Résumé des résultats : Web scraping & NLP

Objectif du projet

Le projet consiste à automatiser l'extraction et l'analyse textuelle d'articles provenant de **Le Monde** et **Midi Madagasikara**. L'objectif est de comparer leurs rythmes de publication, leurs thématiques dominantes et leurs priorités éditoriales à travers des analyses statistiques et des nuages de mots.

Web Scraping : Adaptation aux deux sites

Les deux journaux présentent des structures HTML très différentes, ce qui a nécessité une stratégie personnalisée :

❖ **Le Monde**

- Architecture claire, balises CSS bien définies.
- Extraction fiable des **titres, descriptions et URL**.
- Limitation volontaire à **10 articles par catégorie** pour garantir un échantillon cohérent.

❖ **Midi Madagasikara**

- Structure moins homogène, peu de descriptions.
- Extraction centrée sur les balises les plus stables.
- Limitation à **4 articles par catégorie** car :
    - peu d'articles publiés dans certaines sections
    - aucune publication le dimanche ni certains jours fériés

Données générées

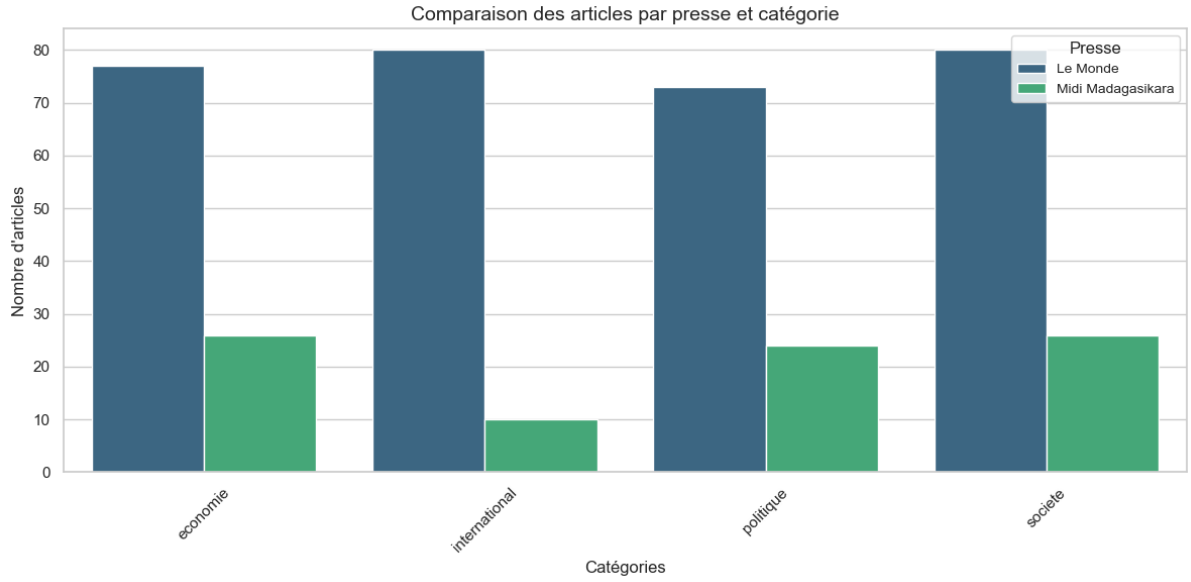
Un fichier global regroupe tous les articles, avec les colonnes :

- **Titre**
- **URL**
- **Description** (Le Monde) / **500 premiers caractères** (Midi Madagasikara)
- **Date de collecte**
- **Presse** et **Catégorie**

Chaque article est également sauvegardé dans un fichier individuel, organisé par presse et catégorie.

Analyse statistique

Principaux résultats :



- **Le Monde** publie un volume d'articles nettement plus élevé et plus équilibré sur toutes les catégories (Politique, International, Société, Économie).
- **Midi Madagasikara** publie moins fréquemment, avec un accent sur **Politique** et **Société**, mais très peu d'articles en **Monde/International**.
- Les différences reflètent la réalité éditoriale :
    - *Le Monde* → ressources élevées, couverture large, rythme quotidien.
    - *Midi Madagasikara* → ressources limitées, irrégularité, contraintes locales.

Analyse textuelle (NLP) (voir dossier « visual » pour les nuages de mots)

Des nuages de mots ont été générés à partir des titres pour identifier les thèmes dominants.

❖ **Le Monde**

- **Société** : termes liés à l'actualité judiciaire (procès, viols, verdict).
- **International** : Syrie, Ukraine, États-Unis, guerre → contexte géopolitique tendu.
- **Économie** : Noël, cadeaux, crise → période de fêtes + contexte d'inflation.
- **Politique** : Macron, Bayrou, gouvernement → actualité politique nationale.

❖ **Midi Madagasikara**

- **Politique** : Rajoelina, municipales, région → élections locales.
- **Économie** : partenariat, commerce, développement, fêtes → initiatives locales.

- **Société** : enfants, Noël, atelier → actions communautaires et évènements sociaux.

Les deux presses reflètent leurs contextes nationaux et leurs priorités :

- Le Monde → actualité internationale + enjeux politiques et économiques nationaux.
- Midi Madagasikara → politique locale, initiatives de développement, société civile.

Difficultés rencontrées

- Structures HTML très différentes → logique de scraping spécifique pour chaque site.
- Volume d'articles limité chez Midi Madagasikara.
- Problèmes d'encodage (HTML entities comme ’,) nécessitant un nettoyage supplémentaire.

Conclusion

Ce projet met en évidence les écarts structurels et éditoriaux entre les deux journaux.
Le Monde publie plus régulièrement et de manière plus homogène, tandis que Midi Madagasikara présente une fréquence réduite et des contenus centrés sur des enjeux locaux.

Les techniques développées pour le scraping, le nettoyage et l'analyse textuelle peuvent être réutilisées pour d'autres médias ou des périodes plus longues, permettant d'élargir l'étude des dynamiques médiatiques.