65, rue des Grands Moulins, 75013 PARIS

Méthodologie du projet d'extraction et de la valorisation d'informations issues de deux plateformes

AUFFRET FABIENNE

Institut National des Langues et Civilisations Orientales (I.NA.L.C.O)

Master 2 TAL IM- avril 2021

Sommaire

Présentation du projet

Le projet porte sur la présentation de données pertinentes pour permettre au client de décider de l'intérêt de s'impliquer à la fois dans le développement d'hôtels respectueux de l'environnement, ainsi que dans une plateforme offrant des ressources linguistiques et culturelles sur les nombreuses langues du continent africain.

Les accords de Paris sur le climat reconnaissent le rôle du secteur privé concernant les actions visant à réduire les émissions.

Il est donc éthique et d'avenir d'investir dans un secteur non polluant, et pourquoi pas les hôtels ? Le tourisme éthique est en pleine expansion et on peut s'attendre à un effet rebond post-covid, avec une prise de conscience de l'impact concret du changement climatique sur tous.

Ils reconnaissent aussi qu'il faut soutenir et promouvoir la coopération régionale et internationale : ce qui n'est pas possible sans communication et outils modernes du digital (pour apprendre les langues par exemple), ce à quoi peut contribuer une plateforme de type Ntealan. Elle peut aussi rentrer dans le cadre de l'initiative sur la protection du patrimoine culturel pouvant être mis en danger à cause du réchauffement climatique (http://climateaction.unfccc.int/views/cooperative-initiative-details.html?id=133).

Méthodologie utilisée

Accès aux données

Avant d'extraire le contenu d'un site, il convient de vérifier qu'on y est autorisé. En premier lieu, nous avons donc consulté les fichiers robots.txt des deux sites :

https://www.nh-hotels.fr/robots.txt :

Ni le user agent, ni les pages que nous comptons consulter (qui commencent par :https://www.nh-hotels.fr/hotels/) ne sont interdites.

Nous consultons également la partie « mentions légales » du site (https://www.nh-hotels.fr/legal-notice). Pour ce qui concerne le point 6 (conditions d'utilisation du site), nous respectons ces conditions. Le point 4 qui concerne la propriété intellectuelle touche la diffusion publique du contenu du site, ce que nous ne faisons pas. Il s'agit ici de communication privée avec le client.

Pour le site Ntealan, le robtos.txt ne contient que :

```
ser-agent: CuteStat
Disallow: /
```

Nous pouvons donc récolter des articles.

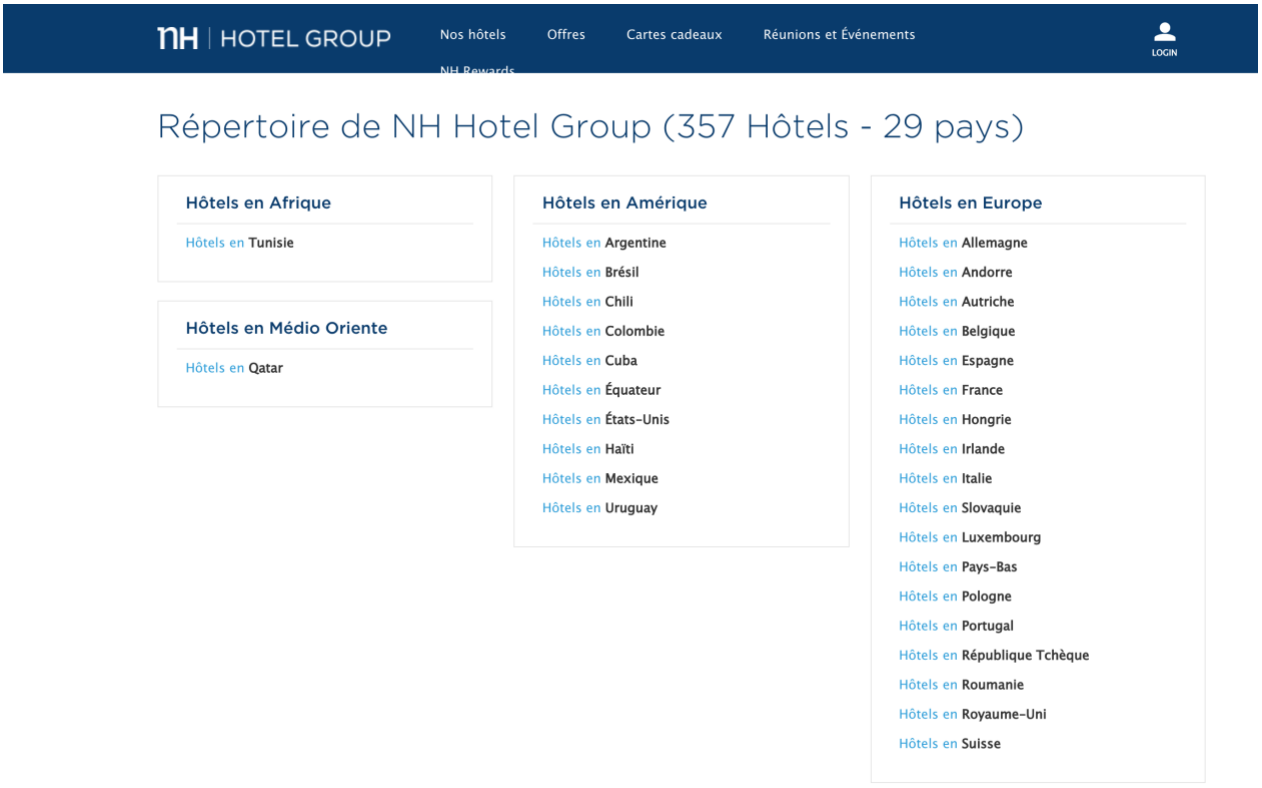
Observation des sites

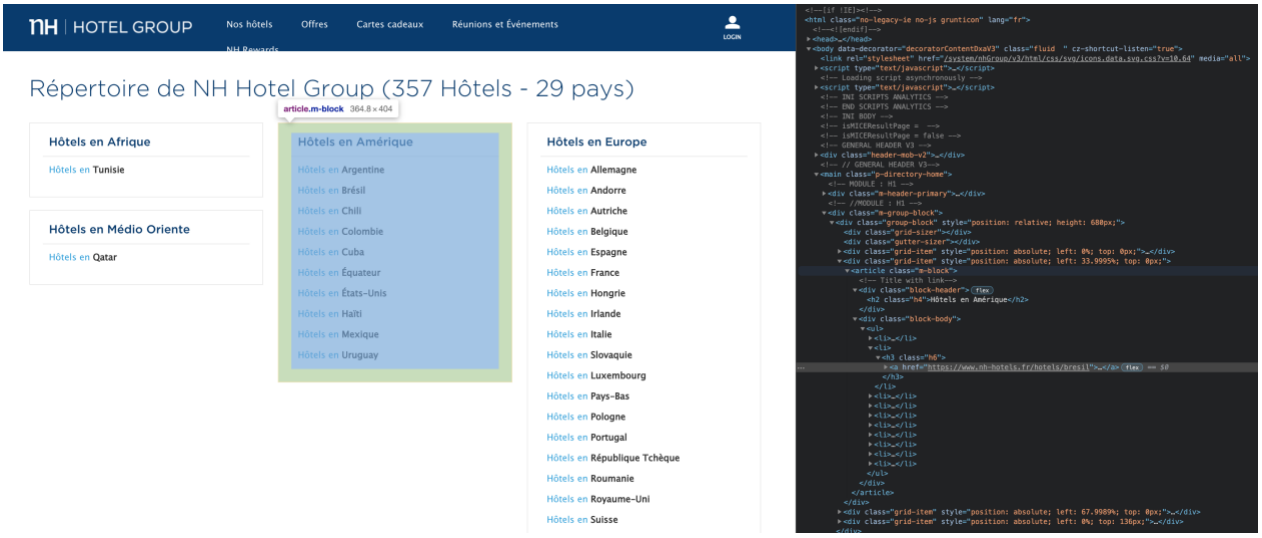
Nous avons consulté les deux sites, noté quelles informations nous pouvions extraire, les actions à faire pour y accéder et observé leurs comportements lors de cette séquence d'actions. Puis nous avons inspecté chaque page pour repérer les éléments à extraire, soit avec leur classe, leur id, leur Xpath parfois quand c'était plus pratique.

NHhôtels :

C'est un site MPA, on peut donc extraire l'information aussi bien avec Beautiful Soup que Sélénium, les principes sont équivalents. Nous avons choisi Sélénium.

La page qui sera notre point de départ est : https://www.nh-hotels.fr/hotels. On va donc commencer par noter tous les liens présents pour les parcourir un à un :



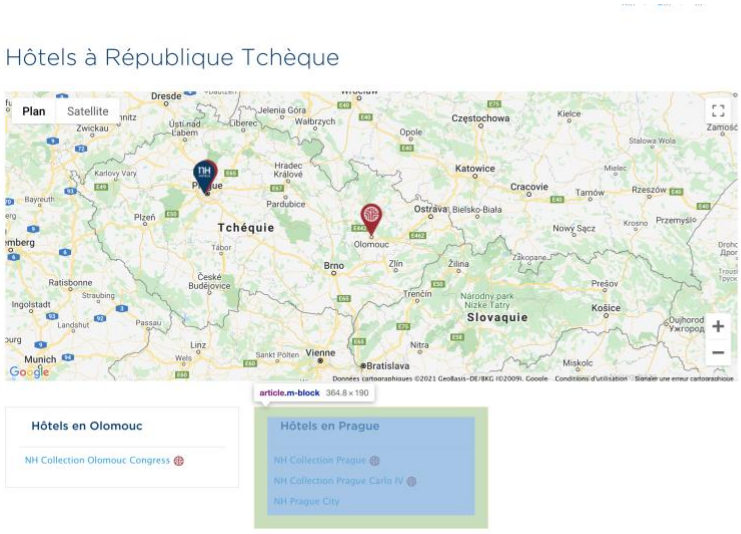


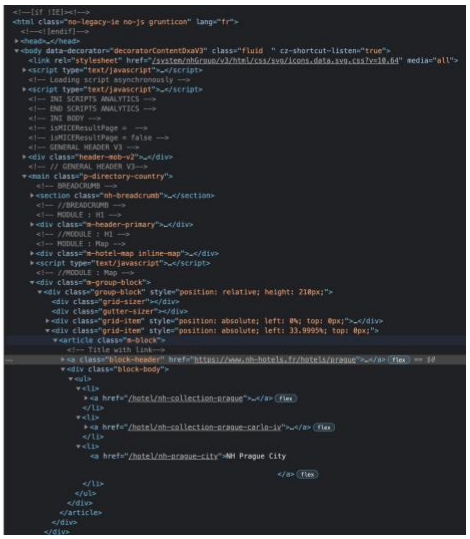
On voit que chaque zone (le Moyen-Orient n'est pas un continent) sont dans un div de class « mblock » que le nom de la zone est dans un tag « h4 » de cet élément et les liens dans un tag « a ». On les sélectionne et stocke une liste de liens par zone, tout en stockant le nom du pays qui est dans un span « strong » dans le texte, mais nous choisissons simplement de construire une liste

avec le texte et d'extraire le dernier élément (le pays). À ce stade nous avons un dictionnaire avec les liens par pays et par zones.

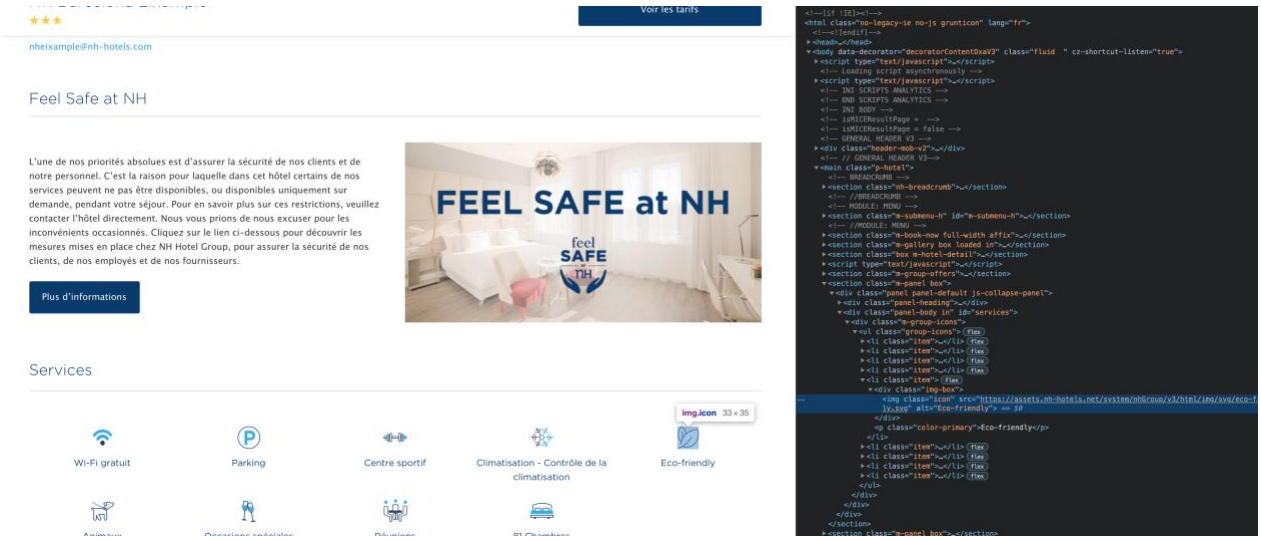
Nous les parcourons donc par pays pour collecter les liens des villes.

Nous parcourons cette liste de liens pour récupérer de la même façon les liens conduisant aux villes :

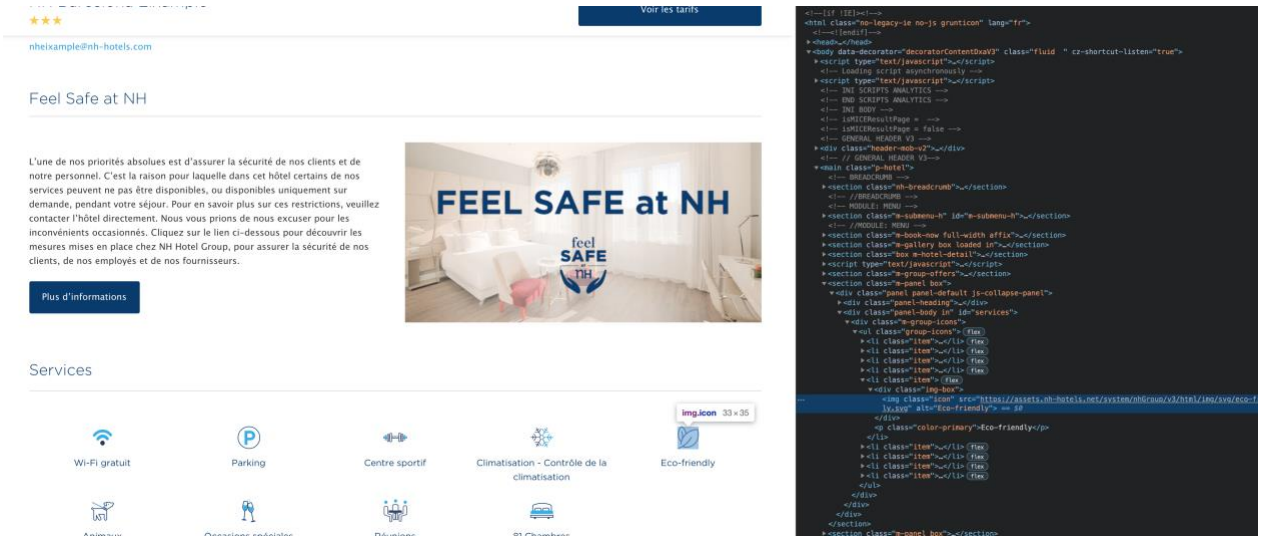




Ils sont eux aussi dans un élément de classe « block -body » contenu dans un élément de classe « m-block » : on extrait les liens (attribut href) pour les stocker et les parcourir à l'étape suivante :



Ici on peut trouver la mention eco-friendly et le nom de l'hôtel :



On récupère donc si l'hôtel (son nom est dans le tag « h1 ») est éco-friendly ou non (si dans les textes de « servies » il y a un « eco-friendly » et on peut enfin construire le dictionnaire des hôtels. On a choisi de d'abord faire un dictionnaire arborescent qui peut être évolutif (il y a d'autres données que l'on peut prendre en compte, le nombre d'étoiles de l'hôtel, sa capacité d'accueil, la présence de certains services ou

d'un label (iso par exemple). Sa structure est {zone : pays : {ville :[nom_hotel1, éco ou non], :[nom_hotel2, éco ou non]]}}}. La liste finale peut être donc être modulée à loisir. Lors de la dernière étape

À partir de là, on peut construire toute forme de json pour sauvegarde et une exploitation ultérieure en dataframes pour la visualisation.

Plateforme Ntealan :

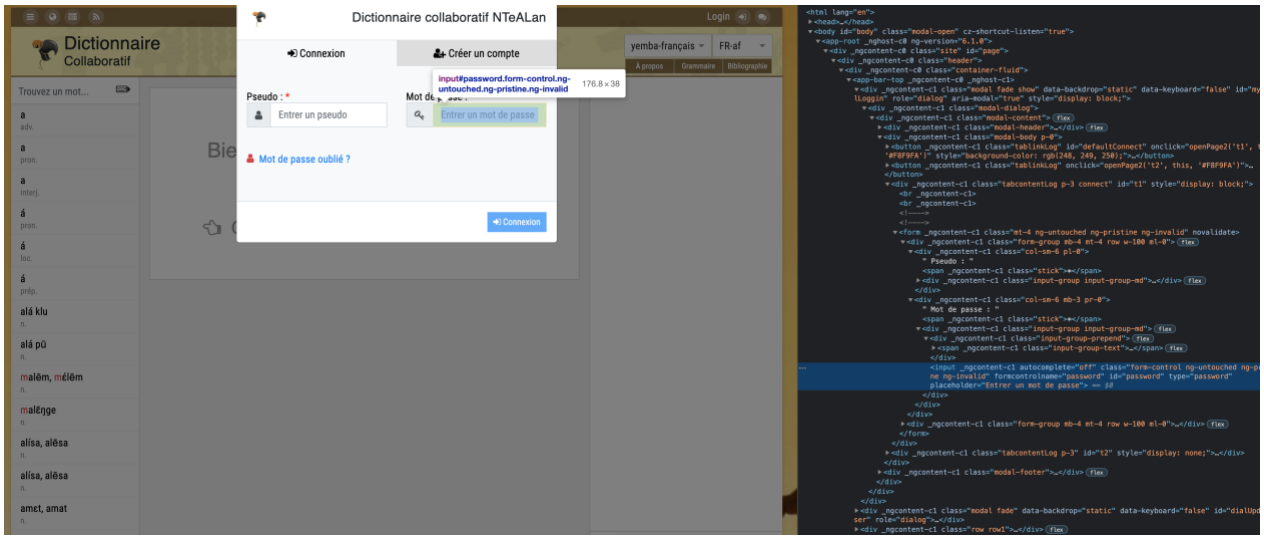
Nous partons de la page : https://ntealan.net/dictionaries/content/fr-af/yb_fr_3031

C'est un site SPA. Il faut utiliser selenium pour pouvoir le parcourir.

Le dictionnaire sélectionné par défaut est yemba-français. Il y a 24 dictionnaires dans le menu déroulant (en haut à droite), mais seuls 17 sont actuellement actifs. Lors du scraping, j'ai finalement choisi de faire une liste des dictionnaires actifs et de la parcourir (évite les try/except avec des attentes à chaque fois).

Il faut en premier lieu fermer le pop-up covid (il se rouvre aussi quand on change de dictionnaire) et se loguer, car le site limite le nombre d'articles consultés si on n'est pas enregistré.

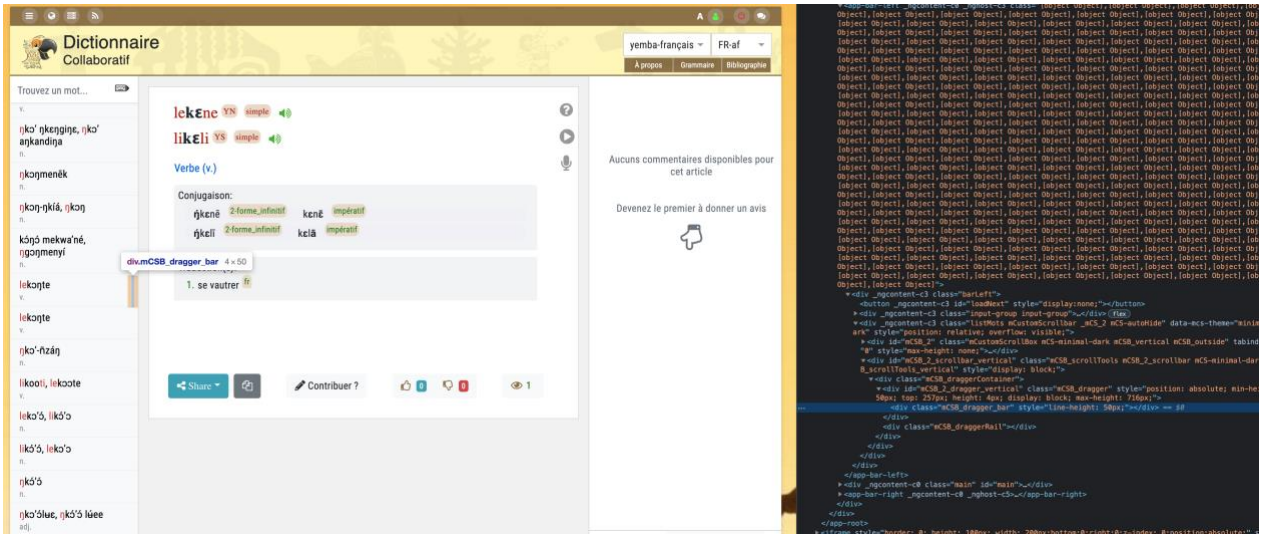
Exemple de détermination des id à choisir (pseudo et password) après avoir sélection l'élément « float-right » et cliqué dessus (login) :



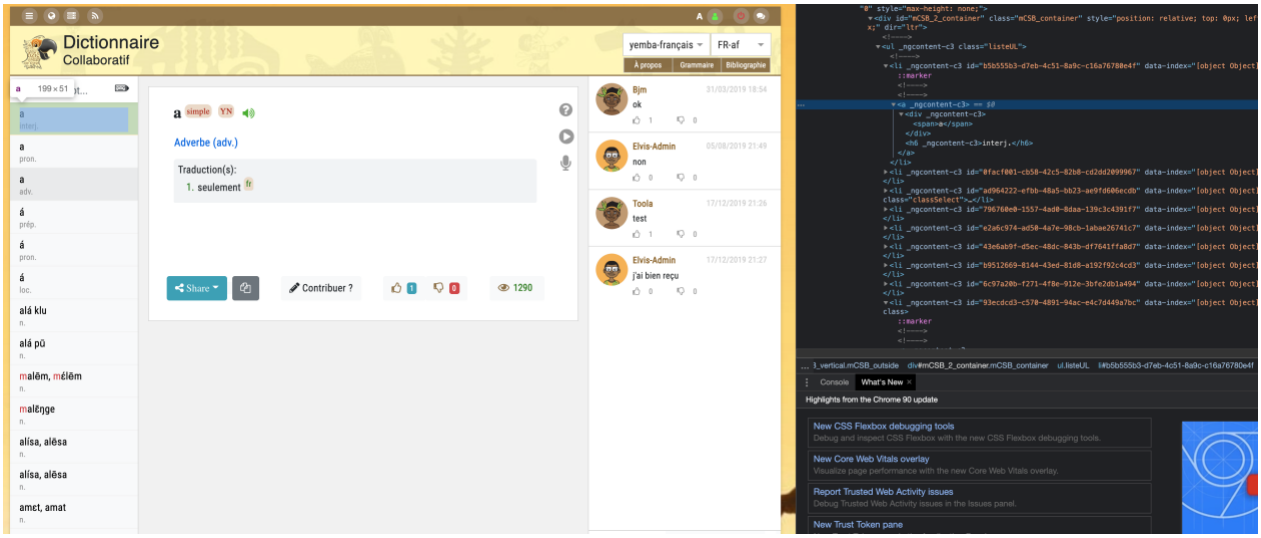
Note : parfois le click s'exectue mieux avec javascript : driver.execute_script("arguments[0].click();", element)

Ensuite il s'agit de scroller à gauche de mot en mot, il faut aller dessus, puis cliquer. Comme sélénium ne « voit » pas l'élément s'il n'est pas dans la fenêtre, nous utilisons : driver.execute_script("arguments[0].scrollIntoView(true);", element).

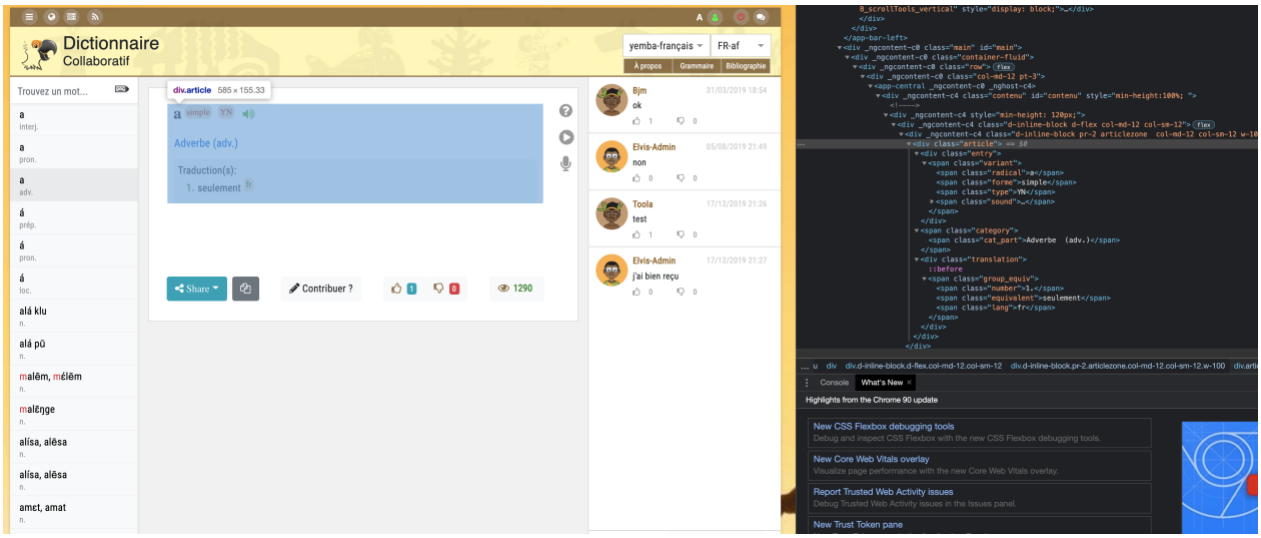
Note : le scroll jusqu'en bas (ou par sauts) aboutissent à un message d'erreur (ce dictionnaire n'a pas encore d'ari-ticles) aussi bien avec des send_keys(Keys.PAGE_DOWN) qu'avec une tentative par drag and drop :



On détermine comment accéder à chaque élément en inspectant (ici les zones cliquables du mot, son contenu dans le div et son pos abrégé dans le h6) :



On choisit de récupérer certaines les informations (pos in extenso, préfixe, suffixe, audio si présents, dans la partie centrale (article). :



Il est à noter un comportement erratique de sélénium lors du scraping (aléatoirement message « stale element reference» alors qu'au test précédent et au suivant on ne rencontre aucun problème de ce type) et que stocker un élément dans une variable avant de cliquer dessus ou le traiter (find) change aussi sa

réaction…Il faut parfois lancer le script plusieurs fois pour récupérer un dictionnaire qui n'a pas pu être collecté.

Visualisation

Utilisation

Prérequis : une machine avec Python installée. Si vous voulez visualiser les données sur streamlit en local, il faut installer streamlit puis lancer l'application avec la commande : streamlit run app.py. Cela ouvrira un navigateur en local.

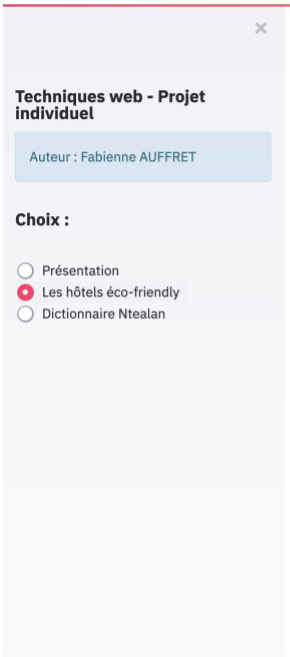
Si besoin, lancer les deux scripts de scraping : python scrap-NH-hotels.py et scrap-ntealan.py. La récolte prend un peu de temps.

Vous pouvez aussi consulter une version en ligne : https://visu-invest-eco.herokuapp.com/ ou le notebook « visualisation » qui est à la racine avec les scripts (lenteurs d'affichage parfois).

Visualisation des données

Nous proposons pour le site NH, une analyse des nombres d'hôtels par pays, par ville, par zone, par continent, en nombre absolu/ pourcentage et comparons les pourcentages d'hôtels éco-friendly et non éco-friendly.

Exemples :



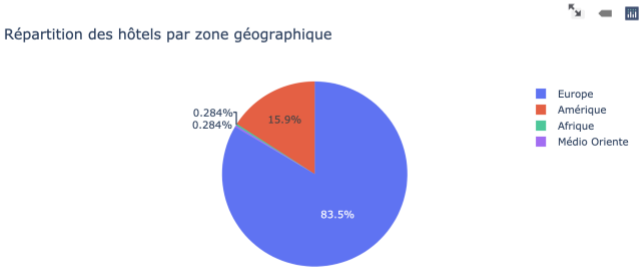
NH Collection Barranquil…	Amérique	Colombie	Barranquilla	0
NH Bogotá Boheme Royal	Amérique	Colombie	Bogotá	1
NH Bogotá Pavillon Royal	Amérique	Colombie	Bogotá	0
NH Bogotá Urban 26 Royal	Amérique	Colombie	Bogotá	0

Il y a 56.53 % hôtels eco-friendly, soit 199 en tout.

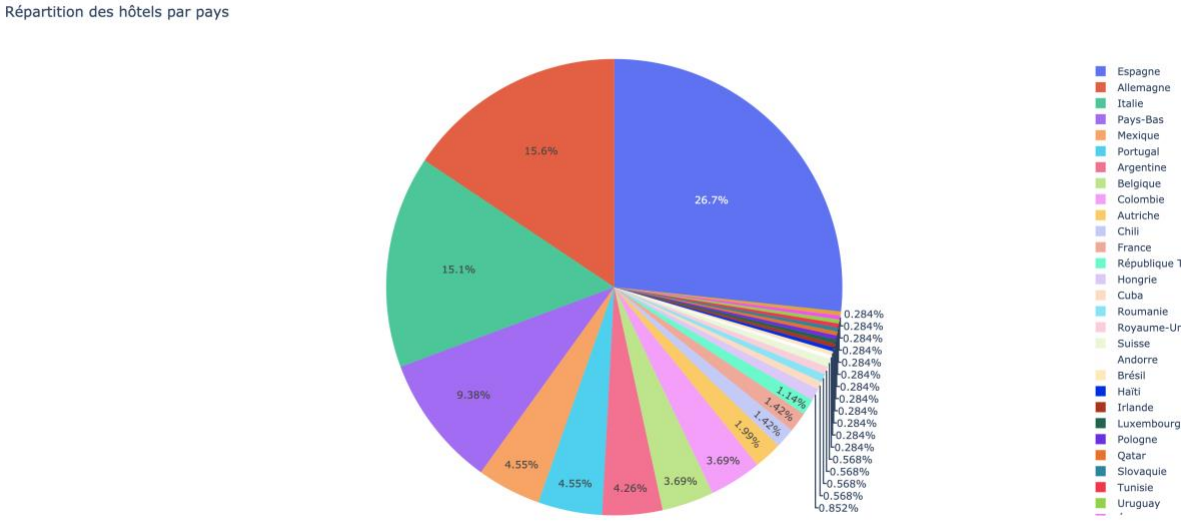
On voit que plus de la moitié des hôtels de NH Hotels ont la mention eco-friendly : ils utilisent de l'eléctricité provenant d'énergies renouvelables, optimisent leur utilisation d'eau, favorisent les déplacment en vélo ou en vég-hicules éléctriques.

(Vous pouvez interagir avec les graphiques à l'aide des boutons situés à leur coin haut droit)

Répartition des hôtels par lieux :

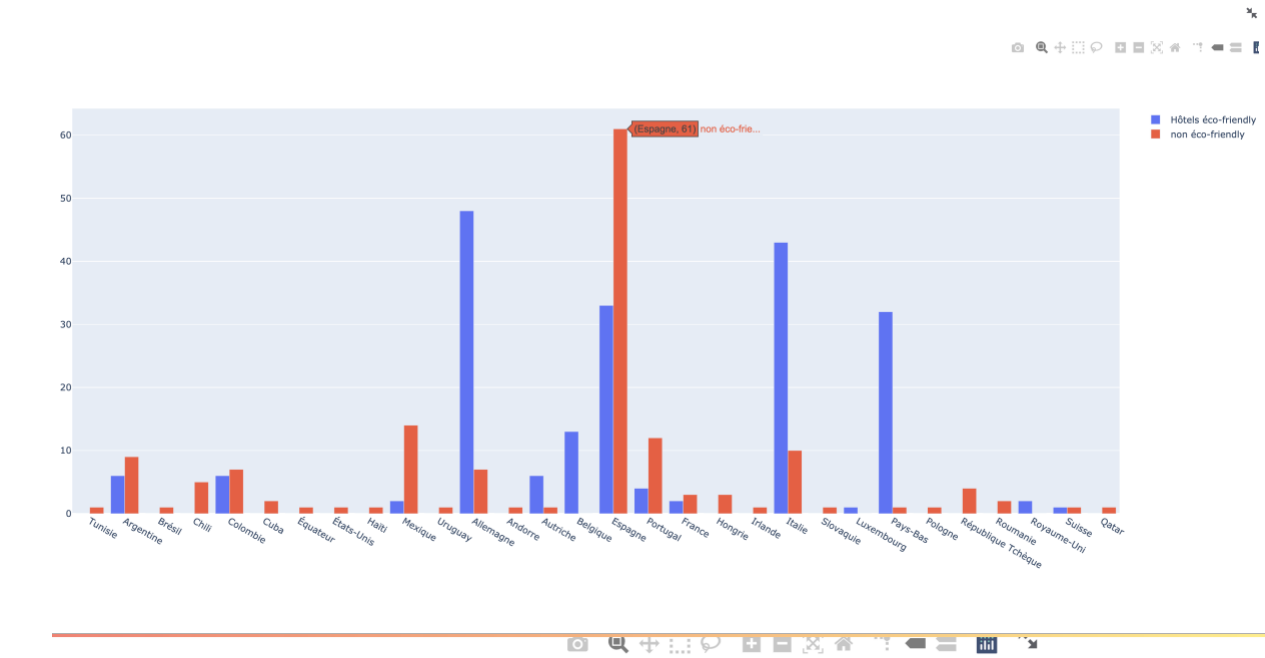


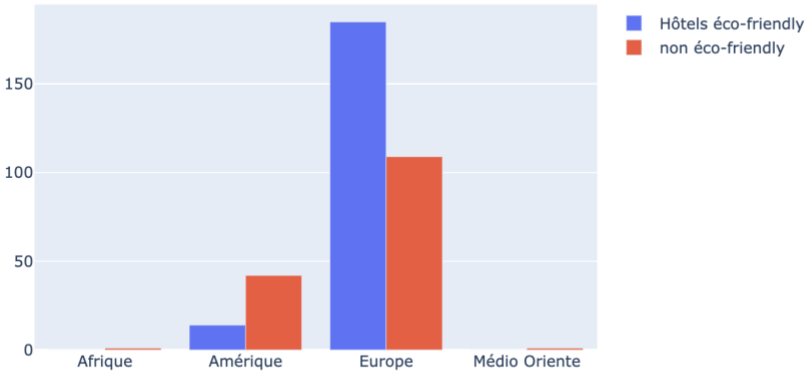
Ainsi que :

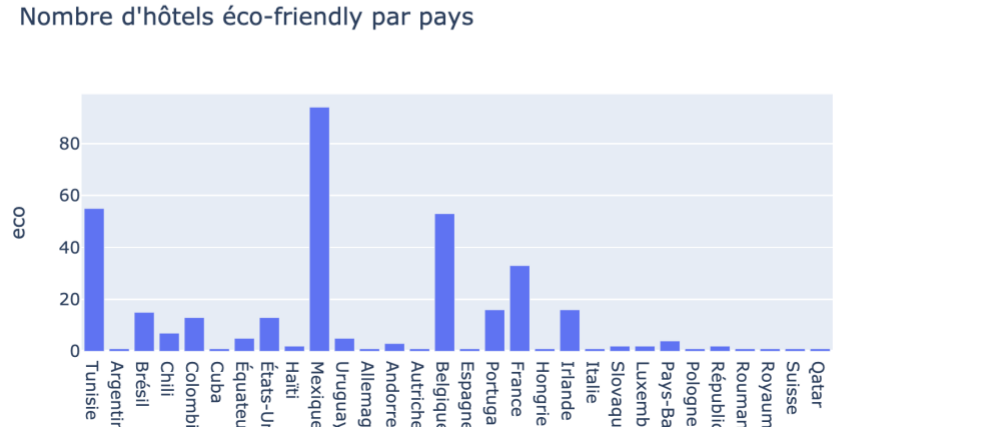


Nous avons utilisé la librairie plotly pour visualiser les données, car c'est interactif (on peut agrandir, voir les données en survolant par exemple)

Exemples de comparaisons :

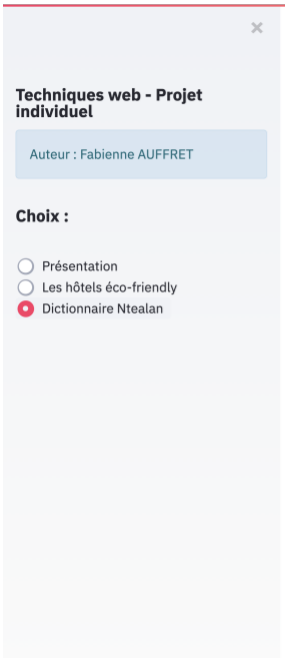






Pour le site Ntealan nous offrons la possibilité de voir des extraits des dictionnaires et de se rendre compte du potentiel énorme de développement de la plateforme si elle est soutenue.

Exemple d'affichage :



	pos	prefix	suffix	audio	traduction-fr
malēm	Nom (n.)	m		0	['sorte d'épice']
malēnge	Nom (n.)	m		1	['balafon avec neuf tons']

Extrait du dictionnaire medumba-francais :

	pos	prefix	suffix	audio	traduction-fr
ā	Verbe (v.)			0	['être']
ba	Nom (n.)			0	['biclyclette']
bág	Nom (n.)			0	['devinette']
bàkələ	Nom (n.)			0	['banane']
bam	Nom (n.)			0	['ventre']
bambwogo	Nom (n.)			0	['motocyclette']
bànbà	Verbe (v.)			0	['envouloir']
bānbán	Adjectif (adj.)			0	['zéro']
bəbwə	Verbe (v.)			0	['naître']
bəntu	Nom (n.)			0	['gens', 'personne']

Extrait du dictionnaire ejagham-francais :

	pos	prefix	suffix	audio	traduction-fr
ŋa	Nom (n.)			0	['viande']
àfònè	Nom (n.)			0	['proprietaire']
a fóɣō	Nom (n.)			0	['urine']
ájé	Nom (n.)			0	['les feuilles']
àjiŋ	Nom (n.)			0	['le foie']
àkāŋ	Nom (n.)			0	['le sel']
àkàp	Nom (n.)			0	['les fruits']
àkát	Nom (n.)			0	['les pieds']
ake	Conjonction (conj.)			0	['mais']
aku	Nom (n.)			0	["l'huile"]
akūk	Nom (n.)			0	['le porc']

Extrait du dictionnaire hotels2 :