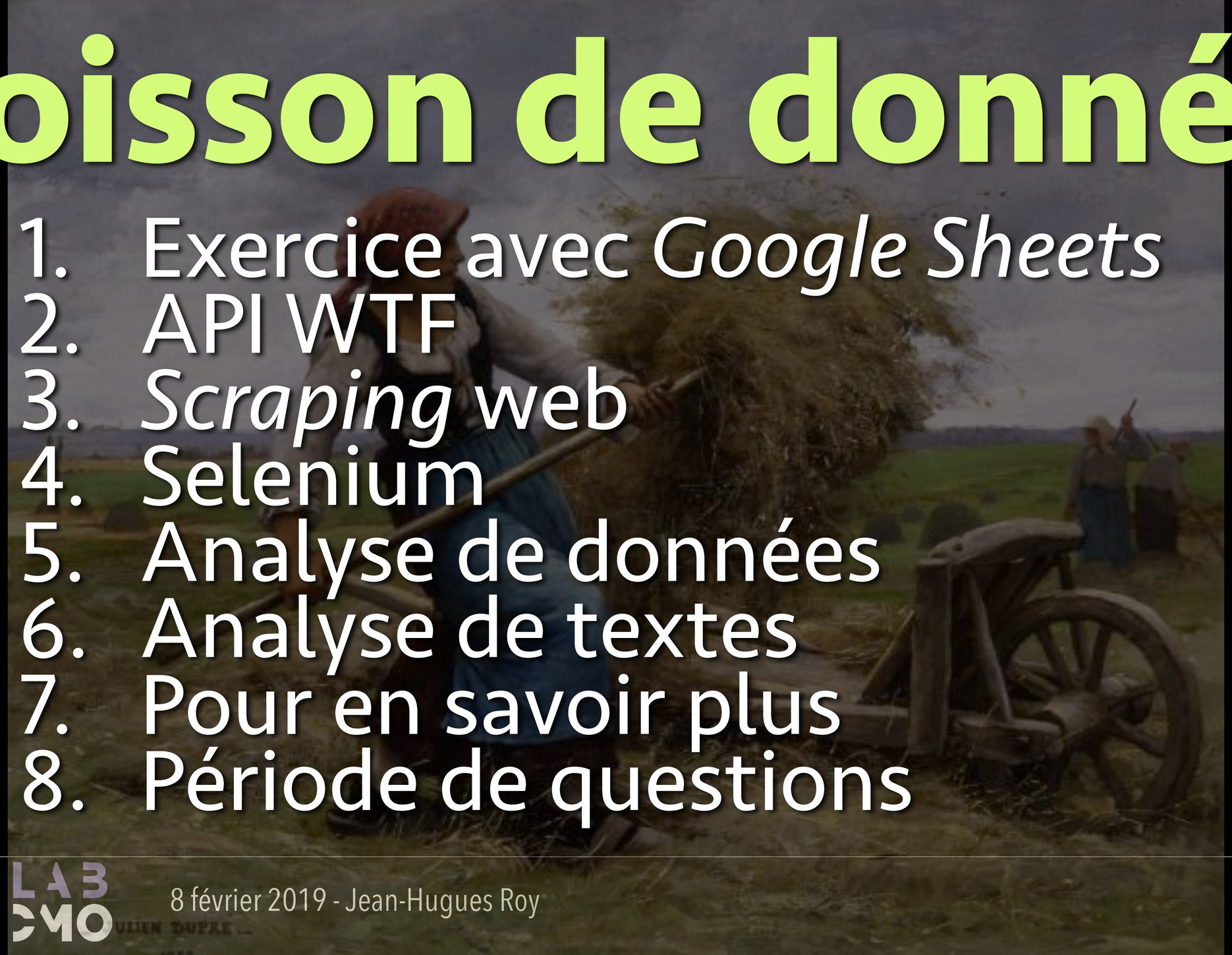
Moisson de données

UQÀM

Moisson de données

1. Exercice avec *Google Sheets*
2. API WTF
3. *Scraping* web
4. Selenium
5. Analyse de données
6. Analyse de textes
7. Pour en savoir plus
8. Période de questions



Google Sheets

Fonctions uniques
N'existent pas dans *OO, LO, Excel, Numbers*

=IMPORTHTML

Rechercher

Jean-Hugues

Jean-Hugues Roy

Fil d'actualités

Messenger

Watch

Marketplace

Raccourcis

Atelier de journalis… 9

Mon Fan Club

Voir plus...

Explorer

Interface pour humains

LEDROIT.CO

«Qui a décidé que la Terre est ronde?» - Nathalie Lemieux

Dans une publication récente, sur Facebook, la conseillère Nathalie…

```
{
  "data": [
    {
      "message": "La nomination d'une proche de la Coalition avenir Québec (CAQ) à la tête de la délégation générale du Québec à New York n'est pas partisane, a a
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "id": "152738728079652_2276881095665394"
    },
    {
      "message": "Un homme a été enlevé par cinq individus dans le rang Saint-Marc, à Saint-Ambroise, et a été battu sévèrement pour une dette de drogue.",
      "caption": "lequotidien.com",
      "status_type": "shared_story",
      "id": "152738728079652_2276866562333514"
    },
    {
      "message": "MONTRÉAL – Félix Auger-Aliassime et Leylah Annie Fernandez ont lancé leur saison 2019 de brillante façon. Les deux jeunes joueurs de tennis cana
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "id": "152738728079652_2276830112337159"
    },
    {
      "message": "OTTAWA – Les frais de garderie ont baissé – ou ont peu augmenté – dans certaines villes canadiennes, ce qui pourrait être un signe que l'argent
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "shares": {
        "count": 1
      },
      "id": "152738728079652_2276800845673419"
    },
    {
      "comments": {
        "data": [
          {
            "message": "Elle est tellement bonne, c'est dommage.",
            "id": "2276771809009656_2276790839007753"
          },
          {
            "message": "Mme Bédard va nous manquer. Merci pour les beaux reportages.",
            "id": "2276771809009656_2276782459008591"
          },
          {
```


Interface pour ordinateurs

Interface pour ordinateurs

REST API

.json


```
{
  "data": [
    {
      "message": "La nomination d'une proche de la Coalition avenir Québec (CAQ) à la tête de la délégation générale du Québec à New York n'est pas partisane, a a
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "id": "152738728079652_2276866562333514"
    },
    {
      "message": "MONTRÉAL - Félix Auger-Aliassime et Leylah Annie Fernandez ont lancé leur saison 2019 de brillante façon. Les deux jeunes joueurs de tennis cana
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "id": "152738728079652_2276830112337159"
    },
    {
      "message": "OTTAWA - Les frais de garderie ont baissé ou ont peu augmenté - sans certaines villes canadiennes, ce qui pourrait être un signe que l'argent
      "caption": "lesoleil.com",
      "status_type": "shared_story",
      "shares": {
        "count": 1
      },
      "id": "152738728079652_2276800845673419"
    },
    {
      "comments": {
        "data": [
          {
            "message": "C'est malheureux.    elle excellait dans tous ses  reportages, triste nouvelle.",
            "id": "2276771809009656_2276799199006917"
          },
          {
            "message": "Elle est tellement bonne, c'est dommage.",
            "id": "2276771809009656_2276790839007753"
          },
          {
            "message": "Mme Bédard va nous manquer. Merci pour les beaux reportages.",
            "id": "2276771809009656_2276782459008591"
          },
          {
```

API

- Facebook
- Twitter
- Google
  - Drive
  - Maps ($)
  - Search ($)
  - Youtube...

- Instagram...
- ~~WhatsApp~~
- Twitch...
- Spotify...
- Uber...
- ~~AirBnb~~
- CanLII
- LCBO
- etc...

API Twitter

D’abord, se créer une «app»

Application under review.

Thanks! We've received your request for API access and are in the process of reviewing it.

Be sure to watch the email address associated with this Twitter account at the time of application, as we may request more information to facilitate the review process in the coming days (be sure to check your spam folder as well).

apps, please edit each of your apps and add a description of your app's use case where it says "Tell us how this app will be used".

We know that this application process delays getting started with Twitter's APIs. This information helps us protect our platform and serve the health of the public conversation on Twitter. It also informs product investments and helps us better support our developer community. For more information about our policies please see our Terms of Service and our Developer Terms.

You'll receive an email when the review is complete. In the meantime, check out our documentation, explore our tutorials, or check out our community forums.

API Twitter

D'abord, se créer une «app»
Intégrer les permissions dans un script.

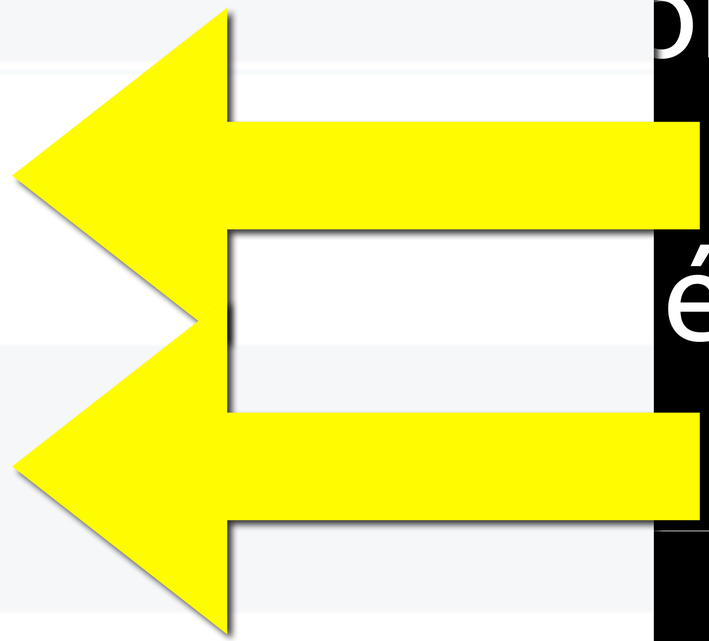
`api-twitter.py`

Utiliser plusieurs mots/expressions
Répéter recherches avec «cron»
Enr. résultats dans base de données
Mais il y a des limites...

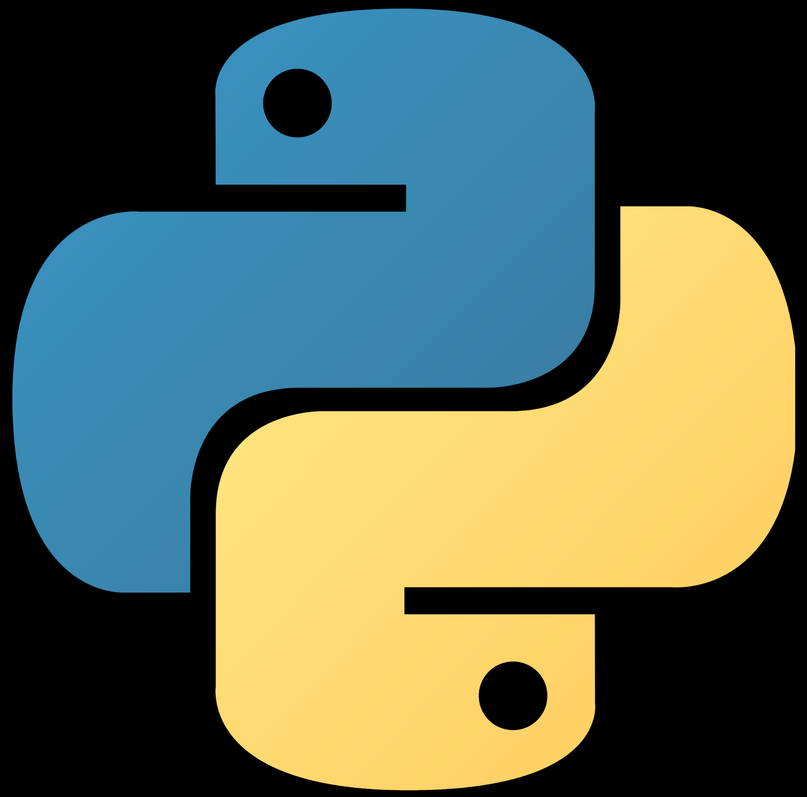
MySQL®

Resource Information

Response formats	JSON
Requires authentication?	Yes
Rate limited?	Yes
Requests / 15-min window (user auth)	180
Requests / 15-min window (app auth)	450



Scraping



+

Python

BeautifulSoup

Scraping

Exemple 1

Objectif :
Ramasser le texte de toutes les lois du Québec en français et en anglais

1re étape :
Recueillir les URLs des lois

`lois01.py`

Scraping

Exemple 1

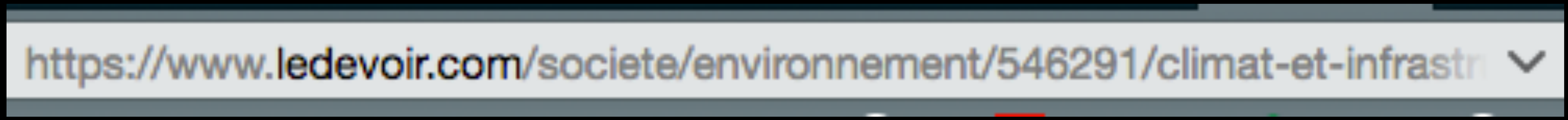
2e étape :
Télécharger les 1042 lois (521 dans chaque langue)

Fichiers HTML (pas PDF 💩) `lois02.py`

`tutoriel`

Scraping

Conseils :
Examinez les URL



Scraping

Conseils :
Examinez les URL

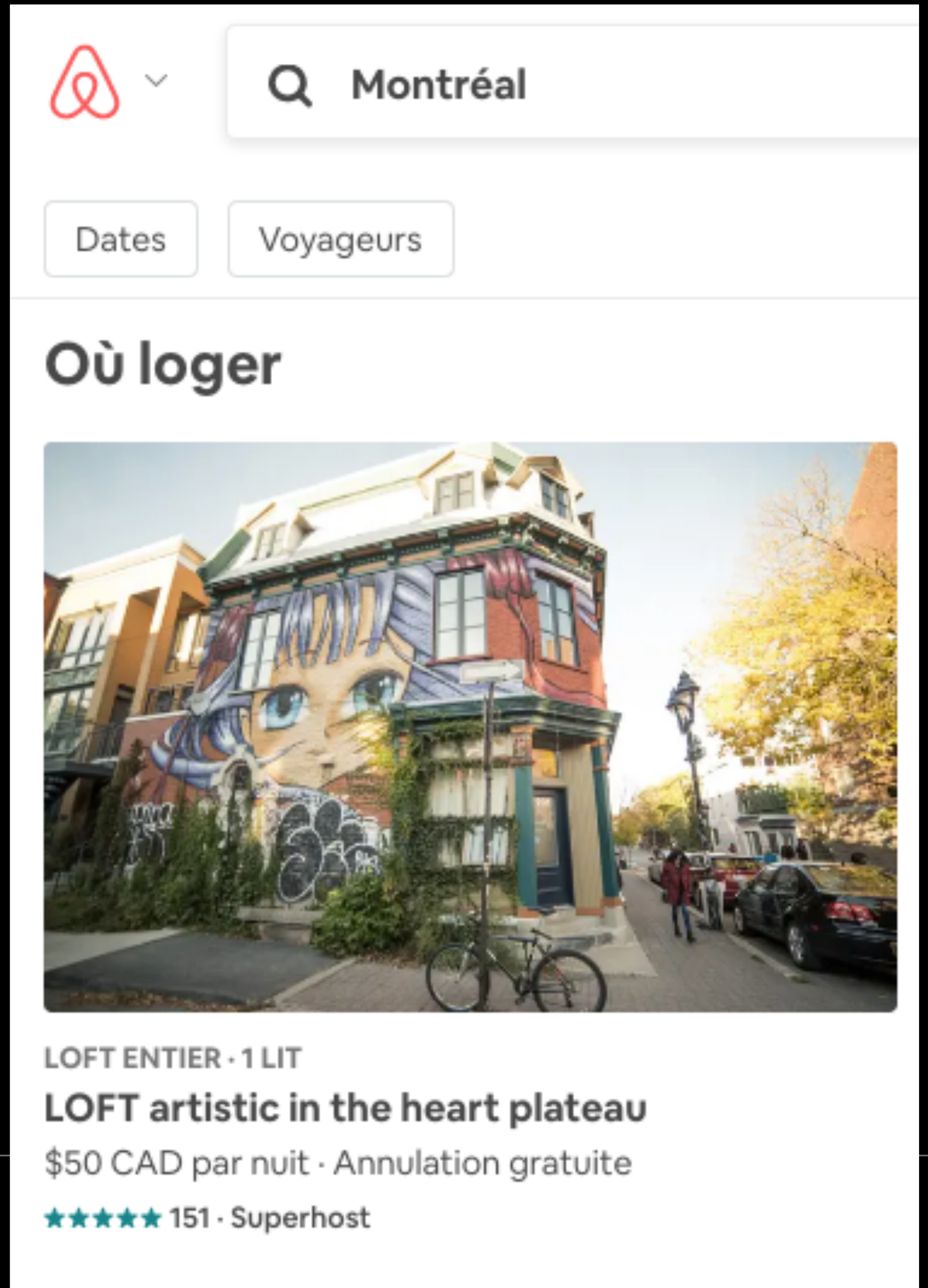
Scraping

Conseils :
Dans le code HTML,
examinez les balises `<meta>`

The New York Times

Scraping

Conseils :
Dans le code HTML, examinez le contenu de certains scripts



Scraping

Stay tuned

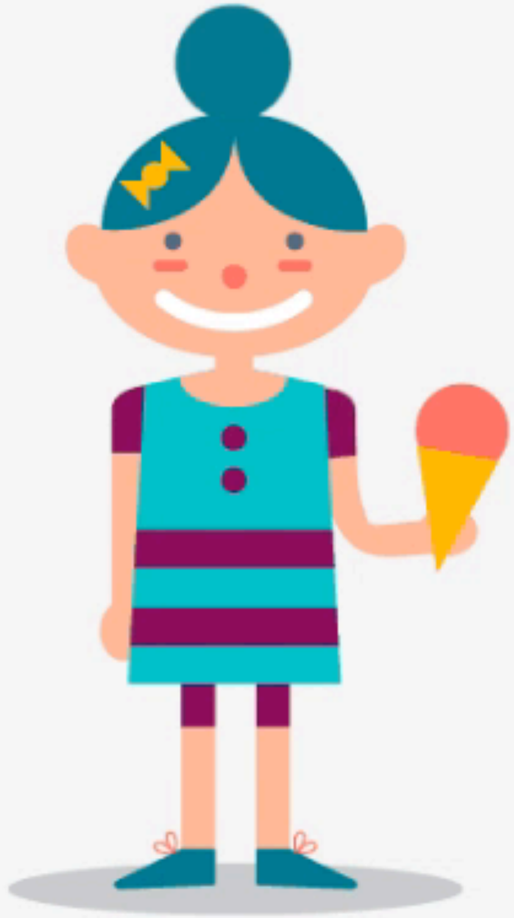
Error code: 503

Airbnb is temporarily unavailable, but we're working hard to fix the problem. We'll be up and running soon! Keep an eye on our Twitter account for updates.

If you need help with an ongoing reservation or for urgent issues, tweet us @AirbnbHelp or call us.

Please note, during site downtime, our response times may be longer than usual.

Thanks for your patience!



Scraping

Difficultés :
Et on fait quoi quand ce qu'on cherche du contenu issu du « deep web »?

COLLÈGE DES MÉDECINS
DU QUÉBEC

Scraping avec Selenium

`md.py`
`bachir.py`

Analyse de données

Tableurs pour journalistes

chercheuse.eur.s

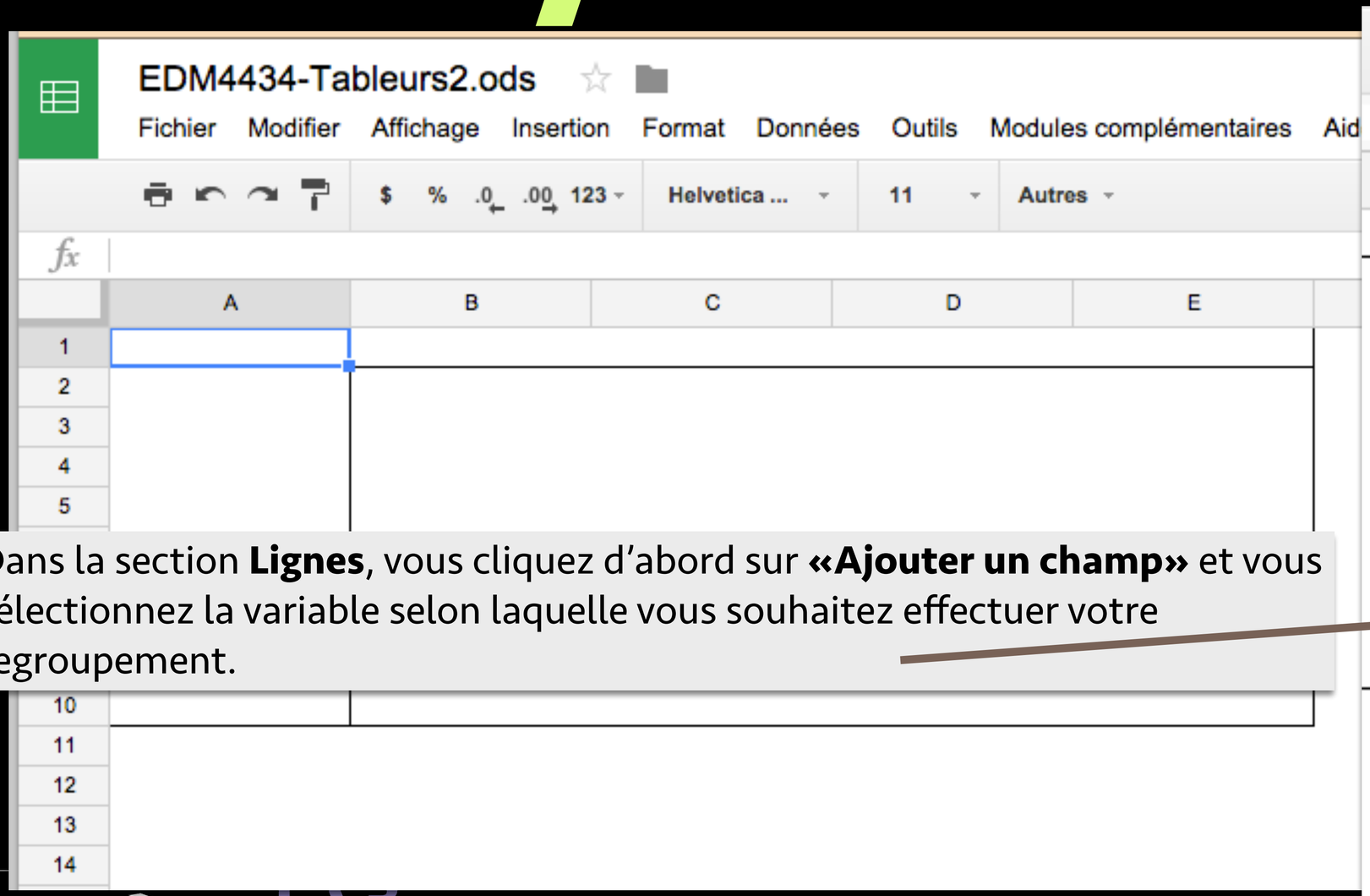
Excel

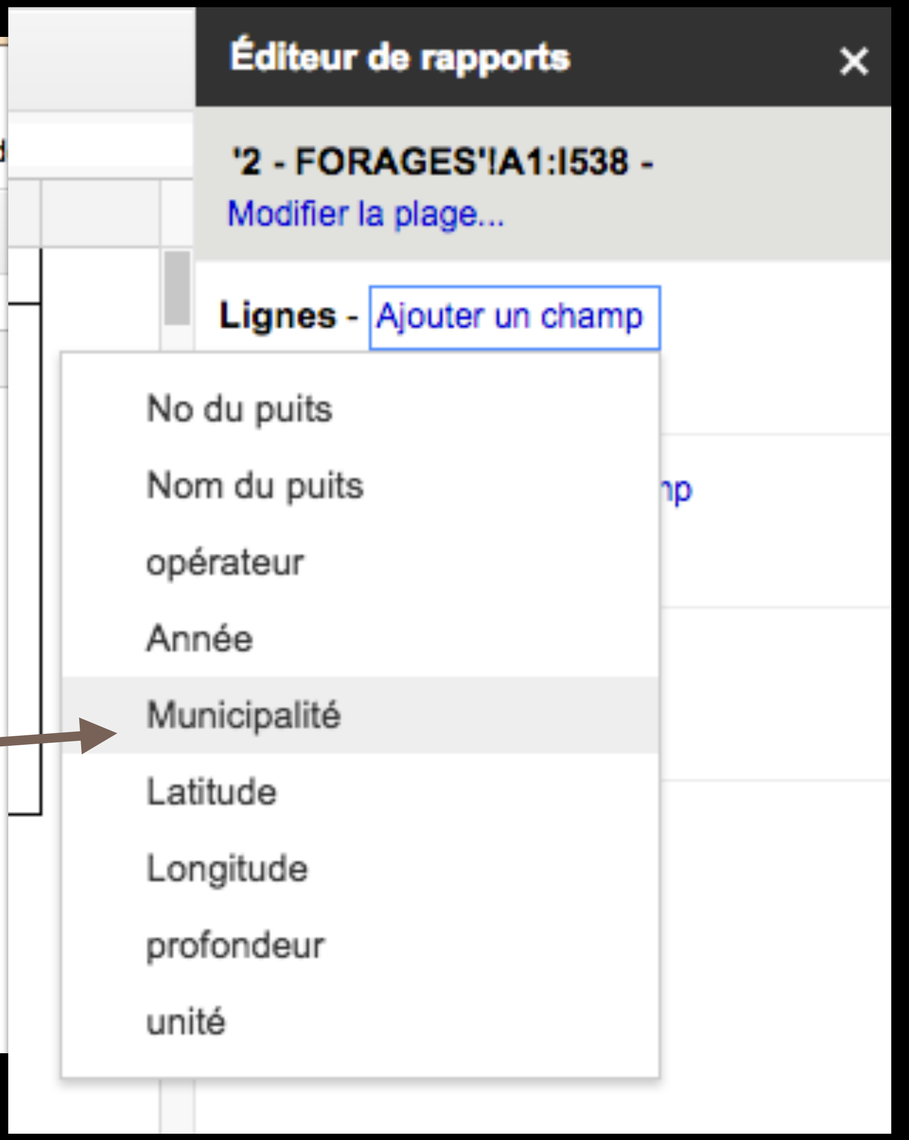
Calc

Numbers

Feuilles de calcul Google

Analyse de données





Dans la section **Lignes**, vous cliquez d'abord sur **«Ajouter un champ»** et vous sélectionnez la variable selon laquelle vous souhaitez effectuer votre regroupement.

Analyse de données

Ouvrir avec LibreOffice

scraping-nettoye.csv

687,8 Mo

Dernière modification le 2017-08-28 22:14:43

`wc dans Terminal`

Analyse de données

MySQL®

UQÀM LAB CMO

8 février 2019 - Jean-Hugues Roy

(MySQL 5.5.30) Hugo serveur/facebook/posts

facebook

Sélectionner une base de données | Structure | Contenu | Relations | Déclencheurs | Info table | Requête | Historique | Utilisateurs

TABLES

- medias
- medias_europe
- posts
- posts_europe
- RCpages
- RCposts

```
sum(partages) as somme_partages,
(sum(partages)/count(titre)) as moyenne_partages,
sum(reactions) as somme_reactions,
(sum(reactions)/count(titre)) as moyenne_reactions,
sum(commentaires) as somme_commentaires,
sum(likes_commentaires) as somme_likes_commentaires,
sum(commentaires_commentaires) as somme_commentaires_commentaires,
((sum(commentaires)+sum(likes_commentaires)+sum(commentaires_commentaires))/count(titre)) as moyenne_commentaires,
(sum(partages)+sum(reactions)+(sum(commentaires)+sum(likes_commentaires)+sum(commentaires_commentaires))) as engagement_total,
((sum(partages)+sum(reactions)+(sum(commentaires)+sum(likes_commentaires)+sum(commentaires_commentaires)))/count(titre)) as
 engagement_moyen,
left(dateqc,7) as mois
from posts
where left(dateqc,4) = "2018"
group by titre, mois
order by engagement_moyen asc
```

Favoris | Historique | Exécuter la requête

titre	publications	somme_partages	moyenne_partages	somme_reactions	moyenne_reactions	somme_commentaires	somme_likes_commentai
Le Journal de Québec	130	6254	48.1077	10661	82.0077	4147	2
URBANIA	217	11883	54.7604	28053	129.2765	3819	1
Protégez-Vous	46	6582	143.0870	2648	57.5652	667	
CTV Montreal	128	12681	99.0703	9831	76.8047	2147	1
MétéoMédia	321	23533	73.3115	35819	111.5857	6389	2
Maurais Live – Radio X	285	17380	60.9825	26432	92.7439	9430	3
Télé-Québec	77	5850	75.9740	8902	115.6104	1369	
Magazine Échos Vedettes	43	565	13.1395	7327	170.3953	2015	
Naître et grandir	72	7380	102.5000	6810	94.5833	1596	
BLVD 102.1	63	3870	61.4286	4493	71.3175	3503	1
BLVD 102.1	76	3941	51.8553	4674	61.5000	6297	1
RDI Économie	59	7388	125.2203	5004	84.8136	701	
TVA Gatineau-Ottawa	264	33903	128.4205	14214	53.8409	6392	3
RDI Économie	91	9155	100.6044	8493	93.3297	1788	1
TVA Sports	590	8222	13.9356	91105	154.4153	15653	12
Jeff Fillion	20	1072	53.6000	2241	112.0500	513	
Le Journal de Québec	537	25918	48.2644	46783	87.1192	22265	16
CTV Montreal	74	5588	75.5135	7961	107.5811	1795	

INFORMATIONS SUR LA TABLE

- créée : 2018-03-04
- moteur : InnoDB
- lignes : ~546 211
- taille : 366,0 MiB
- encodage : utf8mb4

(MySQL 5.5.30) Hugo serveur/facebook/posts

Structure | Contenu | Relations | Déclencheurs | Info table | Requête

```
sum(partages) as somme_partages,
(sum(partages)/count(titre)) as moyenne_partages,
sum(reactions) as somme_reactions,
(sum(reactions)/count(titre)) as moyenne_reactions,
sum(commentaires) as somme_commentaires,
sum(likes_commentaires) as somme_likes_commentaires,
sum(commentaires_commentaires) as somme_commentaires_commentaires,
((sum(commentaires)+sum(likes_commentaires)+sum(commentaires_commentaires))/count(titre)) a
(sum(partages)+sum(reactions)+(sum(commentaires)+sum(likes_commentaires)+sum(commentaires_c
((sum(partages)+sum(reactions)+(sum(commentaires)+sum(likes_commentaires)+sum(commentaires_
 engagement_moyen,
left(dateqc,7) as mois
from posts
where left(dateqc,4) = "2018"
group by titre, mois
order by engagement_moyen asc
```

Favoris | Historique

titre	publications	somme_partages	moyenne_partages	somme_reactions	moyenne_react
Le Journal de Québec	130	6254	48.1077	10661	82
URBANIA	217	11883	54.7604	28053	129

Analyse de données

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

jupyter

jupyter notebook

Analyse de textes

Traitement du langage nat

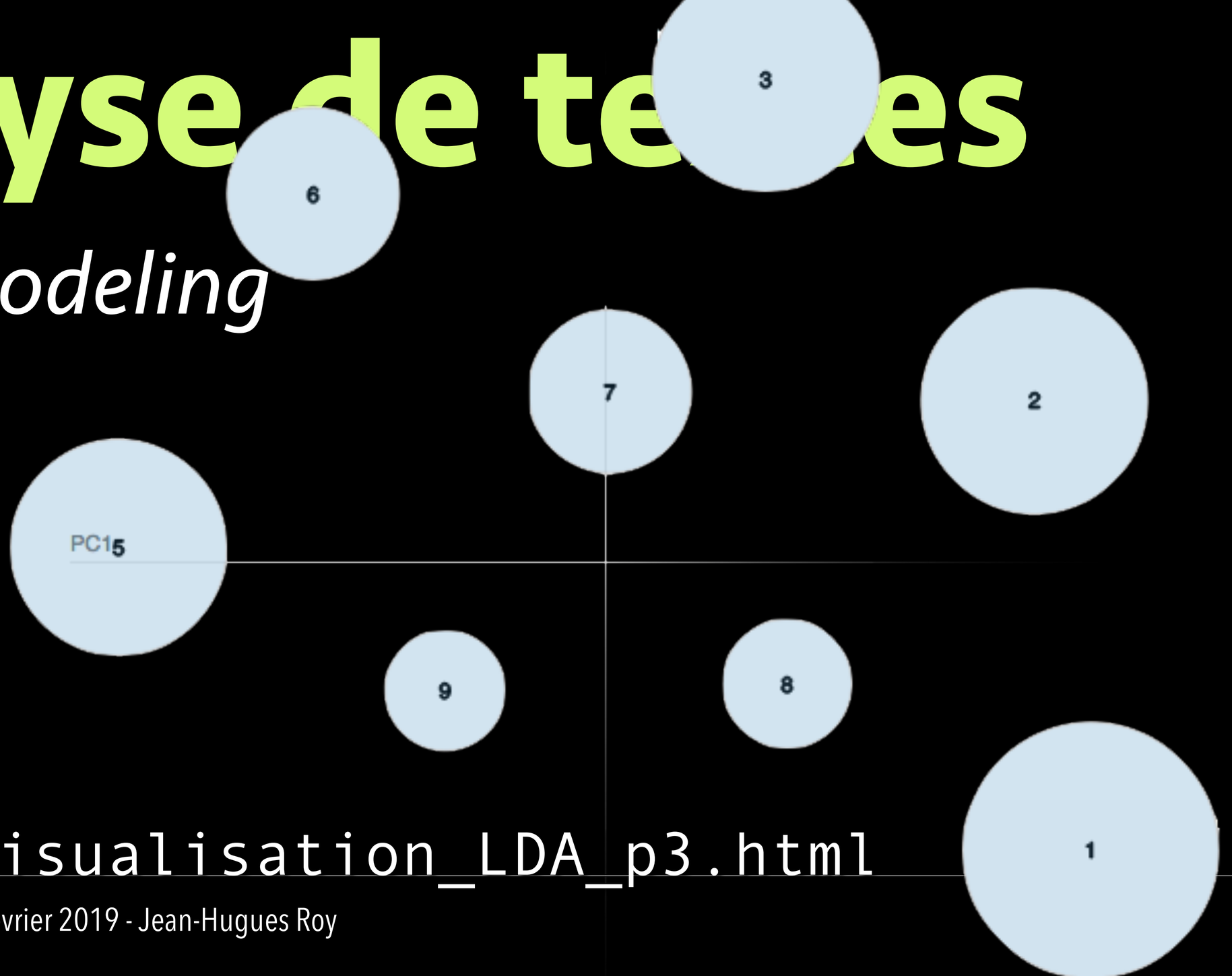
nltk

Trois opérations :

- *Tokenization*
- Traitement des mots-vides
- Lemmatisation

Analyse de textes

Topic modeling



`Visualisation_LDA_p3.html`

UQÀM LAB CMO

8 février 2019 - Jean-Hugues Roy

Ça vous tente?

ANACONDA

anaconda.com

Ça vous tente?

« Écode l'été » :

- Programmation
- Analyse de données massives
- Apprentissage automatique

2019? Sinon 2020.

Merci!

bit.ly/
labcmojhroy