Methodes Numeriques (T410006)

Tommaso Venturini

Semestre, horaire salles

Semestre de printemps
8 séances de 4h (généralement une semaine sur deux, mais avec des exceptions)
Mardi de 14h15-18h, salle M 4220

Résumé

Dans ce cours les étudiants apprendront une série de méthodes d'enquête basée sur le *repurposing* (= réutilisation à des fins de recherche) d'une variété de données, outils et techniques computationnelles propres au Web moderne et librement disponibles en ligne. Le cours introduit les étudiants aux différents avantages de ces méthodes, mais aussi à leurs nombreuses limites et contraintes.

Le cours prévoit une partie importante de travaux pratiques en classe et hors classe, sous la forme d'une série d'expérimentations collectives avec les outils et les méthodes introduits dans le cours. Les étudiants travailleront en groupe de 3 ou 4 pour tester les ressources proposées par les enseignants. Les groupes produiront enfin un « rapport en 5/10 figures » présentant les résultats de leurs l'utilisation de méthodes numériques.

Il est conseillé aux groupes de choisir un sujet qui les intéresse dans les premières séances et de réaliser les différentes expériences sur ce même sujet tout au long du cours. Les étudiants suivant le cours *Subcultures Internet* (T410004) sont encouragés à travailler dans le même groupe et exploiter les techniques introduites dans les cours de méthodes numériques pour l'exploration de leur subculture.

Structure du cours et approche pédagogique

Le cours se déroule en 7 séances de 4h (généralement une semaine sur deux).

Pendant les séances du cours, différentes méthodes/outils/techniques d'enquête seront d'abord introduites par les enseignants et puis testées par les étudiants en groupe, avec l'aide et la supervision de ces enseignants. L'expérimentation en classe constitue le début d'un travail plus approfondi sur chaque méthode que les groupes développeront entre deux séances.

Le cours avance de manière progressive et souvent les résultats de techniques introduites dans une séance deviennent la base pour le travail de la séance suivante. Il est donc essentiel que les groupes ne prennent pas du retard sur les exercices et qu'ils s'investissent diligemment dans le travail entre les séances.

Toutes les séances commencent avec une présentation du travail effectué par un ou deux des groupes en relation aux questions de la séance précédente (selon un calendrier établi aux début cours). Le groupe présente à toute la classe les résultats de son investigation et les points forts et limites de ces résultats sont discutés collectivement.

Objectifs du cours et résultats d'apprentissage

- Apprécier la richesse et la diversité des sources de données et des outils de monitoring, d'analyse et de visualisation disponibles en ligne
- Développer une réflexion critique sur les biais et les angles morts des méthodes numériques
- Devenir capable d'intégrer différentes méthodes dans un protocole de recherche complexe pour exploiter leur potentiel tout en respectant leurs limites
- Développer un imaginaire méthodologique au-delà des ressources proposées dans le cours

Programme des séances

Séance 1 (21/02/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- Mini-théorie des méthodes de recherche
- Méthodes numériques et quali-quantitatives
- Outils prêts à l'emploi
- Outils d'auto-analyse
- Formation des groupes
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Testez vos sujets en les investiguant avec un ou plusieurs outils prêts à l'emploi
- Téléchargez et analysez vos données avec au moins un des outils d'auto-analyse

Lectures conseillées

- Madsen, A. K. (2015). Between technical features and analytic capabilities: Charting a relational affordance space for digital social analytics. *Big Data and Society*, 2(1), 1–15. https://doi.org/10.1177/2053951714568727
- Boullier, D. (2015). Les sciences sociales face aux traces du big data. *Revue Française de Science Politique*, 65(5), 805. https://doi.org/10.3917/rfsp.655.0805
- D'Ignazio, C., & Klein, L. (2018). Chapter 3: What Gets Counted Counts. In *Data Feminism*. https://mitpressonpubpub.mitpress.mit.edu/pub/rykaknh1/release/4

Séance 2 (07/03/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- L'éthique des méthodes numériques
- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- L'approche pédagogique de ce cours
- L'art d'interroger les moteurs de recherche
- *Travail en groupe*
- Les bases de données (ouvertes)
- *Travail en groupe*
- Interroger un API avec une feuille de calcul (Google Sheet & Google App Script)
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Développez un projet de recherche comparative avec un ou plusieurs moteurs de recherche
- Réalisez une extraction d'un (ou plusieurs) jeu des données d'une base (par interface ou API)

Lectures conseillées

- Rogers, R. (2017). Foundations of Digital Methods: Query Design. In M. Schaefer & K. van Es (Eds.), *The Datafied Society* (pp. 75–94). Amsterdam: University Press. www.researchgate.net/publication/314868441_Foundations_of_Digital_Methods_Query_Design

Séance 3 (21/03/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- 4CAT
- *Travail en groupe*
- Scraper avec un outil avec interface
- Scraper avec une librairie de code et un notebook en ligne (Google Collab)
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Finalisez votre extractions et analyse avec 4CAT

- Finalisez votre projet de scraping avec Data Miner ou Beautiful Soup

Lectures conseillées

- Marres, N., & Weltevrede, E. (2013). Scraping the Social? Issues in real-time social research. Journal of Cultural Economy, 6(3), 313–335. https://doi.org/10.1080/17530350.2013.772070
- Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research*, 4(2), 571–589. https://doi.org/10.5117/ccr2022.2.007.hage

Séance 4 (28/03/23, 14h15-18h00, 14h-16h salle M 2150 et 16h-18h salle M 2140)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- Feuilles de calcul (formules, tableaux croisés dynamiques, jonction per recherche verticale)
- *Travail en groupe*
- Outils pour le data-wrangling (Open Refine - Trifacta)
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Nettoyez et analysez un tableau de données pertinent pour votre sujet

Lectures conseillées

- Campbell-Kelly, M. (2007). Number crunching without programming: The evolution of spreadsheet usability. *IEEE Annals of the History of Computing*, 29(3), 6–19. https://doi.org/10.1109/MAHC.2007.4338438
- Dumit, J., & Nafus, D. (2018). The other ninety per cent: Thinking with data science, creating data studies - an interview with Joseph Dumit. In *Ethnography for a data-saturated world* (pp. 252–274). https://doi.org/10.7765/9781526127600.00020
- Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., … Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288. https://doi.org/10.1177/1473871611415994

Séance 5 (04/04/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (1 groupe)*
- La visualisation de l'information en théorie (erreurs classiques, types de diagrammes, variables visuelles)
- *Travail en groupe*
- La visualisation de l'information en pratique (Raw, Flourish, D3.js)
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Réalisez un projet de visualisation de l'information en produisant au moins deux diagrammes différentes pour investiguer votre sujet/jeu de données :
  - un (ou plusieurs) visualisation statique
  - un (ou plusieurs) visualisation dynamique/interactive

Lectures conseillées

- Wainer, H. (1984). How to Display Data Badly. *The American Statistician*, 38(2), 137–147. https://doi.org/10.2307/2683253
- Behrens, J. T., & Chong-Ho, Y. (2003). Exploratory Data Analysis. In I. B. Weiner (Ed.), *Handbook of Psychology* (pp. 33–64). London: Wiley. https://doi.org/10.1002/0471264385.wei0202
- Cox, V. (2017). Exploratory Data Analysis. *In Translating Statistics to Make Decisions* (pp. 47–74). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-2256-0_3
- Drucker, J. (2011). Humanities Approaches to Graphical Display. *DHQ: Digital Humanities Quarterly*, 5(1), 1–20. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Séance 6 (18/04/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- La théorie de la visualisation des réseaux
- La pratique de la visualisation des réseaux
- *Table2Net*
- *Travail en groupe*
- Six histoires de réseau

- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Réalisez l'analyse visuelle complète d'un réseau de votre choix

Lectures conseillées

- Venturini, T., Jacomy, M., & Jensen, P. (2021). What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts. *Big Data & Society*, 8(1). https://doi.org/10.1177/20539517211018488
- Jacomy, M. (2020). Chapter 2 "What is Visual Network Analysis". In *Situating Visual Network Analysis* (pp. 51-61). Aalborg University. https://vbn.aau.dk/files/435977255/PHD_MJ_E_pdf_1_.pdf

---

Séance 7 (02/05/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- Crawling (réseaux complexes et structure du Web, attachement préférentiel et communautaire, les crawlers)
- *Travail en groupe*
- Analyse des textes (techniques de bases et outils avancés)
- *Travail en groupe*

A commencer pendant la séance et compléter avant la séance suivante :

- Obtenez un réseau d'hyperliens par des techniques de crawling
- Constituez et analysez un corpus de textes

Lectures conseillées

- Moretti, F., & Pestre, D. (2015). Bankspeak. The Language of the World Bank Reports. *New Left Review*, 9 2(March-April). https://newleftreview.org/issues/ii92/articles/franco-moretti-dominique-pestre-bankspeak
- Pincemin, B. (2020). La textométrie en question. *Le Français Moderne - Revue de Linguistique Française*, 88(1), 26–43. https://halshs.archives-ouvertes.fr/halshs-02902088
- Ooghe-Tabanou, B., Jacomy, M., Girard, P., & Plique, G. (2018). Hyperlink is not dead! *ACM International Conference Proceeding Series*, 12–18. https://doi.org/10.1145/3240431.3240434
- Jacomy, M., Girard, P., Ooghe, B., & Venturini, T. (2016). Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences. In *International AAAI Conference on Web and Social Media* (pp. 595–598). https://hal.archives-ouvertes.fr/hal-01293078/

Séance 8 (16/05/23, 14h15-18h00, salle M 4220)

Déroulement de la séance :

- *Présentation du travail effectué avec les techniques de la séance précédente (2 groupes)*
- Un « reality check » pour les méthodes numériques
- Les coté obscur du re-purposing
- *Travail en groupe*

Lectures conseillées

- Pablo Jensen. (2018). *Pourquoi la société ne se laisse pas mettre en équations*. Paris: Seuil. https://doi.org/10.4000/lectures.25155
- Gray, J., Lämmerhirt, D., & Bounegru, L. (2016). Changing what counts. *DataShift*, 52. www.thedatashift.org
- Bruno, I., Didier, E., & Vitale, T. (2014). Statactivism: forms of action between disclosure and affirmation. *The Open Journal of Sociopolitical Studies*, 2(7), 198–220. https://doi.org/10.1285/i20356609v7i2p198
- Venturini, T., & Rogers, R. (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, 7(4), pp. 532-540. https://doi.org/10.1080/21670811.2019.1591927
- Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society*, 20(11), 4195–4217. https://doi.org/10.1177/1461444818769236

Présentation travaux des groupes
23/05/23, 14h15-18h00, salle M 4220
26/05/23, 14h15-18h00, salle M 2160

Dans ces deux séances, le groupes présenterons une version (quasi)définitive des figures constituant leur rapport et recevront un dernier feedback en vue de leur finalisation et de la rédaction de leur didascalies.

Modalités d'évaluation

Les étudiant.e.s vont être évalués sur la base de la qualité et de la quantité du travail qu'elles et ils ont investi dans le cours tout au long du semestre.

La partie la plus importante de l'évaluation est basée sur un « rapport en 5/10 figures » que les groupes soumettent juste après la fin du cours. Dans ce rapport, chaque groupe doit montrer d'avoir développé une réflexion critique sur les potentialités et les limites de méthodes numériques. Le rapport est constitué d'une série de figures (entre 5 et 10) (graphiques / visualisations / diagrammes) qui peuvent être statiques ou dynamiques, simples ou composées de différents panneaux. Chaque figure sera accompagnée d'une didascalie de 1 page expliquant de manière détaillée :

- Comment les données ont été obtenus
- Comment les données ont été nettoyées, préparées et analysées
- Comment lire la figure
- Quel message/résultat peut-on trouver dans la figure

Les rapports seront évalués selon les six critères suivants :

1. Effort investi dans l'enquête (quantité et intérêt de donnés traitées, nettoyage, enrichissement…)
2. Qualité des figures (originalité et complexité de l'analyse, lisibilité, capacité de générer des conclusions…)
3. Qualité des didascalies (justesse des explication, qualité du texte, lien avec les figures…)
4. Intégration des résultats dans un récit cohérent (enchainement, comparaison, approfondissement…)
5. Originalité et intérêt des résultats
6. Maitrise avancée d'outils simples ou exploration d'outils avancés

Chaque critère est évalué sur l'échelle suivante : Mauvais (0 points) ; Insuffisant (0,25 points) ; Passable (0,5 points) ; Très bien (0,75 points) ; Excellent (1 point). L'évaluation du rapport est tout simplement la somme des notes des six critères.

Cette note peut être augmentée ou diminuée de 1 point pour les étudiant.e.s et les groupes en fonction de leur investissement dans la discussion et le travail en classe et entre les séances. Ce bonus/malus peut produire une différence entre la note individuelle et celle du groupe pour les étudiant.e.s qui auraient investi sensiblement plus ou moins de travail par rapport au reste de leur groupe.