ÉCOLE CENTRALE LYON

MOS 4.4 Nouvelles Technologies de l'Information et de la Communication

Rapport méthodologique de veille NLP

Élève :
Salaheddine MESDAR

Enseignants :
Daniel Muller
Nicolas Jardin

Tuteur :
Mohsen Ardabilian

12 mars 2021

Table des matières

1 Introduction

Dans le cadre du MOS4.4 « Nouvelles technologies de l'Information et de la Communication » l'école Centrale de Lyon vise à former ses élèves ingénieurs à développer des techniques de recherche d'informations et d'en choisir les plus pertinentes vue le grand nombre d'informations publiées sur Internet chaque jour. Les outils présentés dans ce MOS permettent de faire une bonne veille technologique concernant un sujet particulier. Le sujet que j'ai choisi est le traitement du langage naturel ou Natural Language Processing (NLP) en anglais. Le NLP est l'ensemble des techniques visant à modéliser et à reproduire la capacité humaine à produire et à comprendre des énoncés linguistiques. L'intérêt d'une telle discipline se manifeste dans la possibilité d'avoir de nouvelles informations grâce à l'explosion des données disponibles en ligne. Mais cette tâche est loin d'être facile vue que le NLP nécessite de grandes quantités de données, et celles-ci doivent être de bonne qualité et non biaisées en plus de la complexité des règles puisque certaines de ces règles peuvent être très abstraites. Par exemple, dans le cas d'utilisation d'une remarque sarcastique pour faire passer un message subtil, c'est très difficile pour une machine actuelle de percevoir de telles nuances. Le NLP regroupe l'analyse lexicale, l'analyse syntaxique, l'analyse sémantique et l'analyse pragmatique et il a plusieurs applications dont l'utilisation dans les assistants vocaux comme Google assistant, Alexa d'Amazon et Siri d'Apple ainsi que l'utilisation dans les moteurs de recherche et pour la traduction automatique.
Ce rapport présente la démarche suivie pour réaliser ma veille sur le NLP et qui passe par plusieurs étapes y compris le ciblage, la collection et la sélection des informations pertinentes, la curation et la diffusion des ses informations. Le dispositif de veille mis en place qui illustre cette démarche est présenté vers la fin du rapport.

2 Sources surveillées

Durant cette étude de veille, j'ai essayé de varier les sources d'informations pour pouvoir collecter le maximum d'informations pertinentes. Les différentes sources d'informations que j'ai utilisé sont :

- Google Alertes
- Google scholar
- Twitter
- Medium
- Scoop.it

3 Mots-clés et requêtes utilisés

Pour avoir des résultats pertinents concernant ma recherche, j'ai essayé de faire des recherches par mots-clés dans les différents outils utilisés. "NLP", Natural language processing et "Traitement automatique du langage naturel" sont les principaux mots-clés qui ont donné des résultats significatifs et pertinents sur Twitter, Google scholar et Medium et ils ont donnée lieu à plusieurs alertes aussi. Ainsi, pour avoir des Tweets pertinentes j'ai essayé de faire des recherches en utilisant ces mots-clés et filtrer par nombre de "j'aimes" ou de "retweets" et choisir celles qui ont plus de 4 "retweets" par exemple.
Une autre méthode intéressante permettant de collecter les information est l'utilisation

des requêtes. J'ai essayé d'utiliser des requêtes comme "allinurl :Twitter lists NLP" pour avoir des listes qui regroupent des tweets sur le NLP ou des requêtes comme "«NLP » OR « Natural language processing» OR «Traitement automatique du langage naturel»" sur Google.

4 Outils de collecte d'informations

Pour ne pas passer par toutes les sources possibles dans la collecte des informations, j'ai utilisé Twitter et des alertes.

4.1 Alertes

J'ai choisi de recevoir des alertes comme premières sources des mes informations , car elles permettent de suivre l'actualité et d'en choisir les plus pertinentes en utilisant des filtres. J'ai utilisé pour cela des alertes de Google Scholar étant donné la quantité d'articles scientifiques publiés régulièrement dans le domaine scientifique. J'ai crée donc une alerte avec le mot clés NLP et j'ai choisi de recevoir les notifications relatives à cette alerte à mon adresse mail comme indiqué dans la figure suivante :

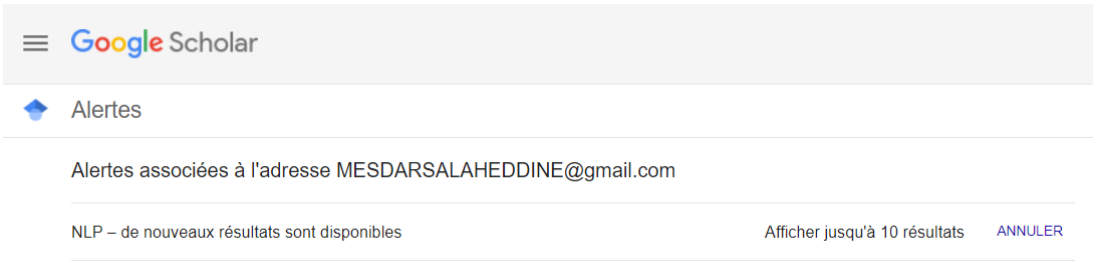

FIGURE 1 – Alerte Google Scholar

Google propose aussi le service Google Alertes qui permet d'avoir des notifications concernant un sujet en envoyant ces notifications à une adresse mail. J'ai crée donc deux alertes avec les mot-clés "NLP" et "natural language processing" et j'ai configuré l'alerte pour avoir les meilleurs résultats des publications concernant le NLP et de les recevoir dès qu'ils sont présents comme indiqué dans la figure suivante :

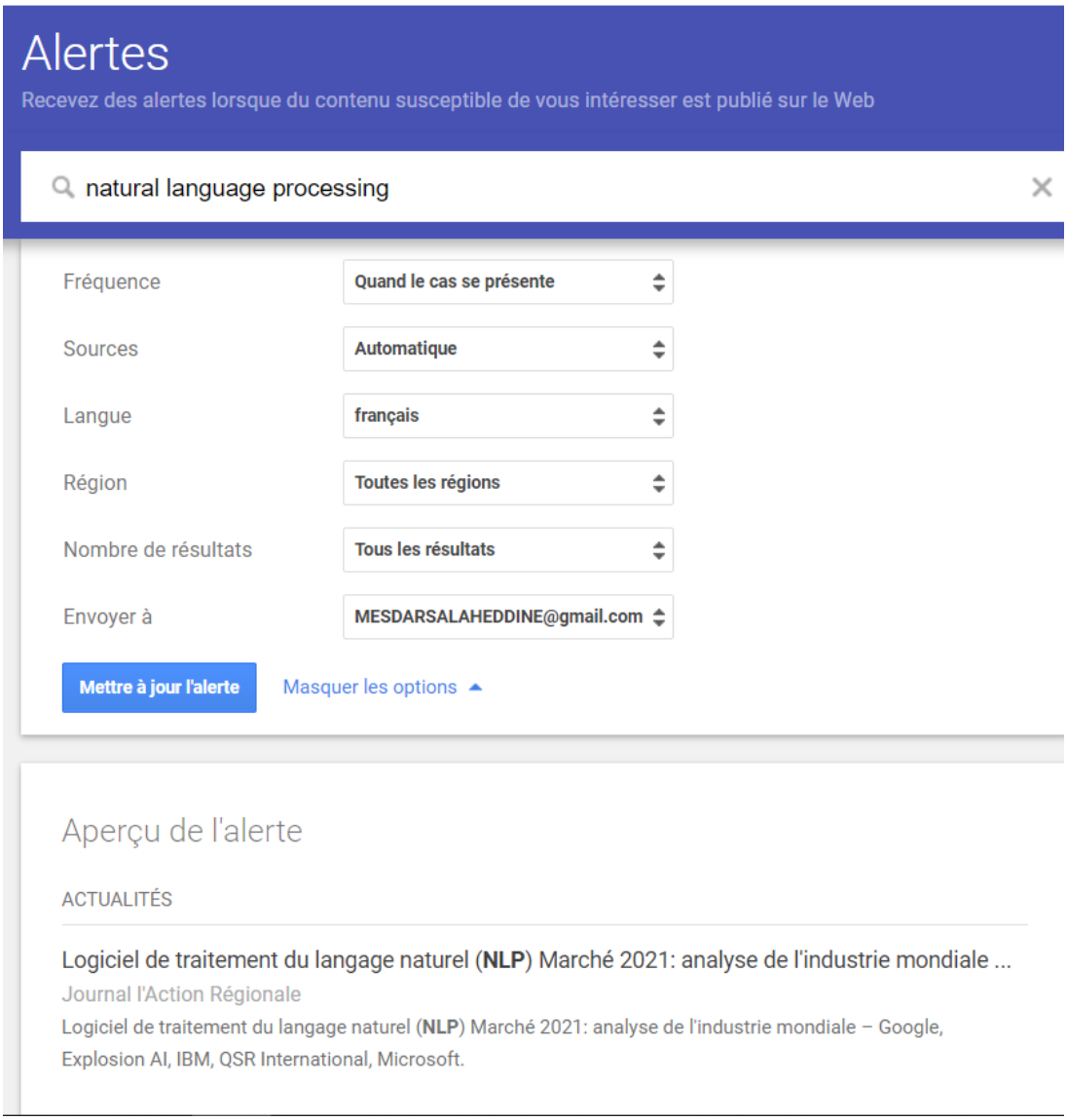

FIGURE 2 – Alerte Googl Alertes

4.2 Twitter

J'ai utilisé Twitter aussi pour ma collecte d'informations et j'ai essayé de filtrer les résultats par nombre de "j'aimes" et de "retweet" pour sélectionner les meilleurs résultats. J'ai suivi aussi des listes sur Twitter qui regroupent des gens intéressés par le NLP et qui publient régulièrement sur le sujet.

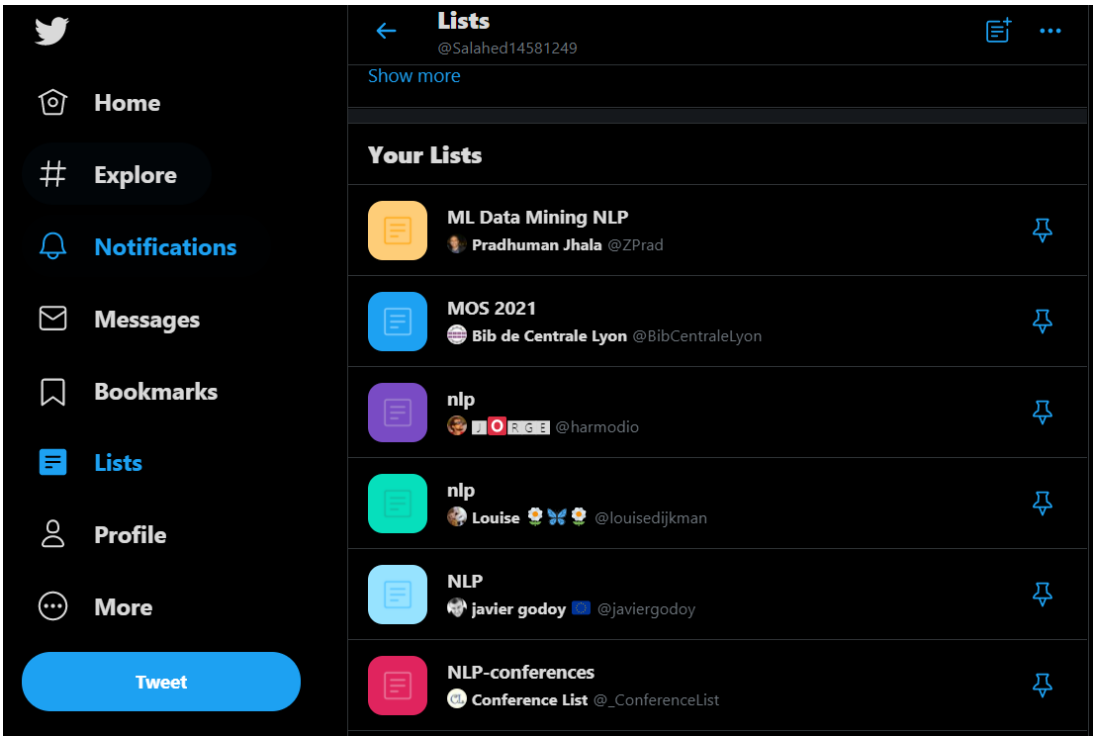

FIGURE 3 – Listes Twitter

J'ai utilisé aussi Tweetdeck qui permet d'avoir plusieurs colonnes et donc faire plusieurs recherches et avoir des résultats dans le même écran pour pouvoir les comparer.

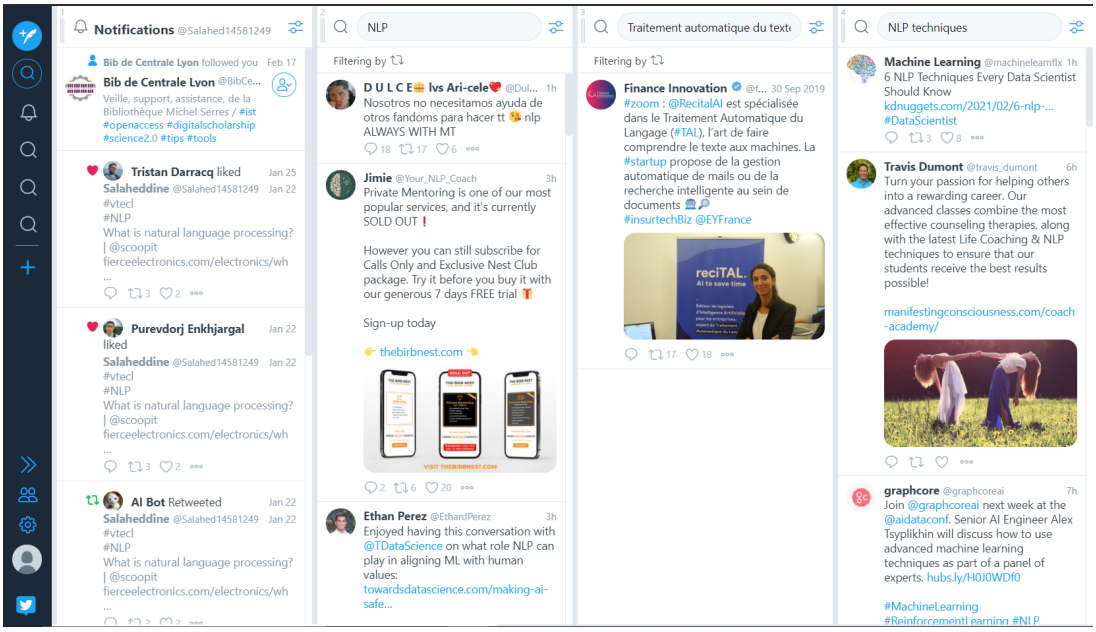

FIGURE 4 – Tweetdeck

4.3 Medium

J'ai décidé de chercher dans la presse spécialisée via Medium pour avoir des blogs contenant des informations qui pourraient être intéressantes dans le domaine du NLP.

5 Outils de curation

Pour cette étape de curation et d'analyse de données collectées, j'ai travaillé avec deux outils : Diigo et Scoop.it.

5.1 Diigo

Mon principal outil de curation et d'analyse de données collectées est Diigo. Cet outil permet d'avoir une base d'articles et de publications dans des sites web, des pdf, des images et des liens qui contiennent des informations correspondant à ma recherche. Il offre aussi la possibilité de regrouper les documents par tag et d'ajouter des titres, des descriptions et des commentaires aux documents pour les retrouver facilement.

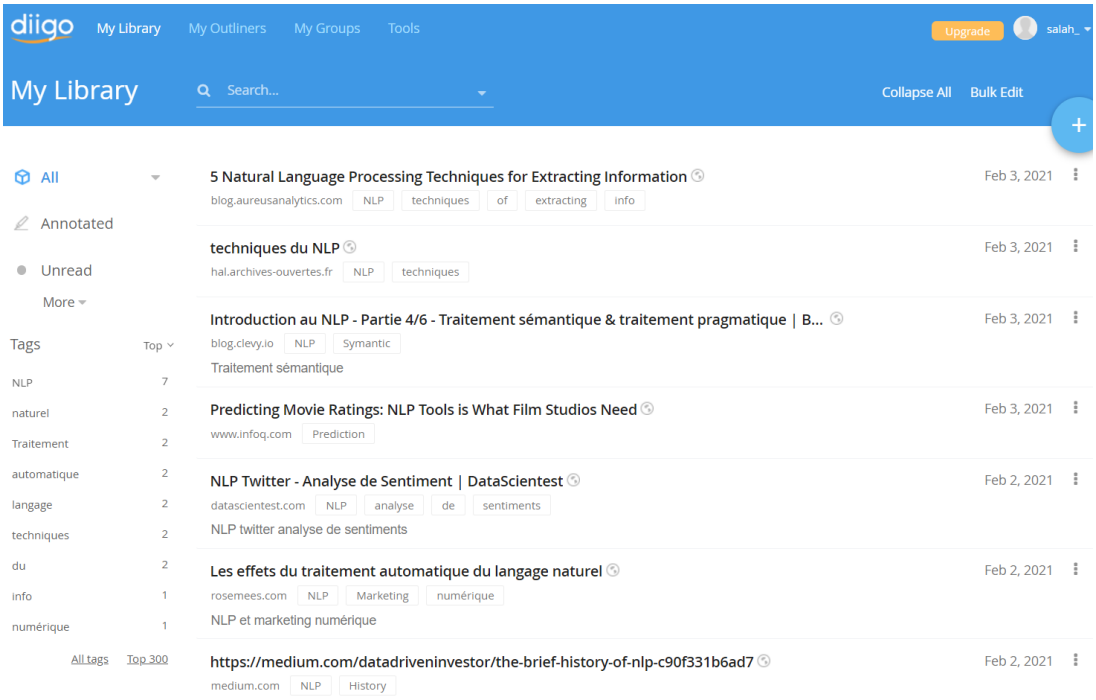

FIGURE 5 – Ma bibliothèque sur Diigo

5.2 Scoop.it

J'ai choisi de travaillé avec Scoop.it en plus de Diigo car il permet d'avoir un tableau de bord regroupant les différentes publications qu'on trouve sur le site et parce qu'il permet aussi de partager le contenu directement sur Twitter.

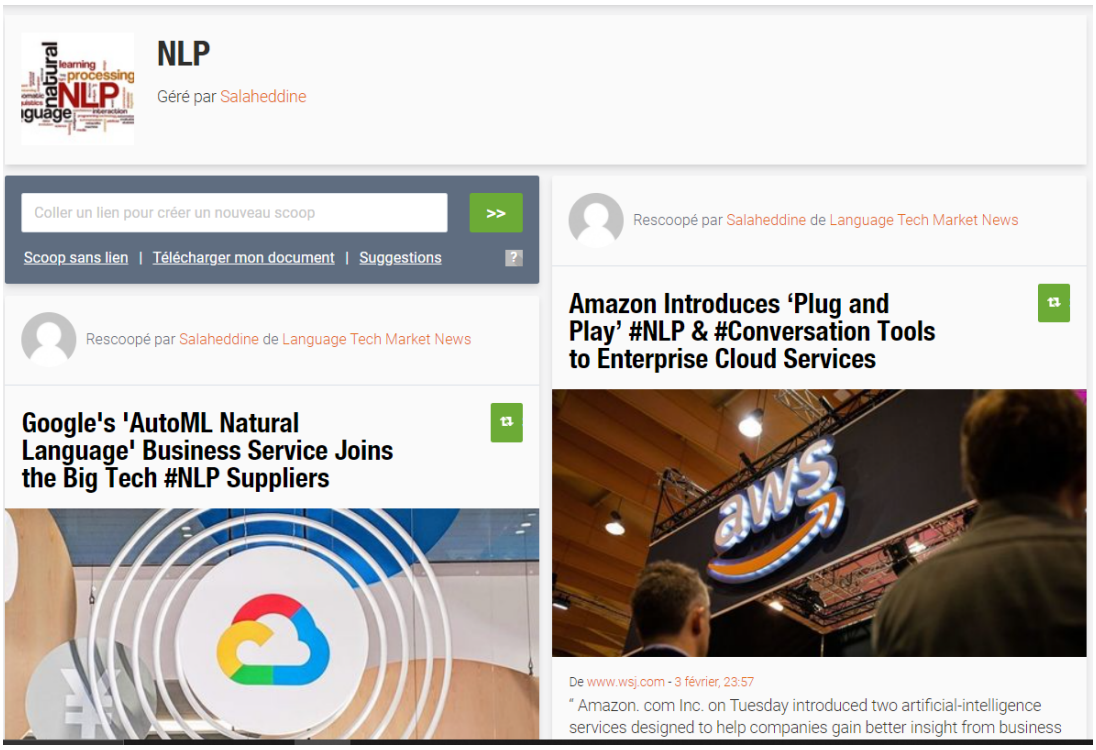

FIGURE 6 – Mon tableau de bord sur Scoop.it

6 Diffusion

Après avoir collecter les informations issues des différentes sources citées et les analyser par les outils de curation, vient l'étape de diffusion. Pour cela j'ai utilisé Twitter(@Salahed14581249) pour publier des retweets ou des tweets contenant des informations pertinentes sur le sujet du NLP.

7 Conclusion

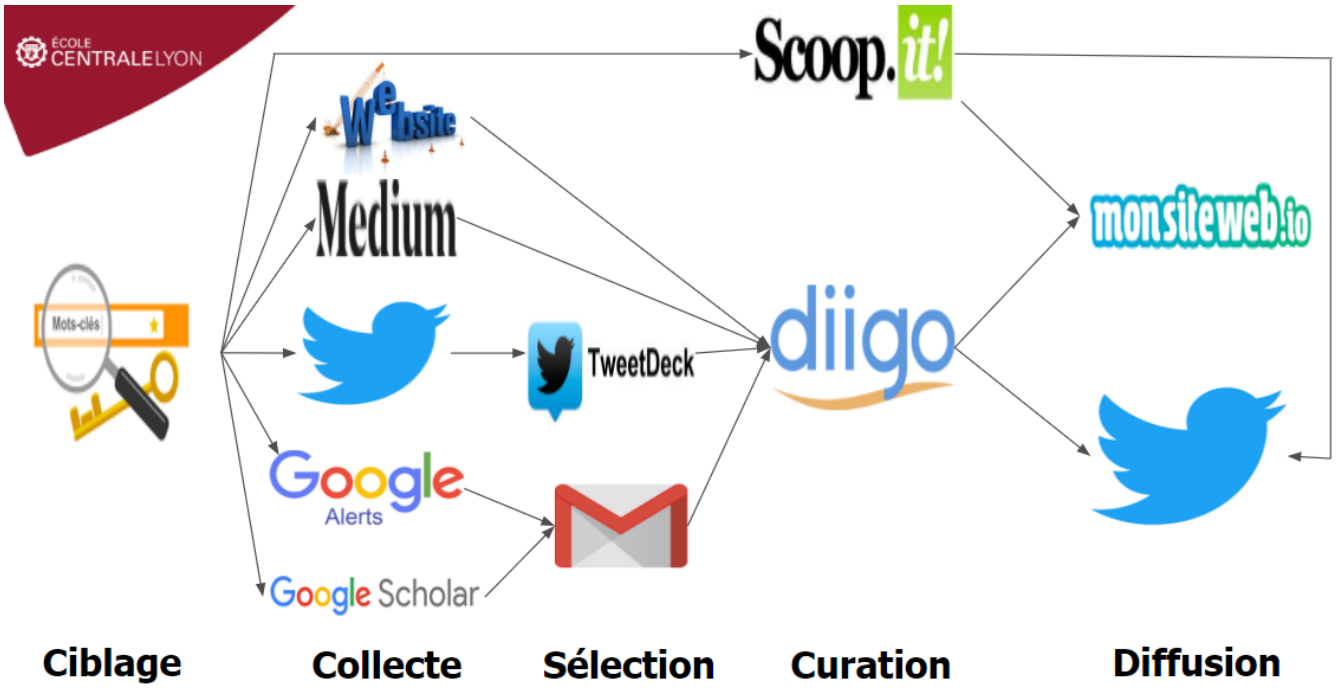

FIGURE 7 – Mon dispositif de veille

Ce dispositif présente les différentes sources surveillées pour la collecte des données et les outils de curation utilisés pour l'analyse des ces données dans le cadre de ma veille sur le NLP. Le dispositif mis en place permet d'avoir des informations pertinentes sur le sujet étudié et en se basant sur les méthodes présentées j'ai pu développé un site Web résumant les résultats trouvés et qui est accessible via ce lien : `https://mesto00.github.io/Veille_NLP/`