MATH 60621: Traitement automatique du language naturel, Hiver 2025

Jeudis 8:30 – 11:30, Édifice Decelles, Salle Luc-Poirier

Professeurs (Jusqu'au 10 mars)
Dr. Eva Portelance (elle)
eva.portelance@hec.ca
Bureau - CSC 4.834 Heure de rencontre: Jeudis 13:00 – 14:00

(Après le 10 mars)
Dr. Marcus Mosbach (il)
marius.mosbach@mila.quebec
Bureau - CSC 4.834 Heure de rencontre: TBD

1 Aperçu

Ce cours couvrira les outils et modèles fondamentaux utilisés pour analyser et générer le langage naturel dans un contexte informatique. Nous apprendrons les principes de base des méthodes contemporaines de traitement automatique du langage naturel (TALN - NLP).

Les sujets couverts incluront la structure des données textuelles et la manière dont elles peuvent être utilisées pour des tâches telles que la recherche et classification de documents, la génération de textes et le résumé de textes. Nous couvrirons les modèles de langage statistique et ceux basé sur les réseaux neuronaux.

Les sujets couvert incluent:

- La nature des données textuelles
- Le prétraitement du texte: tokenisation et lemmatisation
- Les modèles sac-de-mots et la classification naïve
- Les modèles de langage N-grams (modèles de Markov)
- Les modèles de Markov cachés et l'étiquetage grammatical
- Les représentations distribuées, la sémantique vectorielle et les modèles thématiques
- Les modèles de langage neuronaux récurrents, LSTM et la génération de texte
- Les modèles de language Transformers et la modélisation du langage masqué
- Les modèles encodeurs et la recherche sémantique
- Les modèles encodeurs-décodeurs, le résumé de texte et la traduction

A l'issue de ce cours, les étudiants seront capables de :

- Comprendre la nature des données textuelles et la manière dont les algorithmes statistiques peuvent leur être appliqués.
- Comprendre comment les modèles de langage statistiques et basés sur les réseaux neuronaux sont structurés et comment les mettre en œuvre.
- Apprendre à réfléchir de manière critique sur les méthodologies en TALN actuelles et à identifier les applications raisonnables de chaque algorithme couvert.

2 Prérequis

Le cours suppose une expérience préalable du langage de programmation Python, mais aucune expérience préalable en apprentissage automatique n'est requise.

3 Structure des cours

Les cours seront en présentiel et ne seront pas enregistrés. Les diapos seront toujours mis à la disposition des étudiants avant le cours et les étudiants doivent prendrent leurs propres notes.

La première moitié de chaque cours consistera en une présentation du nouveau matériel de classe de façon magistrale. Durant la deuxième moitié, les étudiants travailleront ensemble en classe sur des exercises pratiques sur les sujets couverts durant le cours.

De même, il y aura deux types d'évaluations: (1) des devoirs pratiques qui consisterons seulement de question de codage appliqué sur lesquels les étudiants pourons travailler soit seul ou en équipe de 2-3 personnes (chacun évalué individuellement); (2) des examens qui couvrirons des questions théoriques et éthiques écrites.

Les étudiants sont encouragés à venir aux heures de rencontre pour poser des questions en personne ou pour discuter du matériel si nécéssaire.

4 Politique de travail de groupe

Les étudiants peuvent travailler seul ou en équipe de 2-3 personnes sur les devoirs (ils sont même encouragés à le faire). S'ils choisissent de travailler en groupe, ils doivent inscrire les noms des étudiants avec lesquels ils ont collaboré sur devoir dans le champ indiqué. Ils doivent aussi toute fois chacun soumettre leur travail individuellement et seront évalués individuellement.

5 Politique de courriel

Veuillez compter deux jours ouvrables pour toute réponse à un courriel. Je n'enverrai ou ne répondrai jamais à des courriels en dehors des heures de travail (8:00-17:00)

6 Politique de travaux soumis en retard

Les étudiants se verront enlever 15% de la note final du devoir par jour de retard, sans exceptions. Les devoirs seront mis à la disposition des étudiants 3 semaines avant leur date de remise – planifiez bien votre temps, il n'y a pas de raison d'être en retard.

7 Évaluation

- **Devoirs (50%)**: 2 devoirs, chacun pour 25%
- **Examen d'intra (25%)**
- **Examen final (25%)**

8 Resources

Les lectures de chapitres recommandés viendront du livre suivant, accessible en ligne:

- Dan Jurafsky and James H. Martin. (2024). *Speech and Language Processing (3rd Edition draft)*. https://web.stanford.edu/ jurafsky/slp3/

9 Plan de la session

Notez que les dates exactes ci-dessous sont susceptibles d'être modifiées, en fonction de la rapidité avec laquelle nous avançons dans les sujets du cours.

Date	Contenu
9 janvier	Introduction du cours et la nature des données textuelles

16 janvier	Le prétraitement du texte: tokenisation et lemmatisation *Lectures suggérées: Chapter 2, up to section 2.7*
23 janvier	Les modèles sac-de-mots et la classification naïve *Lectures suggérées: Chapter 4* *Notes:* Devoir 1 publié
30 janvier	Les modèles de langage N-grams, les modèles de Markov cachés et l'étiquetage grammatical *Lectures suggérées: Chapter 3 and Appendix A*
6 février	Les représentations distribuées, la sémantique vectorielle et les modèles thématiques *Lectures suggérées: Chapter 6*
13 février	La regression logistique et les réseaux neuronaux de base *Lectures suggérées: Chapter 5 and Chapter 7*
20 février	Les modèles de langage neuronaux récurrents et les LSTMs *Lectures suggérées: Chapter 8* **Devoir 1 à remettre**
28 février	[Relâche]
6 mars	**Examen d'intra**
13 mars	Les modèles de language Transformers et la modélisation du langage masqué *Lectures suggérées: Chapter 9 and Chapter 11* *Notes:* Devoir 2 publié
20 mars	Les modèles encodeurs et la recherche sémantique *Lectures suggérées: Chapter 14*
28 mars	Les modèles encodeurs-décodeurs, le résumé de texte et la traduction *Lectures suggérées: Chapter 13*
3 avril	Le contexte actuel: les LLMs, Chatbots, et le prompting *Lectures suggérées: Chapter 10 and Chapter 12* **Devoir 2 à remettre**
10 avril	Présentation invitée ou rattrappage
TBD	**Examen final**