; CTC architecture; survey of existing frameworks and pretrained models; notions of speech synthesis.

8. Speech processing tasks, benchmark data and evaluation methods; topic detection, information extraction, and speech translation; multilingual systems.

Part IV: Interaction [ca. 25%]

9. Large language models: survey and emerging capabilities; instruction tuning and reinforcement learning from human feedback (RLHF); prompt engineering.

10. Applications of generative AI; benchmarks with multiple tasks for evaluating foundation models and LLMs; limitations and risks, alignment with human preferences.

11. Spoken and written human-computer interaction: chatbots and dialogue systems.

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, 2023.

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 standard final exams are conducted in written form. For resit exams, lecturers will communicate the exam format (written/oral) together with the exam schedule.

Standard final exam for a module and written resit exam

Kind of exam

Written exam

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 exam

Duration of exam

30 minutes

Permissible aids

No aids permitted

Description du module, disponible en: EN, FR

Traitement automatique des langues avancé

Informations générales

Nombre de crédits ECTS

3

Code du module

TSM_AdvNLP

Valable pour l'année académique

2024-25

Dernière modification

2023-09-22

Coordinateur/coordinatrice du module

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

Explications concernant les langues d'enseignement par site :

- 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** E 100%		**X** E 100%		

Catégorie de module

TSM approfondissement technico-scientifique

Recommandations de profil

Ce module est recommandé pour les profils suivants :

	Lausanne	Zurich
Profil	DS	DS

Leçons

2 leçons et 1 leçon de pratique par semaine

Compétences préalables

Connaissances préalables, compétences initiales

- Mathématiques : algèbre linéaire de base, théorie des probabilités (par exemple, théorème de Bayes), statistiques descriptives et tests d'hypothèses.
- Apprentissage automatique et *deep learning* (p. ex., classifieurs, réseaux de neurones), notions de base de traitement du langage naturel et de recherche d'information (p. ex., prétraitement et manipulation de données textuelles, tokenisation, balisage, TF-IDF, recherche documentaire).

- Programmation pour la science des données : bonne maîtrise de Python, capacité à gérer l'ensemble du pipeline de traitement des données (acquisition et analyse des données, conception et entraînement de modèles ML, évaluation et interprétation des résultats).

Brève description du contenu et des objectifs

Ce module présente les principaux concepts théoriques relatifs au traitement automatique du langage (TAL) écrit et parlé. Ceux-ci permettent par la suite de concevoir des applications qui, d'une part, trouvent, classent ou extraient des informations à partir de textes écrits ou oraux, et d'autre part, génèrent du texte ou de la parole, par exemple pour résumer des textes ou répondre à des instructions des utilisatrices. Le module récapitule les principes fondamentaux du TAL dans la perspective de la science des données, en mettant l'accent sur les méthodes provenant du *deep learning*. Le module s'intéresse aux modèles de langage fondamentaux, qui peuvent être affinés ou instruits pour réaliser de nombreuses tâches, relevant ainsi de l'intelligence artificielle générative. Le module présente également des modèles multitâches pour les textes écrits ou parlés, des modèles multilingues et des systèmes interactifs. Le module offre des compétences pratiques par le biais d'exercices utilisant des bibliothèques et des modèles *open source*, en se concentrant sur le prototypage rapide de solutions à des problèmes typiques.

Le module est divisé en quatre parties. La première partie passe en revue les principaux concepts du TAL puis s'intéresse à la représentation des mots, des modèles vectoriels aux plongements (*embeddings*) non contextuels utilisant les réseaux de neurones, avec application à la recherche de documents et à la similarité de textes. Dans la deuxième partie, les modèles de *deep learning* pour les séquences de mots sont discutés en détail, précédés d'un examen des modèles statistiques de séquences ; les applications concernent par exemple l'étiquetage morpho-syntaxique et la reconnaissance d'entités nommées. L'approche basée sur les Transformers – encodeurs, décodeurs, ou les deux – est également présentée, y compris leur adaptation et le cas non supervisé. La troisième partie aborde les modèles neuronaux pour l'analyse et la synthèse de la parole, avec des tâches, données et métriques typiques. Enfin, le module montre comment atteindre une interaction naturelle avec les systèmes d'IA générative, via des *chatbots* oraux ou écrits, y compris par apprentissage par renforcement à partir de *feed-back* humain, et se termine par une discussion sur les limites et les risques de tels systèmes.

Objectifs, contenus, méthodes

Objectifs d'apprentissage, compétences à acquérir

- Les étudiant·es sont capables de formuler un problème dans le domaine de l'analyse et de la génération du langage écrit ou oral, et peuvent relier un nouveau problème à des cas connus et adapter des solutions à leurs besoins.
- Les étudiant·es sont en mesure de spécifier les caractéristiques des données nécessaires à l'entraînement et au test des modèles, ainsi que les mesures d'évaluation appropriées. À partir d'un problème de TAL, les étudiant·es peuvent concevoir des évaluations comparatives afin d'identifier la meilleure solution.
- Les étudiant·es savent comment sélectionner, parmi des modèles statistiques et neuronaux, les plus efficaces pour une tâche de TAL donnée. De même, entre les bibliothèques et les modèles pré-entraînés, les étudiant·es savent comment sélectionner ceux qui conviennent le mieux à une tâche donnée.
- Les étudiant·es sont conscient·es des capacités des modèles fondamentaux et peuvent les adapter à une tâche spécifique, par le biais de couches supplémentaires, du *fine-tuning*, ou de la conception de nouveaux *prompts*.

Contenu des modules avec pondération du contenu des cours

Partie I : Mots [ca. 20%]

1. Rappels des notions de base du TAL : propriétés du langage écrit ou parlé ; tokenisation en sous-mots avec BPE et SentencePiece ; principales étapes de traitement ; tâches, mesures d'évaluation et applications typiques.

2. Classification de textes et analyse des sentiments basées sur l'apprentissage statistique avec une représentation de type sac-de-mots ; mesures d'évaluation pour ces tâches.

3. Les vecteurs de mots et leur utilisation : (a) vecteurs en grande dimension, le modèle VSM et l'application à la recherche de documents ; (b) vecteurs en faible dimension, plongements de mots non contextuels, LSA, word2vec, FastText et applications à la similarité de textes.

Partie II : Séquences de mots [ca. 35 %]

4. Modélisation statistique de séquences de mots pour des traitements au niveau des mots ou des phrases ; application à l'étiquetage morpho-syntaxique, à la reconnaissance d'entités nommées et à l'analyse syntaxique ; méthodes d'évaluation pour ces tâches.

5. Modèles de langage, des n-grammes aux réseaux de neurones ; modèles de séquence utilisant des réseaux de neurones profonds (RNN et Transformers) ; application à la traduction automatique et au résumé de textes ; méthodes d'évaluation pour ces tâches.

6. Modèles fondamentaux : encodeurs, décodeurs et leur combinaison ; tâches d'entraînement ; adaptation à d'autres tâches à l'aide de couches

supplémentaires ; *fine-tuning* de modèles pré-entraînés ; apprentissage basé sur des exemples.

Partie III : Langage parlé [ca. 20%]

7. Représentation et traitement de la parole avec des réseaux de neurones ; modèles statistiques vs. architectures neuronales basées sur les RNN et les Transformers ; l'architecture CTC ; aperçu des boîtes à outils et des modèles pré-entraînés ; notions sur la synthèse vocale.

8. Tâches de traitement automatique de la parole, données et méthodes pour l'évaluation ; détection de thèmes, extraction d'information et traduction vocale ; systèmes multilingues.

Partie IV : Interaction [ca. 25 %]

9. Grands modèles de langage : présentation de leurs capacités émergentes ; *instruction tuning* et apprentissage par renforcement (RLHF) ; conception de *prompts* ; limites et risques pour la société.

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, 2023.

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 réguliers de fin de module se déroulent sous forme écrite. Concernant les examens de répétition, leur format (écrit ou oral) sera communiqué par l'enseignant-e en même temps que le calendrier des examens.

Examen de fin de module régulier et examen écrit de répétition

Type de l'examen

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

Examen oral

Durée de l'examen

30 minutes

Aides autorisées

Sans aides