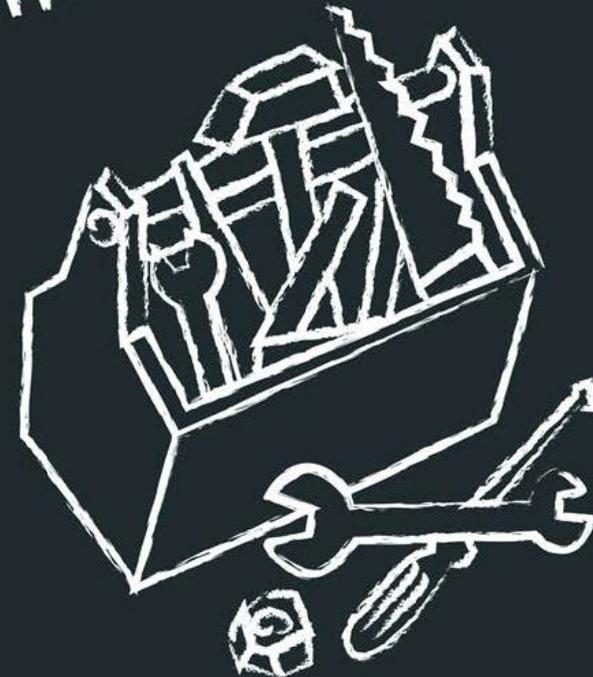


LES TRUCS ET ASTUCES DE LA PLATE-FORME TECHNOLOGIQUE



Web scraping

Virginie Lethier

virginie.lethier@univ-fcomte.fr

PROGRAMME SUR
[HTTPS://MSHE.UNIV-FCOMTE.FR](https://mshE.univ-fcomte.fr)



Pourquoi cette formation? Pour qui?

- Acquérir des données textuelles pour une recherche, un mémoire, pour organiser un projet (adresses mail, numéros de téléphone, liste d'articles) et/ou une veille informationnelle, pour archiver...
- Aucune compétence en programmation requise



Web scraper ? Harvesting ?

- Utiliser des « robots » pour « aspirer » des sites web
- Un **web scraper** est un **programme informatique** qui lit le code html des pages web pour **extraire** des données qui sont présentées sur des sites **web**.



Web scraper, c'est hacker ?

- ❑ Scraper n'est pas une activité illégale en soi
- ❑ Mais une utilisation des données webscrapées peut aller contre l'éthique ou être véritablement illégale...

Rappel : toujours lire et respecter les conditions générales d'utilisation (CGU) et les conditions générales de vente (CGV) avant de web scraper .

- ❑ Rappel des bonnes pratiques et des problématiques liées aux données de la recherche dans un contexte d'open data :

<http://www.oecd.org/fr/sti/inno/38500823.pdf>

<http://www.donneesdelarecherche.fr/>



Image : [UnderNews Actu](#)



De nombreux outils disponibles

- ❑ Parmi lesquels : Import.io, Scrapy, Outwit Hub, Gromoteur
- ❑ **Webscraper** est une **extension** disponible sous **Google Chrome** qui permet d'extraire les données d'un site internet très rapidement.



(1) Je télécharge sur mon ordinateur l'extension de Google Chrome : <http://webscraper.io/>

Web Scraper

Web Scraper is a company specializing in data extraction from web pages. We offer 2 great options: Google Chrome Web Scraper Extension, and cloud based Web Scraper.

Web Scraper Extension (Free!)

Using our extension you can create a plan (sitemap) how a web site should be traversed and what should be extracted. Using these sitemaps the Web Scraper will navigate the site accordingly and extract all data. Scraped data later can be exported as CSV.

[Download Free on Chrome Store »](#)

Cloud Web Scraper

Cloud Web Scraper offers top quality results. This option allows you to extract large amounts of data at once, and even run them on a set schedule.

[Read more about the Cloud Web Scraper](#)



(2) Je vais sur le site dont je souhaite aspirer le contenu

← → ↻ 🏠 ⓘ <https://www.fdesouche.com> ☆ 📅 11

SE CONNECTER PROPOSER UN ARTICLE CONNECTÉS: 1544 VISITEURS UNIQUES: 🔄

 **FDESOUCHE.COM EST UNE REVUE DE PRESSE**

ACCUEIL SOCIÉTÉ SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INSOLITE FDÉSINTOX

BÉZIERS : UNE ESCLAVE « BLANCHE » EXPLOITÉE, HUMILIÉE ET MALTRAITÉE PENDANT DES ANNÉES PAR UNE GUINÉENNE

Migrants en Méditerranée : Castaner reconnaît à son tour la complicité de certaines ONG avec les passeurs (MàJ)

Meurtres d'Élise un Rwandais int territoire français 2011 condamné réclusion crimin perpétuité (MàJ)

Royaume-Uni : après avoir infiltré des agences publiques, des musulmans ont détourné des milliards dont une partie pour

#JusticepourAng serait décédé ap ingéré de la dro grande quantité, des « violences



FDESOUCHE.COM EST UNE REVUE DE PRESSE

ACCUEIL SOCIÉTÉ SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INSOLITE FDÉSINTOX

Accueil / Mentions légales

Mentions légales

Ce blog est édité par Tilak RAJ.

1203/10 Govind Puri, *kalkaji*, New Delhi 110019, Inde

Pour toutes questions, requêtes ou réclamations vous pouvez contacter l'éditeur par courriel à l'adresse suivante : fdesouche@india.com

Ce site participe au Programme Partenaires d'Amazon EU, un programme d'affiliation conçu pour permettre à des sites de percevoir une rémunération grâce à la création de liens vers Amazon.fr.



(3) J'analyse la présentation du site

← → ↻ 🏠 ⓘ <https://www.fdesouche.com/category/societe>

ACCUEIL **SOCIÉTÉ** SÉCURITÉ POLITIQUE ÉCONOMIE MONDE ÉCOLOGIE CULTURE SPORT

Comment le détournement satirique d'une affiche du gouvernement devient une « provocation à un crime ou délit » Le 7 mars dernier, le compte twitter...

Gilets Jaunes : Philippe Pascot appelle à la «Guérilla urbaine» pour réagir au Grand Débat
Accrochage en direct dans « Morandini Live » quand Philippe Pascot, Gilet Jaune, appelle « à la guérilla urbaine » la semaine prochaine pour réagir au Grand Débat Morandini

Tous les journalistes de «Libération» sont-ils des enfants de CSP+ ?
Après la publication d'une enquête sur la défiance des journalistes, vous avez été nombreux à nous demander si les journalistes de «Libé» étaient tous...

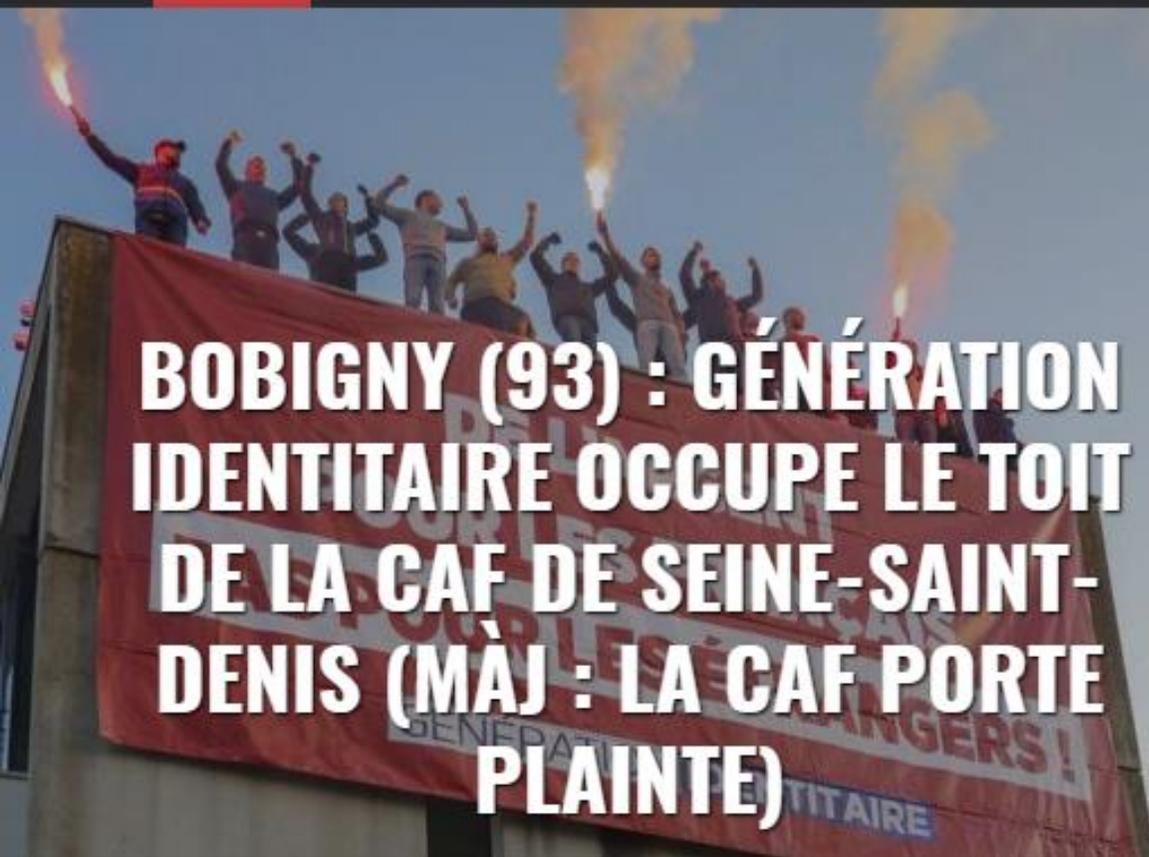
Page 1 / 2083 1 2 3 4 5 Suivant > Dernier >

J'observe la variation de l'adresse liée à la rubrique « Société »

J'observe le nombre de pages...

FDESOUCHE.COM EST UNE REVUE DE PRESSE

ACCUEIL **SOCIÉTÉ** SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INS



La bataille de Ver... n'apparaît plus da... nouveaux program... lycée annoncés p... rentrée 2019 (Mà... Blanquer dément)

« Pendez les Blan... rappeur Nick Con... condamné à 5000... d'amende avec su

J'analyse l'adresse de la page 5 de la rubrique « Société »



Nice (06) : 14 prévenus d'un trafic de drogue remis en liberté après une erreur de la justice

L'affaire semble incroyable et pourtant. 14 prévenus, dont un trentenaire soupçonné d'être un poids lourd du trafic de drogue à Nice, ont obtenu leur remise...



Ivry (94) : harcelé, le lycéen se venge à coups de marteau en pleine classe

Il a assuré aux policiers qu'il vivait un enfer depuis le début de l'année scolaire. Harcelé mais aussi humilié par d'autres élèves à travers des...



Saint-Etienne (42) : des sans-abri, en majorité des migrants, occupent la Bourse du Travail, la CGT dénonce un « fait accompli »

Des sans-abri, en majorité des migrants, ont trouvé refuge à la Bourse du Travail de Saint-Etienne depuis hier soir après la fin de la trêve...

Chaque **PAGE** contient **PLUSIEURS LIENS** vers des articles. Le lien correspond au titre de l'article.

La mise en forme de chaque lien apparaît rigoureusement identique.



(4) J'ouvre Web Scraper

- Windows, Linux: Ctrl+Shift+I or F12
- Mac: Cmd+Opt+I
- Any OS: open Tools / Developer tools

(a) Je clic sur les points verticaux

The image shows a browser window with a menu open. The address bar shows 'souche.com/category/societe/page/5'. The page content includes a headline 'condamné à 5 d'amende avec' and another 'té après une erreur de la justice'. The 'Outils de développement' menu is open, showing options like 'Enregistrer la page sous...', 'Créer un raccourci...', 'Effacer les données de navigation...', 'Extensions', 'Gestionnaire de tâches', and 'Outils de développement'. A sub-menu is also visible, showing 'Modifier', 'Couper', 'Copier', and 'Coller'.

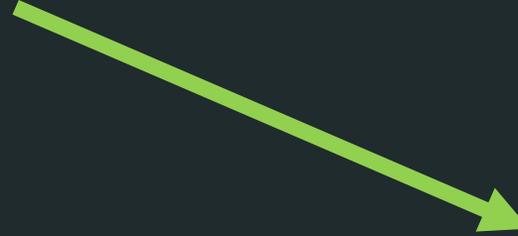
(b) Je clic sur « plus d'outils »

(c) Je clic sur « Outils de développement »



Une « fenêtre » s'ouvre en bas de votre écran

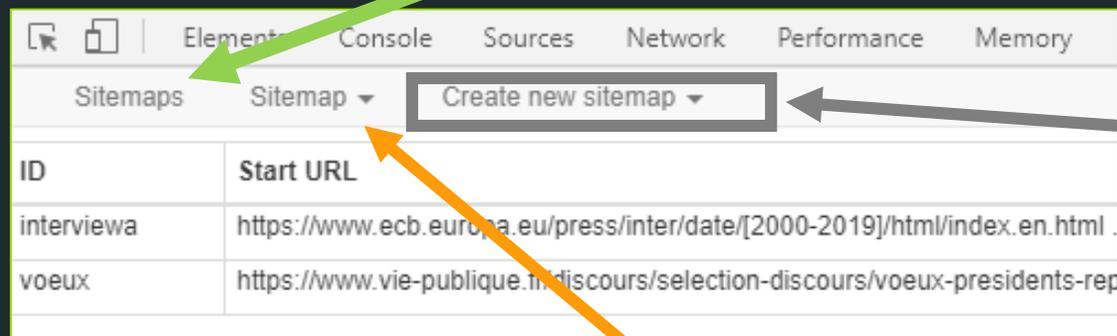
Je clic sur Web Scraper



ID	Start URL
interviewa	https://www.ecb.europa.eu/press/inter/date/[2000-2019]/html/index.en.html ...
voeux	https://www.vie-publique.fr/discours/selection-discours/voeux-presidents-republique-depuis-1974-2.html ...

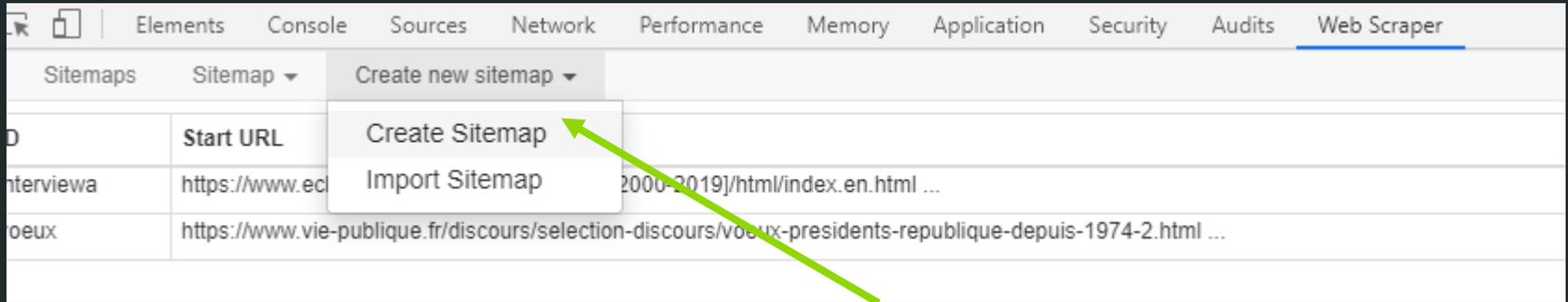


Sous l'onglet « SITEMAPS », seront rangés mes « programmes » spécifiques à un site

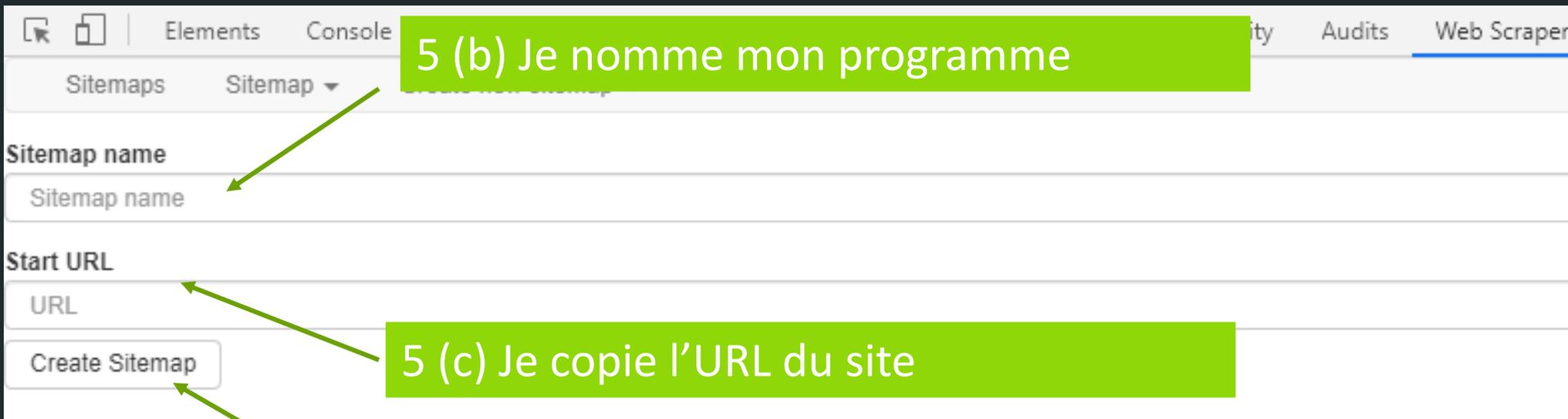


(5) Je clique ici pour créer un nouveau programme

Sous l'onglet « Sitemap » se trouvent mes commandes pour un programme ouvert



5 (a) Je clic sur « create Sitemap »



5 (b) Je nomme mon programme

5 (c) Je copie l'URL du site

5 (d) Je valide pour créer le programme



Exemple : je veux aspirer les titres des articles de la rubrique « Société » du site. Cette rubrique s'étend sur 2083 pages.

Elements Console Sources Network Performance Memory Application Security Audits Web S

Sitemaps Sitemap Create new sitemap

Sitemap name

titrefds

Start URL

https://www.fdesouche.com/category/societe/page/[1-2083]

Create Sitemap

Mon programme tournera les pages de 1 à 2083 et appliquera systématiquement mes commandes sur chaque page



(6) Je clic sur « add selectors »

ID	Selector	type	Multiple	Parent selectors	Actions
Add new selector					



(a) Je nomme l'objet à sélectionner

Sitemaps Sitemap titrefds Create new sitemap

Id
Selector Id

Type
Text

Selector
Select Element preview Data preview

Multiple

(b) Je choisis dans le menu déroulant le type de l'objet

(c) Je coche « multiple » si mon objet est présent plusieurs fois sur une page.

Exemple : sélectionner les liens des articles



Id

Titre

Type

Text

Text

Link

Popup Link

Image

Table

Element attribute

HTML

Element

Element scroll down

Element click

Grouped

0

Id

Titre

Type

Text

Selector

Select	Element preview	Data preview
<input checked="" type="checkbox"/> Multiple		



(d) Je clique sur
SELECT puis je
clique sur les TITRES
du site.

puis je clique sur les
TITRES du site.

Enfin, je coche et
je valide

div.archive-text a S P C Done selecting! ail la CGT

https://www.fdesouche.com/1185843-saint-etienne-42-des-sans-abri-en-majorite-des-migrants-occupent-la-bourse-de...

Text

Selector

Select	Element preview	Data preview
<input checked="" type="checkbox"/> Multiple		



(7) Je valide en cliquant sur « Save selector »

Type

Text

Selector

Select Element preview Data preview `div.archive-text a`

Multiple

Regex

regex

Delay (ms) ⓘ

0

Parent Selectors

_root

Save selector Cancel

Je peux pré-visualiser le résultat en cliquant sur « Data Preview »



ID	Selector	type	Multiple	Parent selectors	Actions
Titre	d	SelectorText	yes	_root	Element preview

(8) Je peux maintenant lancer mon programme en cliquant sur SCRAPE



(8) Je confirme en cliquant sur « start »

Sitemaps Sitemap titrefds ▼ Create new sitemap ▼

Request interval (ms)

2000

Page load delay (ms)

2000

Start scraping

Si besoin, augmenter la durée de l'intervalle (1^{ère} ligne)



Le fichier .csv est prêt !

Sitemaps Sitemap titrefds Create new sitemap

Refresh Data

Web Scrapers User Interface (UI)? **Sure!**

Télécharger le fichier en cliquant sur « export data »

web-scrapers-order	URL	Titre
1554898321-68	https://www.fdesouche.com/category/societe/page/8	Migrants : le pape François exhorte les Romains à utiliser leurs énergies « pour transformer les tensions et les problèmes en opportunités »
1554898306-24	https://www.fdesouche.com/category/societe/page/10	Saïda se bat pour faire « rapatrier » sa petite-fille de 3 ans détenue en Irak
1554898340-128	https://www.fdesouche.com/category/societe/page/5	Ivry (94) : harcelé, le lycéen se venge à coups de marteau en pleine classe
1554898348-175	https://www.fdesouche.com/category/societe/page/4	400 nouveaux radars tourelles débarquent: pourquoi il sera bien plus difficile de leur échapper
1554898356-200	https://www.fdesouche.com/category/societe/page/3	Des « taux choquants de viols » dans les pays nordiques, dénonce Amnesty International
1554898376-232	https://www.fdesouche.com/category/societe/page/1	L'artiste toulousain Colomina expose son « migrant » à Paris
1554898321-74	https://www.fdesouche.com/category/societe/page/8	Abdelkader Merah explique que ses parents n'ont pas été très vigilants sur la scolarité de leurs enfants: « C'est dans la culture des Arabes »



Vue sur le fichier texte après sélection de la colonne titre

“Le prophète lui a dit que les Roms travaillent pour les envoyés du diable et eux ils mangent les enfants »”

"Prostitution à Rome : « Les prêtres ne veulent pas d’Italiens (...) ils veulent des migrants, c’est plus discret »
(Sodoma)”

"Saïda se bat pour faire « rapatrier » sa petite-fille de 3 ans détenue en Irak”

"Le Royaume-Uni refuse l’asile à un Iranien converti au christianisme sous prétexte que la Bible contient des passages violents”

"Gilets jaune : Maxime Nicolle convoqué par la police, Eric Drouet entendu en audition libre mercredi”,



Nouvel exemple : je souhaite
télécharger tous les articles de la
rubrique « société » du site
FDSOUCHE



(1) J'analyse le site...

POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT

« Comportements inappropriés » à l'ancienneté - l'école de journalisme de Sciences Po va ouvrir une enquête interne

L'école de journalisme de Sciences Po a annoncé samedi 6 avril, sur sa page Facebook, l'ouverture d'une enquête interne. Une décision qui fait suite aux...

Vannes (56) : elle devait s'occuper d'un cochon sauvé par une association, elle en fait du pâté

Babe avait évité l'abattoir de justesse, mais il n'a pas échappé à un destin tragique. Pour avoir réalisé des conserves avec la viande d'un...

Ennery (95) : un projet « secret » de centre d'accueil provoque la colère des habitants

L'association Espérer 95 projette à Ennery une résidence d'accueil pour 80 personnes en situation de précarité. Le village a contesté fortement ce projet lors...

https://www.fdesouche.com/1188177-vannes-elle-devait-soccuper-dun-cochon-sauve-par-une-ass...

SE CONNECTER PROPOSER UN ARTICLE CONNECTÉS: 1620 VISITEURS UNIQUES

FDESOUCHE.COM EST UNE REVUE DE PRESSE

ACCUEIL SOCIÉTÉ SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT INSOLITE FDÉSINTOX

Accueil / Insolite / Vannes (56) : elle devait s'occuper d'un cochon sauvé par une association, elle en fait du pâté

Vannes (56) : elle devait s'occuper d'un cochon sauvé par une association, elle en fait du pâté

Par Francois le 07/04/2019




LES POILUS

BONNE D

LA BOUT



(1) J'analyse l'article



Il s'agit de TEXTE

Plusieurs paragraphes par page
Plusieurs niveaux de mise en
forme...

Babe avait évité l'abattoir de justesse, mais il n'a pas échappé à un destin tragique. Pour avoir réalisé des conserves avec la viande d'un cochon qu'elle était censée garder pour le compte d'une association de défense des animaux, une femme de 40 ans a été condamnée, à Vannes (Morbihan), à trois mois de prison avec sursis et 500 euros d'amende, a rapporté *Le Télégramme*, jeudi 4 avril.

Elle était poursuivie pour « abus de confiance et complicité d'abattage hors d'un abattoir dans des conditions illicites ». Le tribunal a par ailleurs fait état d'un risque sanitaire en cas de vente du pâté.



2) Je réitère la procédure de l'exemple précédent jusqu'à l'étape 7

The screenshot shows the Web Scraper interface with a table of selectors. The table has columns for ID, Selector, type, Multiple, Parent selectors, and Actions. A single selector is listed with ID 'Titre', Selector 'div.archive-text a', type 'SelectorText', Multiple 'yes', and Parent selectors '_root'. There are buttons for 'Element preview' and 'Data preview' next to the selector. An 'Add new selector' button is visible at the bottom left.

ID	Selector	type	Multiple	Parent selectors	Actions
Titre	div.archive-text a	SelectorText	yes	_root	Element preview Data preview

[Add new selector](#)

3) Je clique sur titre



Ma requête doit dépendre de celle formulée sur le niveau TITRE...

Sitemaps

Sitemap titrefds ▾

Create new sitemap ▾

[_root / Titre](#)

ID	Selector	type	Multiple	Parent selectors	A
----	----------	------	----------	------------------	---

[Add new selector](#)

(4) Je clique sur « Add new selector »



(5) Je remplis le formulaire

Elements Console Sources Network Performance Memory Application Security Audits **Web Scraper**

Sitemaps Sitemap titrefds Create new sitemap

Id A –Je nomme mon objet

Selector Id

Type B –Je sélectionne sa nature dans le menu déroulant

Text

Selector

Select Element preview Data preview

Multiple C –Je sélectionne sa nature dans le menu déroulant

Regex

regex

Delay (ms)

0

Parent Selectors

_root

(6) Je sélectionne le texte



← → ↻ 🏠 ⓘ <https://www.fdesouche.com/1187987-un-collectif-dassociations-lgbti-appel>

ACCUEIL **SOCIÉTÉ** SÉCURITÉ POLITIQUE ECONOMIE MONDE ECOLOGIE CULTURE SPORT

A l'occasion du « Printemps des assocés », l'un des plus grands salons LGBTI de France qui se déroule ce week-end à Paris, un collectif appelle à rejoindre la liste de Ian Brossat (PC), créditée de 2 à 3% de s intentions de vote, qui prône une «Europe des gens».

En Europe, les forces conservatrices, populistes, racistes et xénophobes ont le vent en poupe. Elles mettent en avant un projet de société rétrograde et prônent un retour à un système patriarcal, s'attaquant aux droits des femmes et des personnes LGBTI. Leurs politiques visent non seulement à aggraver les inégalités existantes et à empêcher toute conquête de droits nouveaux, mais aussi à utiliser les minorités comme boucs émissaires, à l'image de ce qui se pratique dans la Hongrie de

div#content-area S P C Done selecting!

Nous souhaitons aujourd'hui lutter contre cette vague réactionnaire et décomplexée qui

🔍 📄 Elements Console Sources Network Performance Memory Application Security

type

Text

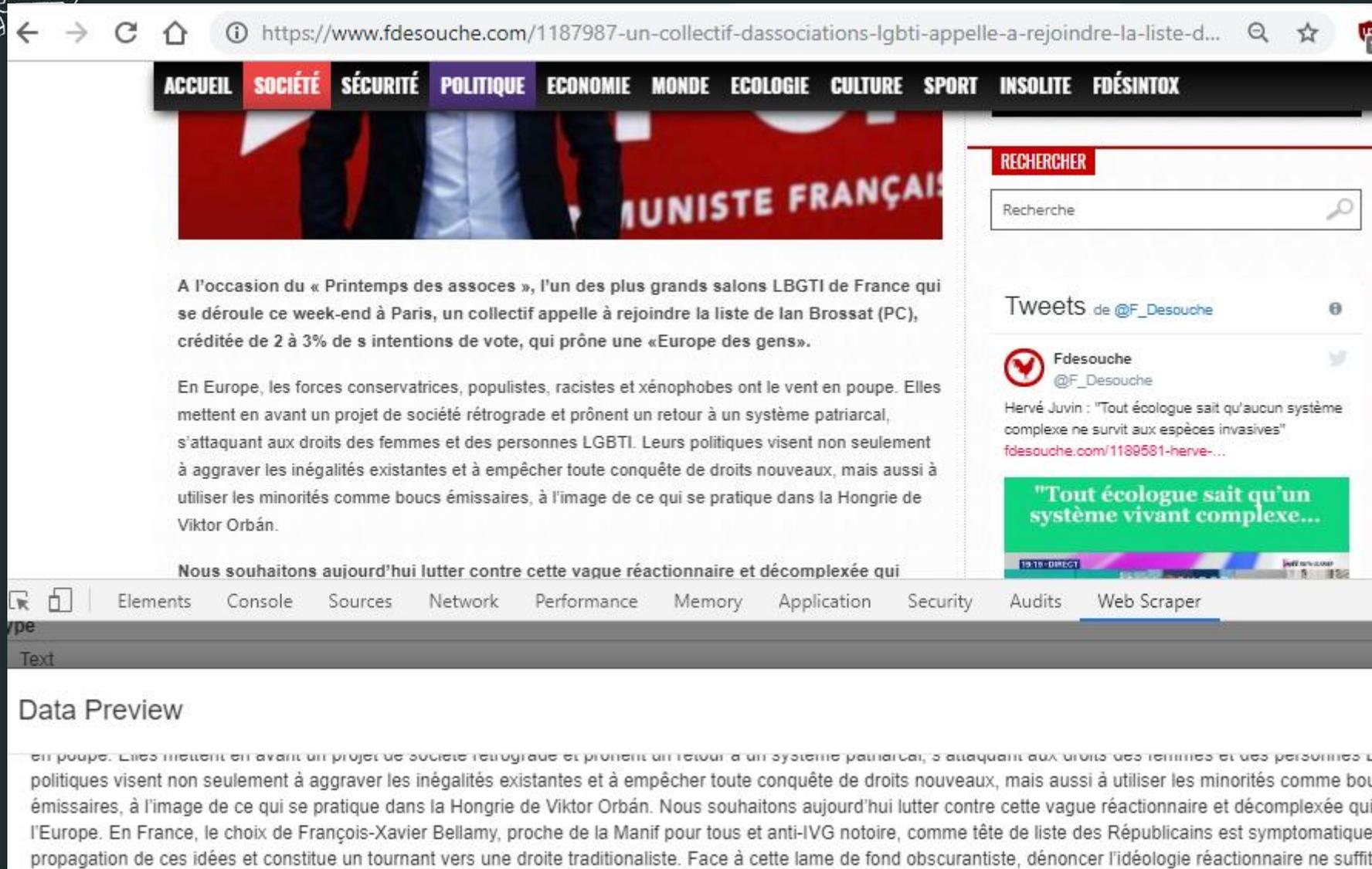
Selector

Select Element preview Data preview div#content-area

Multiple

Je coche la validité de la structuration et je clique sur « Done selecting »

Je prévisualise le résultat

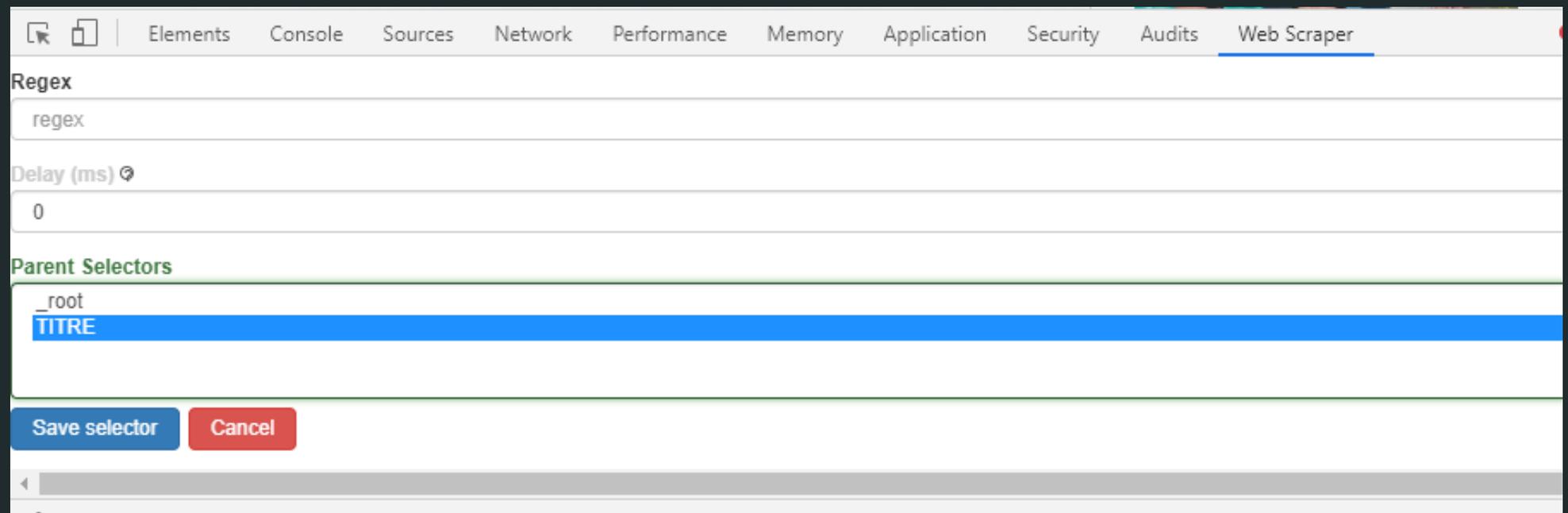


The image shows a browser window displaying a news article on the website fdesouche.com. The article title is "1187987-un-collectif-dassociations-lgbti-appelle-a-rejoindre-la-liste-d...". The navigation bar includes categories like ACCUEIL, SOCIÉTÉ, SÉCURITÉ, POLITIQUE, ÉCONOMIE, MONDE, ÉCOLOGIE, CULTURE, SPORT, INSOLITE, and FDÉSINTOX. The article text discusses the "Printemps des assocs" event in Paris and mentions a collective calling for support for the list of Ian Brossat (PC), which has 2 to 3% of the vote intentions. It also mentions a reaction against a "Europe des gens" and a "vague réactionnaire et décomplexée".

Below the browser window, a web scraper interface is visible. It has a menu with options: Elements, Console, Sources, Network, Performance, Memory, Application, Security, Audits, and Web Scraper. The "Web Scraper" option is selected. Below the menu, there is a "Text" field and a "Data Preview" section. The preview shows a snippet of the article text: "en poupe. Elles mettent en avant un projet de société rétrograde et prônent un retour à un système patriarcal, s'attaquant aux droits des femmes et des personnes LGBTI. Leurs politiques visent non seulement à aggraver les inégalités existantes et à empêcher toute conquête de droits nouveaux, mais aussi à utiliser les minorités comme boucs émissaires, à l'image de ce qui se pratique dans la Hongrie de Viktor Orbán. Nous souhaitons aujourd'hui lutter contre cette vague réactionnaire et décomplexée qui l'Europe. En France, le choix de François-Xavier Bellamy, proche de la Manif pour tous et anti-IVG notoire, comme tête de liste des Républicains est symptomatique propagation de ces idées et constitue un tournant vers une droite traditionaliste. Face à cette lame de fond obscurantiste, dénoncer l'idéologie réactionnaire ne suffit".



(7) Je choisis le sélecteur dont dépend ma requête



Ici, je choisis TITRE puis je confirme en cliquant sur « Save Selector »



(8) Je lance le scrap

ID	Selector	type	Multiple	Parent selectors	Actions
Titre	d	SelectorText	yes	_root	Element preview

(8) Je peux maintenant lancer mon programme en cliquant sur SCRAPE



C'est prêt !

				Web Scaper
				The Local Education : Depuis trente ans, le niveau de math des élèves de CM2 est en baisse, selon une étude L'association openDemocracy accuse les fundamentalistes chrétiens américains d'avoir financé l'extrême droite en Europe
900507-	https://www.fdesouche.com/category/societe/page/5	Beauvais (60) : un Afghan condamné pour avoir harcelé sexuellement des jeunes filles et agressé la gérante d'un bar	https://www.fdesouche.com/1185437-beauvais-un-afghan-condamne-pour-avoir-harcele-sexuellement-des-jeunes-filles-et-agresse-la-gerante-dun-bar-le-migrant-est-en-france-que-depuis-un-mois	Le tribunal de Beauvais a condamné ce lundi un homme pour violences et outrage sexiste. Cette dernière infraction, entrée en vigueur l'été dernier, se caractérise par le fait d'imposer à une personne tout propos ou comportement à connotation sexuelle ou sexiste qui porte atteinte à sa dignité. En clair, c'est ce qu'on appelle le harcèlement de rue, qui a été décrit par deux jeunes Beauvaisiennes devant le tribunal. Vendredi soir, elles sortaient d'un bar du centre-ville lorsqu'elles ont été interpellées par un homme de 28 ans, complètement

web-scrapers-order,web-scrapers-start-url,TITRE,TITRE-href,Article"155490066-361","https://www.fdesouche.com/category/societe/page/1","La Chine a rasé plusieurs grandes mosquées en région ouïgoure, montrent des images satellites","https://www.fdesouche.com/1188803-la-chine-a-detruit-plusieurs-grandes-mosques-en-region-ouigoure-montrent-des-images-satellites",|

"« Où est passée la mosquée ? » Shawn Zhang s'interroge. Cet activiste, étudiant à l'université de Vancouver au Canada, a pris l'habitude de dénoncer sur Twitter les exactions du régime chinois, notamment au Tibet et dans la région du Xinjiang, habitée par la minorité musulmane ouïgoure. En observant des images satellites sur Google Earth Pro, le jeune homme est arrivé à la conclusion que la grande mosquée Aitika de Keriya, une ville de 30.000 habitants dans cette province de l'ouest de la Chine, avait tout simplement disparu, au printemps 2018. Les spéculations autour du sort de la grande mosquée ont rapidement germé sur internet, d'autant que ce monument du XIIIe siècle avait été classé au patrimoine culturel national chinois : toute destruction ne pouvait être décidée que par le pouvoir central à Pékin. Comme le rapporte « Libération », l'étudiant d'origine chinoise a comparé la même vue satellite à différentes périodes, pour arriver à la conclusion que la mosquée avait été rasée entre mars et mai 2018. D'autres services de cartographie, tels que HERE, Planet Labs ou TerraServer, confirment la disparition du monument, selon « Libération » et l'associati



Pour aller plus loin et vous approprier l'outil...

- <https://www.webscraper.io/documentation>
- <https://www.webscraper.io/tutorials>

Pour un exemple sur une récupération d'adresses (tuto en langue anglaise) :

<https://www.youtube.com/watch?v=ZdOrH50WqEo>

