Année universitaire 2023-2024

Introduction à l'Analyse de Données Culturelles
Introduction to Cultural Data Analysis

Responsable du cours : Valentin Thouzeau

Autre(s) enseignante(s) / enseignant(s) : Mathieu Tiret (INRA), Félix Beroud (INEE)

Descriptif du cours :

L'analyse de données est une discipline qui vise à extraire des connaissances exploitables à partir de grandes quantités de données. Nous étudierons comment l'analyse de données s'applique aux données culturelles et comment elle peut fournir un angle d'approche unique pour aborder des questions interdisciplinaires.

Nous aborderons les méthodes fondamentales de l'analyse de données, notamment l'utilisation d'outils statistiques tels que les représentations graphiques, les tests de significativité et les modèles de régression. Nous verrons comment ces techniques peuvent être utilisées pour décrypter les relations complexes au cœur des données culturelles, et comment elles peuvent aider à mieux comprendre l'histoire, la littérature et l'art.

L'objectif de ce cours est d'aider les étudiants à maîtriser les compétences nécessaires pour explorer les données culturelles de manière rigoureuse et approfondie, et être ainsi en mesure de découvrir des informations cachées dans les données qui seraient autrement inaccessibles. Le cours sera centré sur la mise en pratique des compétences de l'analyse de données culturelles à travers la réalisation de projets interdisciplinaires.

Objectifs pédagogiques et compétences développées :

Ce cours a pour objectif de doter les étudiants des méthodes et des outils centraux de l'analyse de données. Ils pourront acquérir ces compétences par la pratique en réalisant l'ensemble des étapes d'un projet interdisciplinaire :

1. Construction d'une question de recherche
2. Collecte de données culturelles
3. Nettoyage et formatage des données
4. Représentations graphiques des données
5. Analyse statistique des données
6. Interprétation des résultats
7. Communication des résultats

Contenu détaillé du cours :

1. **Introduction aux méthodes d'analyse de données**
   - Concepts et outils de base de l'analyse de données (population, échantillon, distribution, moyenne, variance, hypothèses, test, p-value, D.A.G)
   - Types de données et méthodes d'analyses correspondantes
   - Initiation à la programmation pour l'analyse de données

2. **Collecte de données**
   - Sources de données historiques, littéraires et artistiques
   - Création de questionnaires
   - Initiation aux langages *SQL* et *SPARQL*
   - Éthique de la collecte de données

3. **Préparation des données**
   - Nettoyage des données
   - Transformation des données
   - Normalisation des données

4. **Analyse de données I**
   - Visualisation graphique de données univariées :
     Tableau simple, diagramme en barres, digramme en boîte, histogramme
   - Visualisation graphique de données bivariées :
     Tableau double, digrammes en boîte, digramme de dispersion
   - Tests simple : Correlation, Student, Chi-deux
   - Régression linéaire simple et multiple

5. **Analyse de données II**
   - Régression logistique
   - Effet d'interaction
   - Effet aléatoire
   - Classification et prédiction
   - Analyse Factorielle, Analyse en Composantes Principales

6. **Applications interdisciplinaires**
   - Élaboration de questions de recherche
   - Spécificités de l'analyse données appliquée à l'étude de données historiques, littéraires et artistiques
   - Initiation au traitement automatique du langage

7. **Réalisation du projet interdisciplinaire**
   - Choix d'un sujet de recherche interdisciplinaire
   - Collecte et préparation des données
   - Représentation graphique des données
   - Analyse exploratoire et avancée des données
   - Interprétation des résultats
   - Rédaction d'un court rapport de projet
   - Présentation orale du projet

Exemples de projets interdisciplinaires

Question : *Quels sont les facteurs qui influencent l'appréciation des différentes catégories de musique ?*
Données : Liste et description des musiques disponibles sur Spotify
Méthodes :
- Catégorisation des musiques selon leurs propriétés acoustiques (valence, énergie, rythme...) et leurs genres (rap, électro, classique...) afin d'identifier les grandes tendances.
- Identification des caractéristiques musicales qui peuvent prédire les préférences pour chaque catégorie de musique à l'aide d'une modélisation statistique.

Question : *Comment la représentation des femmes dans la littérature chinoise a-t-elle évolué au cours de l'histoire ?*
Données : Résumés de fictions chinoises anciennes et modernes
Méthodes :
- Comptage des occurrences de mots pour mesurer la présence et la proportion de personnages féminins et identifier les thèmes associés à la présence ou l'absence de femmes.
- Comparaison des résultats avec la littérature en anthropologie de la famille et des relations de genre pour analyser les évolutions observées.

Questions : *Comment les protagonistes de films sont-ils représentés sur les affiches ?*
Données : Collection des affiches de films diffusés aux États-Unis et en France
Méthodes :

- Extraction automatique des éléments de chaque affiche (présence des personnages, émotions des personnages, couleurs et contrastes).
- Construction de modèles statistiques pour relier ces éléments aux genres de films, et pour étudier l'évolution des affiches au cours du XXe et du début du XXIe siècle en fonction du pays.

Langue d'enseignement : Français et Anglais

Type de cours :

L'enseignement sera composé de 3 séquences par semestre de 15 heures chacune, modulé selon l'avancement de la classe. Chaque séquence sera divisée entre :

- **Cours magistraux** (30 %) : présentation des notions centrales des statistiques, de la programmation et de l'analyse de données.
- **Travaux pratiques** (35 %) : mise en application des notions vues en cours magistraux avec des cas simples, par le biais d'un langage de programmation.
- **Notions avancées** (10 %) : initiation aux enjeux actuels des sciences des données culturelles dans les industries, les institutions publiques et laboratoires de recherche, par le biais de lectures spécialisées ou d'interventions extérieures.
- **Projets interdisciplinaires** (25 %) : accompagnement à la réalisation d'un projet complet d'analyse de données historiques, littéraires ou artistiques.

La dernière séquence de chaque semestre se déroulera pendant une semaine complète.

Modalités d'évaluation :

L'évaluation portera sur les productions liées au projet interdisciplinaire. Un rendu intermédiaire sera demandé à mi-parcours de chaque séquence, suivi en fin de séquence d'un court rapport écrit accompagné d'une présentation orale du projet. Les coefficients de notations seront les suivants :

Rendu intermédiaire : 25 %
Rapport écrit : 25 %
Présentation orale : 50 %

Année : L1

Semestre : Semestre 1 et Semestre 2

Lectures obligatoires :

Lectures recommandées :

Les étudiantes et étudiants sont invités à lire la première partie (« Background ») de :
Learning Statistics with Python, Ethan Weed (2021)
Ce livre est accessible gratuitement en ligne à l'adresse suivante :
https://ethanweed.github.io/pythonbook/01.01-intro.html