Introduction aux mégadonnées en sciences sociales (FAS 1001)

Hiver 2025

Horaire du cours: Jeudi, 8h30 - 11h29

Local de cours: pavillon Marie-Victorin local G-415_555B

Disponibilité: Sur rendez-vous

Courriel: lomf0@ulaval.ca

Prérequis

Ce cours ne demande aucun prérequis en statistiques avancées ou en programmation, mais avoir des connaissances au préalable demeure un atout. Il est fortement conseillé de suivre en parallèle le cours FAS 1003 - *Visualisation des données*, car la visualisation graphique demeurera un élément important dans la présentation de vos travaux. Ce cours priorise un enseignement avec des ateliers afin de permettre aux étudiant-es de se pratiquer à la programmation, de poser des questions au chargé de cours et d'effectuer leur travail de mi-session et de fin de session.

Description du cours et objectifs

Jamais autant de données n'ont été disponibles pour comprendre les comportements humains. De nouveaux outils de recherche nous permettent désormais de quantifier des données difficilement analysables auparavant, tels que de larges corpus de textes, des images ou des audios en provenance de vidéos. Néanmoins, comment collecter, traiter et analyser ces nouvelles données? Comment combiner ces nouvelles données avec celles déjà largement utilisées en sciences sociales comme les sondages? Quels sont les enjeux techniques et éthiques que

soulèvent l'utilisation de ces données en recherche dans le contexte d'un développement important de l'intelligence artificielle et de l'application croissante de l'apprentissage par machines en sciences sociales? Ce cours aborde ces nombreux enjeux avec une ***approche pratique*** à l'utilisation des mégadonnées en sciences sociales.

À la fin de ce cours, les objectifs suivants seront remplis:

- Avoir une connaissance globale des différentes sources de données disponibles pour étudier les phénomènes humains.
- Développer l'autonomie nécessaire pour collecter, gérer et analyser quantitativement des bases de données et les présenter avec un projet de recherche les mobilisant.
- Faire preuve d'une compréhension des enjeux entourant la mobilisation et l'utilisation de volumineuses bases de données en sciences sociales.

Pédagogie

Le langage de programmation utilisé pour ce cours est `R`. Ce dernier est téléchargeable gratuitement ici. Bien que plusieurs options soient possibles, il vous est également demandé de télécharger `RStudio` ici. RStudio est l'environnement de développement intégré de prédilection pour coder en `R`. Il vous permettra d'éditer et de déboguer plus facilement votre code en plus de vous donner les outils nécessaires pour transformer, prévisualiser et analyser vos données efficacement. Certains étudiant-es ont signalé être plus à l'aise d'utiliser le langage de programmation `Python`. Si vous vous sentez plus à l'aise d'utiliser ce dernier, il est possible de le faire lors de vos travaux de session. Cependant, les capacités du chargé de cours à vous aider seront plus limitées. Cela est à prendre en considération. Les étudiant-es devront, si ce n'est pas déjà fait, créer un compte `GitHub` lors du premier cours, afin de remettre leurs codes de travaux pratiques et de travaux de session. `GitHub` est un service web d'hébergement et de gestion de développement de logiciels qui est largement utilisé en industrie pour le partage de codes. Cela permettra de socialiser les étudiant-es à son utilisation.

- Le bon déroulement de ce cours nécessite que vous ayez un ordinateur, notamment pour les nombreux ateliers pratiques lors des classes. Si vous n'avez pas d'ordinateur, une solution sera trouvée, afin que vous puissiez participer aux activités en classe.
- Généralement, le cours comporte une heure et demie d'enseignement magistral, une pause de 15 minutes et une heure d'atelier en classe. Le cours magistral abordera les lectures et autres ressources obligatoires à consulter avant le cours ainsi que le contenu du jour. Il est attendu une présence des étudiant-es aux cours, car nos ateliers en classe seront importants pour comprendre la matière puis pour avancer dans ses travaux de session.

- Il y aura également la présence de plusieurs personnes invitées ayant des expériences concrètes de recherche avec des mégadonnées à l'intérieur et à l'extérieur du milieu académique. Les présentations seront discutées au début du cours en plus des lectures/vidéos et autres sources que vous devez consulter avant chacun d'entre eux. Lorsqu'un-e invité-e sera présent-e, la partie magistrale du cours sera légèrement moins longue. Les présentations des invité-es dureront environ 30 minutes avec une période de questions de 15 minutes.

Évaluations

Sur un total de 100 points:

Participation en classe (10/100)

10% des points seront alloués à **la participation en classe**. L'entraide entre les étudiant-es est encouragée (évaluation et correction de code sur `GitHub` ou lors des ateliers).

Travaux pratiques (20/100)

20% des points seront alloués **aux travaux pratiques**. Les étudiant-es devront remettre 4 travaux pratiques qui permettront d'évaluer leur évolution à différentes étapes de la session. Les travaux pratiques devront être mis sur votre compte `GitHub` **une journée avant le cours suivant, soit lundi 23h59**. Vous pouvez commencer ou même terminer vos travaux pratiques lors des ateliers qui sont également des périodes pour avancer vos travaux de session. Les jours de remise seront dans le calendrier du cours.

- Introduction à `GitHub` (**5 points**)
- Combiner des données de sondages (**5 points**)
- Analyse de données textuelles (**5 points**)
- Webscraping (**5 points**)

Travaux de session (70/100)

Travail de mi-session (15/100)

15% des points seront alloués à la remise d'un **plan de recherche de votre travail de session de 5 pages**. Le plan de recherche permettra d'avoir un regard du chargé de cours sur l'évolution du travail de session et de corriger la trajectoire de ce dernier si nécessaire. Le plan de recherche consistera principalement en la présentation de votre question de recherche, les raisons motivant votre recherche, les données que vous comptez utiliser et la ou les méthode-s mobilisée-s. Ce travail peut également être la base de votre travail de session final.

Travail de session (40/100)

40% des points seront alloués à **votre travail de session individuel pouvant aller de 10 à 15 pages**. Ce dernier consistera en une recherche complète avec présentation de la question de recherche, les raisons motivant cette dernière, la présentation des données utilisées, la ou les méthode-s mobilisée-s, l'analyse des données et de la présentation des conclusions de l'étude. La ou les méthode-s utilisée-s pour analyser les données n'ont pas besoin d'être poussée-s, une plus grande attention sera portée sur les données collectées, la transformation effectuée sur ces dernières et leur présentation.

Présentation orale (15/1000)

15% des points seront alloués à la présentation orale de **votre travail de session**. La présentation de votre travail de session vous amènera à parler un peu plus de vos données, des conclusions de vos recherches, mais également du cheminement par lequel vous êtes passé, afin d'arriver aux dernières étapes du projet. Les présentations seront de 10 à 15 minutes avec 10 minutes de questions de l'audience et s'échelonneront sur deux périodes du cours. Un support visuel est requis pour les présentations orales. Il est attendu que tous les étudiant-es soient présent-es aux présentations orales.

Ouvrage obligatoire

Le livre de référence du cours sera celui du Professeur Rohan Alexander de l'Université de Toronto *Telling Stories with Data* et disponible gratuitement sur son site web.

Ouvrage ressource

Pour un livre ressource en méthodes quantitatives, il est conseillé de consulter le livre du Professeur Vincent Arel Bundock *Analyse causale et méthodes quantitatives* disponible gratuitement sur le site des Presses de l'Université de Montréal.

Calendrier

Cours 1 (9 janvier 2024) : Introduction et présentation du plan de cours.

- Lectures pour le prochain cours: Lire les chapitres 3 et 4.
- Bonus: **Vous pouvez lire si intéressé(e)** les chapitres 1 et 2.

Cours 2 (16 janvier 2024) : Outils de communication et de collaboration en recherche.

- Description: Introduction à `GitHub` et `Quarto`.
- Lectures pour le prochain cours: Lire la section 8.3 du chapitre 8. Lire également Breton, Cutler, Lachance et Mierke-Zatwarnicki (2017), *Telephone versus online survey modes for election studies: Comparing Canadian public opinion and vote choice in the 2015 federal election*, dans la *Revue Canadienne de Science Politique*. Visionner ce vidéo qui montre un exemple cocasse d'une recherche avec sondage qui souligne les forces et les limites de ce type de données.
- Bonus: **Vous pouvez lire si intéressé(e)** Zaller et Feldman (1992), *A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences*, dans *l'American Journal of Political Science* sur les réponses données lors des sondages. Vous pouvez également lire Ansolabehere, Rodden et Snyder (2008), *The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting*, sur les échelles de mesure pour réduire les erreurs en analyse de sondages. Ces lectures seront utiles pour comprendre la matière du prochain cours.

Cours 3 (23 janvier 2024) : Données de sondages.

- Description: Utilisation, avantages et inconvénients, structuration de bases de données, transformation de variables, variables latentes, expériences par sondage et quasi-expériences.

- Lecture pour le prochain cours: Lire Grimmer, Roberts, et Stewart (2021). *Machine Learning for Social Science: An Agnostic Approach.*, dans *Annual Review of Political Science*.

Cours 4 (30 janvier 2024) : Machine Learning.

- Description: Entraînement de modèles, algorithmes et place de l'IA en sciences sociales.
- Invitée: Professeure Catherine Ouellet du département de science politique de l'Université de Montréal et co-créatrice de l'application Datagotchi.
- Lectures pour le prochain cours: Lire le texte de Grimmer, et Stewart (2013) *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts* dans *Political Analysis*. Lire également le chapitre 17.

Cours 5 (6 février 2024) : Analyses textuelles automatisées partie 1.

- Description: Données tirées de texte, analyses du dictionnaires, analyses supervisées.

Cours 6 (13 février 2024) : Analyses textuelles automatisées partie 2.

- Description: Analyses non-supervisées, word embedding, machine learning et autres.
- Lectures pour le prochain cours: Lire le texte de Chetty et collègues (2022), *Social capital II: determinants of economic connectedness* dans *Nature*. Bien important de lire ***Social capital II*** et non I.
- Bonus: **Vous pouvez lire si intéressé(e)** Barbera (2015), *Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data*, dans *Political Analysis*; Chetty et collègues (2022), *Social capital I: measurement and associations with economic mobility* dans *Nature*; Guinaudeau, Munger et Votta (2022), *Fifteen seconds of fame: TikTok and the supply side of social video*, dans *Computational Communication Research*; Nyhan et collègues (2023), *Like-minded sources on Facebook are prevalent but not polarizing* dans *Nature*.