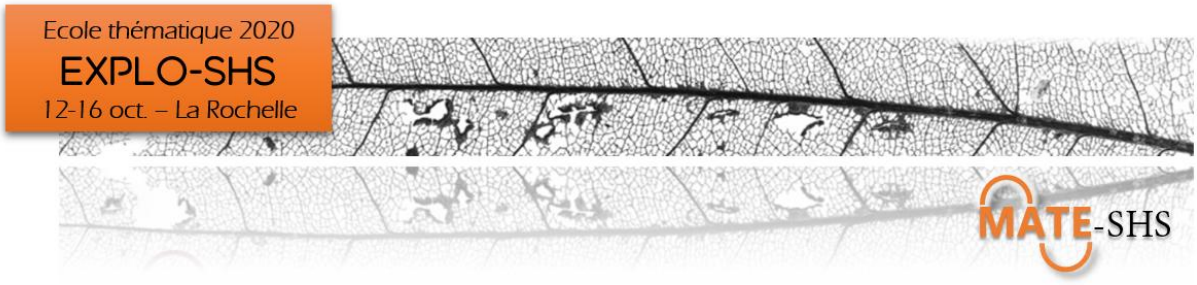
Le scraping de données structurées sur le web à l'aide d'Extractify. Focus sur les données conversationnelles.

Frédéric Vergnaud



Le scraping de données structurées sur le web à l'aide d'Extractify.

Focus sur les données conversationnelles.

Frédéric Vergnaud
CSI – Centre de Sociologie de l'Innovation, i3 UMR CNRS 9217
Mines ParisTech, PSL Research University
frederic.vergnaud(at)mines-paristech.fr

École Thématique « EXPLO-SHS »
12-16 Octobre 2020
La Rochelle, France

Présentation

Extractify est une extension libre pour chrome, dont le but est de scraper des données structurées sur le web. Ce « plugin » est particulièrement conçu pour la récolte de commentaires ou de conversations en ligne tels que les forums.

Il permet de :

1) Sélectionner sur une page web des informations structurées sous forme de tableau (lignes x colonnes) de manière automatique, par sélection directe sur la page web, ou manuelle, en utilisant les sélecteurs CSS simples ou combinés.
2) Sélectionner la pagination de pages de structure identique et de même niveau
3) Recommencer le processus autant de fois que l'on veut pour des niveaux inférieurs
4) Scraper l'ensemble de la sélection
5) Obtenir finalement un fichier au format json facilement exportable vers d'autres logiciels

Installation manuelle pour Chrome

- Rendez-vous ici : https://github.com/fredericvergnaud/extractify
- Allez dans Releases, téléchargez la dernière version, puis décompressez-là dans un répertoire de votre choix
- Ouvrez le navigateur Chrome (ou Chromium) et allez sur la page des extensions : chrome://extensions/
- Sélectionnez « Mode développeur » en haut à droite de la page
- Cliquez sur « Chargez l'extension non empaquetée » et sélectionnez le répertoire « extractify-numero_version » précédemment décompressé.

Utilisation

1) Introduction

D'un point de vue technique, Extractify utilise l'identification d'éléments HTML au sein de la structure d'une page web pour pouvoir en extraire les données. Pour repérer ces éléments, Extractify s'appuie sur les sélecteurs CSS qui utilisent certains des attributs des éléments HTML, principalement la classe et l'identifiant, afin de les repérer facilement dans la structure de la page web. Nous en verrons le détail un peu plus avant dans la présentation.

D'un point de vue conceptuel, pour Extractify, le niveau (level) est le degré de profondeur d'une page. La page sur laquelle vous allez exécuter pour la première fois le plugin constituera le niveau 0. Les pages plus profondes atteignables par hyperliens seront donc aux niveaux -1, puis -2, puis -3 …etc.

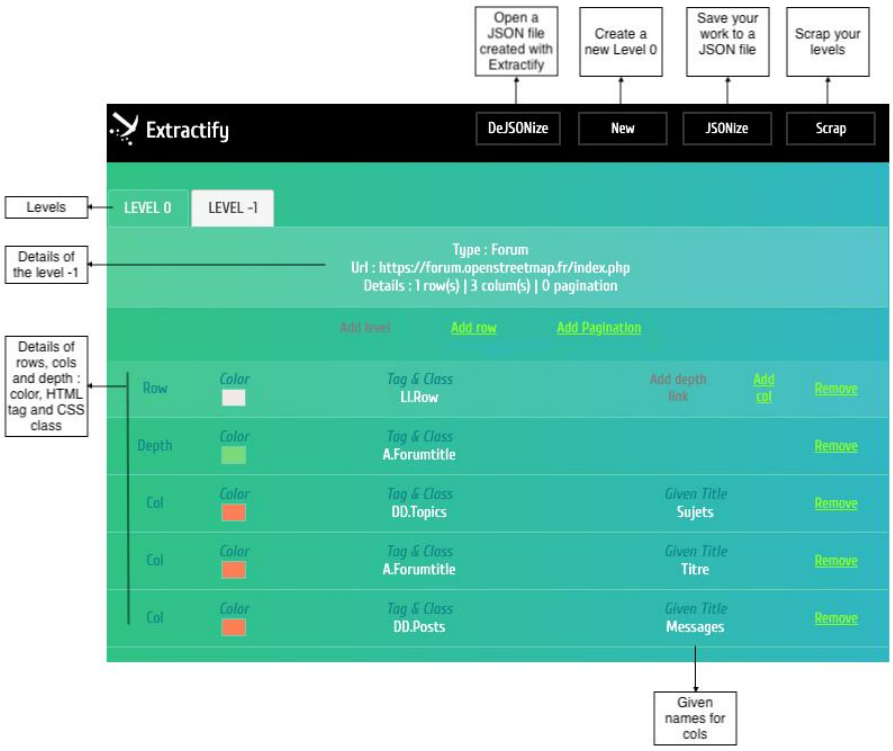
Sur chaque page, Extractify considère que les données que vous voulez extraire sont structurées en lignes et en colonnes, c'est à dire sous la forme d'un tableau. Vous allez donc devoir d'abord sélectionner des lignes, puis des colonnes à l'intérieur de ces lignes.
Ensuite, vous pourrez sélectionner des liens de pagination, c'est à dire des liens vers des pages de profondeur équivalente à la page observée et structurées de la même manière que la page observée.

Enfin, vous pourrez ajouter un niveau inférieur en sélectionnant des liens vers ce niveau inférieur qui seront situés à l'intérieur des lignes déjà sélectionnées.

Extractify a été développé initialement pour extraire des données conversationnelles de forums en ligne.
En prenant un tel exemple typique de son utilisation, vous allez commencer par exécuter Extractify sur la page qui affichera l'ensemble des forums à scraper sous forme de liens. Cette page sera considérée comme le niveau 0.
Chaque lien de cette page, menant à chaque forum, sera un lien vers un niveau inférieur (le niveau -1) de ce forum. Ce niveau inférieur affichera les topics.
Puis chaque lien menant à chaque topic sera lui-même un lien vers un niveau inférieur (le niveau -2) de ce topic. Ce niveau inférieur affichera les messages.
Ainsi, scraper une page web affichant des liens vers des forums reviendra à scraper 3 niveaux de profondeur :

1) Niveau 0 : la page affichant les forums, de laquelle, outre le lien du forum, on pourra extraire le titre du forum, son nombre de vues, son nombre de réponses …etc.
2) Niveau -1 : les pages affichant les topics de ces forums, desquelles, outre les liens de topics, on pourra extraire le titre du topic, son nombre de vues, son nombre de réponses …etc.
3) Niveau -2 : les pages affichant les messages de ces topics, desquelles on pourra extraire le nom de l'auteur, la date du message, le message …etc.

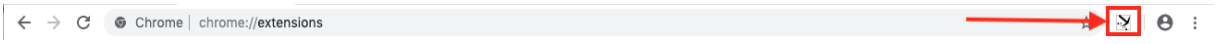
2) Présentation de l'interface



3) Exemple de scraping d'un forum OpenStreetMap

Nous allons procéder à notre premier scraping ensemble sur le forum OpenStreetMap, qui est un forum de discussion en français autour d'OpenStreetMap, qui est un projet de carte ouverte et collaborative du monde.
Ce forum va nous permettre d'utiliser à la fois le mode automatique d'Extractify, mais également le mode manuel, en utilisant les sélecteurs CSS pour cibler certains champs que l'on voudrait récupérer.
Vous pourrez ensuite vous exercer sur d'autres types de données à scraper.

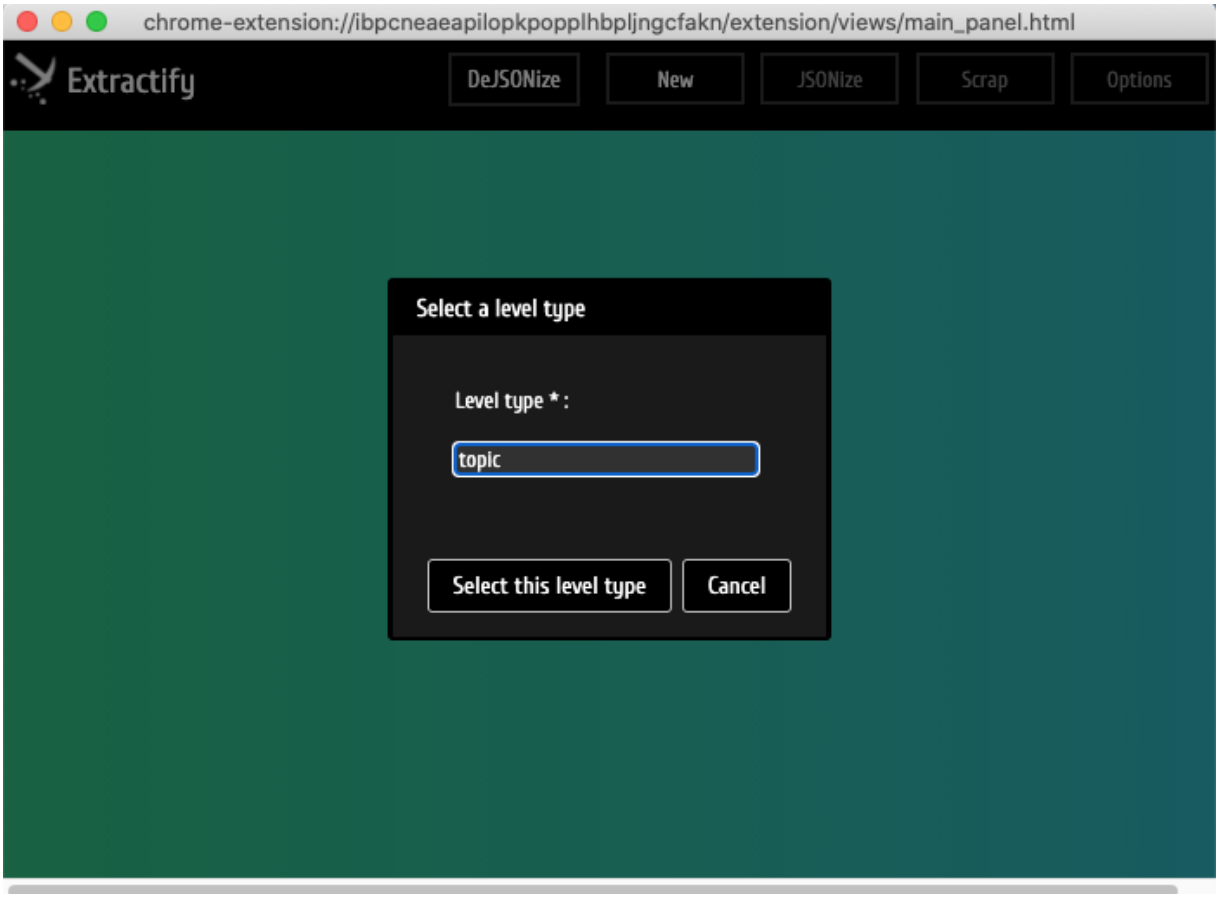
1. **Rendez-vous** à l'adresse https://forum.openstreetmap.fr/
2. **Sélectionner un forum**
   Le forum « Outils avancés pour contribuer » est un bon exemple car il regroupe 250 sujets pour un peu plus de 1450 messages, ce qui va être relativement rapide à scraper.
3. Cliquez sur l'icône du plugin dans la barre des extensions de chrome :



Le plugin s'ouvre sur le dialogue d'ajout du niveau 0.

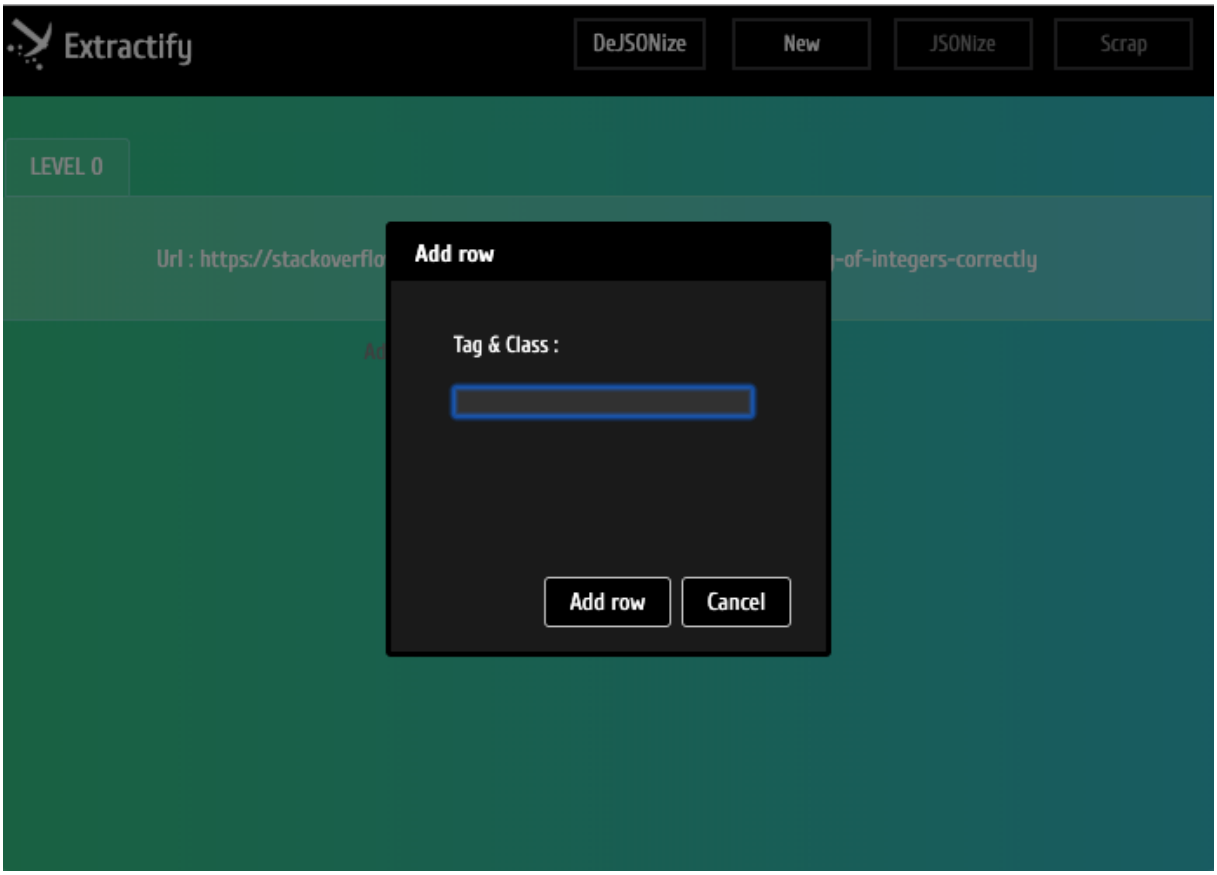
4. **Ajout d'un niveau**

Saisissez un type de niveau (obligatoire) puis cliquez sur « Select this level type ». Pour cet exemple nous allons saisir « topic » comme type de niveau :

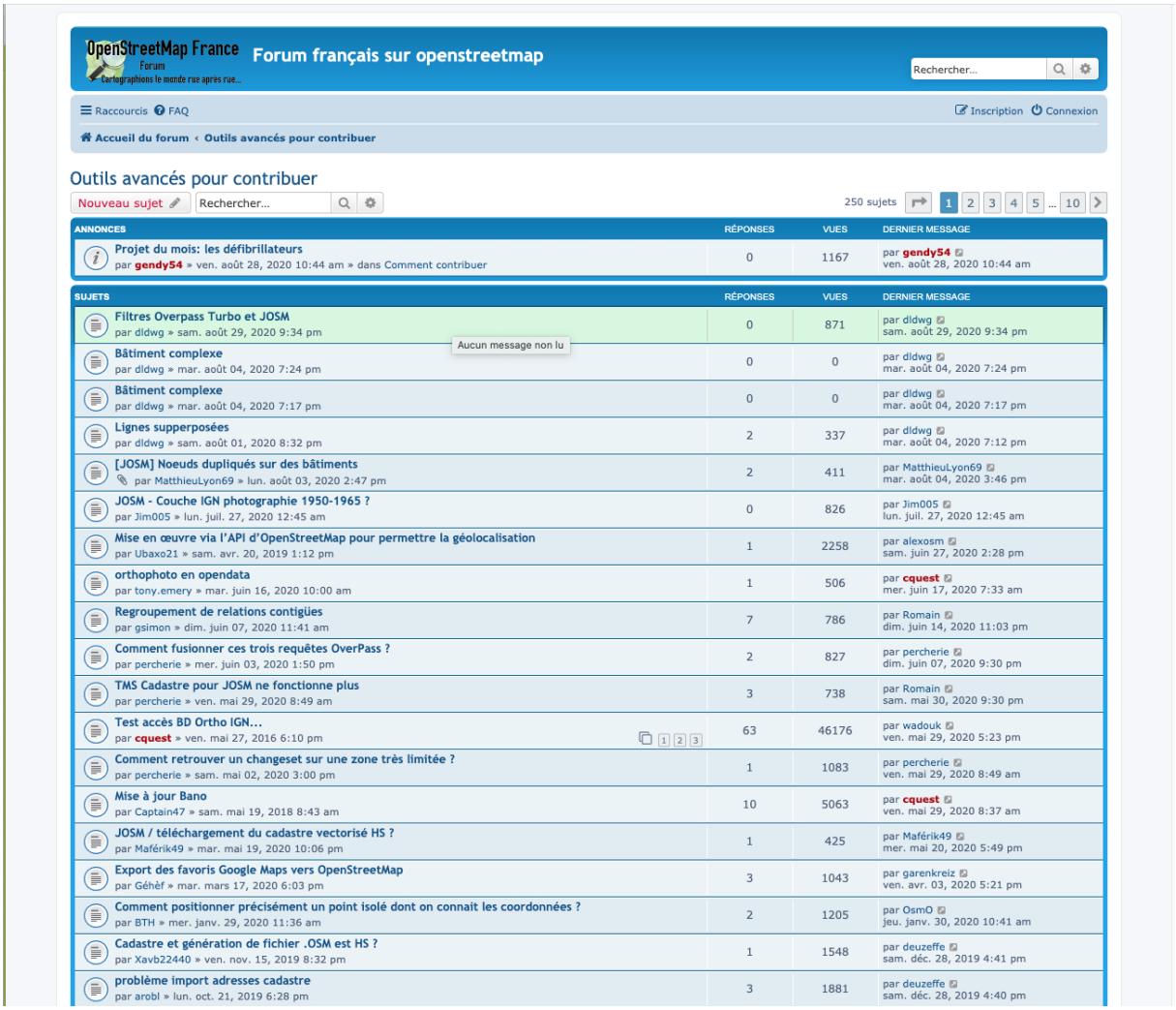


5. Ajout de lignes

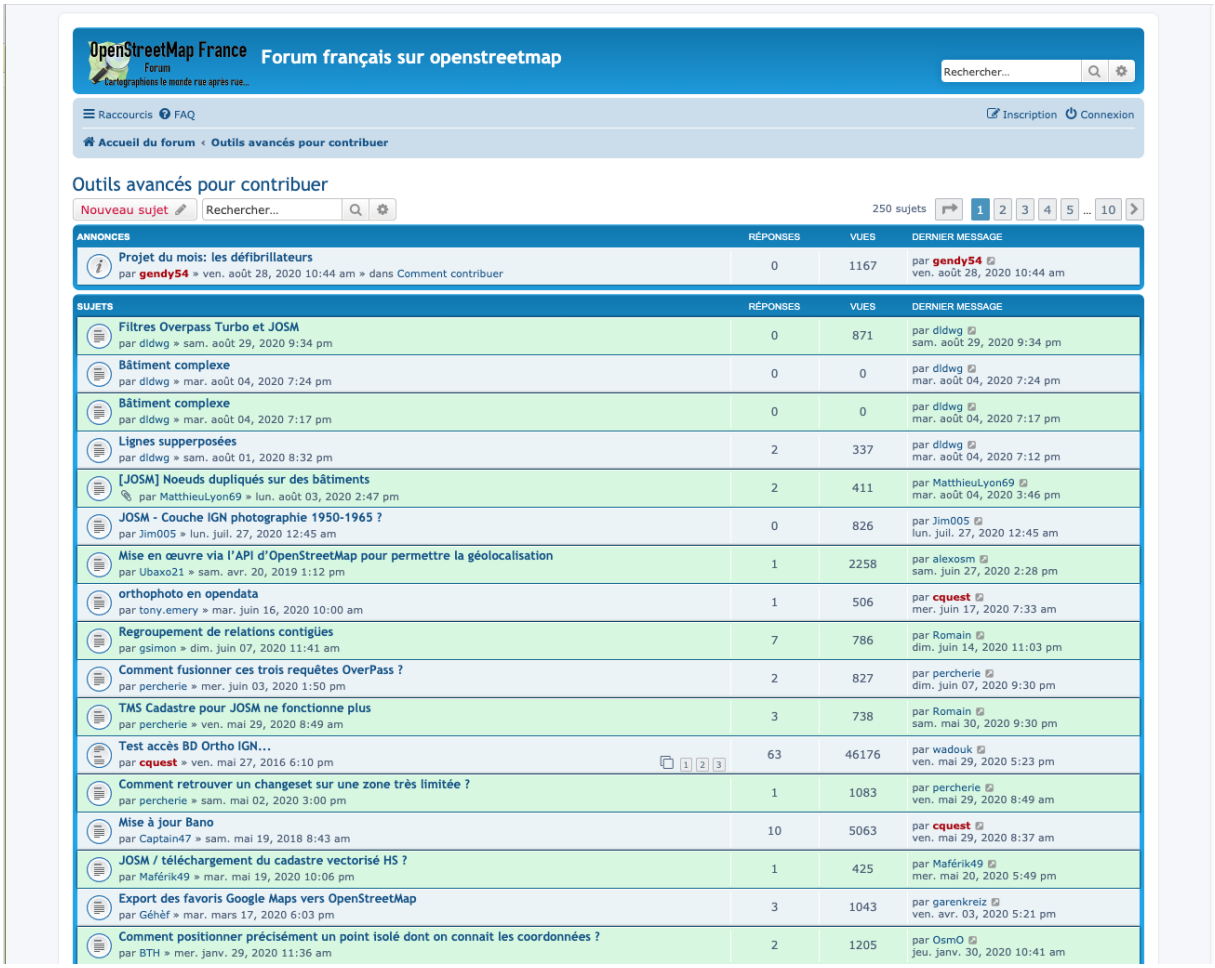
A l'ouverture du dialogue, cliquez directement sur le bouton « Add row » pour sélectionner les lignes **automatiquement**.


Extractify
DeJSONize
New
JSONize
Scrap
LEVEL 0
Url : https://stackoverflo
-of-integers-correctly
Add row
Tag & Class :
Add row
Cancel

Déplacez ensuite votre souris sur la page web de votre navigateur et sélectionnez une ligne en cliquant une zone mise en surbrillance au survol de la structure HTML de la page :



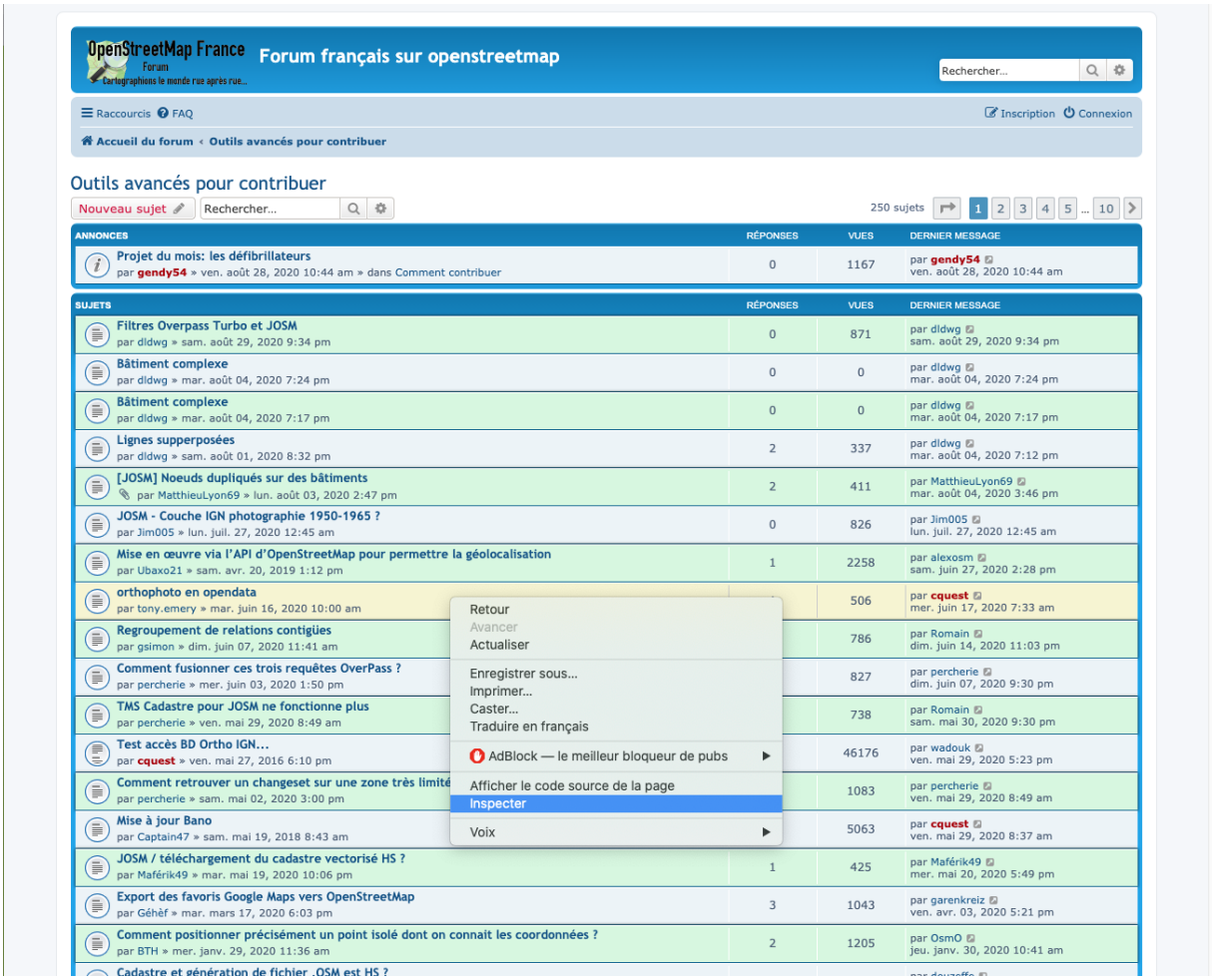
Au clic, les lignes identiques, c'est à dire les lignes qui ont la même structure HTML et CSS, se mettent également en surbrillance :



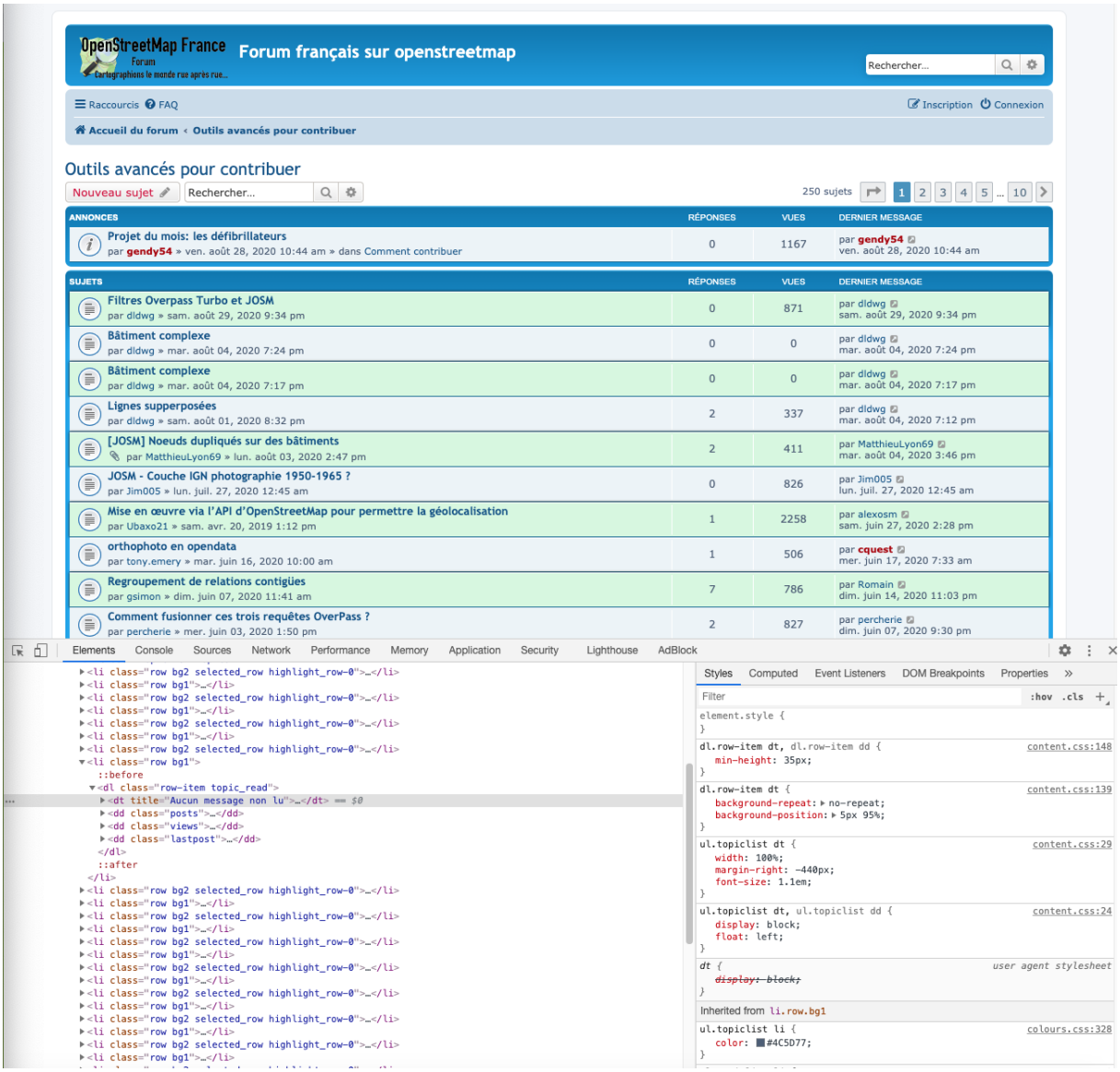
On remarque ici qu'une ligne sur deux a été sélectionnée. Regardons pourquoi à l'aide de l'inspecteur de code de Chrome. Cela nous donne par ailleurs une bonne occasion d'utiliser cet outil qui se révèle indispensable lorsque l'on veut scraper des données web.

Chaque portion de texte sur une page web est encapsulée dans des balises HTML. Ces balises possèdent des attributs qui vont nous permettre de les cibler plus facilement afin d'en extraire le contenu. C'est à l'aide de l'inspecteur interne de Chrome que nous allons pouvoir détailler la structure HTML de la page web, et ainsi repérer les balises et classes à cibler.

Faites un clic-droit sur une ligne non sélectionnée puis choisissez « Inspectez » :

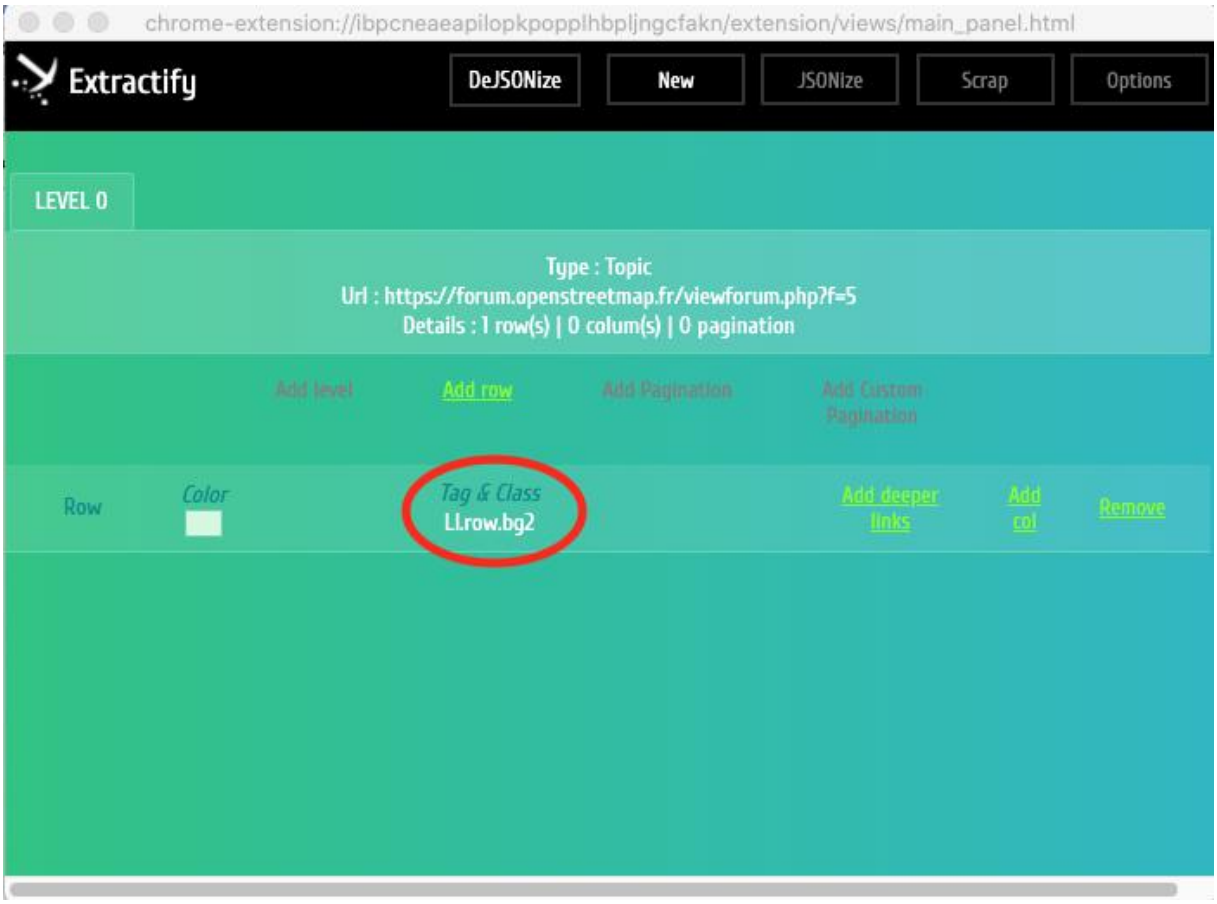


L’inspecteur s’ouvre directement sur la portion HTML que vous venez de sélectionner :

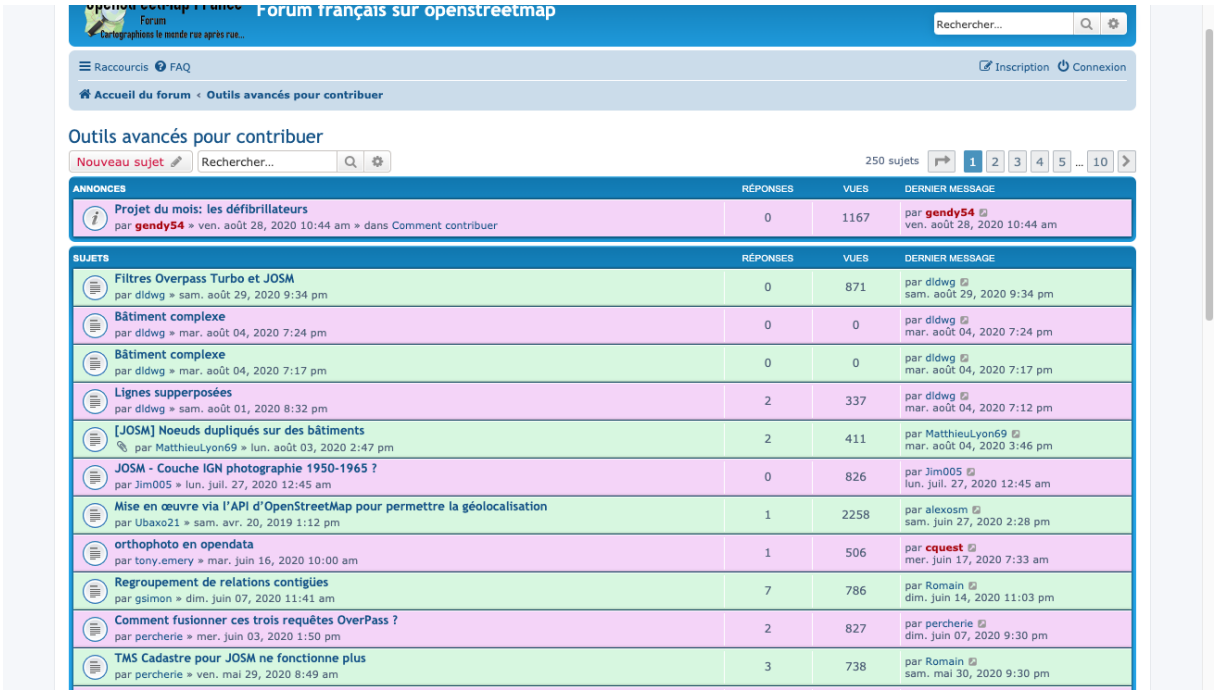


En survolant la structure HTML, on s’aperçoit que les lignes de topics sont structurées par des balises <li>. Chaque balise <li> possède un attribut « class ». Et c’est cet attribut qu’a ciblé Extractify pour sélectionner les lignes. En fonction de la ligne que vous avez survolée puis cliquée, vous avez sélectionné soit une balise <li> de « class » « row » et « bg1 », soit une balise <li> de « class » « row » et « bg2 ».

Et effectivement, on le visualise comme tel sur l'interface d'Extractify :



Si l'on veut la totalité des lignes, il faut donc recommencer l'opération d'ajout en sélectionnant les lignes « li » de classe « row bg2 » non sélectionnées :

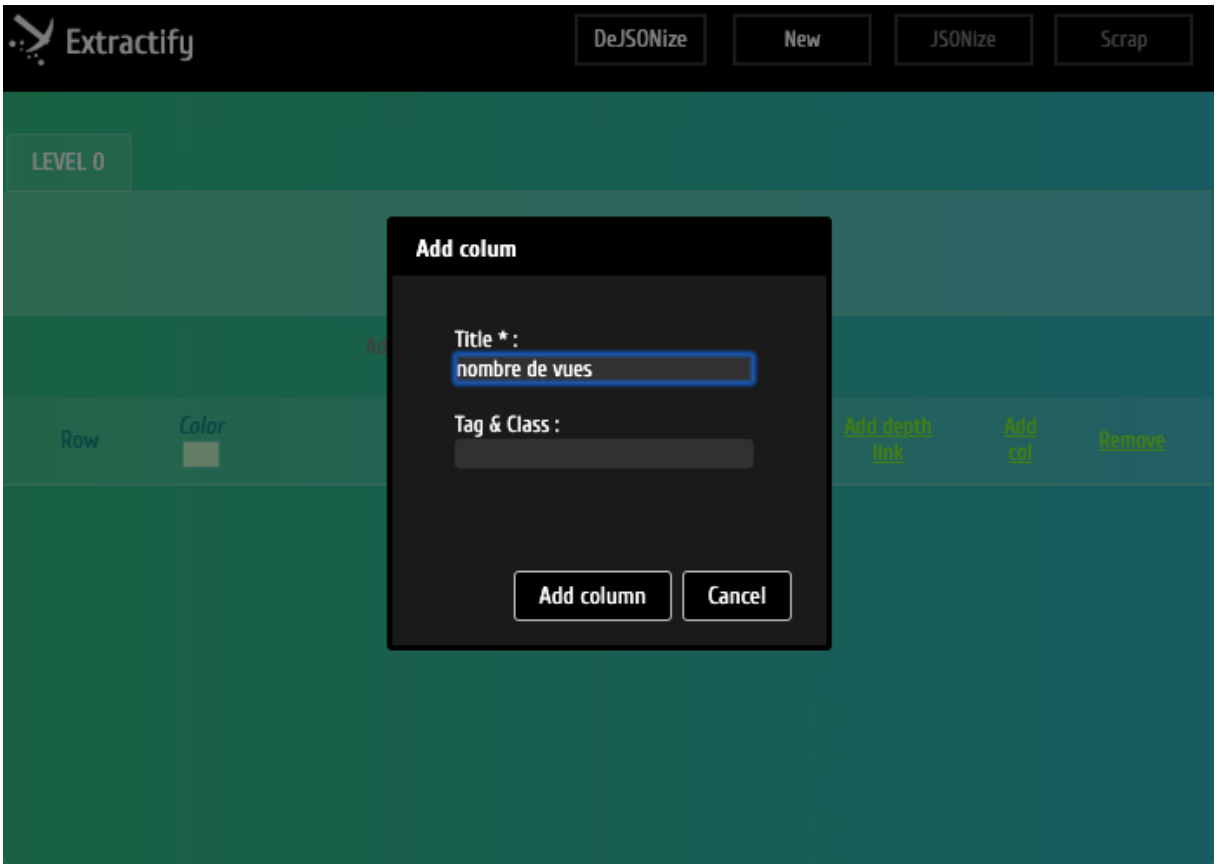


Il est à noter ici qu'une stratégie plus efficace et plus rapide aurait été de sélectionner directement tous les éléments « li » de classe « row ». Ainsi, la totalité des lignes aurait été sélectionnée d'un coup.

6. Ajout de colonnes

L'ajout de colonnes suit le même principe que l'ajout de lignes, à part le fait que vous devez saisir obligatoirement un titre de colonne.
Dans notre exemple, nous n'allons sélectionner que le titre du topic, ainsi que le nombre de vues. Le nombre de réponses étant de toute manière le nombre de messages du topic, ce que nous obtiendrons par le scraping. De la même manière, nous n'avons pas besoin des informations sur le dernier message, car nous retrouverons celui-ci dans le scraping futur de la totalité des messages !

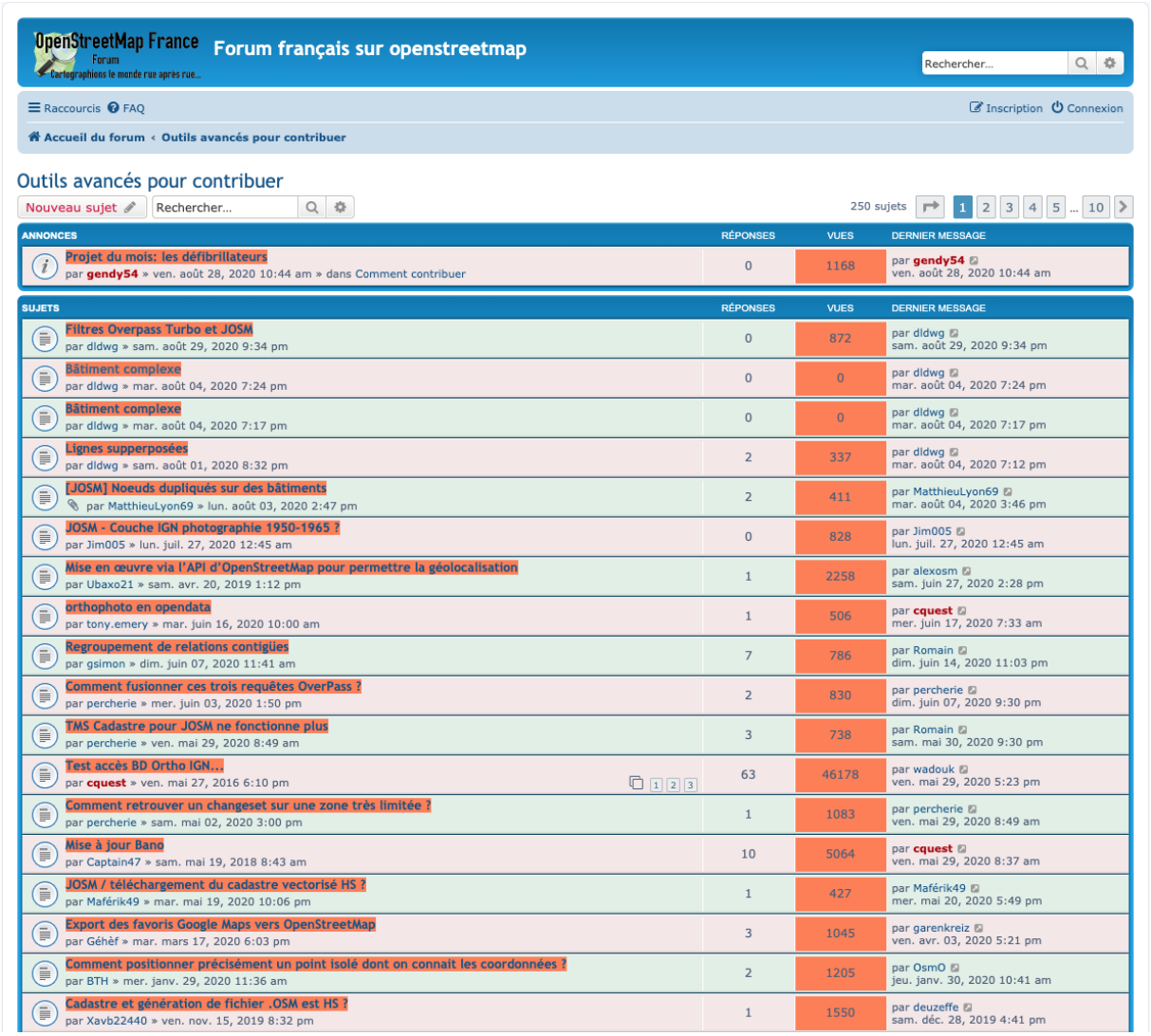
Vous allez donc cliquer sur « Add col » de la première ligne :

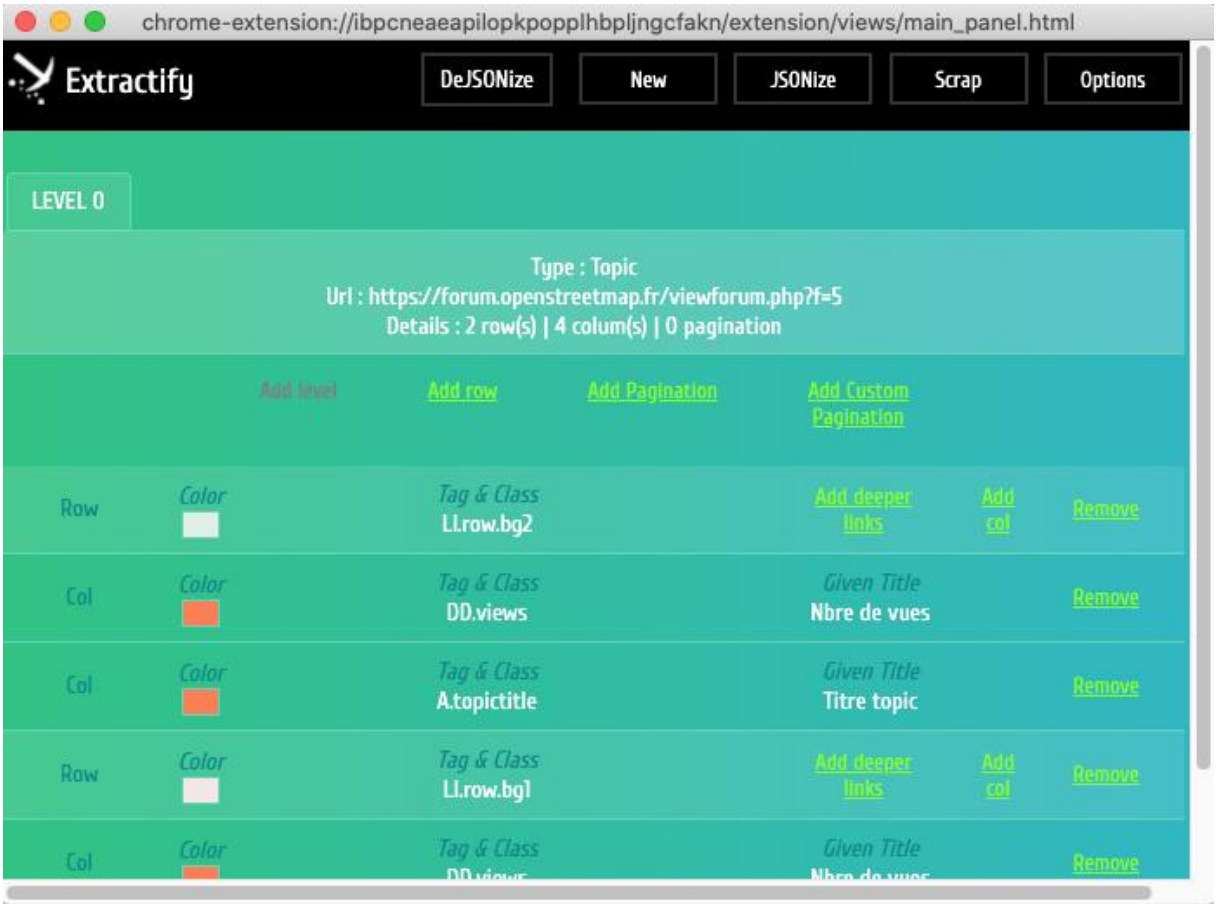


Après avoir donner un titre à votre colonne, cliquez directement sur « Add column » de la boite de dialogue pour exécuter la sélection en mode automatique, et sélectionner sur la page web la case « Vues ».

Comme vous avez 2 types de lignes, il va falloir répéter cette séquence pour l'autre type, car Extractify ne peut sélectionner que les colonnes qui sont situées à l'intérieur des lignes déjà ciblées.

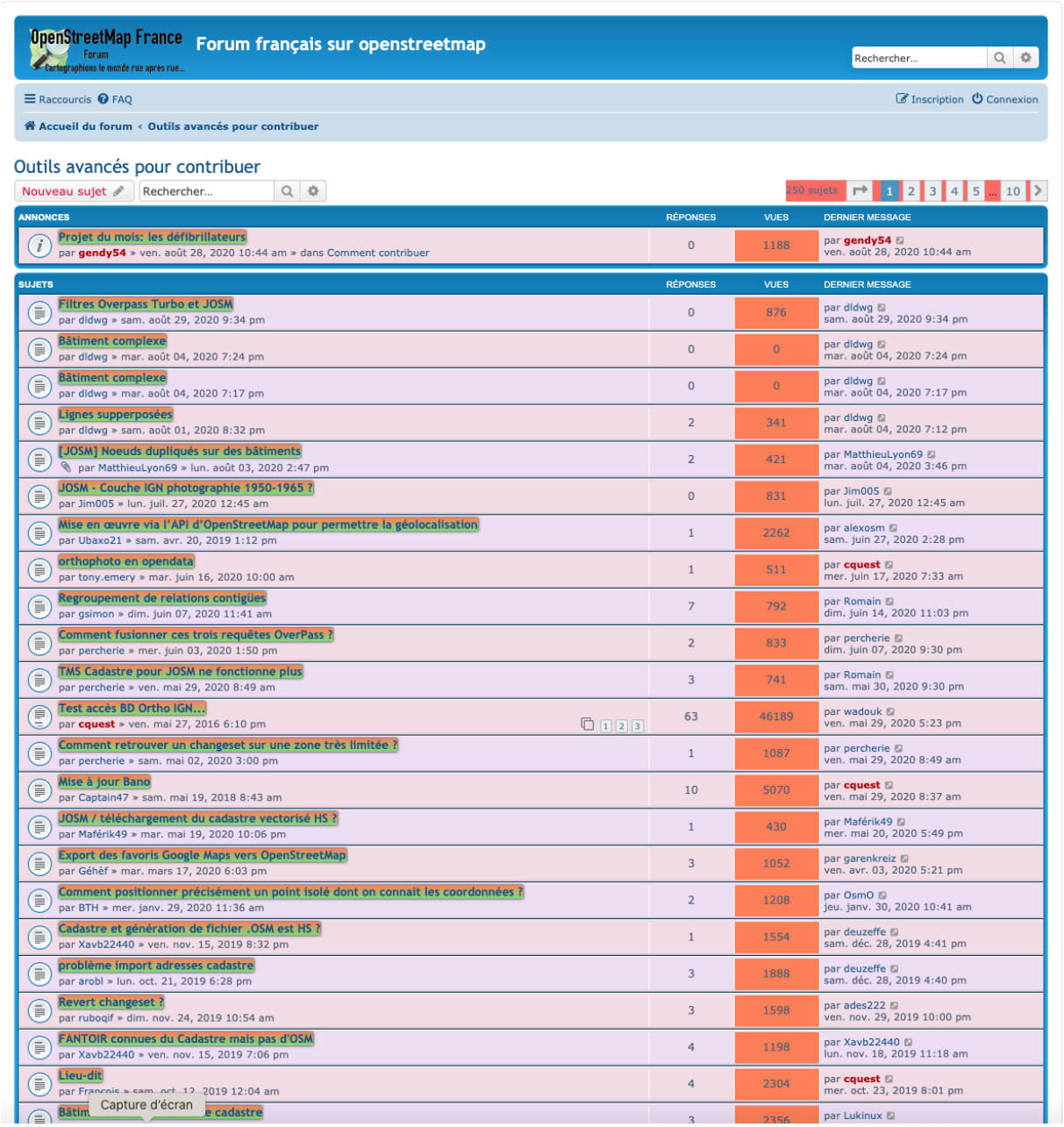
Et vous répétez la sélection pour le titre de topic :




chrome-extension://ibpcneaeapilopkpopplhbpljngcfakn/extension/views/main_panel.html
Extractify
DeJSONize
New
JSONize
Scrap
Options
LEVEL 0
Type : Topic
Url : https://forum.openstreetmap.fr/viewforum.php?f=5
Details : 2 row(s) | 4 colum(s) | 0 pagination
Add level
Add row
Add Pagination
Add Custom Pagination
Row
Color
Tag & Class
LI.row.bg2
Add deeper links
Add col
Remove
Col
Color
Tag & Class
DD.views
Given Title
Nbre de vues
Remove
Col
Color
Tag & Class
A.topictitle
Given Title
Titre topic
Remove
Row
Color
Tag & Class
LI.row.bg1
Add deeper links
Add col
Remove
Col
Color
Tag & Class
Given Title
Remove

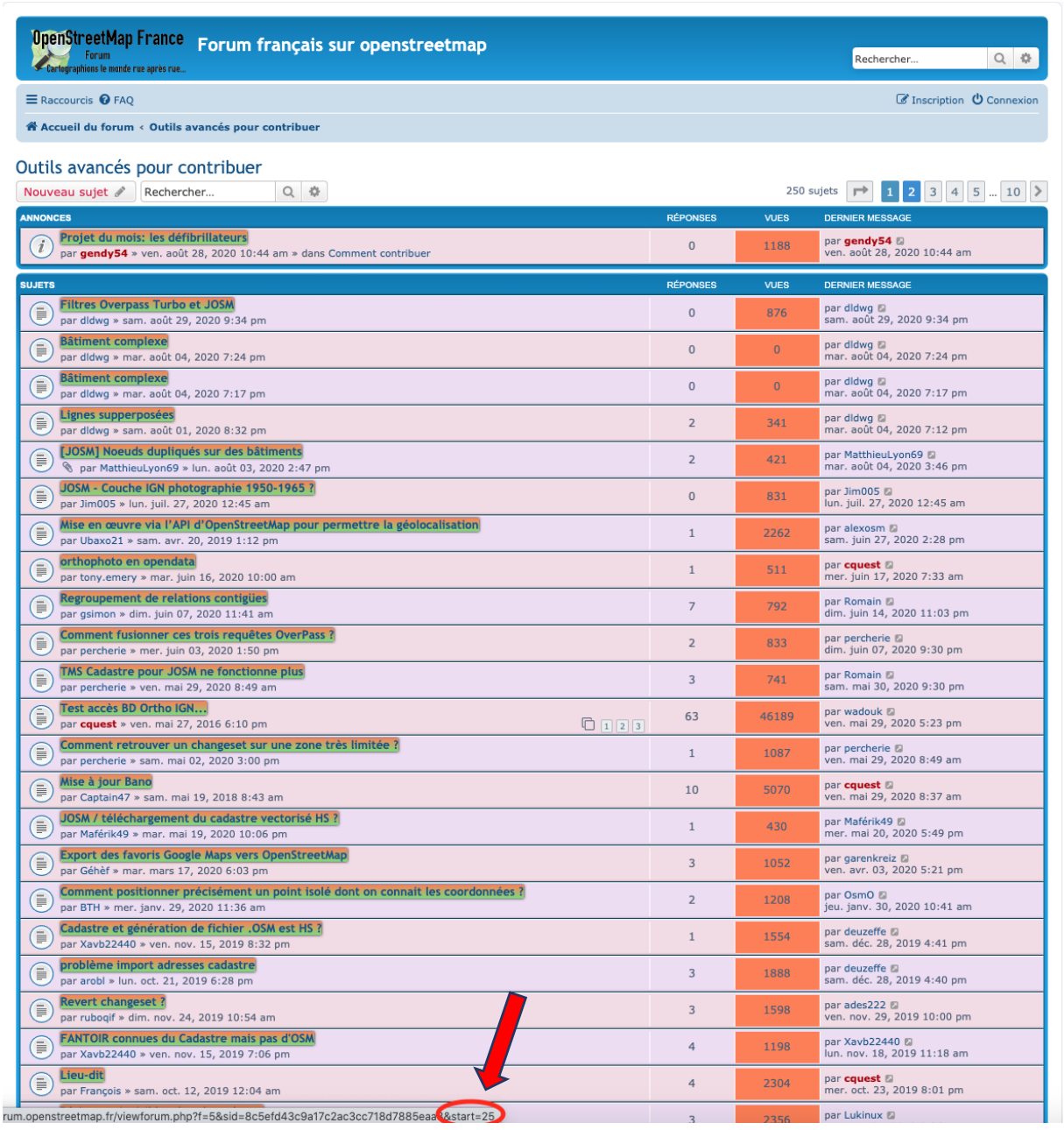
7. Ajout d'une pagination

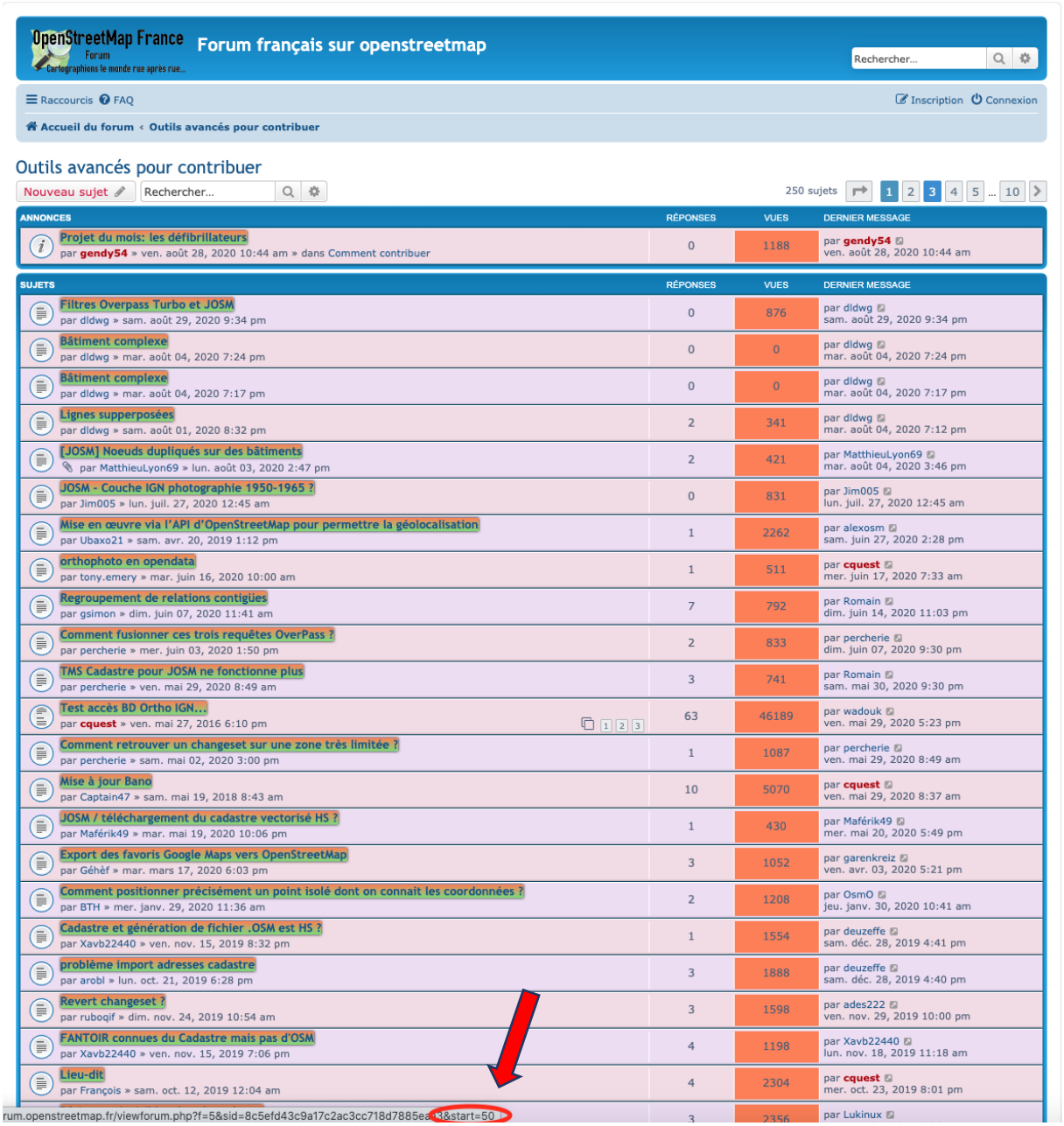
Si vous voulez scraper des pages de même niveau, mais situées à des adresses différentes accessibles *via* des liens de pagination, vous pouvez le faire en cliquant sur « Select Pagination » qui suit le même principe que l'ajout de lignes.



Si le mode de sélection automatique ne fonctionne pas, vous pouvez utiliser le mode manuel en saisissant des sélecteurs. Nous verrons plus avant comment faire.

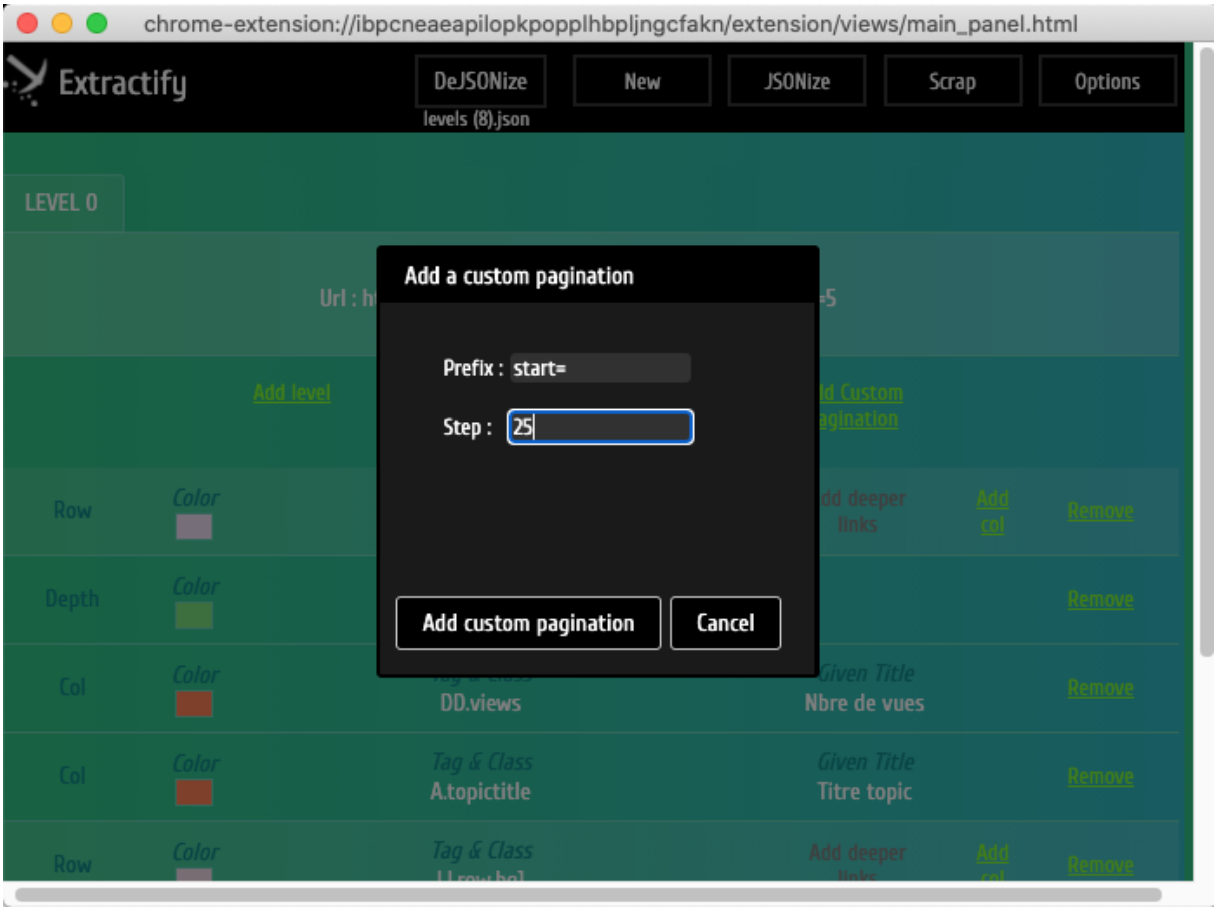
Mais Extractify possède également un mode alternatif se basant sur l'url des pages de même niveau. En effet, l'observation d'un ensemble de liens de pagination par leur survol à la souris peut vous permettre de repérer un préfixe et un palier d'incrémentation :





Ici, nous remarquons que la navigation entre les pages s'effectue à l'aide d'une variable fixe nommée « start » à laquelle on attribue une valeur, par le signe « = », qui s'incrémente de 25 en 25.

Il suffit alors de saisir ces informations dans Extractify pour qu'il repère ces liens. Cliquez sur « Add custom pagination » et saisissez les informations voulues :



Par contre dans ce cas, les liens de pagination ne seront pas surlignés, car à aucun moment nous n'avons indiqué un élément de structure HTML permettant à Extractify de s'appuyer dessus afin d'afficher une identification visuelle.

8. Ajout d'un niveau inférieur *via* des liens de profondeur

Afin d'ajouter un niveau inférieur, vous devez tout d'abord récolter des liens vers ce niveau inférieur : ce sont des liens de profondeur.

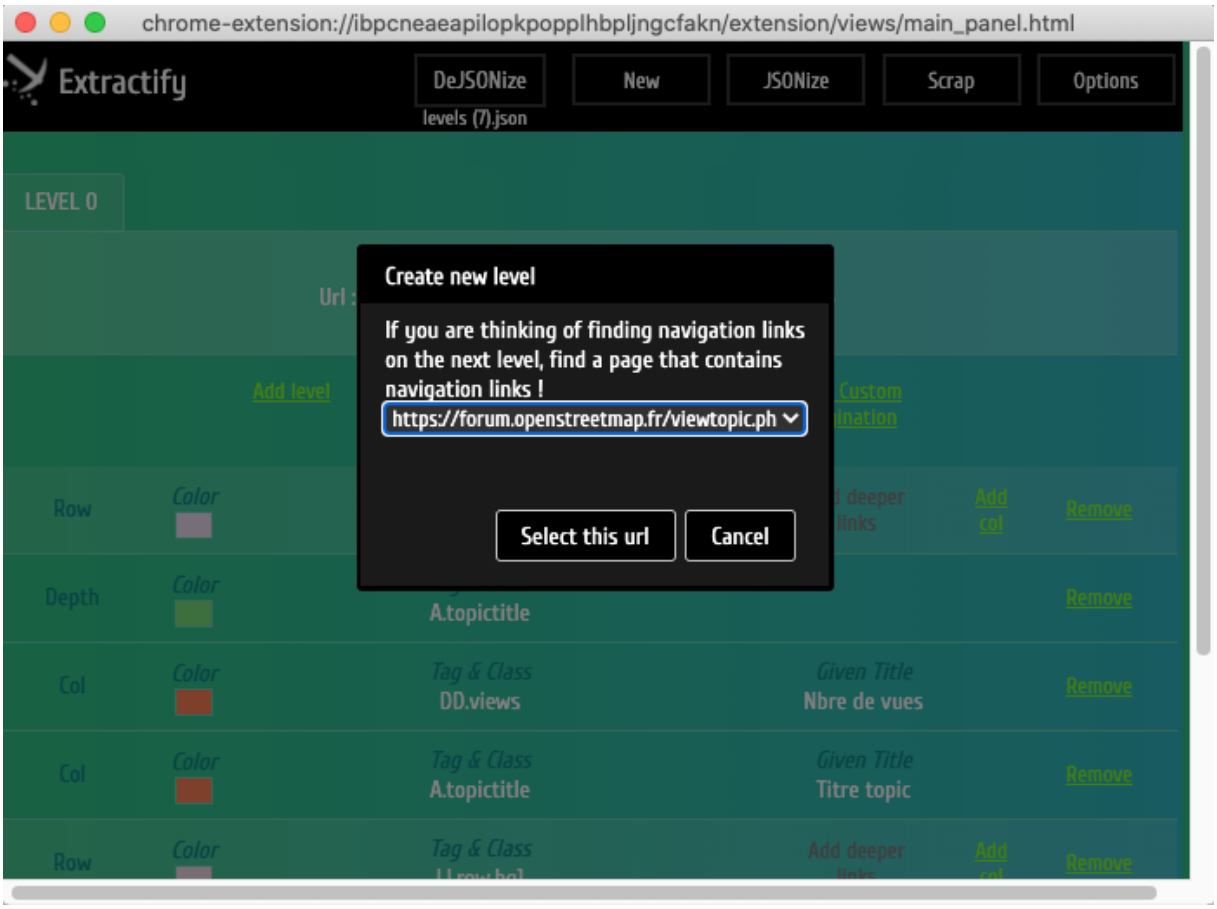
En cliquant sur « Add deeper link », le mode opératoire de sélection est le même que pour une colonne, sauf que vous êtes restreint à la sélection d'un hyperlien. Encore une fois, vous devez répéter la manœuvre pour les 2 types de lignes :

Une fois vos liens de profondeur récoltés pour chaque ligne, un clic sur « Add level » va vous permettre d'ajouter un niveau de profondeur.
Celui-ci doit être le plus exhaustif possible, c'est à dire contenir le plus d'informations possible, dans l'esprit du proverbe « qui peut le plus, peut le moins ». Par exemple dans le cas d'un forum :

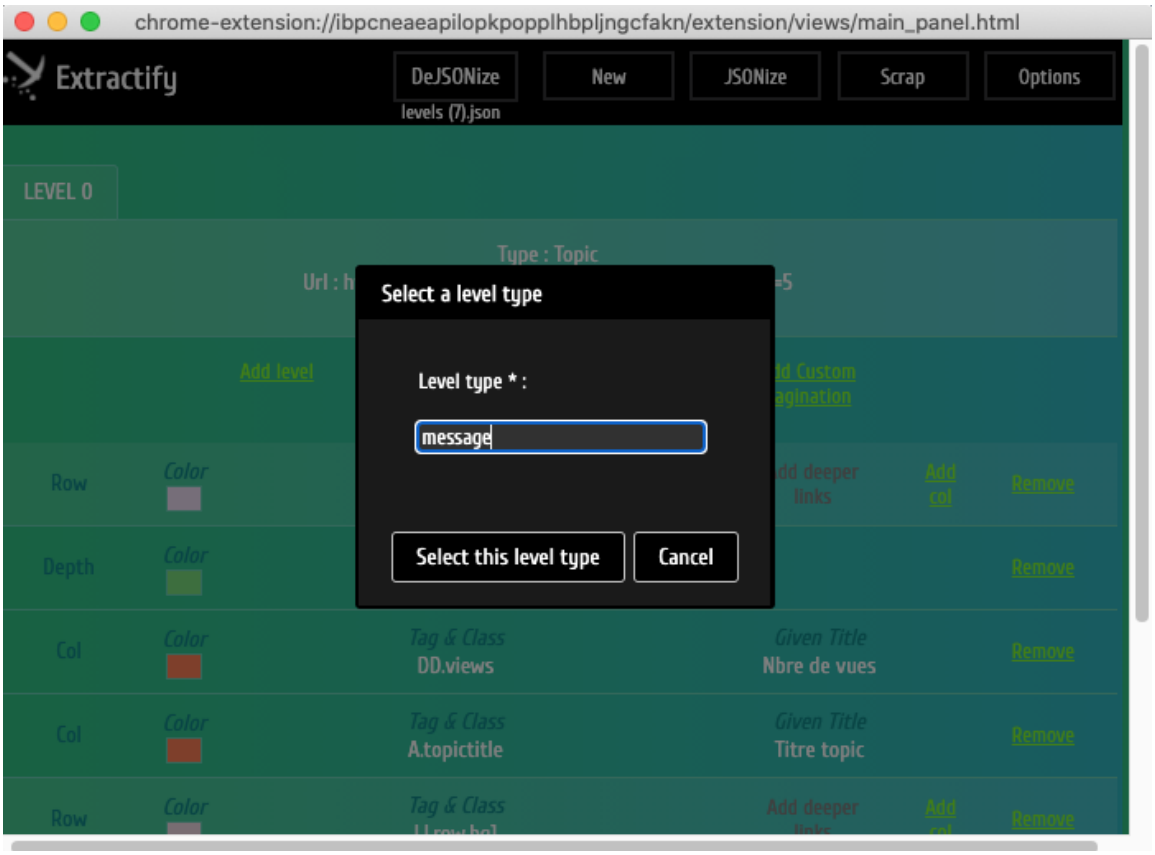
- La page de niveau inférieure est une page de messages. Nous avons vu que la structure de la présentation des topics comportait 2 types différents de lignes : il y a donc fort à parier que la présentation des messages suivra le même schéma. Il nous faudra donc choisir une page avec au moins 2 messages ;
- Toujours sur cette page de messages, nous voulons évidemment récolter le plus d'informations possibles sur les auteurs : leur date d'inscription sur le forum, leur localisation, le nombre de messages qu'ils ont envoyés, un lien vers un site web personnel…etc. : encore une fois, nous aurons besoin d'une page avec suffisamment de messages affichants chacun le maximum d'information ;

- Enfin, l'affichage des messages, comme des topics d'ailleurs, étant limité en nombre pour une meilleure lecture à l'écran, et reporté sur des liens de navigation, si nous voulons obtenir ces liens, il nous faudra là encore cibler une page avec de nombreux messages.

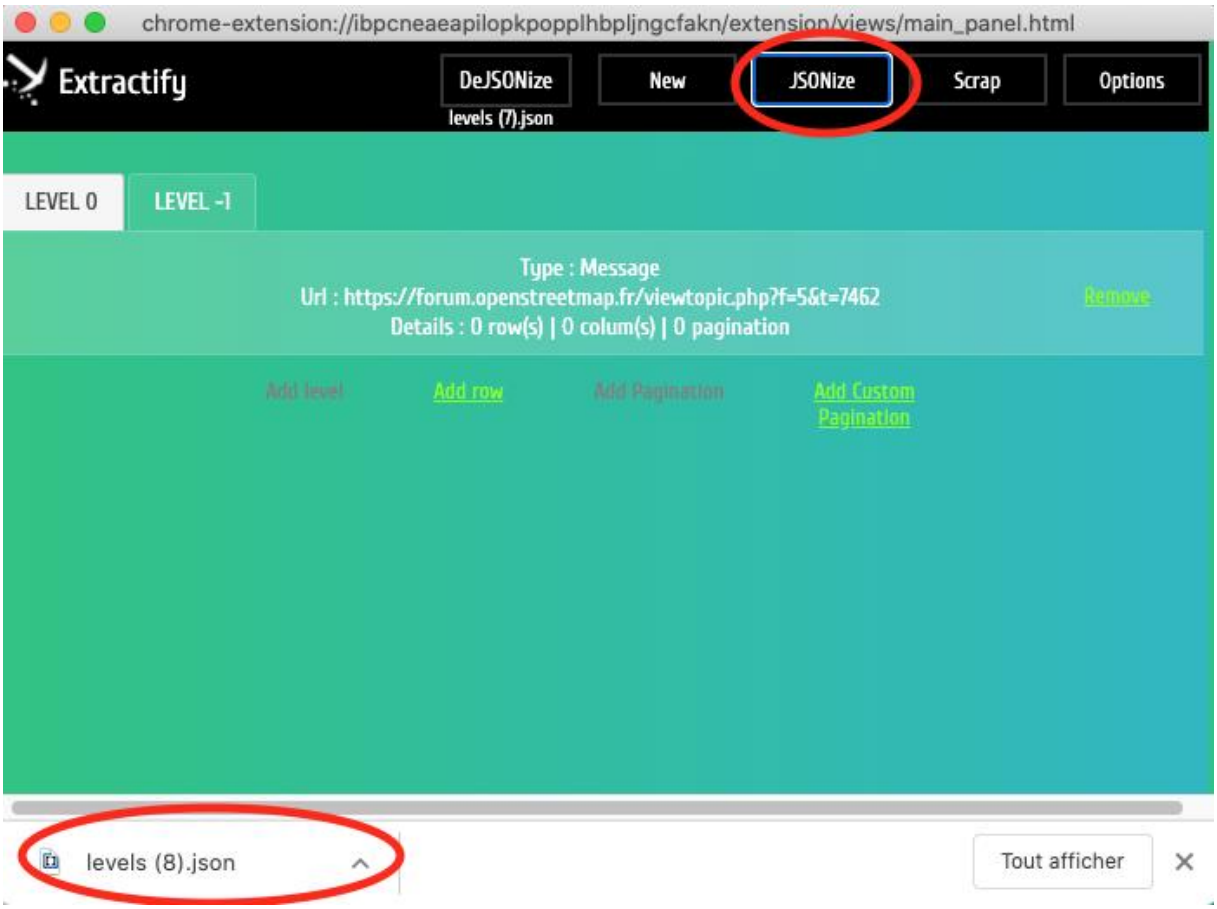
Une première observation visuelle de la page va nous aider à trouver la page qui pourrait remplir tous ces critères de sélection.
Visuellement sur la page web d'affichage des topics, nous pouvons remarquer que le topic « Test Accès BD Ortho IGN » contient un grand nombre de pages de messages. En faisant un clic droit sur la pagination des messages de ce topic puis en sélectionnant « ouvrir le lien dans un nouvel onglet », on peut voir que les 2 premières pages sont vides, mais que la 3ème contient ce que nous recherchons : de nombreux messages et des liens de pagination.
Le problème c'est que cette 3ème page n'est pas référencée par Extractify.
Il faut donc créer un niveau inférieur en cliquant sur « Add level » et sélectionner n'importe quel lien vers une page de niveau inférieur :



Donnez ensuite un nom à ce niveau :



Puis enregistrez votre sélection à l'aide du bonton « JSONize » :



Vous téléchargez ainsi un fichier « levels.json » qui contient la sauvegarde de votre travail de sélection.

A l'aide d'un éditeur de texte, vous pouvez alors modifier l'adresse qui va pointer vers la page de niveau inférieure. Dans notre exemple, il suffit de modifier les variables t et start de façon à cibler la page : « viewtopic.php ?t=4715&start=50 »

Enregistrez ce fichier, puis importez-le dans Extractify en annulant la première boite de dialogue d'ouverture puis en cliquant sur le bouton « DeJsonize ». Vous obtenez bien comme niveau inférieur la 3ème page de messages du topic « Test Accès BD Ortho IGN » qui affiche de nombreux messages et des liens de navigations entre pages de messages.

Au sein de ce nouveau niveau, vous pouvez ainsi recommencer le processus de sélection de lignes et de colonnes.

Les lignes ne posent pas de problèmes.
Les colonnes de date de message et de contenu de message ne posent pas non plus de problèmes particuliers.
Par contre, en ce qui concerne l'auteur du message, il va falloir faire attention et utiliser l'inspecteur pour pouvoir prendre en compte tous les cas de figures :

- Le nom : les noms des auteurs ne sont pas encapsulés de la même manière selon les lignes. Une inspection le confirme : l'attribut « class » peut prendre la valeur « username » ou « username-coloured ». Il faudra le prendre en compte en ciblant les deux types de classes.
- Le rang : certains auteurs en ont un, d'autres pas
- Le nombre de messages et la date d'inscription sont communs à tous, pas de problèmes
- La localisation : là encore, certains l'affichent, d'autres pas
- Pour le contact, malheureusement Extractify n'extrait que le texte, et pas les liens. Ça sera pour une future mise à jour !

Donc pour chaque ligne, nous allons devoir cibler toutes ces colonnes, même si elles n'existent pas pour chaque message. Dans ce dernier cas, nous utiliserons le mode de sélection manuel, soit en simplement copiant-collant, soit en saisissant directement dans l'interface l'élément HTML accompagné de son attribut « class » que nous souhaitons cibler.

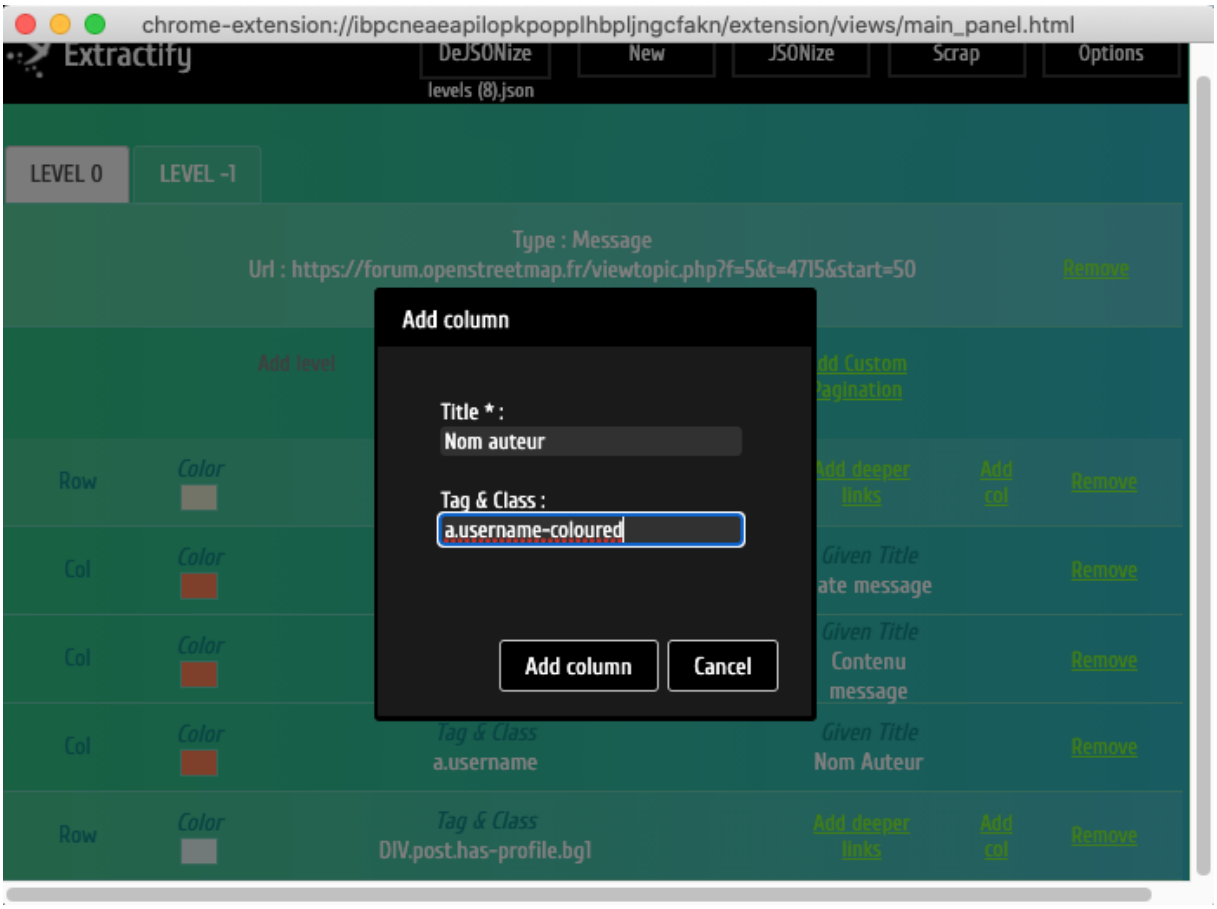
Le nom de l'auteur :

- On commence par sélectionner automatiquement le nom de l'auteur en bleu : pas de soucis
- Pour les noms d'auteurs en rouge, on a repéré dans l'inspecteur qu'ils étaient encapsulés dans une balise <a> accompagnée de la classe « username-coloured ».

C'est à ce moment-là que l'on va utiliser les sélecteurs. Ce sont des mots-clés qui permettent de désigner une catégorie d'éléments de la page.
https://developer.mozilla.org/fr/docs/Web/CSS/S%C3%A9lecteurs_CSS
Comme vous le voyez sur cette page, il y a différent type de sélecteurs, que nous allons pouvoir utiliser pour mieux cibler ce que l'on veut extraire.
Extractify permet aussi d'utiliser des sélecteurs jquery, qui est une bibliothèque javascript, et qui étend en quelques sortes les fonctionnalités des sélecteurs CSS, en permettant par exemple la sélection de textes à partir des mots qu'ils contiennent.
D'une manière générale, lorsque le mode de reconnaissance automatique d'Extractify n'arrive pas à sélectionner ce que vous voulez, le mode manuel par sélecteur vous sauvera !

Dans le cas présent, on va donc utiliser un sélecteur de type, la balise « a », auquel on va ajouter le nom de la classe ciblée, « username-coloured ».

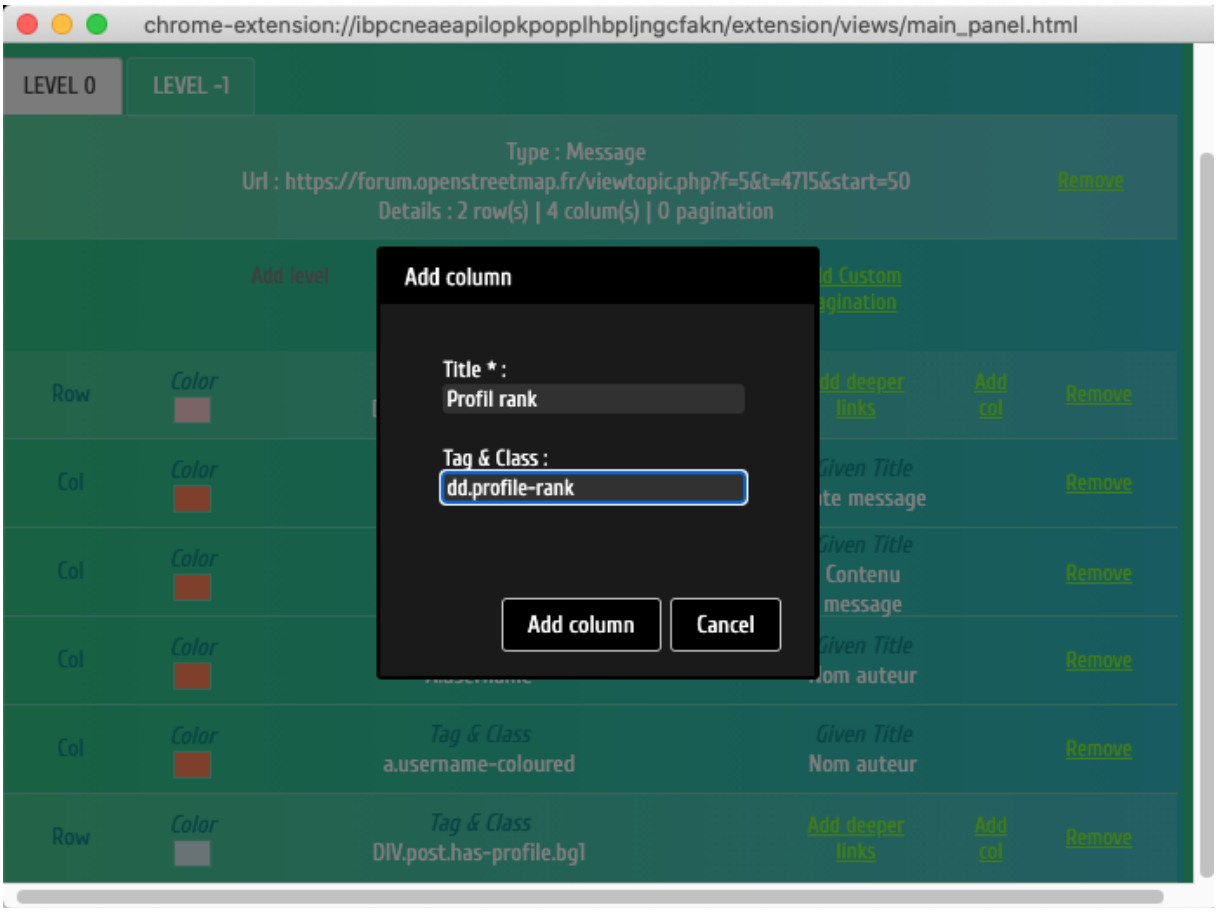
On ajoute donc une colonne, en saisissant le titre, puis en saisissant le nom de la balise accompagnée d'un point et du nom de la classe visée « a.username-coloured », le point étant le caractère signifiant en CSS « class » :



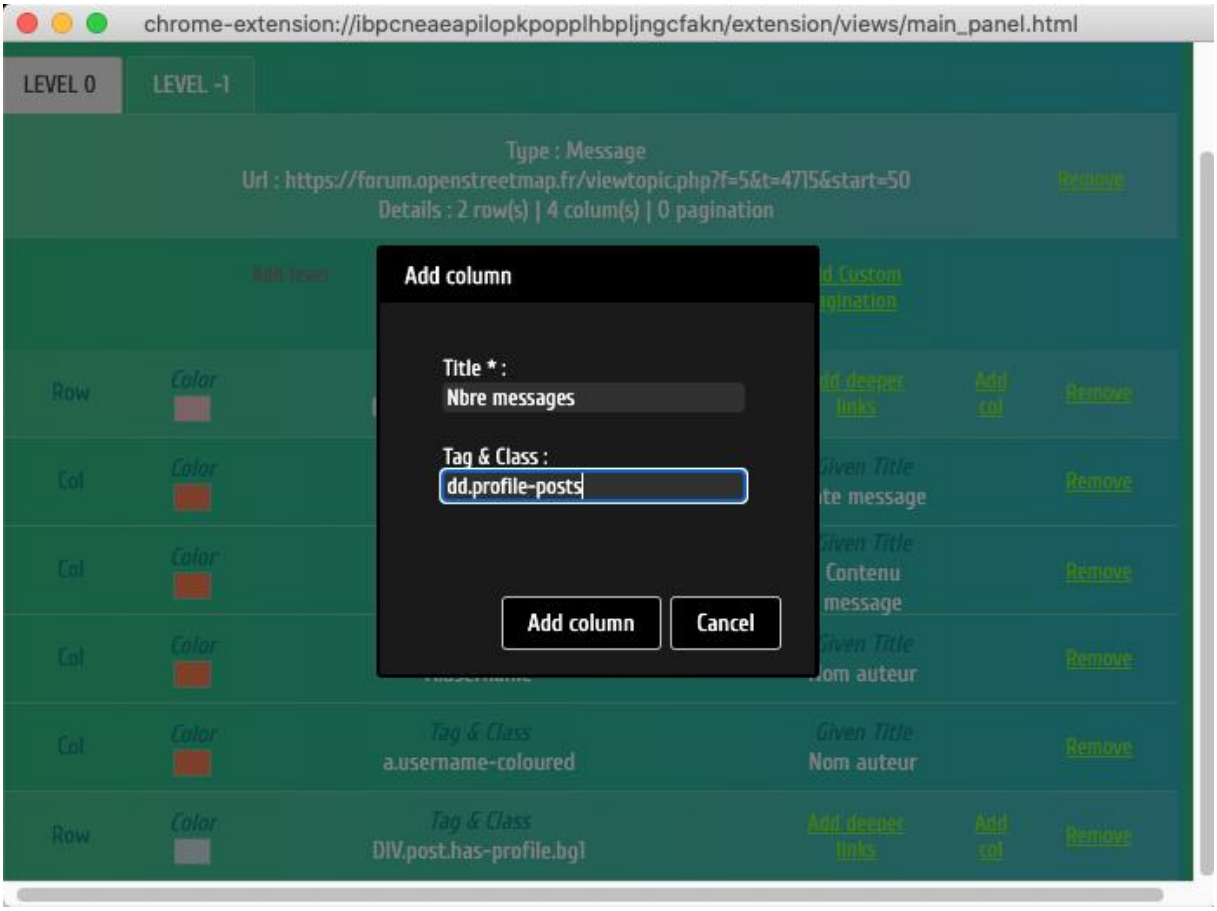
Sur le premier message de la page de messages, la balise <a> accompagnée de la classe « username-coloured » n'existant pas, elle ne sera pas surlignée. Mais si l'on « scroll » plus bas dans la page, on s'aperçoit que l'on a 1 message concerné, voire plusieurs !

Et on recommence de la même manière pour :

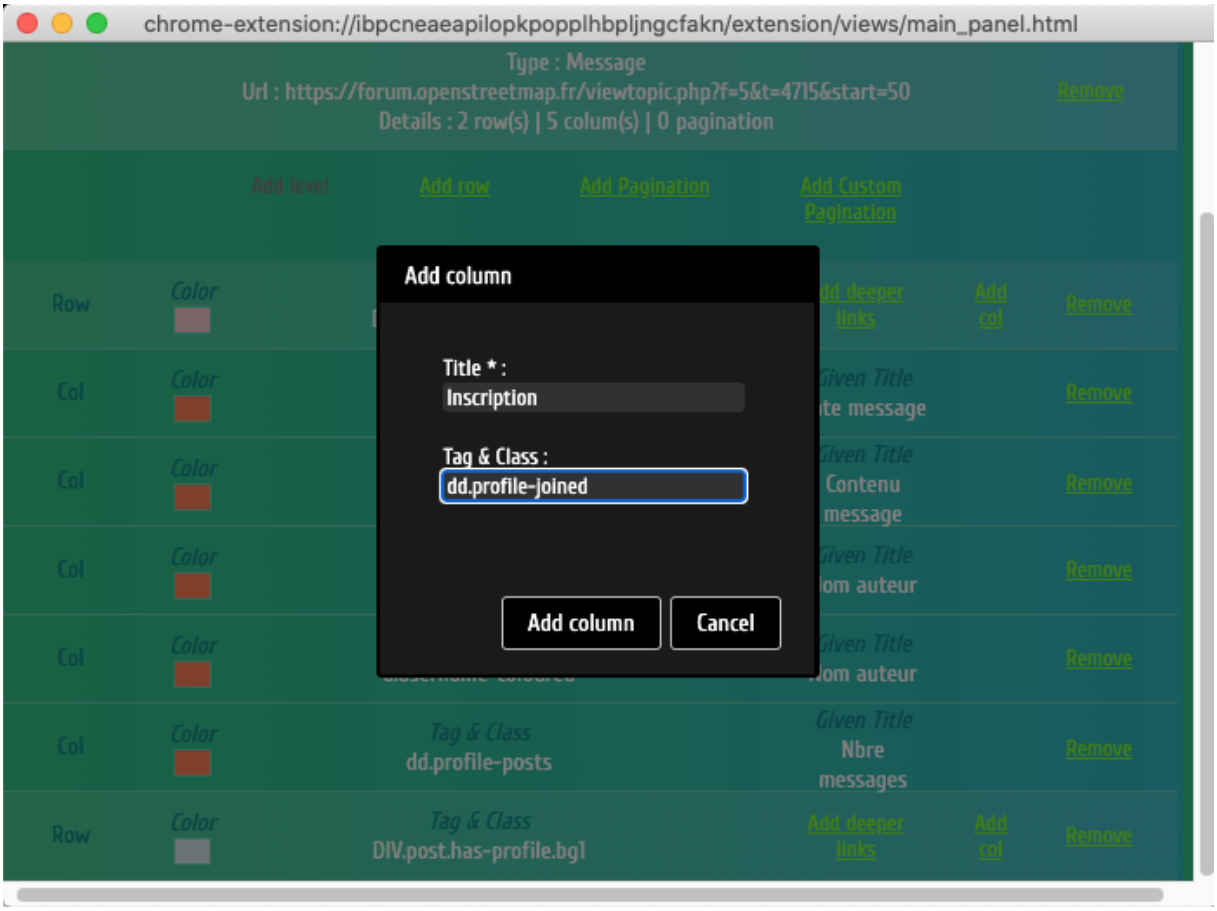
- le rang :



- le nombre de messages



- la date d'inscription



Lorsque vous avez sélectionné toutes les colonnes pour le premier type de ligne, vous faites la même chose pour l'autre type. Vous pouvez à ce moment-là copier/coller les sélecteurs pour aller plus vite !

9. Enregistrement de la sélection

Lorsque vous avez terminé, vous pouvez enregistrer votre travail de sélection en cliquant sur « JSONize » : cela vous permet d'enregistrer les paramètres de votre sélection et de pouvoir le charger ou le modifier ultérieurement.

10. Scraping

Enfin, cliquez sur « Scrap » pour scraper les données encapsulées dans les balises que vous avez sélectionnées. A la fin du processus, le fichier est automatiquement ajouté à vos téléchargements.

11. Format de sortie

Le fichier de sortie est au format json.

Que fait-on avec ce fichier json ?

- La plupart des tableurs vont pouvoir les importer, puis vous pourrez ensuite travailler dessus, les exporter en bdd …etc.
- D'autres logiciels type OpenRefine font aussi très bien ce travail d'importation
- Pour les données strictement conversationnelles, les données de forums de discussion, vous pouvez utiliser un autre outil que j'ai développé, qui s'appelle L@ME : https://github.com/fredericvergnaud/lame

12. Annexes

Extractify supporte les sélecteurs identiques à CSS (ou jquery) :

Aperçu :

- tagname: find elements by tag, e.g. a
- ns|tag: find elements by tag in a namespace, e.g. fb|name finds <fb:name>elements
- #id: find elements by ID, e.g. #logo
- .class: find elements by class name, e.g. .masthead
- [attribute]: elements with attribute, e.g. [href]
- [^attr]: elements with an attribute name prefix, e.g. [^data-] finds elements with HTML5 dataset attributes
- [attr=value]: elements with attribute value, e.g. [width=500] (also quotable, like [data-name='launch sequence'])
- [attr^=value], [attr$=value], [attr*=value]: elements with attributes that start with, end with, or contain the value, e.g. [href*=/path/]
- [attr~=regex]: elements with attribute values that match the regular expression; e.g. img[src~=(?i)\.(png|jpe?g)]

Combinaisons :

- el#id: elements with ID, e.g. div#logo
- el.class: elements with class, e.g. div.masthead
- el[attr]: elements with attribute, e.g. a[href]
- Any combination, e.g. a[href].highlight
- ancestor child: child elements that descend from ancestor, e.g. .body p finds p elements anywhere under a block with class "body"
- parent > child: child elements that descend directly from parent, e.g. div.content > p finds p elements; and body > * finds the direct children of the body tag
- el, el, el: group multiple selectors, find unique elements that match any of the selectors; e.g. div.masthead, div.logo