GenderedNews

Developers : AGUIAR Mathilde - HAJJI Oumaima - SIDIBE Rokiatou dite Rose

Project owners : RICHARD Ange - PORTET François - BASTIN Gilles

Plan :

I. A reminder of the project
II. Old state of the projet
III. Technical contributions
IV. Project management
V. Conclusion

The Team

Mathilde Aguiar

Head of projet

Oumaima Hajji

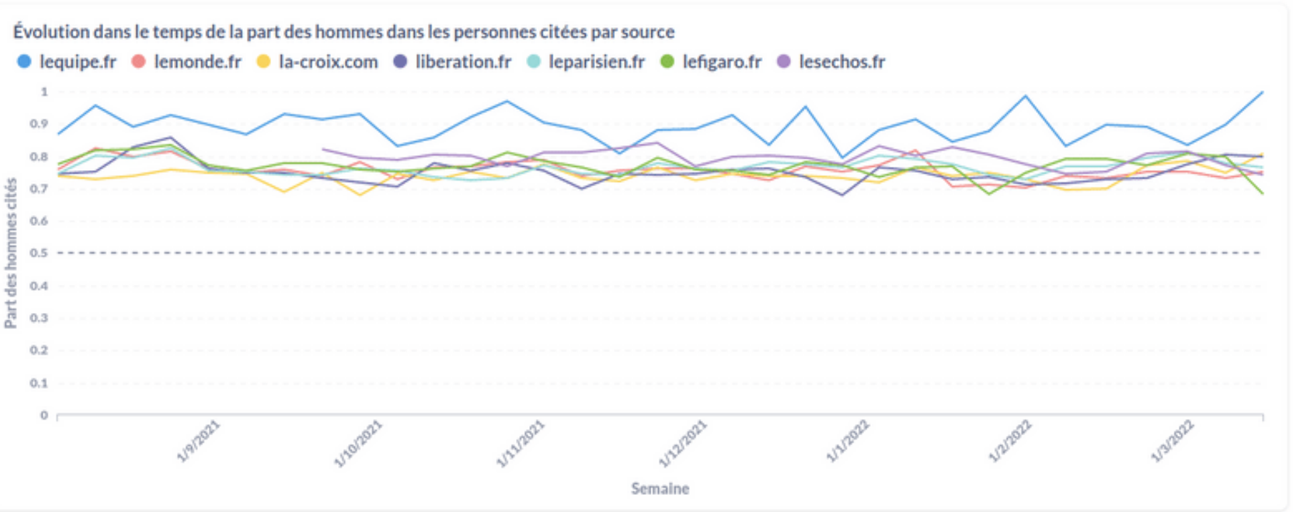
SCRUM Master

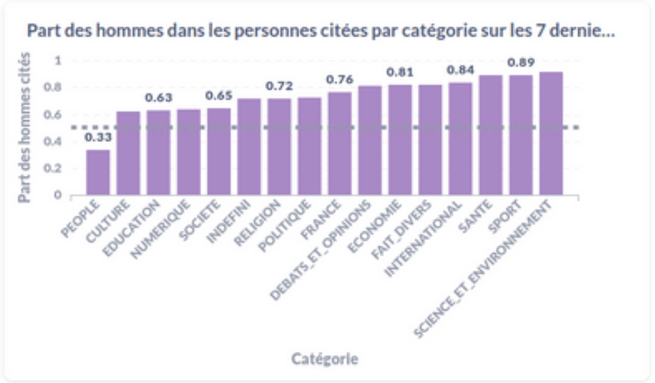
Rokiatou dite Rose Sidibe

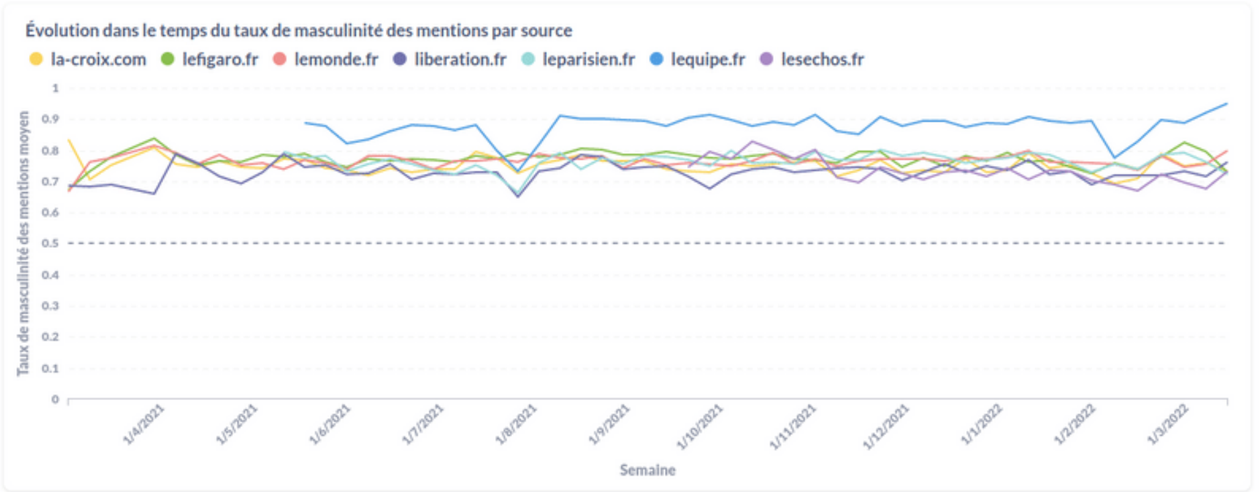
Developer

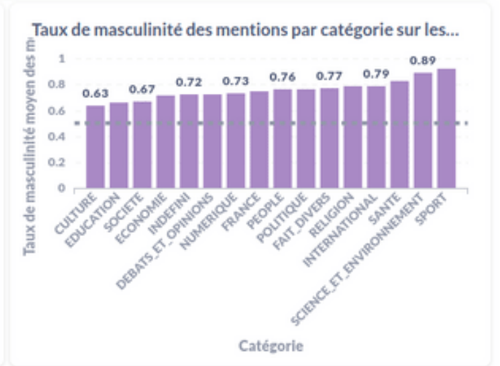
Current website

Mentions









Quotes

Mention versus Quote

Mentions

The number of times the name of a person is used by an other person/journalist within the article

Quotes

The sayings reported, most often by a journalist, that concern a person. They could be written within double quotes (") or introduced by certain expressions such as: "according to, by, etc."

Context and subject

Context

Time allowed for women in medias is still unequal compared to men, lack of women representation in the medias.

Subject

GenderedNews : a Web site that collects masculinity rates calculated from French written medias.

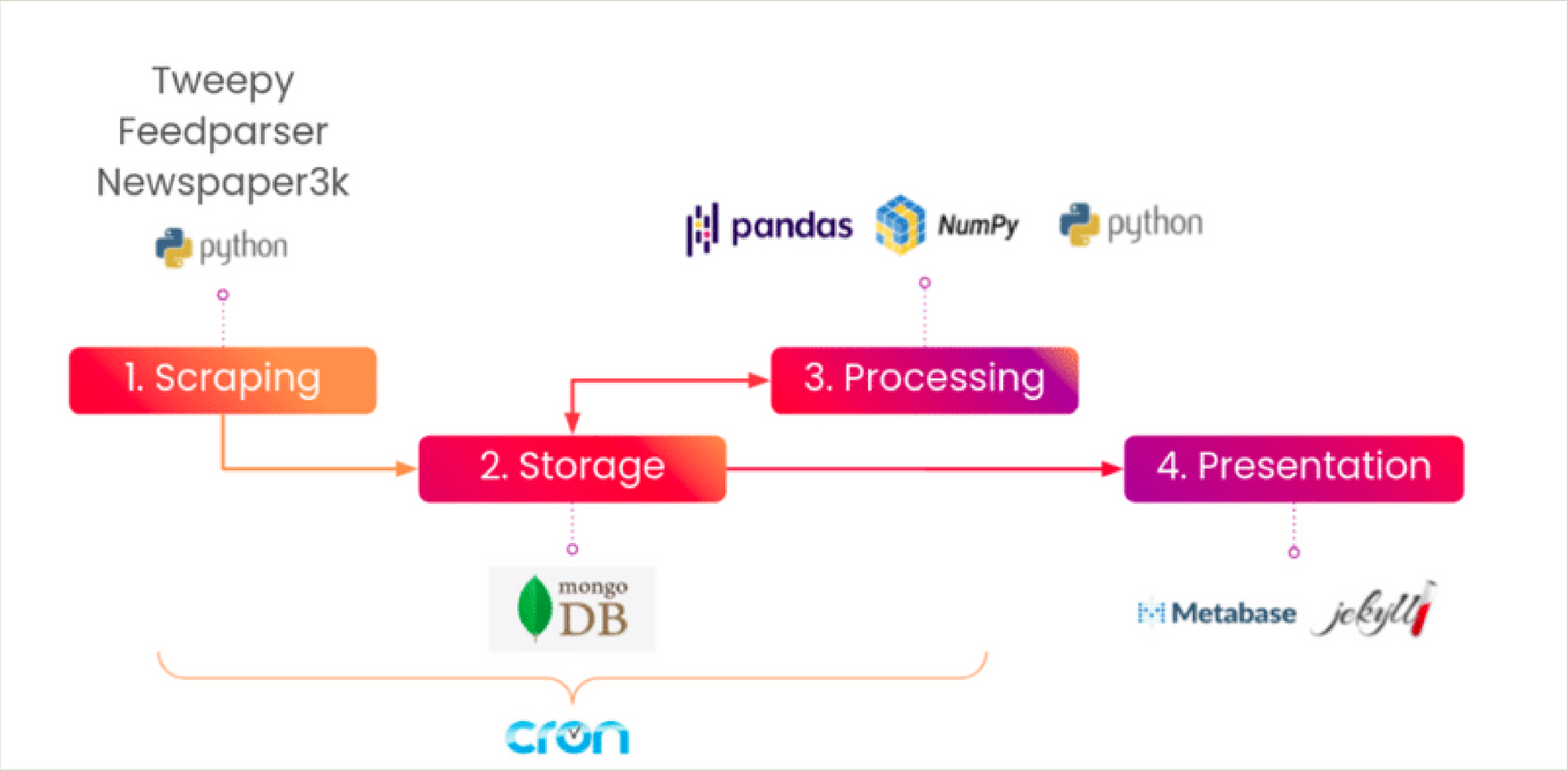
Goals

Replace the old Metabase graphs and work on the natural language processing scripts.

Specifiactions

Replace de Metabase graphs	Metabase suffers from a limitation in sources that we can add.
Use NER to target masculine/feminine names	Ambiguities persists between names of people and homonyms name of locations.
Add new media sources	Limitation of the number of sources exploited.

Old Architecture



New Architecture

Tweepy
Feedparser
Newspaper3k

Scraping

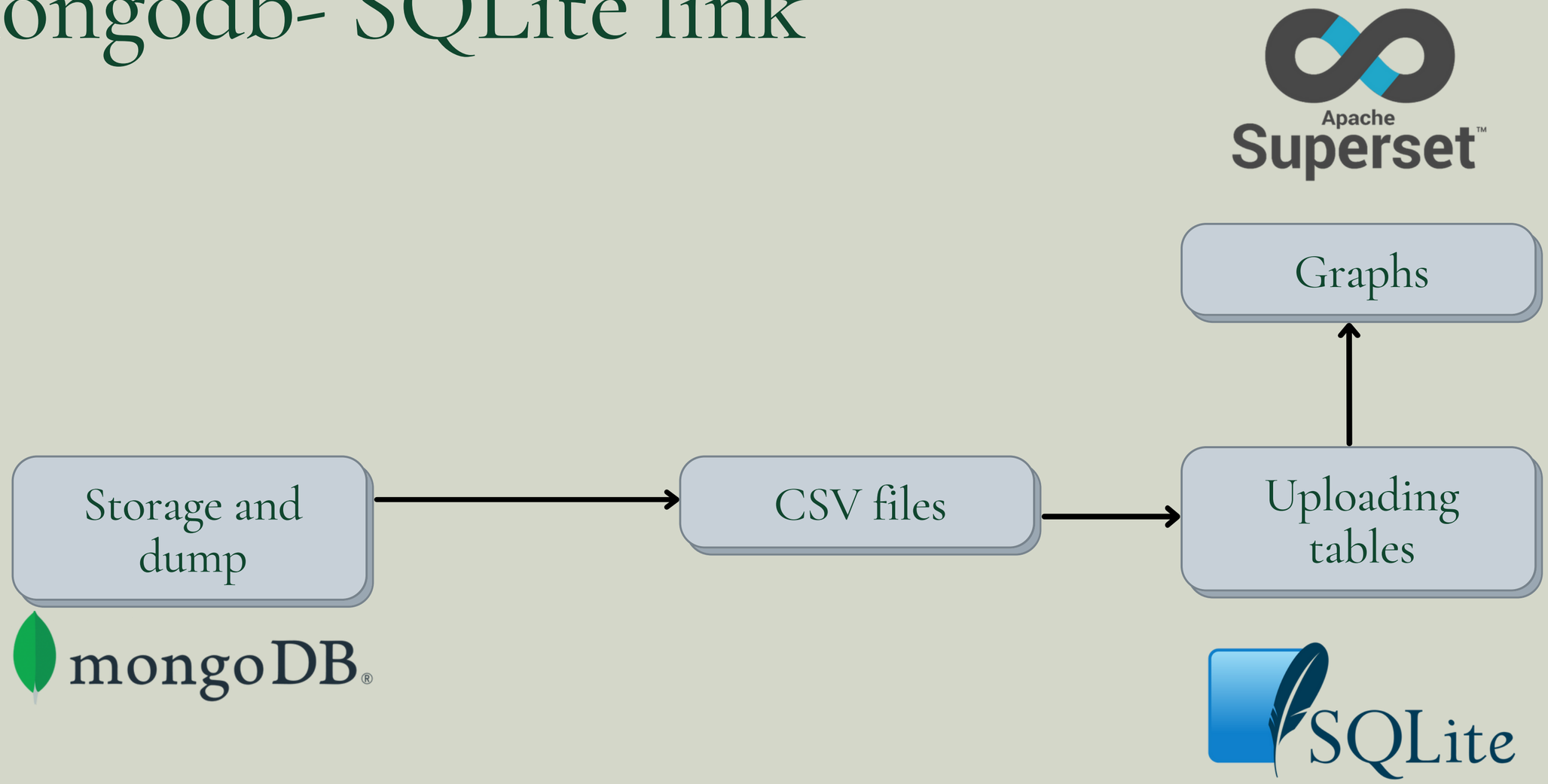
Front-end

Stockage

Processing

Data Warehouse

Mongodb- SQLite link



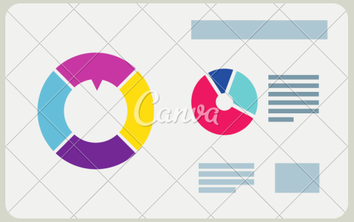
Technologies used

Databases

Mongodb, SQLite

TALN

SpaCy (NER pipelines), docanno



Data vizualisation

Apache Superset, Metabase

Programming languages

Python, SQL

Collaborative tools

Kanban

Gitlab

Google colab -
test the NER scripts

Data Annotator

CI/CD : already displayed in the gitlab repo

Tasks done

1 Migrating from Metabase to Superset to produce new charts

2 Adding new media sources

3 Implementing NER scripts

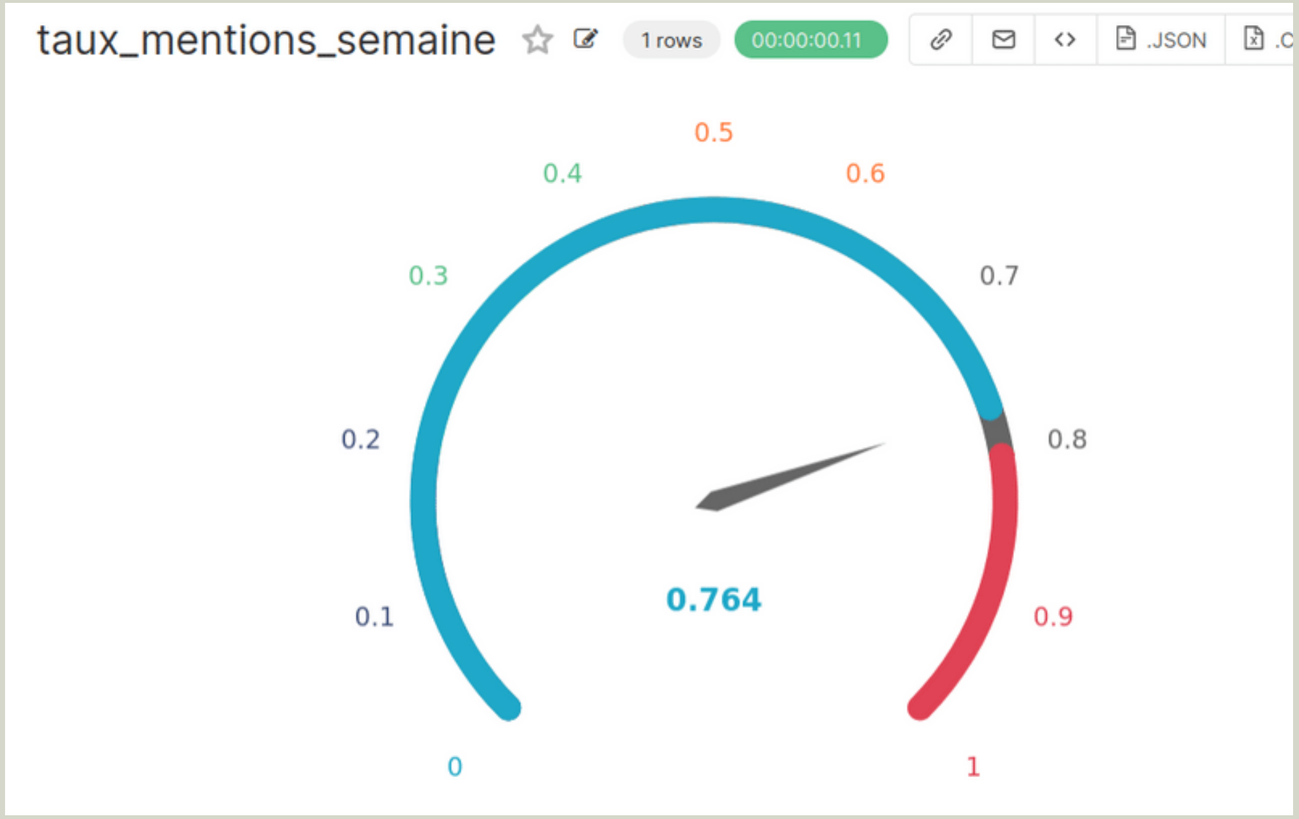
I Technology migration: Metabase to Apache Superset

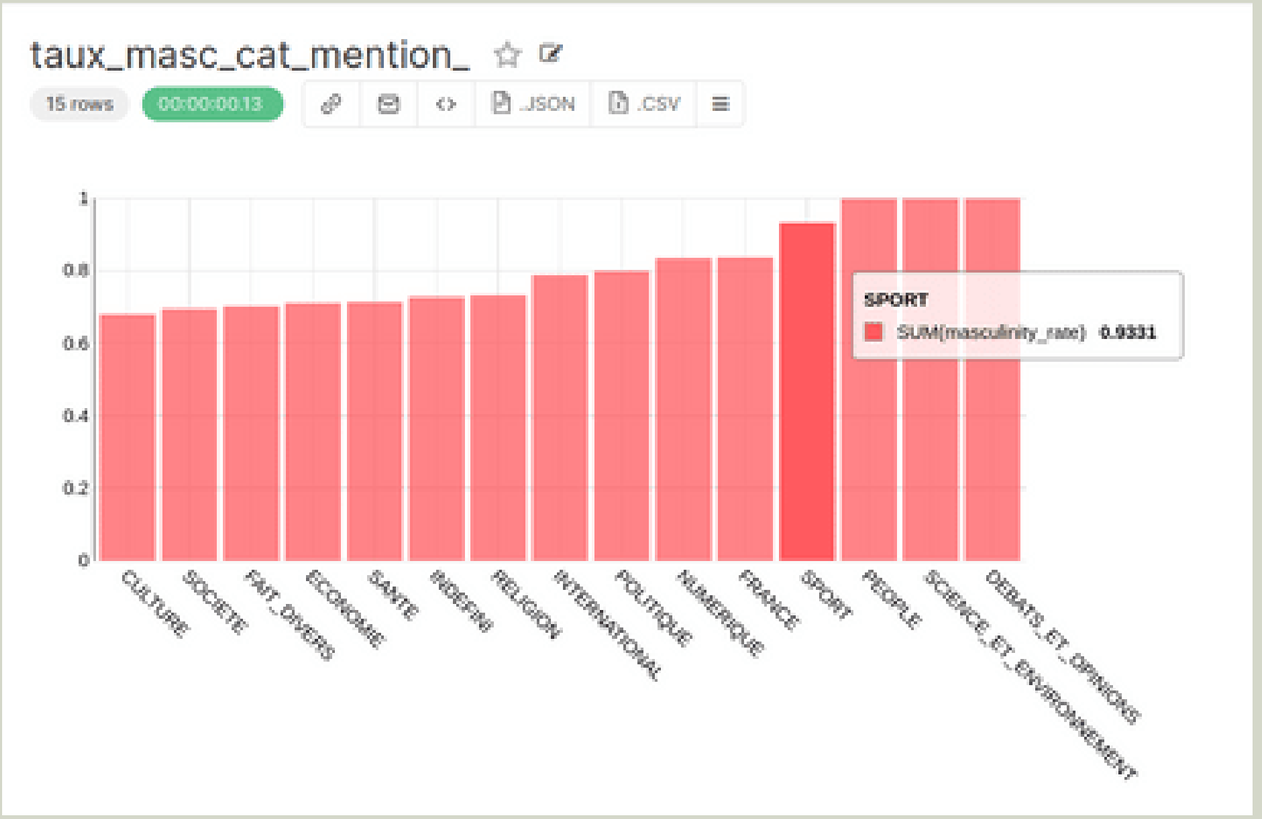
1 The creation of an intermediate SQLite Database

2 The implementation of the link between SQLite and Mongodb

3 The creation of coherant graphs on Apache Superset

Example of charts done in Superset





2 Adding new media sources

1 Adding 16 new media sources

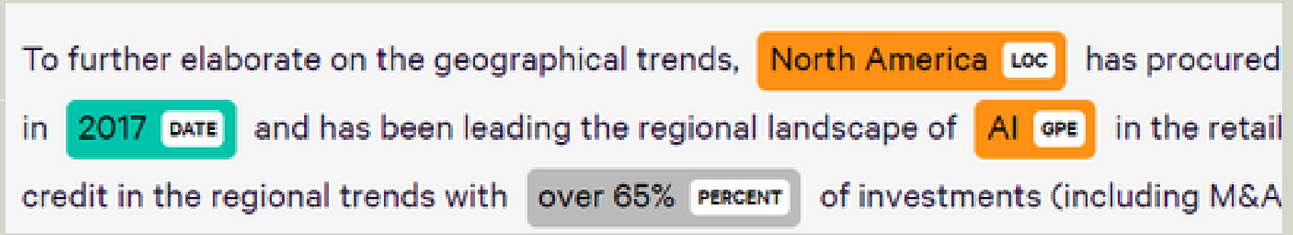
2 Recensement of the main categories of each source by browsing their websites

3 Implementation of an observation script to keep track of the real categories of the articles

Added sources

20minutes.fr
Actu.fr
France Inter
France24.fr
Franceinfo.fr
L'Express
L'Humanité
L'Opinion
La Voix du Nord
Le Dauphine Libéré
Le Monde Diplomatique
Le Point
Le Télégramme
Marianne
Ouest France
Sud Ouest

3 Implementation of the NER scripts



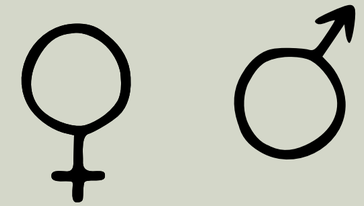
1 Test of multiple existing NER models

2 Experimental approach to explore different trails

3 Evaluation of the performance of the algorithm

NLP Approach

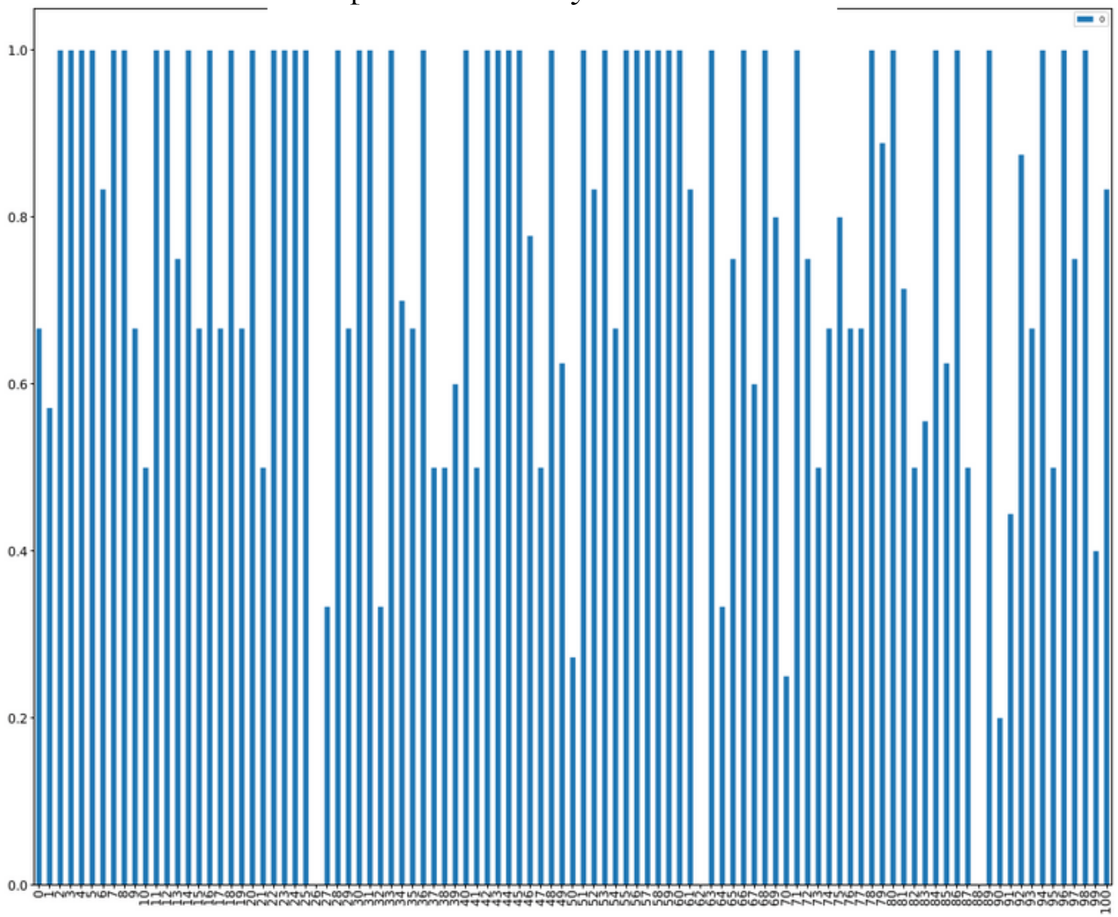
Spacy NER modèle → Filtering the entities: we keep only PER →



Retrieving the gender of the name

Performances

performance by document



total performance for all the documents:
~0.772

{ 0: [],

1: [('alexis', 'Homme')],

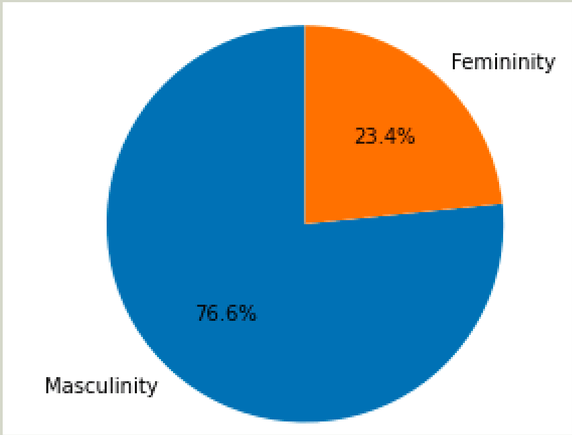
2: [('claude', 'Homme')],

3: [('maxine', 'Femme'), ('maxime', 'Homme')],

4: [('frédérique', 'Femme')],

5: [('sacha', 'Homme')],

6: [('jackie', 'Homme')]}



The error femme -> homme: 0

The error homme -> femme: 1

The error femme -> epicene: 0

The error homme -> epicene: 11

here are the names with the error (label_in_docanno, label_in_algo)

{'sameh': ('Homme', 'Epicene'), 'george': ('Homme', 'Epicene'), 'dany': ('Homme', 'Epicene'), 'billy': ('Epicene', 'Homme'), 'alex': ('Epicene', 'Homme'), 'harmony': ('Homme', 'Femme')}

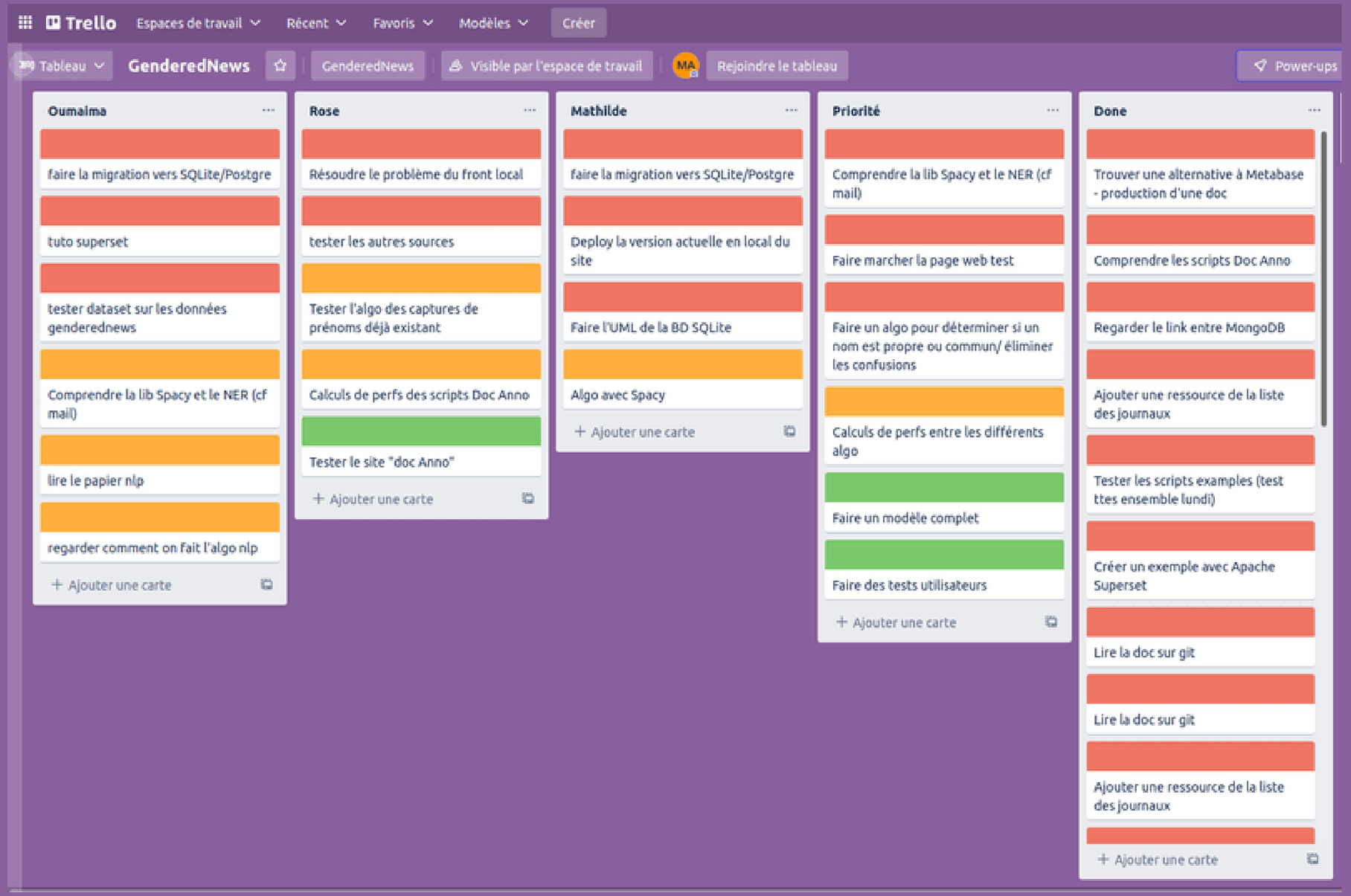
Project management

Kanban

Definition of the tasks by the SCRUM Master and displaying them in the trello Kanban to track their progression.

Weekly meetings

Weekly meeting with the project owners to exchange about the current tasks and convey ideas.


Trello
Espaces de travail
Récent
Favoris
Modèles
Créer
Tableau
GenderedNews
GenderedNews
Visible par l'espace de travail
MA
Rejoindre le tableau
Power-ups
Oumaima
faire la migration vers SQLite/Postgre
tuto superset
tester dataset sur les données genderednews
Comprendre la lib Spacy et le NER (cf mail)
lire le papier nlp
regarder comment on fait l'algo nlp
Ajouter une carte
Rose
Résoudre le problème du front local
tester les autres sources
Tester l'algo des captures de prénoms déjà existant
Calculs de perfs des scripts Doc Anno
Tester le site "doc Anno"
Ajouter une carte
Mathilde
faire la migration vers SQLite/Postgre
Deploy la version actuelle en local du site
Faire l'UML de la BD SQLite
Algo avec Spacy
Ajouter une carte
Priorité
Comprendre la lib Spacy et le NER (cf mail)
Faire marcher la page web test
Faire un algo pour déterminer si un nom est propre ou commun/ éliminer les confusions
Calculs de perfs entre les différents algo
Faire un modèle complet
Faire des tests utilisateurs
Ajouter une carte
Done
Trouver une alternative à Metabase - production d'une doc
Comprendre les scripts Doc Anno
Regarder le link entre MongoDB
Ajouter une ressource de la liste des journaux
Tester les scripts examples (test ttes ensemble lundi)
Créer un exemple avec Apache Superset
Lire la doc sur git
Lire la doc sur git
Ajouter une ressource de la liste des journaux
Ajouter une carte

Metrics

Project length : 6 weeks

Added code lines : ~2500

Mathilde

- Commits : 9
- Tasks : 15

Oumaima

- Commits : 9
- Tasks : 9

Rose

- Commits : 10
- Tasks : 10

Demonstration

https://genderednews-apache.herokuapp.com/superset/welcome/
https://genderednews.herokuapp.com/

Conclusion

- The discovery of new technologies (Spacy, Mongodb, ...)
- The participation to a useful and enriching project
- The Exploration of an important subject (Gender Inequality

Thank you for listening !