

MATH 60621: Traitement automatique du langage naturel, Hiver 2026

Jeudis 8:30 – 11:30, Local TBD

Professeure DR. EVA PORTELANCE (elle)
eva.portelance@hec.ca
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.

Thèmes couverts:

- La nature des données textuelles
- Le prétraitement du texte
- Les modèles de langage N-grams et la classification
- Les embeddings et la sémantique vectorielle
- Les modèles de langage neuronaux récurrents et LSTMs
- Les modèles de langage Transformers
- Les modèles encodeurs et la recherche sémantique
- Les modèles encodeurs-décodeurs et les tâches génératives
- Le peaufinement et l'apprentissage en contexte

À 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 prendre 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 exercices pratiques sur les sujets couverts durant le cours.

De même, il y aura deux types d'évaluations: (1) des devoirs pratiques qui consisteront seulement de question de codage appliqué sur lesquels les étudiants pourront travailler soit seul ou en équipe de 2-3 personnes (chacun évalué individuellement); (2) des examens qui couvriront 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 nécessaire.

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'envverrai 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 Ressources

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

- Dan Jurafsky and James H. Martin. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition*. Online manuscript released August 24, 2025. <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
8 janvier	Introduction du cours et la nature des données textuelles
15 janvier	Le prétraitement du texte <i>Lectures suggérées: Chapter 2, jusqu'à 2.7</i>
22 janvier	Les modèles de langage N-grams <i>Lectures suggérées: Chapter 3</i>

29 janvier	Les embeddings et la sémantique vectorielle <i>Lectures suggérées: Chapter 5</i> <i>Notes: Devoir 1 publié</i>
5 février	La regression logistique et les réseaux neuronaux de base <i>Lectures suggérées: Chapters 4 and 6</i>
12 février	Les modèles de langage neuronaux récurrents et les LSTMs <i>Lectures suggérées: Chapter 13 jusqu'à 13.6</i>
19 février	Les modèles de langage Transformers <i>Lectures suggérées: Chapter 8</i>
26 février	[Relâche] Devoir 1 à remettre
X mars	Examen d'intra
12 mars	Les grands modèles de langage <i>Lectures suggérées: Chapter 7</i>
19 mars	Les modèles encodeurs et la recherche sémantique <i>Lectures suggérées: Chapters 10 and 11</i> <i>Notes: Devoir 2 publié</i>
27 mars	Les modèles encodeurs-décodeurs et les tâches génératives <i>Lectures suggérées: Chapter 13.7-13.9 and Chapter 12</i>
2 avril	Le peaufinement et l'apprentissage en contexte <i>Lectures suggérées: Chapter 9</i>
9 avril	Présentation invité sur la recherche
13 avril	Devoir 2 à remettre
X avril	Examen final