

Journalisme Audiovisuel, Supports Mobiles et Data-Journalisme

Master ProCAN, Semestre 3

Pr. Abdelhak Mahmoudi

abdelhak.mahmoudi@um5.ac.ma

Octobre 2019

16 October 2019

Abdelhak MAHMOUDI

1

02-Recherche de Données

16 October 2019

Abdelhak MAHMOUDI

2

Plan

- Sources des données
- Recherche stratégique - conseils et astuces
- Récupération des données du Web
- Crowdsourcing

16 October 2019

Abdelhak MAHMOUDI

3

Plan

- Sources des données
- Recherche stratégique - conseils et astuces
- Récupération des données du Web
- Crowdsourcing

16 October 2019

Abdelhak MAHMOUDI

4

Sources des données

- Différents types de sources
 - Gouvernementales: statistiques (HCP)
 - Locales, régionales
 - Associations, corporatives, syndicats: rapport annuel
 - Internationales: EU, ONU, OMS, BM, FIFA, etc.
 - Scientifique, Académiques: articles scientifiques, rapports, etc.
 - Plateformes Open data (<http://data.gov.ma>)
- Fonctionnalité d'alerte par email (Newsletter)
- Flux RSS (IFTTT)
- <https://visualping.io>

16 October 2019

Abdelhak MAHMOUDI

5

Sources des données



- Alerte par email
 - Exemple: newsletter de <https://www.finances.gov.ma>

16 October 2019

Abdelhak MAHMOUDI

6

Sources des données



- RSS
 - RDF Site Summary (RDF: Resource Description Framework)
 - Rich Site Summary
 - Really Simple Syndication
- Un type de **flux web** qui permet aux utilisateurs d'accéder aux mises à jour du contenu en ligne dans un format normalisé et lisible par ordinateur
- Document RSS en Format XML
 - Métadonnées
 - Contenu

RSS feed Reader
(Aggregateur)



Sites web
offrant un
service RSS



16 October 2019

Abdelhak MAHMOUDI

7

Sources des données



Document RSS

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">
<channel>
<title>RSS Title</title>
<description>This is an example of an RSS feed</description>
<link>http://www.example.com/main.html</link>
<lastBuildDate>Mon, 06 Sep 2010 00:01:00 +0000 </lastBuildDate>
<pubDate>Sun, 06 Sep 2009 16:20:00 +0000</pubDate>
<ttl>1800</ttl>

<item>
<title>Example entry</title>
<description>Here is some text containing an interesting description.</description>
<link>http://www.example.com/blog/post/1</link>
<guid isPermaLink="false">7bd204c6-1655-4c27-aeee-53f933c5395f</guid>
<pubDate>Sun, 06 Sep 2009 16:20:00 +0000</pubDate>
</item>

</channel>
</rss>
```

16 October 2019

Abdelhak MAHMOUDI

8

Sources des données



- Comment transformer Gmail en un lecteur de flux RSS intelligent?
 - Compte IFTTT : <http://ifttt.com>
 - Chercher le RSS
 - Navigateur,
 - CTRL + U : afficher la source
 - CTRL + F : recherche,
 - Chercher rss
 - Feedly: lecteur RSS
 - RSS FeedFinder: <http://rssfeedfinder.com> (*payant*)

16 October 2019

Abdelhak MAHMOUDI

9

Sources des données



- <https://visualping.io>

16 October 2019

Abdelhak MAHMOUDI

10

Plan

- Sources des données
- Recherche stratégique - conseils et astuces
- Récupération des données du Web
- Crowdsourcing

Recherche stratégique - conseils et astuces

<https://support.google.com/websearch/answer/134479?hl=en>

- **Recherche une phrase:** Mettre la phrase entre le symbole “ ”
- **Combiner les recherches:** Utiliser l'opérateur OR
- **Exclure un terme (ou phrase):** Mettre l'opérateur – avant le mot à exclure
- **Rechercher des phrases avec des mots manquants:** Remplacer les mots manquant par le symbole *
- **Filtrer par date:** Voir le bouton Outils après avoir effectué la recherche
- **Rechercher un terme (ou phrase) dans un seul site web:** Utiliser le format **site:url** où url est une adresse ou un domaine

Recherche stratégique - conseils et astuces

- **Rechercher un terme (ou phrase) dans des sites web similaires:** Utiliser le format **related:url**
- **Trouver un terme (ou phrase) dans une page relié à une autre page:** Utiliser le format **link:url**
- **Rechercher un terme (ou phrase) dans une version en cache:** Utiliser le format **cached:url**. Affichera les résultats avant la modification de l'url.
- **Rechercher un type de fichier spécifique:** Utiliser le format **filetype:type**

16 October 2019

Abdelhak MAHMOUDI

13

Recherche stratégique - conseils et astuces

- **Afficher uniquement les résultats qui incluent le terme dans le titre de la page:** Utiliser le format **intitle:terme**
- **Rechercher une image:** Aller à **images.google.com** et importer une image de votre disque ou bien une url.
- **Rechercher Google sans utiliser Google (et protéger votre vie privée):** Utiliser le site **startpage.com**.
- **Démo**

16 October 2019

Abdelhak MAHMOUDI

14

Plan

- Sources des données
- Recherche stratégique - conseils et astuces
- **Récupération des données du Web**
- Crowdsourcing

Récupération des données du Web

- Comment les récupérer?
 - **Structurées**: à traiter par **Machine** (CSV, XML, JSON, Excel, etc.)
 - Demander à une **institution**
 - **Non-Structurées**: à présenter à un **Humain** (doc, html, pdf, png, etc.)
 - **API Web** (Twitter, Facebook, etc.) pour obtenir des données gouvernementales ou commerciales ou provenant des médias sociaux.
 - **PDF, Image**, logiciels pour extraire les données des fichiers pdf et image, etc.
 - **Scraping** des sites Web. Extraction du contenu structuré à partir de n'importe quelle page Web (non structurée). Regarder cette vidéo https://www.youtube.com/watch?v=FzvN0_N_b2w

Récupération des données du Web (API)

- Twitter (package tweepy, Python)
 - Create twitter app (<https://apps.twitter.com>)
 - Create access tokens
 - Create python environment (conda create --name myenv)
 - Activate python environment (conda create --name myenv)
 - Create python file
 - Execute python file
 - **Démo: test_tweepy.py**



16 October 2019

Abdelhak MAHMOUDI

17

Récupération des données du Web (PDF, PNG)

- Tabula: <http://tabula.technology>
- smallPDF : <https://smallpdf.com>
- Online OCR: www.onlineocr.net
 - Capture écran d'une table
 - à partir d'un article scientifique: exemple: <http://www.sciencedirect.com>
 - Conversion à Excel
- **Démo:**
 - Convertir le fichier Pdf: Femmes et Hommes en chiffres 2016.pdf
 - Rechercher un article scientifique, capture un tableau, convertir



16 October 2019

Abdelhak MAHMOUDI

18

Récupération des données du Web (Scraping)

- Ce que vous ne pouvez pas scraper
 - Code HTML mal formaté
 - Les systèmes d'authentification censés empêcher l'accès automatique
 - Les systèmes utilisant les cookies du navigateur
 - Blocage de l'accès en masse par les administrateurs du serveur de BD
 - Restrictions juridiques du téléchargement des données!
 - Etc.

Récupération des données du Web (Scraping)

- Techniques NLP (Natural Language Processing)
- Outils
 - [Google Drive spreadsheet](#) (Gros volumes de données)
 - [OutWit Hub](#) (*payant*)
 - [Import.io](#) (*payant*)
 - [Chrome extension Scraper](#)
 - [Quickcode](#) (ancien ScraperWiki devenu payant!)
- From scratch
 - HTML, protocole HTTP
 - Python (Beautiful Soup), Ruby, Php

Récupération des données du Web (Scraping)

- Avec Google Drive

- Scraping d'une page web avec Google Drive spreadsheet
- Dans une cellule: `=ImportHtml("URL","query",index)`

- **Démo**

- URL: https://www.finances.gov.ma/Pages/TAND_Maroc_Francais.htm
- Ou bien <http://fr.fifa.com/governance/match-agents/association=eng/index.html>
- query: table (sinon liste)
- Index: 1 (première table)

16 October 2019

Abdelhak MAHMOUDI

21

Récupération des données du Web (Scraping)



- ParseHub

- OutWit (www.outwit.com) :

- Scraping de plusieurs pages
- Télécharger la version gratuite
- Exemple: <http://www.missing-you.net/browse/Liverpool-t9096.php>
- **Démo**

- [Chrome extension Scraper](#)

- **Démo** : https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population
- Inspect element: copy raw: `//*[@id="mw-content-text"]/div/table[2]/tbody/tr`
- Ouvrir scraper (clic droit-> scrape similaire)
- Coller dans Xpath et extraire les colonnes: `td[1]`, `td[2]`, ...

- Scraping with python and [Beautiful Soup](#)

- **Démo** [test_bsoup.py](#)

16 October 2019

Abdelhak MAHMOUDI

22

Plan

- Sources des données
- Recherche stratégique - conseils et astuces
- Récupération des données du Web
- Crowdsourcing

16 October 2019

Abdelhak MAHMOUDI

23

Crowdsourcing

- En français: Production participative, l'externalisation ouverte
- Google Forms, Typeform
- **Force**: Bon pour recueillir des informations sur des éléments dont personne ne connaît vraiment.
- **Faiblesse**: il repose sur l'honnêteté des participants
- Vérifiez l'exactitude
 - Demandez la preuve (plus vous demandez, moins participent)
 - Ajoutez une deuxième couche de vérification, une deuxième source
- Transparence: données proviennent d'un nombre N de participants
- Utiliser aussi pour le nettoyage et l'analyse des données (voir après)

16 October 2019

Abdelhak MAHMOUDI

24