DEPARTEMENT MIASHS - UFR SHS GRENOBLE

Année Universitaire 2025-2026

GENDERED NEWS : Cahier des charges

Autrices :

Rachel PERETTI
Aylin GUVENTURK

Contexte du projet

Présentation de Gendered News

Le projet GenderedNews a été conçu afin de mettre en évidence les inégalités de représentation entre les hommes et les femmes dans les sources médiatiques françaises, en particulier au sein des principaux journaux de presse en ligne.

Le site web du projet constitue une vitrine de restitution des résultats obtenus. Il est mis à jour quotidiennement et présente, sous forme de graphiques, les nouvelles mesures d'inégalités observées. Ces mesures reposent principalement sur deux notions :

- les mentions : correspondant au fait de parler d'hommes
- les citations : correspondant au fait de donner la parole à des hommes

Ces représentations graphiques permettent de visualiser clairement la part des médias consacrée aux hommes par rapport à celle réservée aux femmes, la variation de cette part selon les différents sujets des articles et nous permettent aussi de comparer le taux de présence masculine dans les différents journaux et son évolution dans le temps.

Le projet est basé sur des méthodes de traitement du langage naturel (TAL) permettant de quantifier le déséquilibre dans la représentation des genres à travers le calcul de deux indicateurs principaux: le taux de masculinité des mentions et le taux de masculinité des citations. En complément, les articles sont classés par catégorie : éducation, politique, société, santé, économie, etc…

Les sites de presse en ligne sur lesquels les analyses sont faites sont : *Le Monde, Le Figaro, La Croix, Le Parisien, Libération, Les Echos* et *L'Equipe*. Le nombre de ces journaux est susceptible d'augmenter pour avoir des résultats globaux plus significatifs.

GenderedNews est un projet de recherche développé au sein de l'Université Grenoble Alpes, dont l'objectif est de s'inscrire dans une démarche de transfert vers le monde industriel. Une collaboration avec l'Agence France-Presse (AFP) est en cours. Cette collaboration souligne l'intérêt du projet et renforce sa crédibilité en tant qu'outil potentiellement intégrable dans des environnements professionnels.

Présentation de l'équipe GETALP

Le projet s'inscrit dans les travaux de l'équipe GETALP (Groupe d'Étude et de Traitement Automatique de la Langue Parlée et Écrite) du Laboratoire d'Informatique de Grenoble (LIG). Cette équipe de recherche est spécialisée dans le traitement automatique du langage naturel, domaine central du projet GenderedNews, notamment pour l'analyse automatique de corpus journalistiques.

Le projet GenderedNews est porté par Gilles Bastin (Laboratoire Pacte et Sciences Po Grenoble / UGA), François Portet (Laboratoire d'Informatique de Grenoble et UGA), Ange Richard (Pacte et LIG) et Patrick Juen (Pacte et Gricad).

Problématique

Le système actuel de GenderedNews repose sur un ensemble de scrapers permettant de collecter quotidiennement des articles issus de plusieurs sites de presse en ligne. Le scraping est un procédé automatisé permettant d'extraire des informations à partir de pages web, en analysant leur structure afin de collecter des données spécifiques, telles que le contenu textuel d'articles de presse.

Or, ces sites sont soumis à des évolutions fréquentes de leurs structures, que ce soit au niveau du code HTML ou des mécanismes de chargement des contenus. Certains sites mettent également en place des protections pour lutter contre le scraping automatisé. Ces changements rendent le système particulièrement sensible, des données incorrectes pouvant le rendre instable, et nécessitent des ajustements.

Par ailleurs, le projet possède un historique technique important, ce qui complexifie la compréhension globale du code et sa maintenance. Les scrapers et configurations sont répartis dans différentes parties du projet. Certaines dépendances sont devenues obsolètes et les tests unitaires existants ne sont plus fonctionnels. Cette situation rend difficile l'anticipation des pannes, la détection des erreurs et l'évolution du système.

Présentation du projet existant

Architecture actuelle

L'architecture actuelle du projet GenderedNews repose sur une séparation claire entre le backend, chargé de la collecte et du traitement des données, et le frontend, dédié à leur présentation. Le système passe par plusieurs étapes :

- **Collecte des données (Scraping)**
  Collecte automatique des articles de presse à partir de différents sites de journaux en ligne, tels que LeMonde.fr. Cette étape est réalisée via des scripts de scraping développés en Python, s'appuyant notamment sur les bibliothèques Feedparser pour l'analyse des flux RSS et BeautifulSoup pour l'extraction du contenu HTML des pages. Les scrapers récupèrent les métadonnées des articles (titre, date, auteur, URL, etc...)

- **Stockage des données**
  Les données collectées sont ensuite stockées dans une base de données MongoDB. L'exécution régulière de la collecte et de l'enregistrement des données est automatisée à l'aide de tâches planifiées via cron, pour une mise à jour périodique du corpus.

- **Processing** :
  Une fois ces informations stockées, les calculs de taux de masculinité sont réalisés selon les mentions et les citations.
  Pour cela la partie processing se divise en 3 grandes parties :

- Les mentions : elles correspondent au nombre de fois où un prénom désignant une personne est employé par une autre personne tierce. Par exemple ”Nous allons effectuer la critique du livre écrit par madame Virginia Woolf”, représente une mention d’une personne de sexe féminin.
- Les citations sont les propos rapportés, le plus souvent par un journaliste, qui concernent une personne. Ces propos peuvent être à la fois entre guillemets ou bien introduits par certains termes tels que ”selon, d'après, etc.”. Là où les mentions se basent plus sur de la détection puis analyse de prénoms, les citations requièrent une compréhension de la phrase plus complexe car la sémantique ainsi que le contexte sont souvent plus compliqués à saisir. Dans la solution existante on utilise un modèle entier.
- Catégorisation homogène : Comme chaque source (*Le Monde*, *Libération*, etc.) possède ses propres catégories, ce traitement permet d’harmoniser ces catégories.

- **Présentation (frontend)**
  Cette partie correspond au frontend du site web. Elle repose sur des technologies web telles que Node.js, JavaScript et Jekyll, permettant de générer des pages web présentant les analyses de manière accessible.

Une représentation de l’architecture est la suivante :

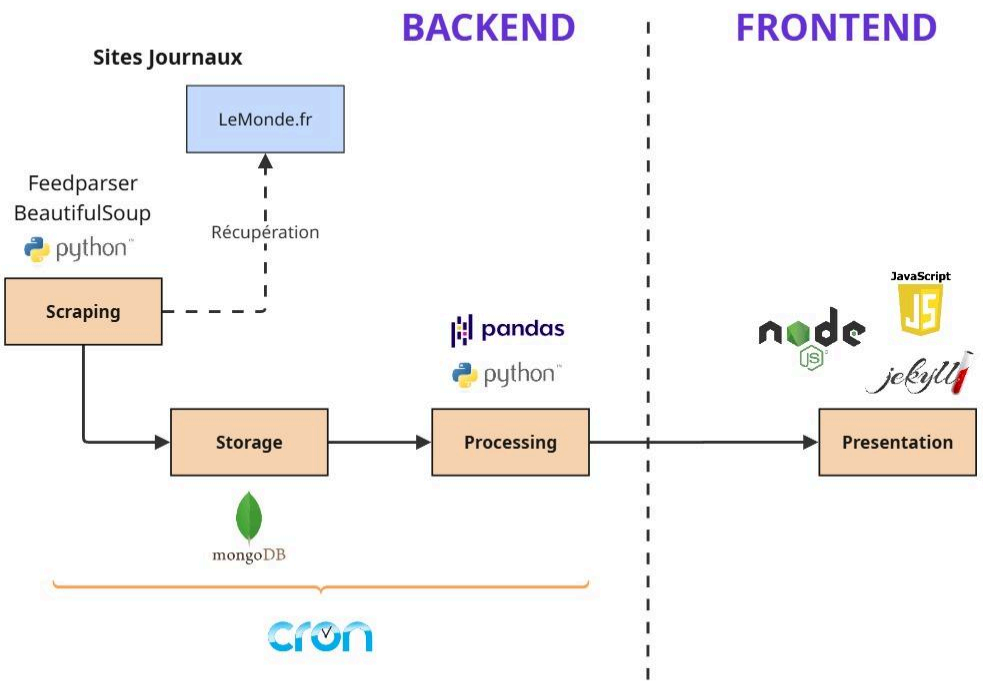

Figure 1. Schéma d’architecture de Gendered News

Contraintes et limites

Le projet est contraint par le choix des technologies existantes, qui doivent être conservées afin d'assurer la stabilité de l'infrastructure. En revanche, une certaine flexibilité est possible concernant les bibliothèques utilisées pour le scraping. Les outils actuels, notamment basés sur BeautifulSoup, présentent une fragilité structurelle liée aux problèmes de compatibilité entre versions de bibliothèques. Il est donc envisageable de remplacer ou compléter ces bibliothèques, avec des alternatives étant maintenues par des équipes fiables.

Le projet repose sur une base de données MongoDB contenant plusieurs années d'articles de presse, donc une quantité importante de données. Dans ce contexte, certaines requêtes peuvent devenir très coûteuses en temps de calcul si elles ne s'appuient pas sur des index, particulièrement lorsqu'elles portent sur des périodes longues ou sur l'ensemble des périodiques. Toute opération de traitement doit donc être pensée en fonction de ces index afin de raccourcir au mieux les temps de réponse.

Enfin, le scraping constitue une composante centrale du projet mais également l'une de ses parties les plus instables. Certaines bibliothèques ne sont plus maintenues ou deviennent incompatibles au fil du temps, ce qui oblige parfois à contourner les problèmes plutôt qu'à les corriger de manière pérenne. Cette fragilité rend impossible toute garantie de stabilité à long terme et impose une maintenance régulière, ainsi que la mise en place de mécanismes de détection rapide des dysfonctionnements.

Plan de développement

Refactoring et nettoyage

Le développement du projet s'articulera autour de plusieurs axes principaux. Dans un premier temps, un travail de refactoring et de nettoyage du code sera mené afin d'améliorer la lisibilité et la maintenabilité du système existant.

Cette étape consistera notamment à

- Réorganiser l'architecture du projet
- Clarifier le rôle des différents modules, classes et fichiers
- Centraliser les configurations : imports des packages, versionning des packages, éléments d'installation du projet…
- Supprimer les éléments obsolètes ou non utilisés

Mise à jour des scrappers

La mise à jour des scrapers constitue un des axes principaux du développement du projet.

Elle débutera par une étude approfondie du code existant, afin d'identifier les points de fragilité et d'évaluer les possibilités d'amélioration ou de modification. Dans le cas où certaines limites structurelles empêcheraient ces adaptations, une réflexion pourra être menée sur l'utilisation de bibliothèques de scraping plus flexibles et maintenues. Il faudra

étudier le cas d'alternatives à la bibliothèque Beautifulsoup actuellement utilisée sur le projet.

Par ailleurs, il est nécessaire d'améliorer la capacité du programme à détecter automatiquement des anomalies survenant lors du processus de scraping, telles que l'absence de contenu, des erreurs d'extraction ou des changements de structure des pages. Cette détection permettra d'identifier plus rapidement les scrapers défaillants et de limiter l'impact des erreurs sur la chaîne de traitement.

Enfin, la mise à jour des scrapers s'accompagne de la mise en place de nouveaux tests unitaires, conçus pour vérifier le bon fonctionnement du système sur différentes périodes temporelles. Ces tests permettront de s'assurer de la stabilité des scrapers dans le temps et de faciliter la détection de régressions liées aux évolutions des sites sources.

Améliorations complémentaires

En complément des tâches prioritaires dans le cadre de ce projet, plusieurs pistes d'améliorations pourraient être envisagées si le temps le permet. Les tâches sont les suivantes :

- Implémentation d'un versioning temporel des données : pouvoir récupérer et travailler sur des articles datant d'avant 2020.
- Analyse des baisses drastiques : à partir des résultats produits par GenderedNews, identifier et justifier les causes possibles de ces variations importantes pour lesquelles il y a des baisses drastiques à certaines périodes.
- Contribution à l'évolution du site : inclure des ajustements de l'interface, une amélioration de la lisibilité des visualisations, ou encore l'ajout de nouvelles fonctionnalités.