PSL Master Course

Digital Humanities Meet Artificial Intelligence

March 29 to April 2

Home page

IMPORTANT: you need to

1. Subscribe to the **discord server**
2. **Give your preference for the projects** (see details below)

Course list:

- Course 1 (Léa Saint-Raymond): Quantitative data analysis and cartography.
- Course 2 (Mathieu Aubry): Computer vision for the humanities.
- Course 3 (Béatrice Joyeux-Prunel): History of art and AI.
- Course 4 (Jean-Baptiste Camps and Thierry Poibeau): NLP for the humanities.
- Course 5 (Ségolène Albouy and Matthieu Husson): History of astronomy and AI.

Schedule:

Monday March 29th:

- 09:00-10:30 - Course 5 (Ségolène Albouy) #1
- 11:00-12:30 - Course 2 (Mathieu Aubry) #1
- 14:00-15:30 - Course 3 (Béatrice Joyeux-Prunel) #1
- 16:00:17:30 - Projects

Tuesday:

- 09:00-10:30 - Course 3 (Béatrice Joyeux-Prunel) #2
- 11:00-12:30 - Course 4 (Jean-Baptiste Camps) #1
- 14:00-15:30 - Course 1 (Léa Saint-Raymond) #1
- 16:00:17:30 - Projects

Wednesday:

- 09:00-10:30 - Course 1 (Léa Saint-Raymond) #2
- 11:00-12:30 - Course 4 (Thierry Poibeau) #2
- 14:00-15:30 - Projects
- 16:00:17:30 - Projects

Thursday:

- 09:00-10:30 - Course 5 (Matthieu Husson) #2
- 11:00-12:30 - Course 2 (Mathieu Aubry) #2
- 14:00-15:30 - Projects
- 16:00:17:30 - Projects

Friday:

- 09:00-10:30 - Projects
- 11:00-12:30 - Projects
- 14:00-15:30 - Projects - defense
- 16:00:17:30 - Projects - defense

Projects description

Project 1: Being an artist in Toulouse in the early 20th c. (Léa Saint-Raymond, ENS-PSL and Paul JEAN <paul.jean777@gmail.com>, Ecole du Louvre)

Abstract: This project aims at understanding what it meant to be an artist in a French secondary city at the beginning of the 20th century. It will cross historical, sociological and economic approaches, while conducting an analysis of collective biographies (prosopography) and, ultimately, an analysis of the artworks.

The starting corpus will be a transcription of the catalogs of an annual Toulouse exhibition - the "Salon des artistes méridionaux" - between 1907 and 1939 (data compiled by Léa Saint-Raymond). Students will be led to work on this data through the digital humanities: analysis of the biographies of the exhibitors, of their networks, of the addresses (places of birth and places of residence) and, through an analysis of the titles, of the places they represented. Sale prices will also be questioned by econometrics. The result of this analysis will be the subject of a paper on arXiv, signed collectively.

Bibliography and resources:

- Léa Saint-Raymond, “Bordeaux vs. Paris: An Alternative Market for Local and Independent Artists?”, *Arts*, vol. 4, issue 4 : https://www.mdpi.com/2076-0752/9/4/114
- Softwares : R, QGis, Gephi, Palladio and, for the final paper, Inkscape

Project 2: Image search with deep features for art history applications (Xi Shen, xi.shen@enpc.fr)

Abstract: The project will aim at studying different approaches to searching images in an art database. More specific goals will be defined with the group, but could include looking for duplicate images, images with similar style, or images where a small detail has been copied. One of the more open and important issues that could be studied is scaling detail detection, which with current approaches is extremely time consuming.

The project will start using a relatively small (~1,6k images) curated dataset of images from Brueghel's workshop, then potentially move to larger datasets, such as a dataset from Venus' depictions (~100k images), artworks from wikipedia, or museum collection

Bibliography and resources:

- http://imagine.enpc.fr/~shenx/ArtMiner/ : Xi Shen's project (Xi will be advising the project)
- EJ Crowley, A Zisserman, In search of art (2014) In European conference on computer vision Workshops
- Seguin, B., Striolo, C., & Kaplan, F. (2016). Visual link retrieval in a database of paintings. In European conference on computer vision Workshops
- Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Ceci n’est pas une pipe: A deep convolutional network for fine-art paintings classification. In International Conference on Image Processing
- Mensink, T., & Van Gemert, J. (2014). The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*

Project 3: Visual Circulations and Globalization (Adrien Jeanrenaud <adrien.jeanrenaud@chartes.psl.eu>**)**

Abstract: The workshop proposed as part of the course aims to assess, on the basis of a worldwide corpus of illustrated printed matter from the period 1890-1950, the worldwide circulation of images and motifs in the age of printing. At stake is the extent to which images have contributed to globalisation; whether the circulation of images reflects homogenisation or, on the contrary, cultural heterogeneity. The group will work on the basis of a pre-segmented corpus by a team of students from the University of Geneva, who will have already identified the duplicates in circulation on this corpus. The French team will detect the circulation of certain patterns more than others, and will propose spatio-temporal visualizations of these circulations. If time permits, using statistical methods and data visualization, it will propose to isolate the factors of the circulation of certain images and patterns rather than others. Depending on the results obtained by the group, and from a critical perspective, a prediction experiment could then be tested: predicting a potentially (statistically) "successful" "global" image, which will be launched on social networks.

Project 4: *Come together…* Stylometry revisits the Lennon vs McCartney Debate

(Jean-Baptiste.Camps@chartes.psl.eu, antoinedesacy@gmail.com)

Abstract: The goal of this project is to explore authorship and attribution questions in the corpus of the Beatles songs, with a special focus on the Lennon vs McCartney debate. The idea is both to explore attribution on a per song basis, as well as possible shared authorship. For this, we will analyse linguistic features in the text of the songs, and perhaps in the musical score too.

One of the focus will be on the individual style vs synergy hypothesis. Without ruling out any hypothesis and without having any a priori on the data, we will try to analyze to what extent the songs written by Lennon and McCartney are the result of a four-handed writing process. Is it a true collaborative creative process? Is one of the two lyricists more in a corrective posture? By comparing the songs written exclusively by one or the other and those said to be written by both, we will try to see if it is possible to assign more specific roles to Lennon or McCartney in the writing of these songs and to confront our results with the scientific hypotheses produced by researchers on the subject.

The project will go through all phases of a stylometry endeavour, from corpus building, to model training and discussion of the results.

Topics

0. (install party) Installing and setting up a working environment (Notebook, Github);
1. Data collection, clean-up and preprocessing;
2. Extracting stylometric features (tokenisation, n-grams, etc.);
3. Descriptive and exploratory visualisations ;
4. Data normalisations;
5. Supervised analysis : training and testing (SVM's…) ; leave-one-out
6. Supervised analysis : rolling stylometry ; deep learning;
7. Non stylistic features for cross-validation: named entities in the songs.

The exact content of the final sessions will be adapted according to the participants' appetites and skills, as well as the time remaining.

Bibliography and resources:

- Glickman, Mark, et al. '(A) Data in the Life: Authorship Attribution in Lennon-McCartney Songs'. *Harvard Data Science Review*, vol. 1, no. 1, PubPub, July 2019. *hdsr.mitpress.mit.edu*, doi:10.1162/99608f92.130f856e.
- Karsdorp, Folgert, et al. *Humanities Data Analysis: Case Studies with Python*. Princeton University Press, 2021.
- Eder, Maciej. 'Rolling Stylometry'. *Digital Scholarship in the Humanities*, vol. 31, no. 3, Oxford University Press, 2016, pp. 457–469.
- Stamatatos, E. 'A Survey of Modern Authorship Attribution Methods'. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538–556, doi:10.1002/asi.21001.
- Compton, Todd (2017). *Who Wrote the Beatle Songs? A History of Lennon-McCartney*. San Jose: Pahreah Press. ISBN 978-0-9988997-0-1.
- Rowley, David (2008). *Help! 50 Songwriting, Recording and Career Tips used by the Beatles*. Matador. ISBN 978-1-906221-37-9.
- Jere Xu, jerrytigerxu/Beatles-NLP, Github.com, 2020.
- « Liste des chansons des Beatles », *Wikipedia francophone*, https://fr.wikipedia.org/wiki/Liste_des_chansons_des_Beatles.

Project 5: Natively digital critical edition of astronomical tables with HTR (Tristan Dot, tristan.dot@ens-paris-saclay.fr)

This project will focus on the handwritten numbers recognition in order to produce critical editions assisted with computer vision. The dataset on which this project will be based is a corpus of IIIF scans of manuscripts from the alphonsine tradition (14th and 15th centuries) featuring astronomical tables. Astronomical tables are large "spreadsheets" of numbers that allowed medieval astronomers to calculate various celestial phenomena such as the position of the stars or the occurrence of eclipses. The study of astronomical tables allows contemporary researchers to better understand the evolution of computation techniques, as well as the representation of the cosmos in the Medieval period.

Students will be tasked with producing critical editions of a set of tables on the true position of the Moon. Based on several tables describing the same calculation, they will be able to compile their transcriptions using the automatic critical edition program CATE and integrate it into the DISHAS information system. While very efficient and informative, this process also unifies and standardizes some key features of the original sources, for instance with respect to their graphical organisation. A critical reflection on these aspects will also be important especially as they might provide explanation for variants among sources and influence the way in which the tables could be concretely used for calculations.

In order to carry out these tasks, students will be able to rely on existing tools in the DISHAS platform, as well as on different proof-of-concept algorithms able to segment tabular grid or transcribe their numerical content. Improving these models performance and robustness, integrating them into a pipeline, while reflecting on their potential and limitation for the humanities will be an important part of the project.

Bibliography and resources:

- J. Chabás. *Computational Astronomy in the Middle Ages: Sets of Astronomical Tables in Latin*. Madrid: Consejo Superior de Investigaciones Científicas (2019), esp. 199-206

- M. Husson, “Work cohesion as a test of Manuscript Transmission: The case of John of Lignères *Tabule Magne*”, in Kremer, Husson, Chabás (eds.), *Alfonsine Astronomy: The Written Record*, Brepols (forthcoming, pre-print available on demand)
- M. Husson, “Computing with Manuscripts:Time between mean and true syzygies in John of Lignères *Tabule Magne*”, in Husson, van Dalen, Montelle (eds.), *Editing and Analysing Astronomical Tables,* Brepols (forthcoming, pre-print available on demand)
- http://imagine.enpc.fr/~monniert/docExtractor/ (Tom Monnier’s project)
- L. Gao *et al.*, "ICDAR 2019 Competition on Table Detection and Recognition (cTDaR)," *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, NSW, Australia, 2019, pp. 1510-1515, doi: 10.1109/ICDAR.2019.00243.
- Paliwal, Shubham & D, Vishwanath & Rahul, Rohit & Sharma, Monika & Vig, Lovekesh. (2019). TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images. 10.1109/ICDAR.2019.00029.
- S. Schreiber, S. Agne, I. Wolf, A. Dengel and S. Ahmed, "DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japan, 2017, pp. 1162-1167, doi: 10.1109/ICDAR.2017.192.
- Raja, Sachin & Mondal, Ajoy & Jawahar, C.. (2020). Table Structure Recognition using Top-Down and Bottom-Up Cues.
- Shi, Baoguang & Bai, Xiang & Yao, Cong. (2015). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2016.2646371.

Projet assignments:

First name	Last name	Adresse e-mail	Project assigned
~~Breno~~	~~BALDAS SKUK~~	~~breno.skuk@gmail.com~~	~~1~~
Haythem	Hassayoun	haythemhassayoun@gmail.com	1
Julien	Sentuc	julien.sentuc@dauphine.eu	1
Mathurin	VIDEAU	mathurin.videau@dauphine.eu	1
Pavel	Soriano	pavel.soriano@data.gouv.fr	1
Solène	AMICE	solene.amice@univ-paris1.fr	1
Danyl	Hadjali	danyl.hadjali@ens.fr	1
Yi	Zhao	zhaoeasy@gmail.com	1
Christian	Alaka	christian.alaka@dauphine.eu	2
Pierre	Fernandez	pierre.fernandez@polytechnique.edu	2
Lucas	Gnecco	lucas.gnecco@dauphine.eu	2
~~Anna~~	~~Kukleva~~	~~akukleva@mpi-inf.mpg.de~~	~~2~~
Louis	Leprince	louislpr@gmail.com	2

Danae	Di Salvo	dsd39@protonmail.com	2
Hugo	Scheithauer	hugo.scheithauer@chartes.psl.eu	2
Divya	mathur	divmat010@gmail.com	2
Blaise	Delattre	bldelattre@gmail.com	3
Nathan	Godey	nathan.godey@eleves.enpc.fr	3
Maxime	Reynouard	maxime.reynouard@dauphine.eu	3
Hugo	Sonnery	Hugo.sonnery@student-cs.fr	3
raphael	clouet	raphael_clouet@hotmail.fr	3
~~Alba~~	~~Irollo~~	~~alba.irollo@europeana.eu~~	~~3~~
Krister	Kruusmaa	krister.kruusmaa@chartes.psl.eu	3
~~Luca~~	~~Bernasconi~~	~~ljbernasconi@gmail.com~~	~~3~~
Diana	OSPINA	diana.ospina@chartes.psl.eu	4
Mathieu	Molina	mathieu.molina.1212@gmail.com	4
Sam	Perochon	sam.perochon@ens-paris-saclay.fr	4
Théo	PEROCHON	theo.perochon@etu.emse.fr	4
Matthieu	Serfaty	matthieu.serfaty@dauphine.eu	4
Victor	Rambaud	victor.rambaud@gmail.com	4
Emilien	ARNAUD	emilien.arnaud@chartes.psl.eu	4
betul	kaya	betul.kaya1@gmail.com	4
Gereltuya	Bayanmunkh	qerelt@gmail.com	5
Guillaume	Cayeux	guillaume.cayeux@dgcl.fr	5
Eloi	Massoulié	eloi.massoulie@dauphine.eu	5
Laura	Lavezza	laura.lavezza17@gmail.com	5
Doriane	Hare	doriane.hare@chartes.psl.eu	5
Adrien	Jeanrenaud	adrien.jeanrenaud@chartes.psl.eu	5
Jules	Rostand	jules.rostand@outlook.com	5

Tiziano Barbari <tiziano.barbari@studenti.unipd.it> - Project 3