Université de Montréal

École de bibliothéconomie et des sciences de l'information

Plan du cours SCI6203 - Intelligence artificielle et données textuelles (Automne 2024)

École de bibliothéconomie et des sciences de l'information, Maîtrise en sciences de l'information (*cours optionnel*)

3 crédits

Professeur titulaire : Dominic Forest
Courriel : dominic.forest@umontreal.ca
Téléphone : 514-343-6119
Bureau : C-2046
Disponibilités : Sur rendez-vous

Site Web du cours : https://studium.umontreal.ca/

Les cours se donnent le Lundi, 15h30 à 18h30, C-2046.

Description

Concepts et techniques de l'analyse de données textuelles. Fouille de textes. Extraction et organisation automatiques. Méthodes descriptives et prédictives pour l'analyse de corpus documentaires. Applications d'algorithmes d'IA en sc. de l'info.

Objectifs d'apprentissage

À titre d'objectif général, ce cours entend développer chez les étudiants les habiletés intellectuelles et techniques nécessaires à l'utilisation réfléchie de certains outils de fouille et d'analyse de données textuelles dans le domaine des sciences de l'information. La réalisation de cet objectif implique que l'étudiant soit en mesure, au terme du cours, de témoigner de ses habilités à réaliser de manière concrète un certain nombre de tâches et d'opérations.

Méthodes pédagogiques

Le cours sera composé d'un volet plus théorique, sous forme d'exposés magistraux, de démonstrations de logiciels, de lectures et de discussions en groupe, portant sur les notions et techniques relatives à la matière étudiée. Lors de ces exposés, les étudiants sont invités à intervenir activement pour discuter de la matière au programme. Il sera aussi composé d'un volet plus pratique visant à expérimenter les diverses techniques et les logiciels présentés. Ce second volet a pour objectif d'approfondir certaines notions et de permettre à l'étudiant une prise de contact directe avec les techniques et les logiciels vus en classe.
Pour ce cours, la présentation des exposés magistraux sera soutenue par des documents PowerPoint. Cependant, ces documents ne sauraient en aucun cas remplacer les exposés magistraux. La maîtrise des concepts et des techniques abordés dans le cours repose principalement sur la présence et la participation au cours. Les documents ne constituent donc que le support de diffusion pédagogique du contenu du cours. Dans les jours suivant chaque cours, les étudiants pourront télécharger ces documents à partir du site Web du cours. On y retrouvera aussi plusieurs ressources (textes, logiciels à télécharger, etc.) en lien avec la matière présentée dans le cours. Ce site Web servira aussi de lieu d'interaction entre le professeur et les étudiants. Il est donc fortement recommandé de consulter régulièrement ce site. En plus des documents relatifs à chaque cours, des textes et des informations supplémentaires en rapport avec la matière présentée en classe seront mis à la disposition des étudiants.

Contenu du cours

Ce cours vise à introduire les étudiants au domaine de la fouille de textes et, plus globalement, à celui de l'analyse des données textuelles (text analytics). Parmi les différentes approches abordées dans le cadre du cours, une importance particulière sera accordée à l'intelligence artificielle.
Le cours s'insère dans une problématique en science des données liée à l'analyse et à la gestion informatisées des documents textuels. Au niveau théorique, nous présenterons les concepts fondamentaux, les enjeux, ainsi que les techniques principales du domaine de la fouille de données textuelles. Au niveau pratique, nous exposerons les principes et les fonctionnalités de quelques outils informatiques de fouille, dans leur application aux sciences de l'information. Plusieurs approches seront exposées et discutées (statistique, linguistique, etc.). Toutefois, un volet important du cours sera consacré au traitement numérique des documents textuels.
Le contenu du cours est de nature multidisciplinaire. Les théories et les concepts présentés proviennent de plusieurs disciplines (linguistique, informatique, science de l'information, etc.). Le cours ne présuppose cependant aucune connaissance approfondie dans des domaines autres que celui des sciences de l'information.
Le cours est divisé en trois volets thématiques principaux répartis inégalement durant la session :
1. Introduction à la fouille de données textuelles. La première partie du cours vise à introduire les étudiants au domaine de la fouille de données. À cet égard, nous définirons et exposerons les concepts principaux et les enjeux fondamentaux de ce domaine. En outre, nous en distinguerons les différentes approches théoriques et en délimiterons les contextes d'application.
2. Les principales techniques et méthodes de fouille de données. La deuxième partie du cours présente les principales techniques et méthodes de fouille de données textuelles. C'est dans le contexte que nous présenterons le domaine de l'intelligence artificielle et de l'apprentissage automatique. Nous présenterons et comparerons différentes approches d'extraction et d'organisation automatiques d'informations. Plus spécifiquement, nous exposerons différentes méthodes visant à assister le prétraitement, la transformation, le regroupement et la classification des données textuelles. Dans le cadre de ce volet, nous présenterons les principes généraux de quelques algorithmes d'apprentissage automatique et d'intelligence artificielle pour assister le traitement des données textuelles.

3. Les différentes applications de fouille de documents. Ce volet thématique sera traité dans plusieurs séances. Nous verrons comment les différents processus de fouille et d'analyse textuelle peuvent être employés dans différentes applications. Nous présenterons et distinguerons aussi différentes catégories d'applications (description des documents, analyse thématique et identification automatique de thèmes, indexation automatique et repérage d'informations, extraction et découverte d'informations, analyse d'opinions, etc.).

Calendrier des activités

Date	Thématiques abordées	Travaux ou évaluation
09 septembre 2024	Présentation du plan de cours et des modalités d'évaluation	
16 septembre 2024	Intro. à l'analyse des données textuelles	
23 septembre 2024	Les campagnes d'évaluation, le modèle vectoriel pour le traitement des documents	
30 septembre 2024	Les corpus de documents textuels	
07 octobre 2024	La segmentation des documents textuels, l'extraction du lexique, les statistiques textuelles	
14 octobre 2024	Congé	
21 octobre 2024	Semaine de lecture	Remise de la partie 1
28 octobre 2024	Introduction à l'intelligence artificielle	
04 novembre 2024	Fouille de données textuelles et analyse descriptive : le regroupement automatique des documents	
11 novembre 2024	Fouille de données textuelles et analyse prédictive : la classification automatique des documents	Remise de la partie 2
18 novembre 2024	Fouille de données textuelles et analyse prédictive : la classification automatique des documents	
25 novembre 2024	Exemple d'application, expérimentation	Remise de l'évaluation
02 décembre 2024	Expérimentation	
09 décembre 2024	Expérimentation	Remise de la partie 3

Évaluation

Pour réussir ce cours, il est essentiel d'assister aux exposés magistraux et de participer activement aux laboratoires (lors desquels vous pourrez débuter vos travaux pratiques). L'évaluation du niveau de compréhension des notions et de la maîtrise des habiletés techniques se fera au moyen de plusieurs évaluations.
Description détaillée de l'évaluation proposée
a) Expérimentation en fouille de textes, présentation sous forme d'article scientifique (en équipe de deux) [80%] :
1. Partie 1. Définition de la problématique et présentation de travaux reliés [25%]
2. Partie 2. Constitution d'un corpus [20%]
3. Partie 3. Expérimentation [35%]
b) Évaluation d'un logiciel (évaluation individuelle) [20%].

Politiques, règlements et directives

L'ensemble des politiques, règlements et directives énoncés dans le Règlement pédagogique de la Faculté des études supérieures et postdoctorales s'appliquent. Une attention particulière est à porter aux éléments suivants :
- Règlement disciplinaire sur le plagiat ou la fraude concernant les étudiants. Tous les étudiants doivent prendre connaissance du document « Règlement disciplinaire sur le plagiat ou la fraude concernant les étudiants ».
Remarque : L'utilisation d'une intelligence artificielle (telle que ChatGPT) pour vous accompagner dans la réalisation de vos travaux est encouragée (selon des modalités discutées en classes). Cependant, nous vous demandons de citer ChatGPT comme n'importe quelle autre source d'information. Bien que les façons de citer demeurent à éclaircir (il n'y a pas encore de consensus dans la communauté), vous trouverez sur le site des bibliothèques, des indications et des ressources pour apprendre à citer ChatGPT.
- Délais et dates de remise des travaux. Tout retard non justifié dans la remise d'un travail sera sanctionné : 5 % de la note maximale du travail retranchés par jour calendaire de retard, jusqu'à concurrence de 35 %. Le jour de la date prévue de la remise du travail ne compte pas. Le samedi et le dimanche ainsi que les jours fériés sont comptés. Au-delà de ce délai : note F (échec).

- Enregistrement des cours. La prestation des cours est soumise au droit d’auteur. Il est interdit de faire une captation audio ou vidéo du cours, en tout ou en partie, sans le consentement écrit du professeur. Le non-respect de cette règle peut mener à des sanctions disciplinaires en vertu du Règlement disciplinaire concernant les étudiants. Une autorisation écrite de la part du professeur est requise pour réaliser un enregistrement audio ou vidéo d’un cours, même pour un usage strictement personnel. Les étudiants en situation en handicap doivent présenter au professeur, au début du cours, le formulaire de mesures d’accommodement du SESH qui leur accorde le droit d’enregistrer les cours.
- Qualité de la langue. Un maximum de 10% de la note globale d’un travail pourra être retranché pour la qualité de la langue.

Ressources

Voir site Web du cours.

Page mise à jour le 10 juillet 2024