



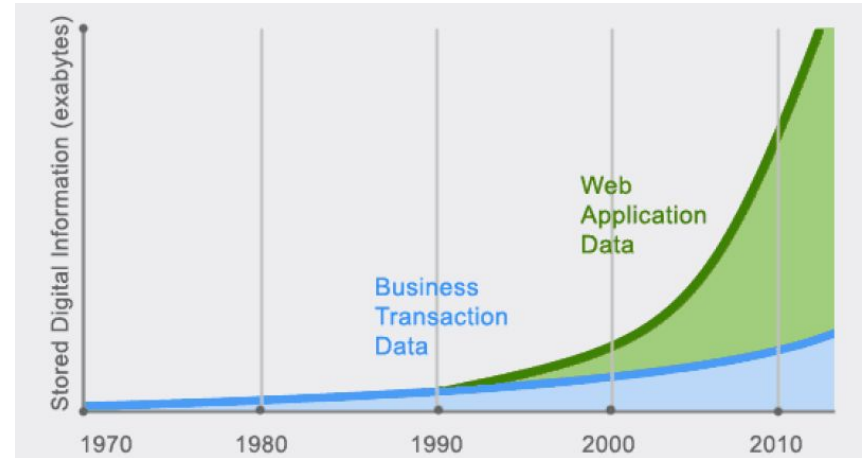
Data Wrangling & Data Cleaning : Project Presentation

**Structured Data Extraction from a social media :
Instagram**

Vincent Gouteux - M2 IASD
25-03-2020

Introduction :

- At the beginning : Create a “**web-database**” containing every information available on the web
- Now : Huge **data business**
- Amount of data generated on the web is growing **exponentially**





Recall on data extraction :

- Arvind Arasu, Hector Garcia-Molina: **Extracting Structured Data from Web Pages.** SIGMOD Conference 2003: 337-348
- Bing Liu, Robert L. Grossman, Yanhong Zhai: **Mining data records in Web pages.** KDD 2003: 601-606
- Michael J. Cafarella, Jayant Madhavan, Alon Y. Halevy: **Web-scale extraction of structured data.** SIGMOD Record 37(4): 55-61 (2008)



Data extraction :

Definition from the paper :

Structured Data is any set of data values conforming to a common *schema* or *type*. A type is defined recursively as follows [1]:

1. The *Basic Type*, denoted by \mathcal{B} , represents a string of *tokens*. A token is some basic unit of text. For the rest of the paper, we define a token to be a word or a HTML tag.



Social media case :

-Every user generates data directly available on the web → Public

Twitter : 500M Tweets are sent each day, **Quite structured**

Instagram : ~100M Posts per day, **structured**

Facebook : 4 petabytes of data created on Facebook, **less structured**

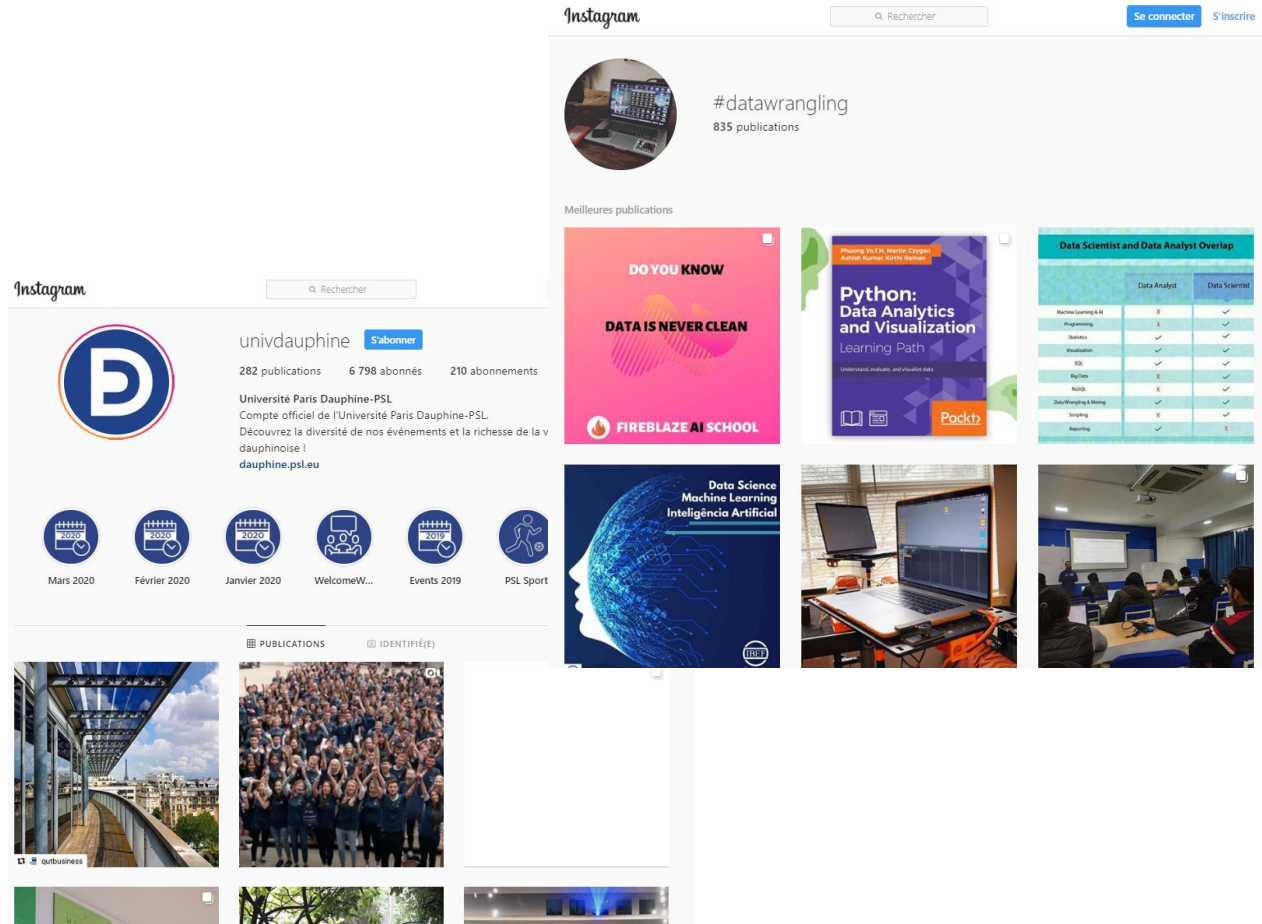
Source : <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>

Instagram :

2 principal ways to retrieve Instagram posts :

- From an **user's page**
- From a **Hashtag**

Examples :
@univdauphine
#datawrangling



Instagram Posts :

Very structured

- User
- Place
- Description
- Number of like
- Comments
- Number of comments
- Date
- Etc.



HTML Code

Instagram

Rechercher

Se connecter

S'inscrire



univdauphine • S'abonner
Université Paris-Dauphine

...



#text 191 x 49

univdauphine Orientation
Dauphine organise sa Journée des
Masters le 07 février 2020 🌍 #JPO

Venez découvrir l'ensemble de l'offre
des Masters (M1 et M2) de Dauphine
dans huit grands domaines : Droit,
Economie, Gestion, Informatique,
Mathématique, Management,
Journalisme, Sciences Sociales 📖

Tout au long de cette journée, les
enseignants, étudiants et diplômés
seront à votre disposition pour
répondre à toutes vos questions sur
les matières enseignées et les
déroulés de ces formations



433 J'aime

24 JANVIER

Connectez-vous pour aimer ou commenter.

qutbusiness

```
<!doctype html>
<html lang="fr" class="js not-logged-in client-root js-focus-visible sDN5V">
  <head>...</head>
  <body class style>
    <div id="react-root">
      <section class="_9eogI E3X2T">
        <div></div>
        <main class="SCxLW o64aR " role="main">
          <div class="Kj7h1_0gdQ3 ">
            <div class="1tEKP">
              <article class="QBxJj M9sTE L_LMM Jyscu ePUX4">
                <header class="PpJfr UE9AK wdOqh">...</header>
                <div class="_97aPb wKhK0">...</div>
                <div class="eo2As ">
                  <section class="ltpMr Slqrh">...</section>
                  <section class="EDFFK ygqzn">...</section>
                  <div class="EtaWk">
                    <ul class="XQXOT pXf-y">
                      <div role="button" class="ZyFrc">
                        <li class="gElp9 rUo9f PpGvg " role="menuitem">
                          <div class="P9vgZ">
                            <div class="C7I1f X7jCj">
                              <div class="RR-M- h5UC0 TKzGu " role="button" tabindex="0">...</div>
                              <div class="C4VMK">
                                <h2 class="_61Ajh ">...</h2>
                                <span class title="Modifié"> == $0
                                  ...
                                <div class="Orientation Dauphine organise sa Journée des Masters le 07 février 2020">
                                  <a class=" xil3i" href="/explore/tags/jpo/"#JPO/>a
                                  <br>
                                  <br>
                                  "Venez découvrir l'ensemble de l'offre des Masters (M1 et M2) de Dauphine
                                  dans huit grands domaines : Droit, Economie, Gestion, Informatique,
                                  Mathématique, Management, Journalisme, Sciences Sociales 📖"
                                  <br>
                                  <br>
                                  "Tout au long de cette journée, les enseignants, étudiants et diplômés sera
                                  à votre disposition pour répondre à toutes vos questions sur les matières
                                  enseignées et les débouchés de ces formations ."
                                  <br>
                                  <br>
                                  ⚠ Entrée libre mais soumise à inscription : www.weezevent.com/journee-
                                  masters-2020-02-07"
```




Tools

Instagram API : Interface that “allows” data collection from instagram.

- Need an access_token
- “Too easy” !

Much mote interesting to scrap via HTML code :

- **BeautifulSoup** : Packages that make scrapping much easier.
- Still too easy !

From scratch with python requests

```
user = univdauphine
r = requests.get("https://www.instagram.com/{0}/?__a=1".format(user))
htmlcode = r.text
data = json.loads(htmlcode.strip(),strict = False)
```

```
r = requests.get("https://www.instagram.com/p/B7s9C8s0icn/?__a=1")
htmlcode = r.text
data = json.loads(htmlcode.strip(),strict = False)
data

{
  'graphql': {
    'shortcode_media': {
      '__typename': 'GraphImage',
      'accessibility_caption': 'Photo by Université Paris Dauphine-PSL in Université Paris Dauphine - PSL. Image may contain: sky, cloud and outdoor',
      'caption_is_edited': True,
      'commenting_disabled_for_viewer': False,
      'comments_disabled': False,
      'dimensions': {
        'height': 1080,
        'width': 1080,
      },
      'display_resources': [
        {
          'config_height': 640,
          'config_width': 640,
          'src': 'https://scontent-iad3-1.cdninstagram.com/v/t51.2885-15/sh0.08/e35/p640x640/82767862_153372909436658_287435489823473576_n.jpg?nc_ht=scontent-iad3-1.cdninstagram.com&nc_cat=102&nc_ohc=N56sXDMoPIAAX8xzngG&oh=f0808...',
        },
        {
          'config_height': 750,
          'config_width': 750,
          'src': 'https://scontent-iad3-1.cdninstagram.com/v/t51.2885-15/sh0.08/e35/p750x750/82767862_153372909436658_287435489823473576_n.jpg?nc_ht=scontent-iad3-1.cdninstagram.com&nc_cat=102&nc_ohc=N56sXDMoPIAAX8xzngG&oh=ad4d...',
        },
        {
          'config_height': 1080,
          'config_width': 1080,
          'src': 'https://scontent-iad3-1.cdninstagram.com/v/t51.2885-15/e35/p1080x1080/82767862_153372909436658_287435489823473576_n.jpg?nc_ht=scontent-iad3-1.cdninstagram.com&nc_cat=102&nc_ohc=N56sXDMoPIAAX8xzngG&oh=4926a6a82c...',
        }
      ],
      'display_url': 'https://scontent-iad3-1.cdninstagram.com/v/t51.2885-15/e35/p1080x1080/82767862_153372909436658_287435489823473576_n.jpg?nc_ht=scontent-iad3-1.cdninstagram.com&nc_cat=102&nc_ohc=N56sXDMoPIAAX8xzngG&oh=4926a6a82c...',
      'edge_media_preview_comment': {
        'count': 8,
      },
      'edges': [
        {
          'node': {
            'created_at': 1581237399,
            'did_report_as_spam': False,
            'edge_liked_by': {
              'count': 0,
            },
            'id': '1784421854968458',
            'is_restricted_pending': False,
            'owner': {
              'id': '1405010293',
              'is_verified': False,
            },
          },
        },
      ],
    },
  },
}
```



Techniques

- 1) Post by posts \implies Problem : Too many requests, **Kicked** after 20 posts
- 2) From an User's Page \implies Less information but works
- 3) From an Hashtag \implies Less information but works

Problem : HTML codes are different :

3 Methods : `Data_From_Posts`, `Posts_Data_From_User`,
`Posts_Data_From_Tag`

Results

10 firsts posts for “#datawrangling”


	Tag_Id	Tag_Name	Extraction_Date	Posts_W_Hashtag	Owner_Id	Shortcode	Timestamp	Likes	Hashtags	Mentions	Com's	Text	Img_Description
0	17843831179043460	datawrangling	21032020	862	10602884403	B9zqLYug8bL	2020-03-16 20:11:07	39	[datatechcon\nNeed, dashboard, kpi, smallbusin...	[datatechcon]	0	Customers are the foundation of any business'	Video
1	17843831179043460	datawrangling	21032020	862	9299626406	B9yFEDcAKx5	2020-03-16 05:18:48	27	[businessintelligence\n, automateddatawranglin...	[]	4	Why wouldn't you prefer the speed and data pro...	Video
2	17843831179043460	datawrangling	21032020	862	13935806959	B9o5bH1gczi	2020-03-12 15:42:41	12	[datawrangling, businessintelligence, inglés, ...	[]	0	#datawrangling #businessintelligence #inglês #...	1 person, text
3	17843831179043460	datawrangling	21032020	862	9299626406	B9nlb2IAtH_	2020-03-12 03:31:24	14		[]	1	Reduce human resource costs by taking complete...	Video
4	17843831179043460	datawrangling	21032020	862	12763554440	B9gIW5oAt9N	2020-03-09 06:00:01	35	[ml, machinelearningalgorithms, machinelearnin...	[data, data, data]	1	A very subtle and sharp way to understand the ...	possible text that says 'Statistics: Given th...
5	17843831179043460	datawrangling	21032020	862	12763554440	B9eJomnAIOP	2020-03-08 11:32:41	32		[data, data, data]	1	Some very cool and handy R packages which help...	Null
6	17843831179043460	datawrangling	21032020	862	186032245	B9ZVgKvBMRe	2020-03-06 14:40:12	49	[CineColombiano, detrasdecamaras, datawrangling]	[ficcifestival,]	0	Después de un tiempo dando vueltas por el mund...	1 person, standing and text
7	17843831179043460	datawrangling	21032020	862	9299626406	B9TD-UGgv50	2020-03-04 04:11:35	10		[]	1	Change reality into a more desirable state wit...	1 person, beard and outdoor, possible text th...
8	17843831179043460	datawrangling	21032020	862	18075761281	B9RXXdHHKji	2020-03-03 12:22:33	10	[data, datascience, datasavy, datascientist, d...	[]	0	Data visualization tools... #data#datascience#...	Null
9	17843831179043460	datawrangling	21032020	862	27474528070	B9HfkqQH0s-	2020-02-28 16:21:51	15	[fridayfeeling, fridaytreats, ditpic, davincir...	[]	1	Cheeky Friday beer to round off the weeks work...	Null

Results

10 firsts posts of the page “@univdauphine”

User_Id	User_Name	Is_Verified	Followers	Extraction_Date	Nb_Of_Posts	Owner_Id	Shortcode	Timestamp	Likes	Hashtags	Mentions	Com's	Text	Img_Description
257662767	univdauphine	False	6868	21032020	283	257662767	B99yqg6oJxi	2020-03-20 18:27:40	58	['Covid19']		0	#[Covid19] Dauphine est fermée jusqu'à nouvel ...	é Paris Dauphine-PSL on March 20, 2020. Image ...
257662767	univdauphine	False	6868	21032020	283	257662767	B7s9C8soicn	2020-01-24 12:28:31	436	['JPO'\n'nVenez', 'Repost', 'assoetudiante', 'v...]	['qutbusiness', 'univdauphine', 'itsabrina']	8	Orientation 🏠 Dauphine organise sa Journée des...	é Paris Dauphine-PSL in Université Paris Dauph...
257662767	univdauphine	False	6868	21032020	283	257662767	B7YHb_zHKZG	2020-01-16 10:19:19	423		['univdauphine', 'figaroetudiant']	10	🏆 Vous avez classé @univdauphine en 2e place d...	video
257662767	univdauphine	False	6868	21032020	283	257662767	B7EPCLhCon_	2020-01-08 16:56:50	450	['parcoursup']		34	Nouveauté 2020 🏠 Dès le 22 janvier, postulez e...	Null
257662767	univdauphine	False	6868	21032020	283	257662767	B5DM9JxoSb9	2019-11-19 14:16:38	255	['SEEPH2019'], 'SemaineduHandicap', 'DauphineD...]		1	#[SEEPH2019] Hand'icap sur Dauphine ! L'univer...	é Paris Dauphine-PSL on November 19, 2019. Ima...
257662767	univdauphine	False	6868	21032020	283	257662767	B3_s2JCAwhi	2019-10-24 09:06:57	225	['DauphineDurable', 'ODD']		11	Un projet, une idée ou une initiative concern...	é Paris Dauphine-PSL in Université Paris Dauph...
257662767	univdauphine	False	6868	21032020	283	257662767	B3PeofXohs0	2019-10-05 15:39:12	716			4	Félicitations aux près de 1400 diplômées et di...	é Paris Dauphine-PSL in Université Paris Dauph...
257662767	univdauphine	False	6868	21032020	283	257662767	B269Z0ECofG	2019-09-27 16:24:02	445	['HousingByDauphine,']	['dauphine-housing']	5	Hier soir, Isabelle Huault, Présidente de Daup...	é Paris Dauphine-PSL on September 27, 2019. Im...
257662767	univdauphine	False	6868	21032020	283	257662767	B2WpRIPCevD	2019-09-13 13:55:29	606	['RentréeDauphine', 'LSO', 'assoetudiante', 'v...]	['univdauphine', 'psi_univ']	9	#RentréeDauphine des L1 #LSO \nL'occasion de d...	é Paris Dauphine-PSL on September 13, 2019. Im...
257662767	univdauphine	False	6868	21032020	283	257662767	B0YVFkhiESV	2019-07-26 12:34:46	303	['summer', 'paris', 'dauphine', 'psi', 'univer...]	['univdauphine', 'psi_univ']	1	[Fermeture estivale] Le campus parisien de @un...	é Paris Dauphine-PSL in Université Paris Dauph...
								2019-07-26 12:34:46		['10ans', 'Carthage']				é Paris Dauphine...

Problems & Additional information

Problem : Description  Solution

Kicked requests	Can not send a large amount of requests	Send one request that collect every post
Exceptions	Posts without descriptions, Videos ...	Add tests everywhere
End cursor	Collecting through several pages	Search in HTML code what happens while scrolling

Additional options, to make data cleaning easier and collect only relevant posts :

Number of likes minimum, Verified account, Creation of csv file...



Further Work and Applications

Can be used for statistics, quantify trends, follow the behaviour of a brand, a celebrity...

Can be subject to machine learning algorithms :

- 1) Automatization → Scrap the same account every week : Quantify trend, evolution ...
- 2) Deep learning on image → huge amount of images
- 3) NLP descriptions, comments → Sentiment analysis, topic modelling...



Conclusion

- Very “Naive” scrapping methods that decompose HTML code.
- Works well, as instagram posts ar very structured data.
- Does not need a lot of cleaning
- Machine learning or statistical models easy to apply.