GenderedNews

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

Porteurs : RICHARD Ange - PORTET François - BASTIN Gilles

Plan :

I. Rappel du sujet
II. Etat actuel du projet
III. Réalisations techniques
IV. Gestion de projet
V. Conclusion

L'Équipe

Mathilde Aguiar

Cheffe de projet

Oumaima Hajji

SCRUM Master

Rokiatou dite Rose Sidibe

Développeuse

Contexte et sujet

Contexte

Inégalité du temps de parole, représentation des femmes dans les médias.

Sujet

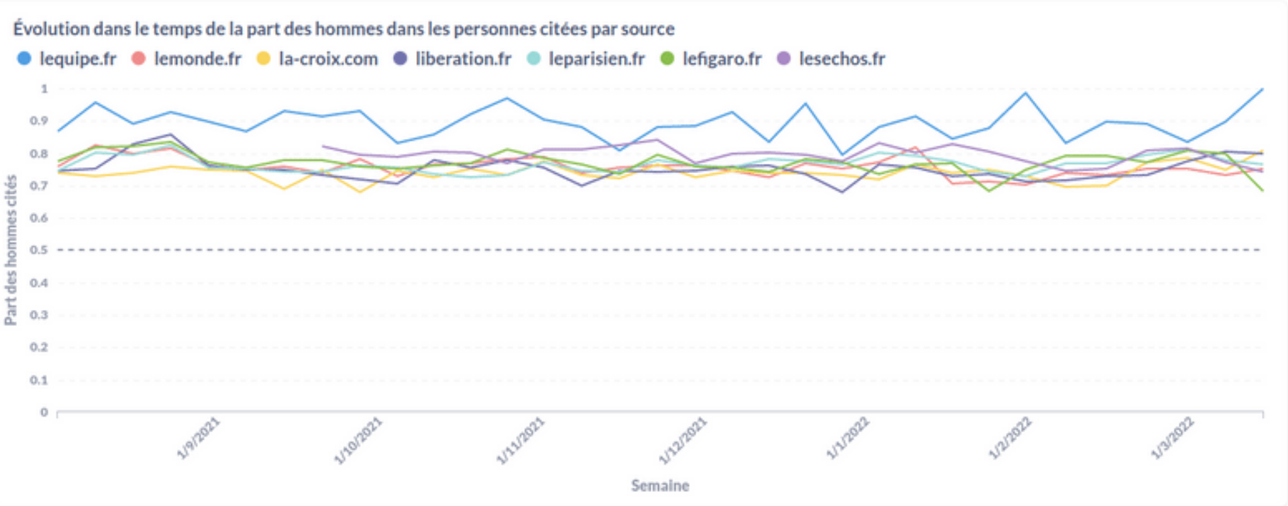
GenderedNews : un site Web qui recense le taux de masculinité présent dans les médias écrits.

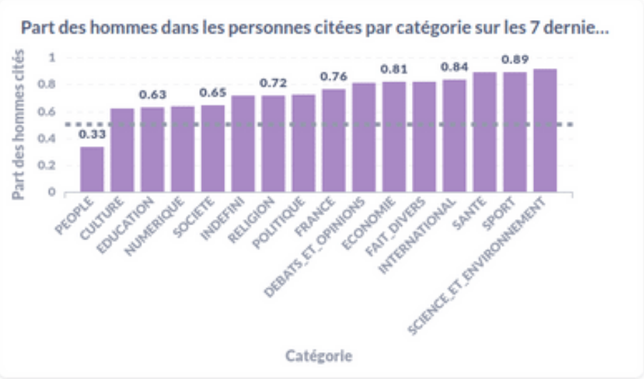
Objectifs

Améliorer la solution déjà existante en remplaçant la solution Métabase et en travaillant sur l'aspect traitement du langage naturel

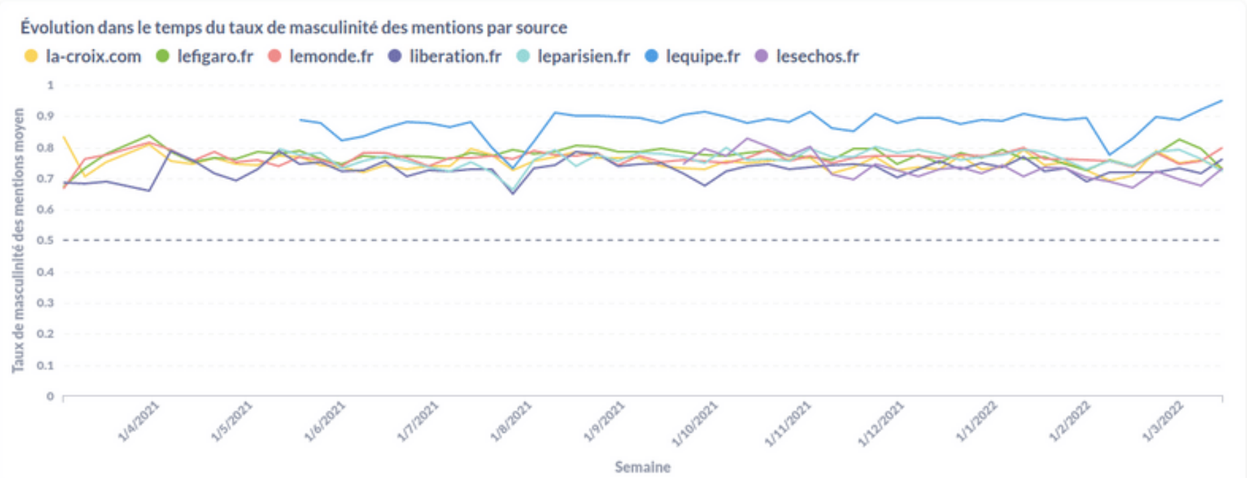
Site Web actuel

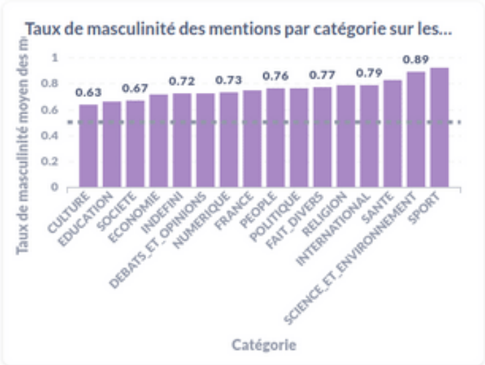
Mentions





Citations





Mention versus Citation

Mentions

le nombre de fois où l'un prénom
désignant une personne est
employé par une autre personne tierce.

Citations

les propos rapportés, le plus souvent par
un journaliste, qui concerne une
personne. Les propos rapportés peuvent
être à la fois entre guillemets ou bien
introduits par certains
termes tels que "selon, d'après, etc.".

Cahier des charges

Remplacer la technologie Metabase	Metabase étant limité en nombre de sources
Améliorer la reconnaissance des noms NER	Problème de confusion des algorithmes TAL entre prénom et nom de ville
Ajouter de nouvelles sources	Limitation du nombre de sources étudiées

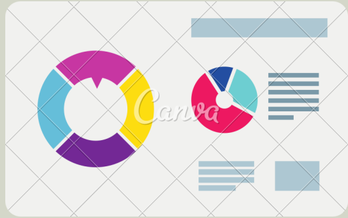
Mots clés

Base de données

Mongodb, SQLite

TALN

SpaCy (pipelines NER), docanno



Visualisation de données

Apache Superset, Metabase

Langage de programmation

Python, SQL

Outils collaboratifs

Trello

Kanban

Gitlab

Google colab - test des scripts de NER

doccano

Annotateur de données

CI/CD : déjà mise en place sur notre repo gitlab

Tâches réalisées

1. Migration de la technologie Metabase vers Apache Superset
2. Ajout de nouvelles sources médiatiques
3. Implémentation de scripts NER

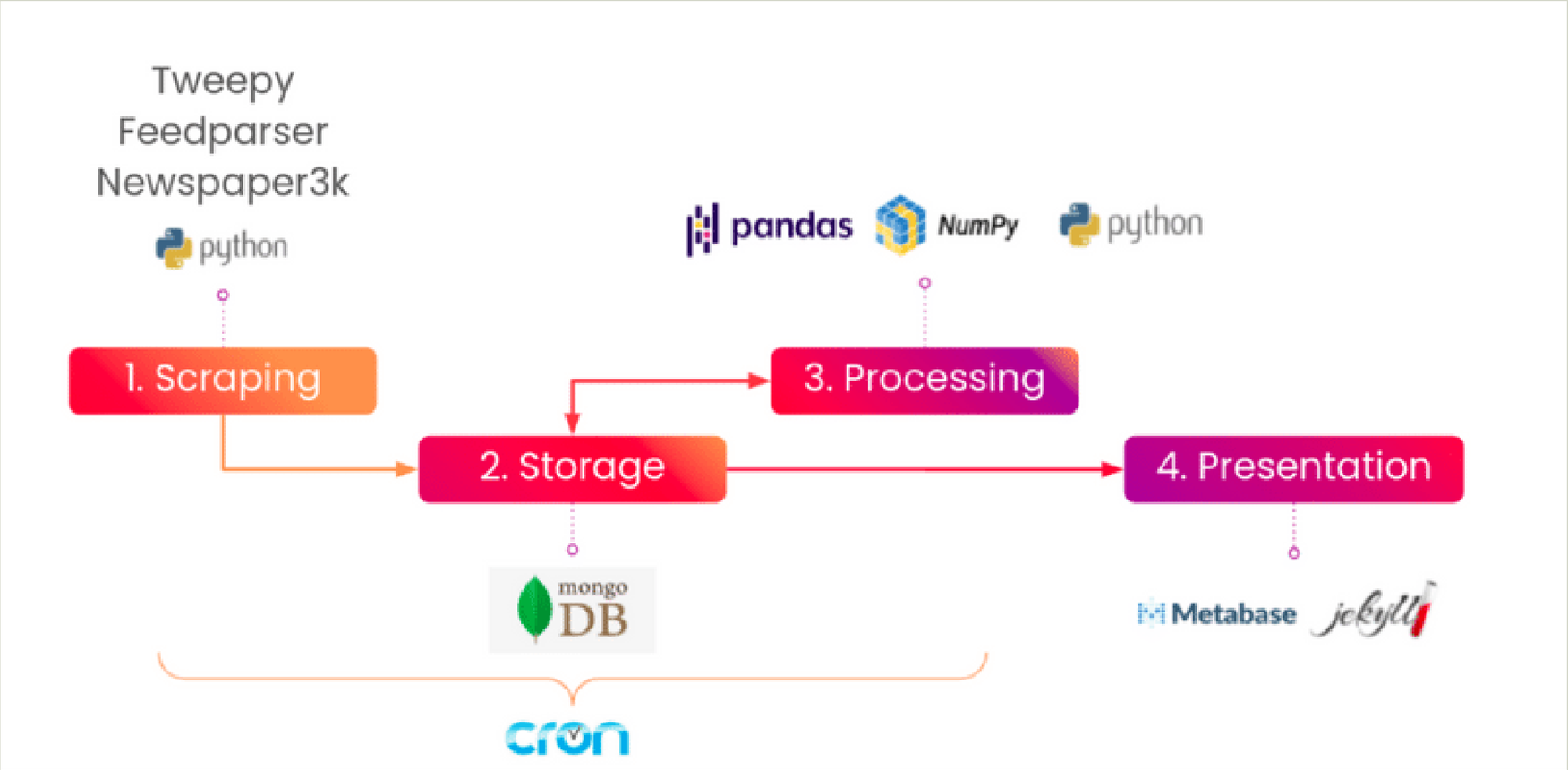
I Migration de la technologie Metabase vers Apache Superset

1 Création d'une base de données SQLite intermédiaire

2 Codage du lien entre SQLite et la base de données Mongodb

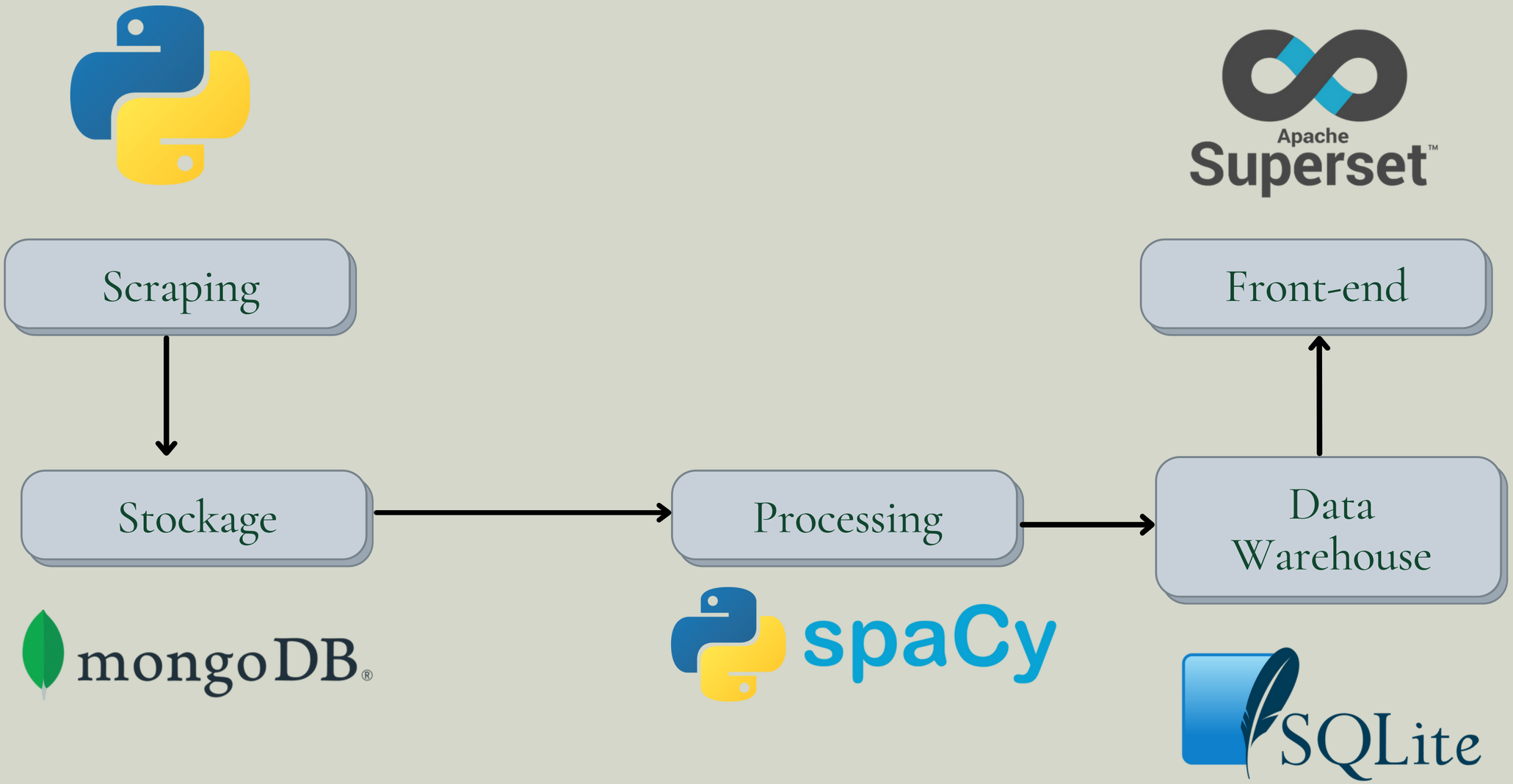
3 Création de graphes cohérents sur Superset

Ancienne Architecture

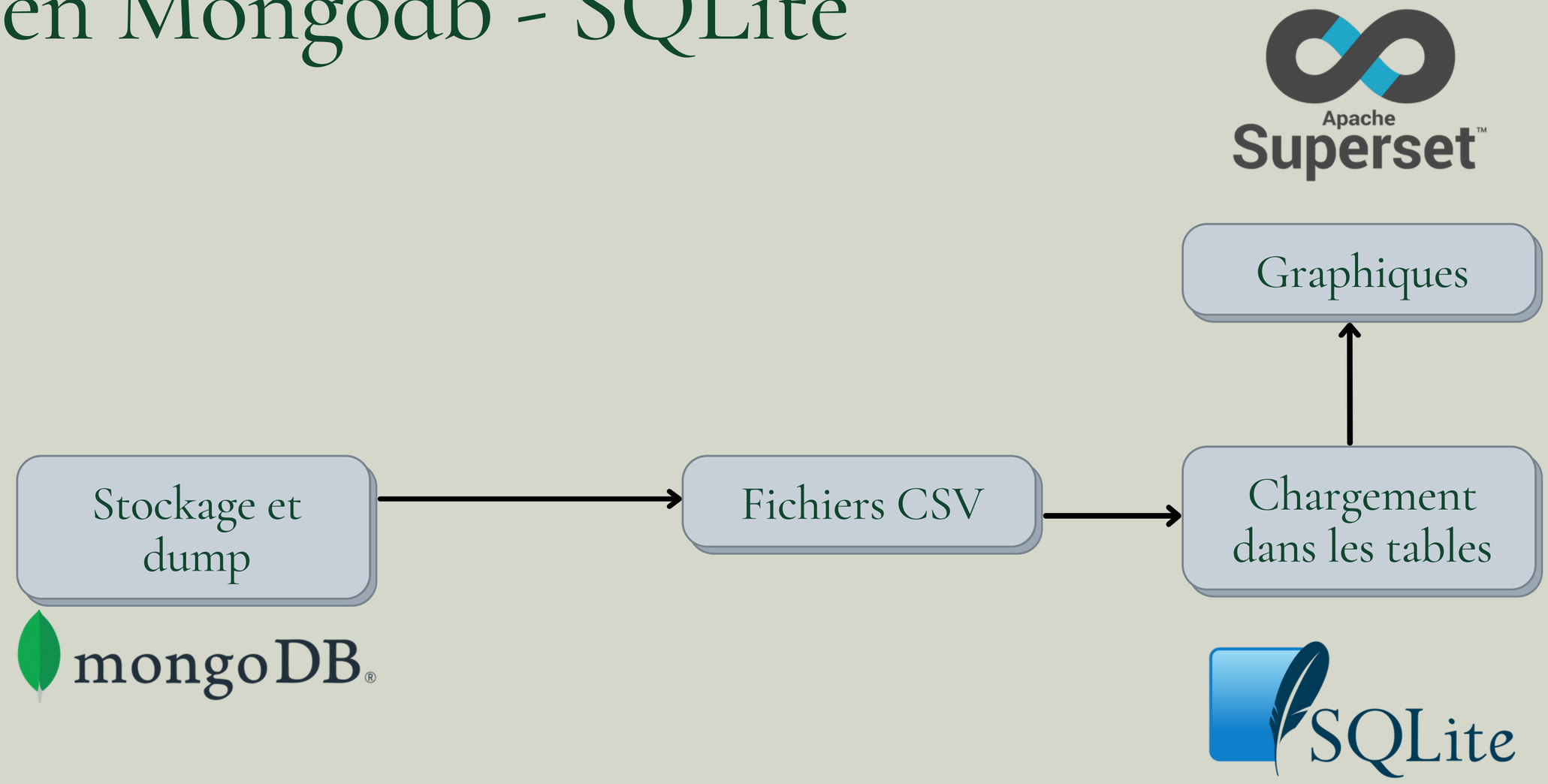


Nouvelle Architecture

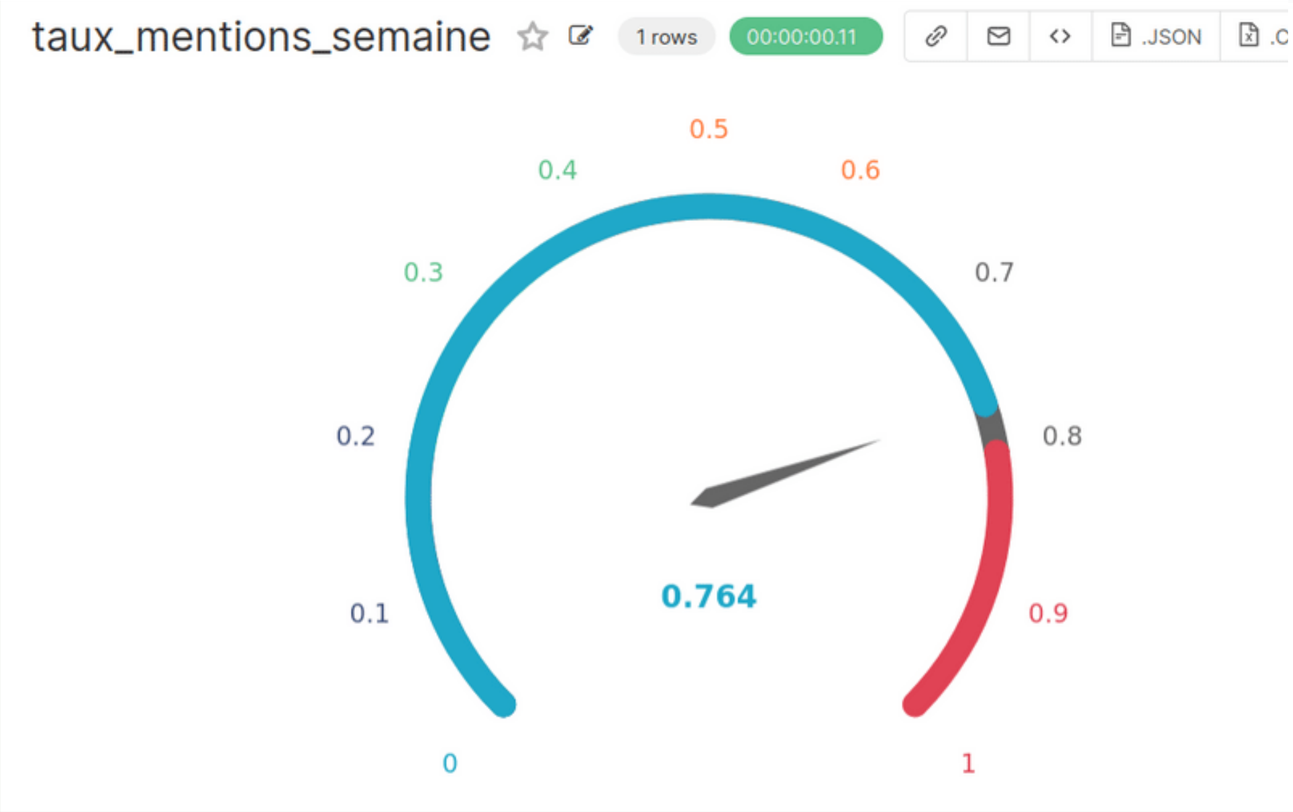
Tweepy
Feedparser
Newspaper3k

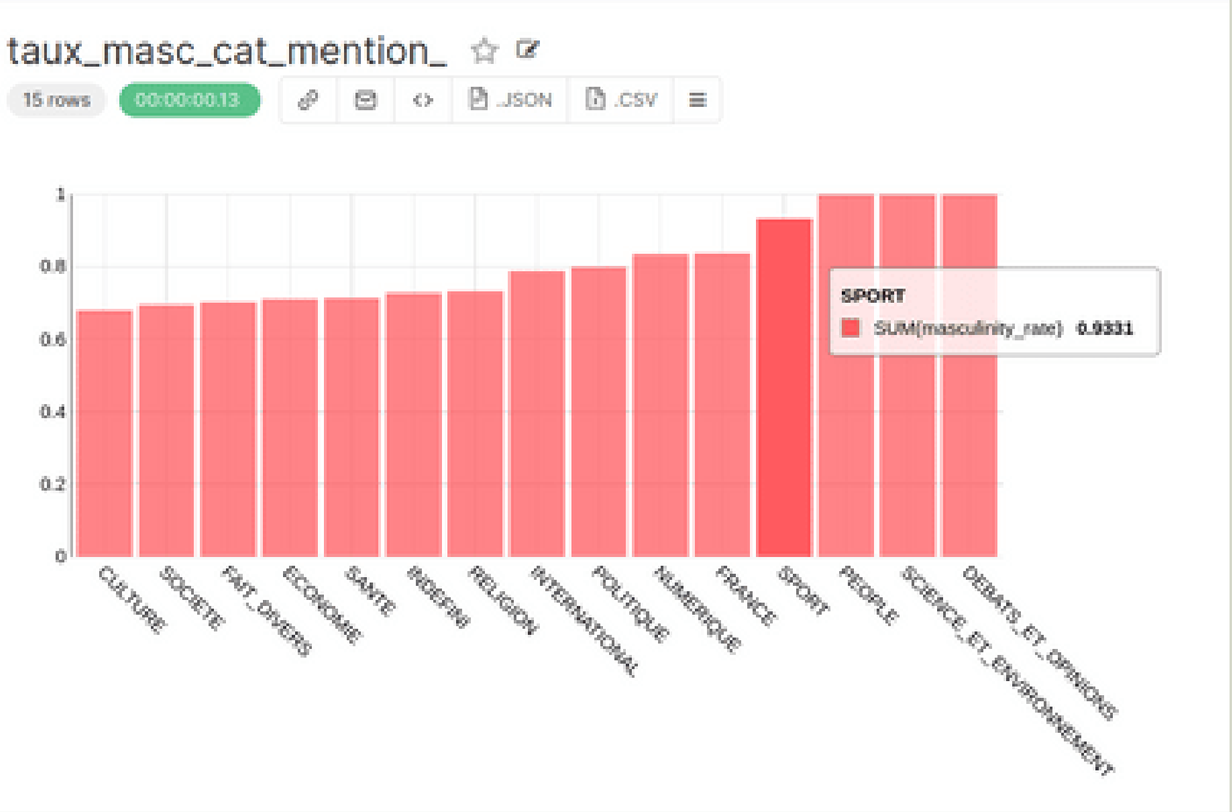


Lien Mongodb - SQLite



Exemple des charts sur Superset





2 Ajout de nouvelles sources médiatiques

1 Ajout de 16 nouvelles sources prioritaires

2 Recencement des catégories principales de chaque source à travers le parcours de leur site

3 Implémentation d'un script d'observation des articles afin de déterminer les catégories réelles

Les sources ajoutées

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 Implémentation de scripts NER

1 Test de plusieurs modèles NER existants

2 Démarche expérimentales en explorant plusieurs pistes différentes

3 Evaluation de la performance de l'algorithme

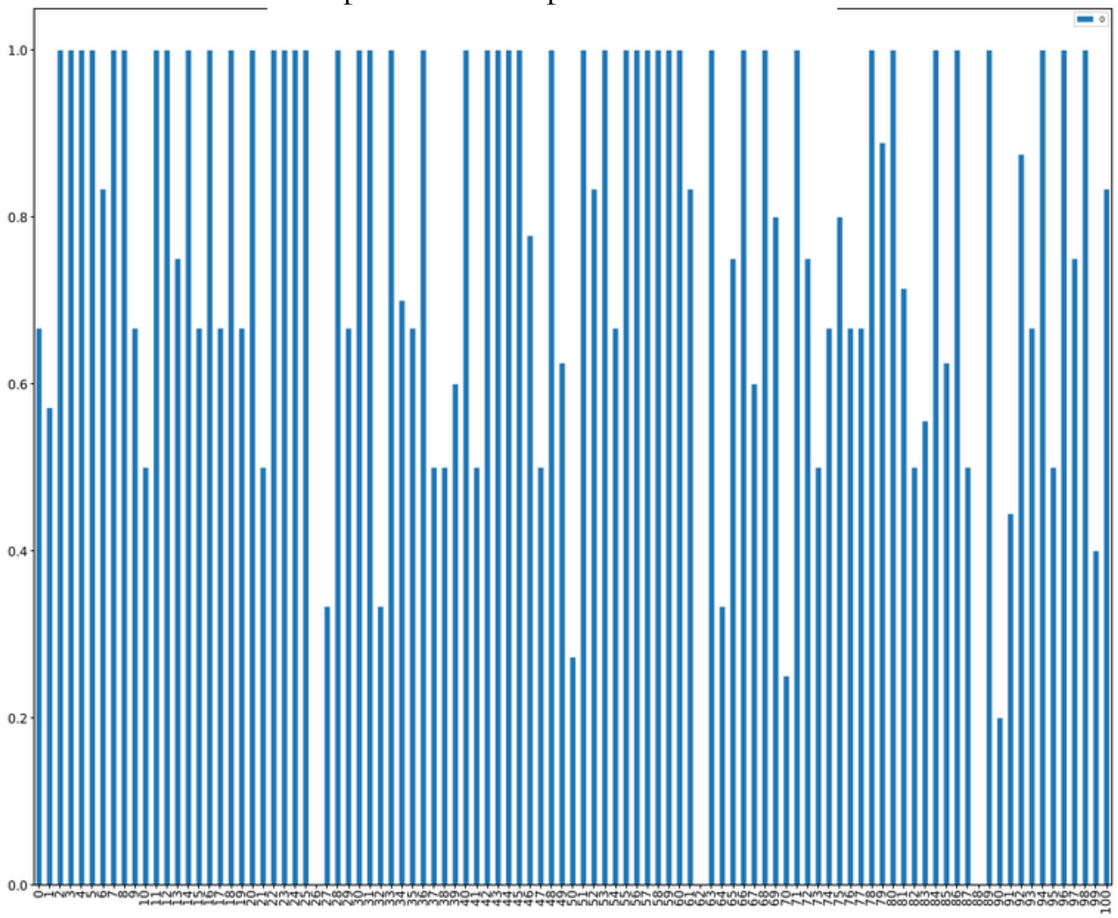
Approche TAL

Spacy NER modèle → Filtrage sur les entités: on ne guarde que PER → Récupération de genre du prénom

♀ ♂

Performances

performance par document



performance totale sur l'ensemble des documents: ~0.772

{ 0: [],

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

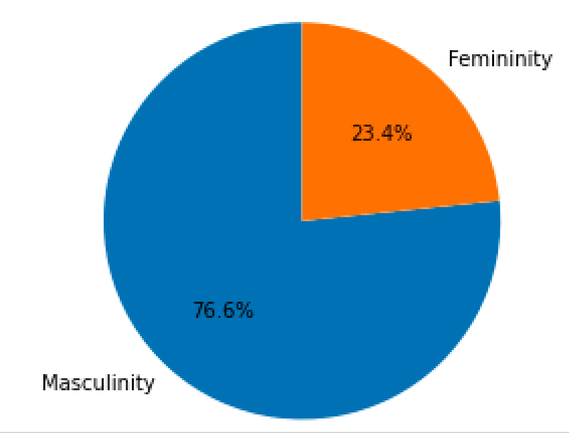
2: [('claude', 'Homme')],

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')}

Gestion de projet

Kanban

Définition des tâches par la SCRUM Master et utilisation d'un trello kanban pour le suivi des tâches.

Réunions hebdomadaires

Réunions hebdomadaires avec les porteurs pour faire les point sur les tâches réalisées et futures.

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

Métriques logicielles

durée de projet : 6 weeks

lignes de code ajoutés : ~2500

Mathilde

- Commits : 9
- Tâches : 15

Oumaima

- Commits : 9
- Tâches : 9

Rose

- Commits : 10
- Tâches : 10

Lien démonstration

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

Conclusion

- Découverte de nouvelles technologies (Spacy, Mongodb, etc.)
- Participation à un projet utile et enrichissant
- Exploration d'un sujet important (l'inégalité Homme-Femme)

Merci de votre attention