Summary of results: Web wcraping & NLP

Objective

This project aims to automate the extraction and textual analysis of news articles from Le Monde and Midi Madagasikara.
The objective is to compare their publication frequency, dominant themes, and editorial priorities through statistical analysis and word cloud visualisation.

Web scraping: adapting to two different websites

The two newspapers present significantly different HTML structures, requiring a tailored scraping strategy for each source.

Le Monde

- Clear and well-structured architecture with well-defined CSS selectors
- Reliable extraction of titles, descriptions, and URLs
- Voluntary limitation to 10 articles per category to ensure a balanced and representative sample

Midi Madagasikara

- Less homogeneous page structure with limited article descriptions
- Extraction focused on the most stable HTML elements
- Limitation to 4 articles per category due to:
    - a low number of published articles in some sections
    - no publications on Sundays and public holidays

Generated dataset

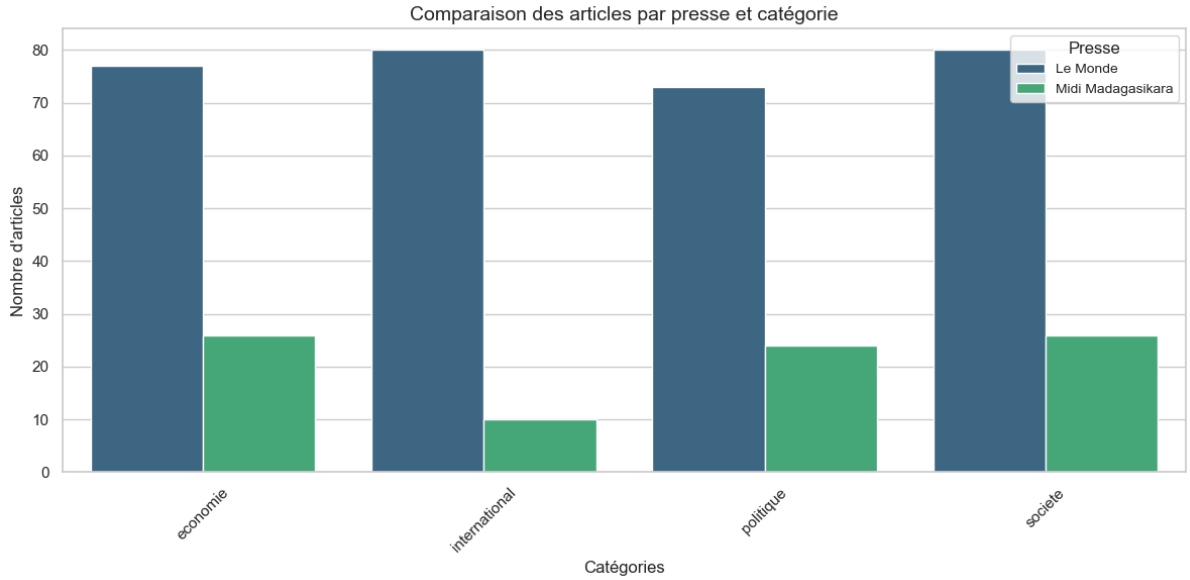
A consolidated dataset was created containing all extracted articles, with the following fields:

- Title
- URL
- Description (Le Monde) / first 500 characters (Midi Madagasikara)
- Collection date
- Newspaper and Category

Each article was also saved as an individual file, organised by newspaper and category to facilitate comparative analysis.

Statistical analysis

Key results:



- ❖ **Le Monde** publishes a significantly higher and more balanced volume of articles across all categories (*Politics, International, Society, Economy*).
- ❖ **Midi Madagasikara** publishes less frequently, with a stronger focus on *Politics* and *Society*, and very limited coverage in *International/World* news.

These differences reflect each newspaper's editorial context:

- **Le Monde**: strong editorial resources, broad coverage, daily publication rhythm.
- **Midi Madagasikara**: limited resources, irregular publication, and local constraints.

Textual Analysis (NLP)

(See the visuals/ folder for word clouds)

Word clouds were generated from article titles to identify dominant themes across newspapers and categories.

Le Monde

- Society: terms related to judicial news (trial, sexual assault, verdict).
- International: Syria, Ukraine, United States, war → reflects a tense geopolitical context.
- Economy: Christmas, gifts, crisis → holiday season combined with inflationary pressures.
- Politics: Macron, Bayrou, government → national political news.

Midi Madagasikara

- Politics: Rajoelina, municipal elections, regions → local political events.

- Economy: partnership, trade, development, holidays → local economic initiatives.
- Society: children, Christmas, workshops → community actions and social events.

Overall, both newspapers reflect their national contexts and editorial priorities:

- Le Monde → international news coverage and national political and economic issues.
- Midi Madagasikara → local politics, development initiatives, and civil society topics.

Challenges Encountered

- Significant differences in HTML structures required custom scraping logic for each website.
- Limited article volume for Midi Madagasikara.
- Encoding issues (HTML entities such as ’,) required additional text cleaning.

Conclusion

This project highlights the structural and editorial differences between the two newspapers.

Le Monde publishes more frequently and in a more consistent manner, while Midi Madagasikara shows a lower publication frequency with content primarily focused on local issues.

The scraping, cleaning, and NLP techniques developed in this project can be extended to other media outlets or longer collection periods, enabling deeper analysis of media dynamics.