# Rapport sur le projet final

L’application “Presse-Sentiment” récupère depuis le web les articles de presse en Français.

Après un traitement analysant les textes, toutes ces données sont enregistrés dans une base de donnés.

Un site internet permet de consulter ces données sous différentes formes (graphiques, statistiques, positivité, articles similaires, nuage de mots, ...) .

## Démarche

Scraping

Le script en charge du scraping récupère les articles de [Google Actu](https://news.google.com/u/1/topstories?hl=fr&gl=FR&ceid=FR%3Afr), en extrait le titre, l’éditeur, le lien vers l’article, ...

Puis scrape le lien de chaque articles pour récupérer le texte complet des articles en question.

Sur la base des infos textuelles, (titre + chapeau + texte), différents traitements sont réalisé:

* Concaténation pour obtenir une chaîne de caractère par article
* Calcul de la positivité
* Calcul de la subjectivité
* Tokénistion, Stémmatisation et Lémmatisation
* Nettoyage du texte à l’aide d’expressions régulières
* Création d’une matrice de termes

L’exécution du script est automatisé et se lance régulièrement

Les logs sont enregistrés dans un fichier

Base de données

L’article ainsi traité est comparé à la base de données existante pour éviter de doublons.

Puis enregistré dans une base de données non relationnelle (Mongodb).

Application Flask

L’application web consulte la base de données pour retourner les données utiles à divers traitement:

- Graphiques

- Classement par positivité

- Présentation des articles similaires

- Nuage de mots

- Statistiques

Docker

Le tout s’exécute en conteneur depuis Docker-Compose.

## Problèmes

Docker

## Idées d’améliorations

* Mise en place de JavaScript pour une actualisation des pages dynamique
* Mise en place d’un système de pagination
* Clusterisation des articles pour essayer de faire apparaître des catégories
* Extraction des Noms propres, classement par positivité et possibilité d’afficher les articles liés