Des notebooks Jupyter pour les SHS

Pourquoi ? Comment ? Pour quel résultat ? Et quelle suite ?

Émilien Schultz (médialab/SESSTIM) - Antoine Blanchard (Datactivist) - Mathieu Morey (Datactivist)

Contexte général : l'arrivée des notebooks en SHS

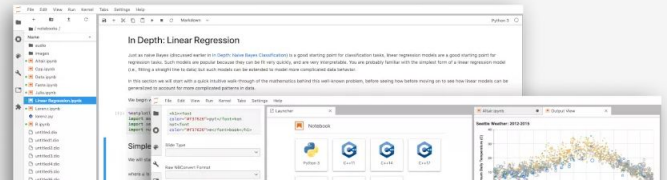
- Les notebooks / écriture exécutable comme nouvel objet de la science ouverte
- Prospective **Huma-Num Lab** (Stéphane Pouyllau ; Nicolas Sauret ; Mélanie Bunel) sur le déploiement d'un service Jupyter Lab + friends
  - Développer des démonstrateurs de Notebooks "Machine Learning"
- Partenariat Datactivist (Antoine Blanchard ; Mathieu Morey) et Émilien Schultz pour définir le besoin/réaliser les notebooks
- **TL;DR :**
  - Des notebooks Jupyter produits sur IO : https://gitlab.huma-num.fr/io
  - Une plateforme CALLISTO arrêtée : https://hnlab.huma-num.fr/blog/2022/09/26/arret-de-Callisto/
  - La suite : (r)assembler des notebooks dans Bibl-io

"L'écriture exécutable vient en effet ouvrir un espace de collaboration d'un type nouveau où s'articulent écriture discursive et écriture programmative. Un tel espace collaboratif permet en fait à des personnes aux compétences complémentaires de véritablement travailler ensemble. Une fois couplés à des répertoires *git*, les notebooks s'inscrivent alors dans un véritable écosystème collaboratif, profitant à la fois d'un espace commun d'écriture (le notebook), d'un espace collaboratif de stockage et de versionning du notebook (git), et d'un espace reproductible d'exécution du notebook." https://hnlab.huma-num.fr/blog/2021/03/23/lancement-du-groupe-de-travail-callisto/

(Parenthèse : un Notebook Jupyter, c’est quoi ?)

Tout le monde a déjà vu un notebook jupyter ? Notion de *literate programming*

Free software, open standards, and web services for interactive computing across all programming languages



JupyterLab: A Next-Generation Notebook Interface

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

Try it in your browser | Install JupyterLab

Pas spécifique à Python…

Petite chronologie

- **2020** : prospective à Huma-Num Lab sur le Deep Learning
  - conférence "Eléments de réflexion sur les enjeux du deep learning en SHS" par N. Sauret et S. Pouyllau https://edunumrech.hypotheses.org/3203
  - projets ModOAP et BaOIA
- **mars 2021** : démonstrateur CALLISTO
  - groupe de travail et lancement du Jupyter Hub
- **fin 2021** : consultation "Création et développement de notebooks Jupyter et de matériaux pour le deep learning en Sciences Humaines et Sociales (SHS)"
- **janvier 2022** : début du travail de Datactivist
- **septembre 2022** : fin de CALLISTO
- **octobre 2022** : fin de la mission

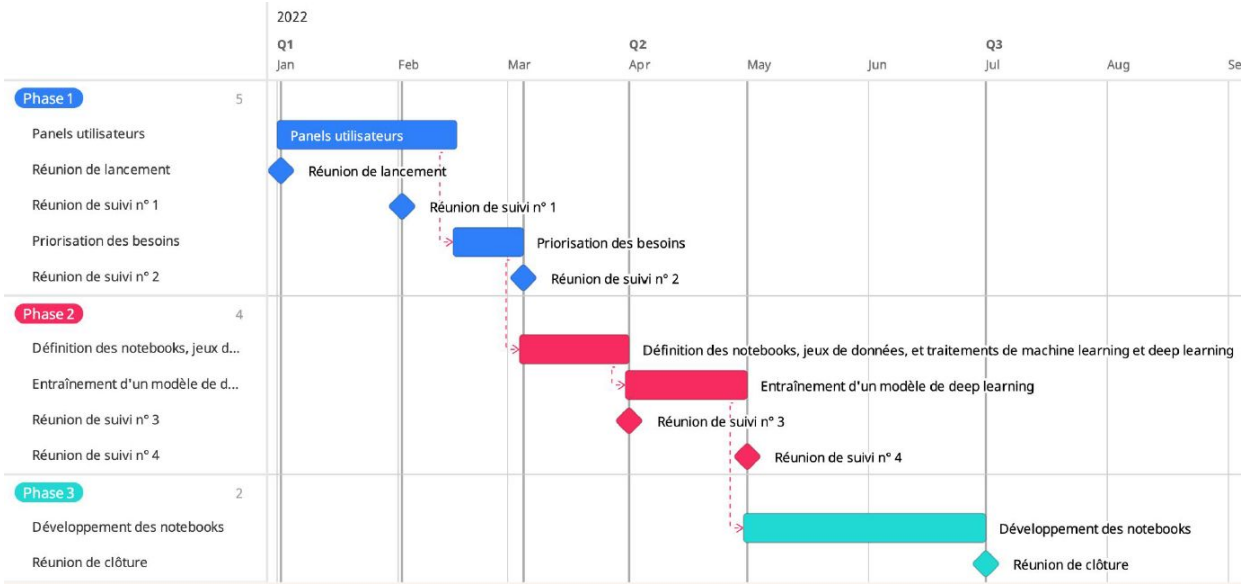
Une réponse contextualisée au contact de la pratique

DATACTIVIST

TGIR Huma-Num

Création et développement de notebooks Jupyter et de matériaux pour le *deep learning* en SHS

Offre en date du 29 novembre 2021

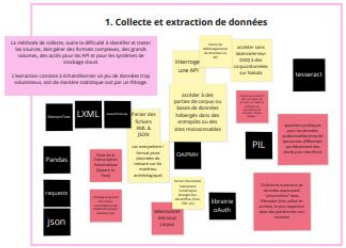


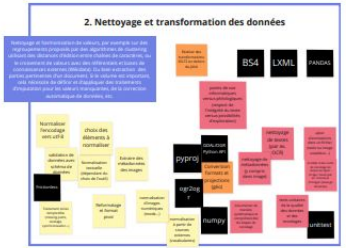
Objectif implicite : permettre le développement des notebooks

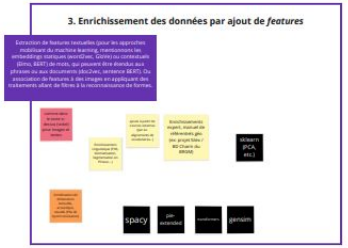
Consultation des usagers : qui utilise des notebooks ?

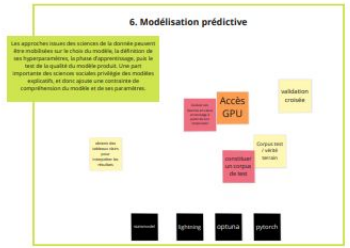
- Un public (encore) mal défini
- Deux grands usages identifiés :
  - Recherche
  - Enseignement
- Constitution de panels : produire une représentativité ?
  - travail en amont de cadrage (grandes thématiques)
  - degré d'usage, de centralité dans la communauté, balance disciplinaire, balance ingénieur/usagers, parité, appartenance institutionnelle
  - trame de la consultation (toute l'ingénierie de Datactivist : gather.town, miro)

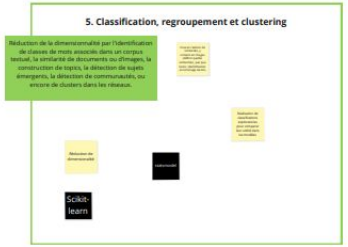
Les notebooks : un support pour une diversité de tâches

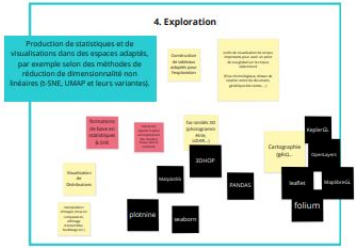


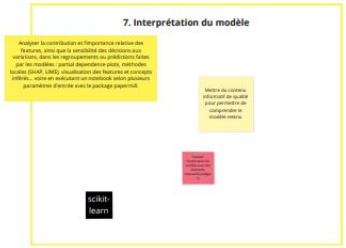


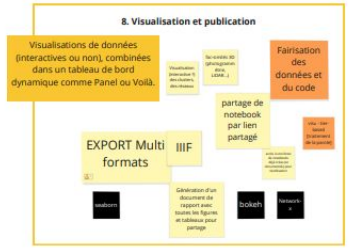


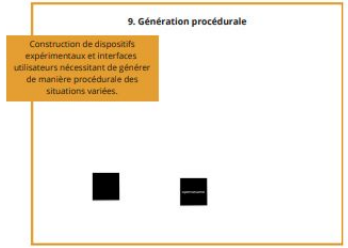












Les opérations qui doivent trouver leurs place

57 besoins priorisés

- **Priorité 1 :** 34 besoins, dont **22 traitements**
- Priorité 2 : 14 besoins
- Priorité 3 : 9 besoins

- Se connecter à des sources de données
- Convertir l'encodage vers UTF-8
- Enrichissement à partir de sources externes
- Construire des représentations adaptées pour l'exploration de données
- Ajout de *features* aux images et textes
- Ajout d'annotations
- Constituer un corpus de test
- Valider des données avec un schéma de données
- Parser des documents numériques
- Reformater des données
- Faciliter l'exploration du modèle par l'interactivité
- Normaliser et nettoyer des données texte
- Parser des fichiers XML et JSON
- Enrichissement linguistique
- Calculer des similarités et faire du clustering
- Convertir des formats
- Expliquer les choix de traitement et le raisonnement scientifique en complément du code
- Tests intégrés de la qualité des données et des recodages
- Conversion des formats de données (incl. projections) géographiques
- Représentation cartographique de données géographiques
- Manipuler des images
- Visualiser des clusters, réseaux… de façon interactive

DATACTIVIST

Des leçons apprises

- Les IE-IR sont un public naturel d'Huma-Num, et de Callisto.
- Toucher les chercheurs et enseignants, et tout le public qui ne connaît pas les notebooks, demandera un effort supplémentaire.
- Les notebooks sont utiles pour apprendre la recherche (étudiants), plus que pour faire de la recherche.
- Il n'y a pas de consensus sur la modularité ou non des notebooks.
- Il y a un gros besoin de notebooks accessibles à un public débutant, et peu de besoins de traitements très avancés (DL / ML).
- Jupyter Lab / Hub et ses extensions proposent de nombreuses fonctionnalités avancées mal connues (cf. aussi leur roadmap) : il faudra être proactif sur les mises à jour, communiquer dessus et "animer" la communauté pour assurer les usages.

Identification de différents positionnements des notebooks

1. un outil pédagogique et facile d'emploi de formation aux notebooks, utilisé par des formateurs avec un public qui apprend la programmation
2. un lieu d'hybridation entre des méthodes de recherche existantes et l'outil notebook, avec un fonctionnement de type bac à sable
3. une collection de traitements de données ancrées dans la recherche et inclus dans un cycle de vie qui va jusqu'à la publication académique, permettant le dépôt, le partage, et la réutilisation
4. un outil de découverte de traitements de données très avancés et encore peu utilisés, de type ML/DL
5. une boîte à outils, ne correspondant pas à des cas d'usages bien définis, pour le "garage" des chercheurs

Au travail ! Mais c’est quoi un bon démonstrateur ?

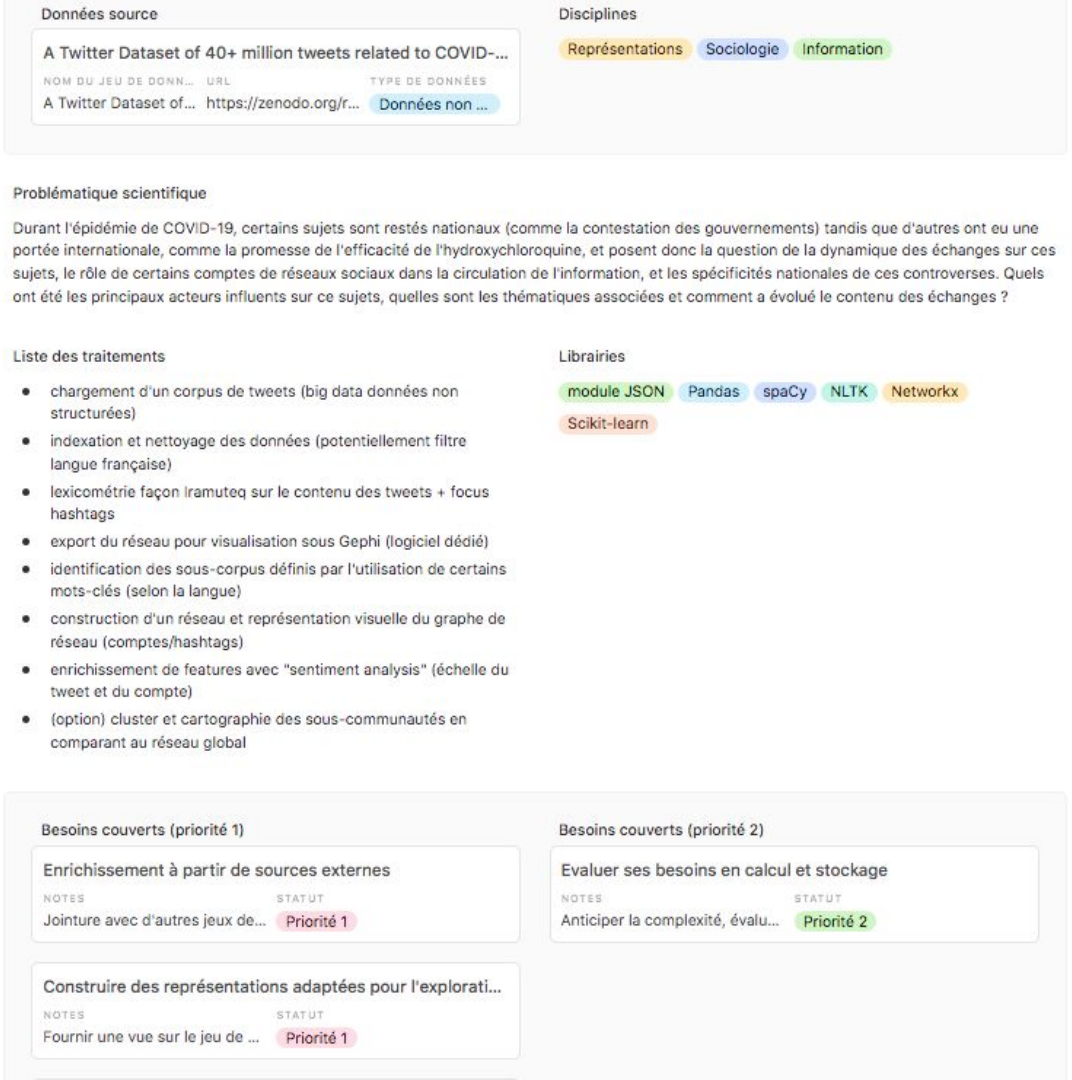
Des questions auxquelles nous avons dû répondre :

- Accessibilité
- Degré d’explicitation
- Degré de reproductibilité (matériel, etc.)
- Compacité de l’analyse
- Ouverture disciplinaire
- Eviter l’artificialité des questions/données
- Des “paillettes” mais réutilisables : modularité

Seule l’épreuve de la réutilisation nous le dira (des idées ?)

Définir 5 notebooks

- Contraintes
  - jeu de données : contraintes d'ouverture, d'accès, etc.
  - couvrir les besoins
  - couvrir différents niveaux
- 5 notebooks :
  - Données d'enquête
  - Données INSEE
  - Données Twitter
  - Apprentissage & Grand Débat
  - Images Gallica



Données source

A Twitter Dataset of 40+ million tweets related to COVID-...

NOM DU JEU DE DONN...	URL	TYPE DE DONNÉES
A Twitter Dataset of...	https://zenodo.org/r...	Données non ...

Disciplines

Représentations | Sociologie | Information

Problématique scientifique

Durant l'épidémie de COVID-19, certains sujets sont restés nationaux (comme la contestation des gouvernements) tandis que d'autres ont eu une portée internationale, comme la promesse de l'efficacité de l'hydroxychloroquine, et posent donc la question de la dynamique des échanges sur ces sujets, le rôle de certains comptes de réseaux sociaux dans la circulation de l'information, et les spécificités nationales de ces controverses. Quels ont été les principaux acteurs influents sur ce sujets, quelles sont les thématiques associées et comment a évolué le contenu des échanges ?

Liste des traitements

- chargement d'un corpus de tweets (big data données non structurées)
- indexation et nettoyage des données (potentiellement filtre langue française)
- lexicométrie façon Iramuteq sur le contenu des tweets + focus hashtags
- export du réseau pour visualisation sous Gephi (logiciel dédié)
- identification des sous-corpus définis par l'utilisation de certains mots-clés (selon la langue)
- construction d'un réseau et représentation visuelle du graphe de réseau (comptes/hashtags)
- enrichissement de features avec "sentiment analysis" (échelle du tweet et du compte)
- (option) cluster et cartographie des sous-communautés en comparant au réseau global

Librairies

module JSON | Pandas | spaCy | NLTK | Networkx | Scikit-learn

Besoins couverts (priorité 1)

Enrichissement à partir de sources externes

NOTES	STATUT
Jointure avec d'autres jeux de...	Priorité 1

Construire des représentations adaptées pour l'explorati...

NOTES	STATUT
Fournir une vue sur le jeu de ...	Priorité 1

Besoins couverts (priorité 2)

Evaluer ses besoins en calcul et stockage

NOTES	STATUT
Anticiper la complexité, évalu...	Priorité 2

Notebooks

Enquête par questionnaire sur les pratiques numériques des ...

Nom du notebook

Enquête par questionnaire sur les pratiques numériques des chercheurs

Données source

Enquête état de la science ouverte en France | https://ze...

NOM DU JEU DE DONN...	URL	TYPE DE DONNÉES
Enquête état de la s...	https://zenodo.org/r...	Données tabu...

Disciplines

Sociologie · Épistémologie et méthodes · Études des sciences

Problématique scientifique

La numérisation de la recherche transforme de manière différenciée les disciplines et les chercheurs. L'enquête "Etat de la science ouverte en France" a conduit un questionnaire auprès d'un échantillon de chercheurs qui détaillent leurs pratiques de science ouverte, notamment les outils numériques utilisés. Une partie des champs de réponse sont des champs libres qui nécessitent des stratégies de recodage pour être analysés, et identifier les profils des chercheurs.

Liste des traitements

- Chargement de données CSV
- Exploration des variables et nettoyage de catégories utilisant des heuristiques de ML
- Statistiques exploratoires et inférentielles
- Production de visualisations publiables

Librairies

Pandas · Statsmodels · Matplotlib · Seaborn · Bokeh · PySHS

Besoins couverts (priorité 1)

Expliquer les choix de traitement et le raisonnement scie...

NOTES	STATUT
Programmation lettrée	Priorité 1

Reformater des données

NOTES	STATUT
En évaluant la qualité et la str...	Priorité 1

Besoins couverts (priorité 2)

Générer un rapport contenant toutes les figures et tablea...

NOTES	STATUT
Cf. cheminement pédagogiqu...	Priorité 2

Réduire la dimensionnalité des données

NOTES	STATUT
Réduction de la dimensionnali...	Priorité 2

Notebooks

Tweets Covid

Nom du notebook

Tweets Covid

Données source

A Twitter Dataset of 40+ million tweets related to COVID-...

NOM DU JEU DE DONN...	URL	TYPE DE DONNÉES
A Twitter Dataset of...	https://zenodo.org/r...	Données non ...

Disciplines

Représentations · Sociologie · Information

Problématique scientifique

Durant l'épidémie de COVID-19, certains sujets sont restés nationaux (comme la contestation des gouvernements) tandis que d'autres ont eu une portée internationale, comme la promesse de l'efficacité de l'hydroxychloroquine, et posent donc la question de la dynamique des échanges sur ces sujets, le rôle de certains comptes de réseaux sociaux dans la circulation de l'information, et les spécificités nationales de ces controverses. Quels ont été les principaux acteurs influents sur ce sujets, quelles sont les thématiques associées et comment a évolué le contenu des échanges ?

Liste des traitements

- chargement d'un corpus de tweets (big data données non structurées)
- indexation et nettoyage des données (potentiellement filtre langue française)
- lexicométrie façon Iramuteq sur le contenu des tweets + focus hashtags
- export du réseau pour visualisation sous Gephi (logiciel dédié)
- identification des sous-corpus définis par l'utilisation de certains mots-clés (selon la langue)
- construction d'un réseau et représentation visuelle du graphe de réseau (comptes/hashtags)
- enrichissement de features avec "sentiment analysis" (échelle du tweet et du compte)
- (option) cluster et cartographie des sous-communautés en comparant au réseau global

Librairies

module JSON · Pandas · spaCy · NLTK · Networkx · Scikit-learn

Besoins couverts (priorité 1)

Enrichissement à partir de sources externes

NOTES	STATUT
Jointure avec d'autres jeux de...	Priorité 1

Construire des représentations adaptées pour l'explorati...

Besoins couverts (priorité 2)

Evaluer ses besoins en calcul et stockage

NOTES	STATUT
Anticiper la complexité, évalu...	Priorité 2

Notebooks

Les références aux lieux dans le Grand débat - l'apport de l'a...

Nom du notebook

Les références aux lieux dans le Grand débat - l'apport de l'apprentissage automatique

Données source

Données du Grand Débat en France | https://www.data.g...

NOM DU JEU DE DONN...	URL	TYPE DE DONNÉES
Données du Grand ...	https://www.data.g...	Données non ...

Disciplines

Sociologie · Études urbaines · Géographie · Histoire · Études du politique · Économie · Droit

Problématique scientifique

Le Grand Débat commandé par le Président Macron en 2019 a été construit comme une consultation nationale centralisée pour couvrir l'ensemble des revendications et analyses des Français. Ces contributions comportent des mentions de lieux, proches ou lointains, qui ancrent les prises de position et les arguments dans l'espace. La détection de ces toponymes, augmentée par les outils issus de l'apprentissage automatique et du traitement du langage, permet d'apporter un ancrage géographique utile aux travaux de différentes disciplines.

Liste des traitements

- chargement de corpus en JSON
- application de modèles spaCy "off the shelf"
- annotation de corpus
- apprentissage de modèle spaCy
- évaluation de modèle
- détection de toponymes
- visualisations finalisées (cartes & distributions)

Librairies

spaCy · Pandas · Matplotlib · module JSON

Besoins couverts (priorité 1)

Ajout d'annotations

NOTES	STATUT
Interventions manuelles dans ...	Priorité 1

Normaliser et nettoyer des données texte

NOTES	STATUT
Cf. cheminement pédagogiqu...	Priorité 1

Représentation cartographique de données géographiques

NOTES	STATUT
HNL: priorité à débattre pour ...	Priorité 1

Besoins couverts (priorité 2)

Réduire la dimensionnalité des données

NOTES	STATUT
Réduction de la dimensionnali...	Priorité 2

Validation (croisée) d'un modèle

NOTES	STATUT
Dans une démarche de modél...	Priorité 2

Notebooks

Evolution des iconographies dans les ouvrages de cuisine (G...

Nom du notebook

Evolution des iconographies dans les ouvrages de cuisine (Gallica)

Données source

Gallica - série de revue historique de cuisine | https://gall...

NOM DU JEU DE DONN...	URL	TYPE DE DONNÉES
Gallica - série de re...	https://gallica.bnf.fr...	Données non ...

Disciplines

Histoire · Épistémologie et méthodes · Langage

Problématique scientifique

La transformation de l'iconographie est intrinsèquement liée aux évolutions culturelles et techniques, mais aussi épistémiques selon les manières de représenter la réalité. Les ouvrages de cuisine moderne font largement usage d'iconographie pour illustrer les recettes. La question se pose de l'intégration progressive de cette iconographie à travers les périodes. Pour mesurer ces évolutions sur de grands corpus par exemple avec un indicateur ratio texte/image, il est nécessaire d'identifier, d'extraire et d'analyser des images dans une page. Ces questions ont été abordées dans le projet BaOIA dont s'inspire ce notebook.

Liste des traitements

- Interfacer avec Gallica pour récupérer les pages de livres/numéros de revue
- Feuilletage de collection IIIF
- Test de différents modèles pré-entraînés de LayoutParser pour identifier la meilleure stratégie d'extraction
- Traitement d'ensemble du corpus avec le modèle choisi
- Calcul des features ratio image/texte et type d'images (densité de noir / couleur, etc.)
- Représentation de l'évolution de l'iconographie

Librairies

Requests · Pandas · LayoutParser · OpenCV · module JSON

Besoins couverts (priorité 1)

Manipuler des images

NOTES	STATUT
Comparaison, affichage d'ens...	Priorité 1

Construire des représentations adaptées pour l'explorati...

NOTES	STATUT
Fournir une vue sur le jeu de ...	Priorité 1

Besoins couverts (priorité 2)

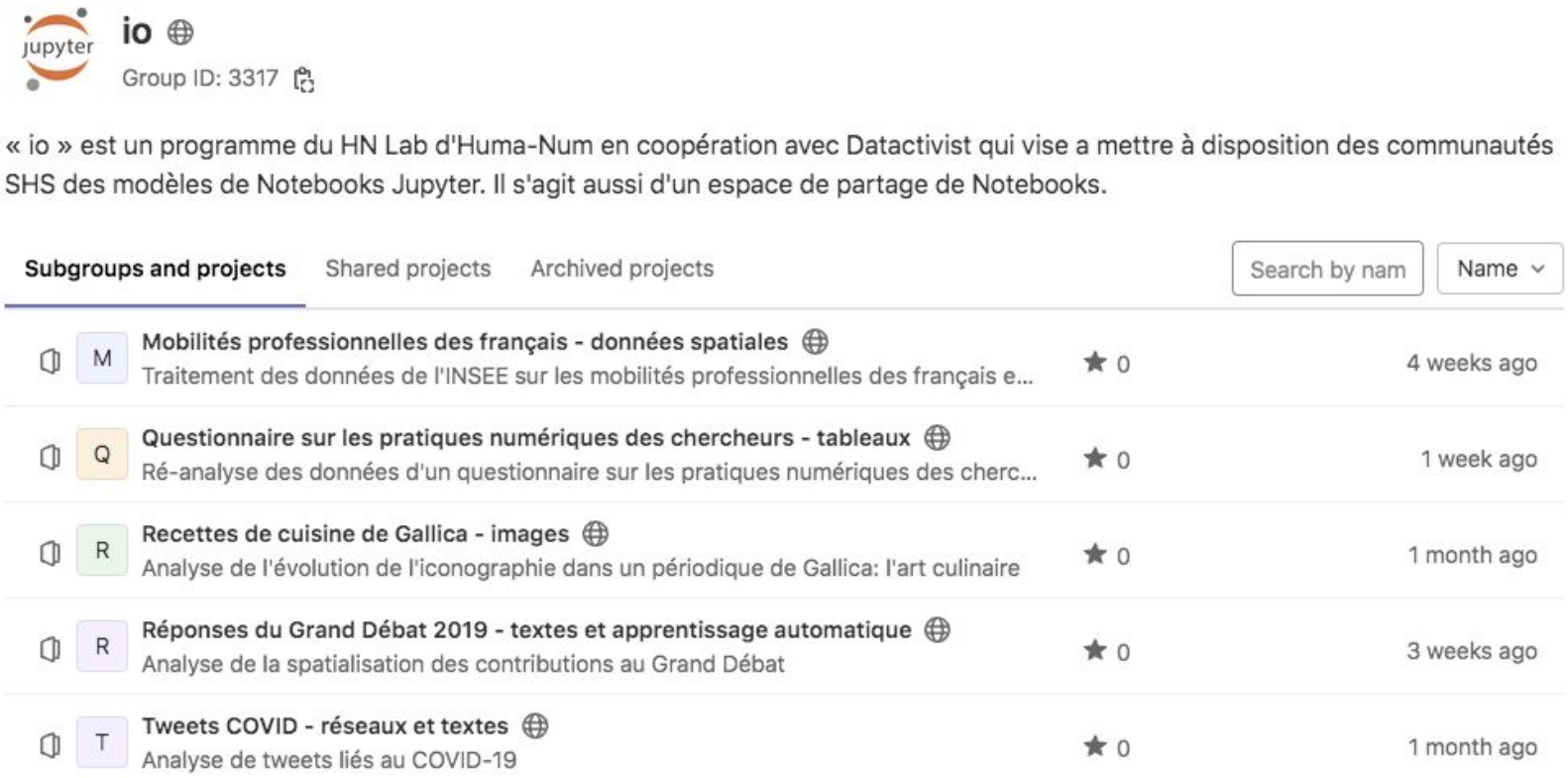
Evaluer ses besoins en calcul et stockage

NOTES	STATUT
Anticiper la complexité, évalu...	Priorité 2

Séparer clairement l'exploration et l'analyse définitive

NOTES	STATUT
Faire cette distinction dans le ...	Priorité 2

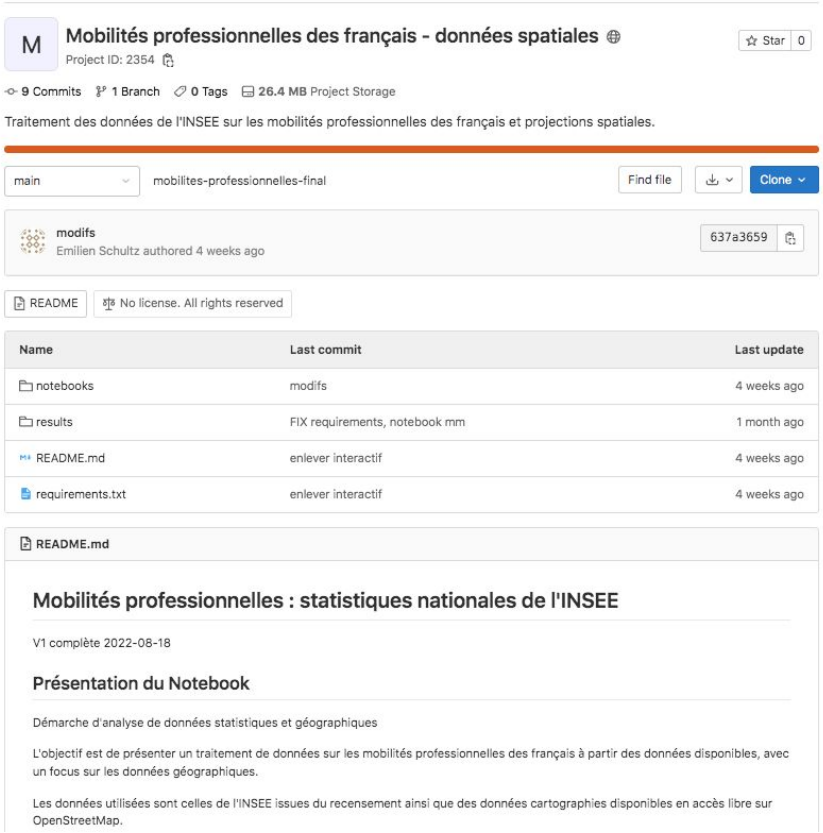
Quelques aller/retours plus tard…



Regardons en détail

- Un dépôt par notebook
- Une philosophie
  - Notebook construit selon un modèle narratif linéaire exécutable ;
  - Priorité donnée à la lisibilité en détaillant les étapes (forte verbosité ; une cellule par étape) ;
  - Application des bonnes pratiques de programmation en langage Python 🐍 (structure avec Cookiecutter Data Science et Black pour la mise en forme du code) ;
  - Séparation données/traitement (entre les données brutes et le code, et dans le code) ;
  - Recherche d'un équilibre entre minimisation du nombre de bibliothèques et facilité d'usage ;
  - Documentation du code respectant les standards (numpydoc docstrings utilisé notamment par scikit-learn);
  - Mention explicite des versions de bibliothèques utilisées ;

https://gitlab.huma-num.fr/io/mobilites-professionnelles-final



Quelques constats de fin de projet

- Distinguer les formats :
  - Notebooks, papiers exécutable, support pédagogique
  - Un notebook utile doit être pédagogique
- Des usages différenciés
  - Chercheurs & ITA : pas forcément le même combat
- Au-delà du notebook : place dans le workflow, licences, versions, etc.
  - Pratiques encore peu stabilisées
- Sur le “Machine learning”
  - Une brique parmi d’autres
- Sur le “Big data”
  - Enjeu de stockage et de manipulation difficile dans le notebook
- Sur les langages :
  - Aller au-delà de Python ?
- Sur la suite de Callisto
  - Une nécessité !
  - Des questions : GPU, noyaux, stockage, etc.
- Des problèmes pour le futur :
  - La maintenance des notebooks & leurs compatibilités

Construire vers l'avant : #Bibl-io

- Constat ancien : rassembler une galerie d'exemples de différents niveaux
- Convergence avec CoCoPySHS
- Partir de l'initiative d'Huma-Num Lab : https://gitlab.huma-num.fr/io
- Etape 1 :
  - Améliorer les notebooks existants
  - Ajouter vos notebooks de recherche "exemplaire"
  - Construire des notebooks dédiés dans le cadre de projets Science Ouverte
- Etape 2 :
  - Construire un index et des mots-clés adaptés
  - Adosser à une plateforme d'exécution
- Des questions ouvertes:
  - Polarité enseignement/recherche
  - Python, R, …