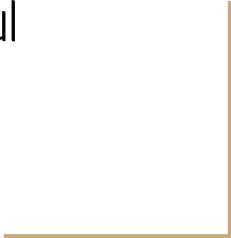




# BOOTCAMP

# DATA ETL

Make Data Powerful



# Work with text data

## Feature extraction

# Feature extraction

## Stat about paragraphs, sentences, words, ...

- Count (Paragraphs, Sentences, Words, Characters, Numbers, ...)
- Average
- ...

## Extract information

- proper noun
- most frequent word
- ...

# Work with text data

## Text Pre-processing

# Text pre-processing

- Lower casing  
This is a new kind of learning experience !
- Remove punctuation/Special characters/Line return/...  
this is a new kind of learning experience
- Stopwords  
new kind learning experience
- Tokenization  
[new, kind, learning, experience]
- Stemming / Lemmatization  
[new, kind, learn, experience]

# Work with text data

## Advanced Text Processing

# Advanced Text Processing : Bag of words

- A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms.
- The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents.
- It is called a “*bag*” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

# Advanced Text Processing : Bag of words

- N-grams
- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- TF-IDF
- Count Vectorizer

# Work with text data

## Tools

# Tools

- Pandas
  - nltk
  - sklearn
- 
- TextBlob
  - Spacy
  - ...



# ETL - Extract

## Live Extraction : Scraping & loading text

- What is “Web Scraping” ?
- What is HTML ?
- How to retrieve an HTML page from website ?
- How to parse, extract and load HTML content ?

## Live Extraction : API

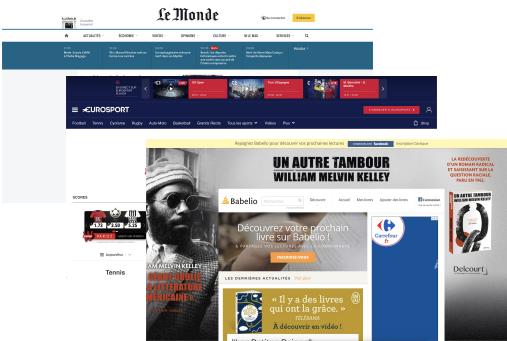
- What is an API ?
- How to interact with an API ?
- Tools to interact with an API

# Live extraction : Web Scraping & Loading Text

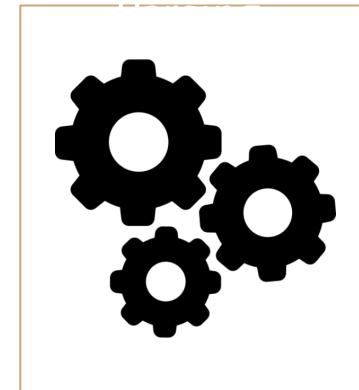
## What is web scraping ?

# What is web scraping ?

Web scraping is a technique for **extracting the content of websites**, via a script or a program, in order to transform it to allow its use in another context.

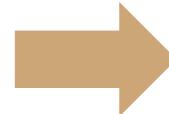
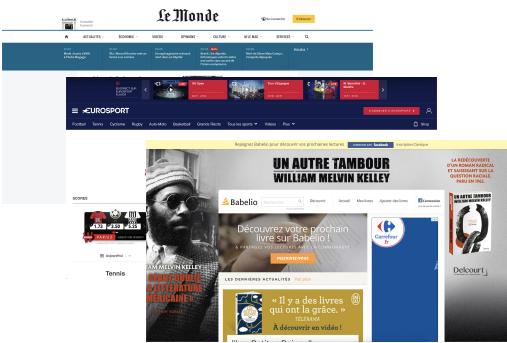


## Scraping

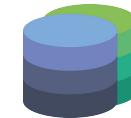
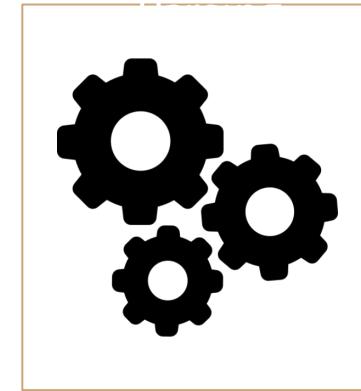


# What is web scraping ?

Web scraping is a technique for **extracting the content of websites**, via a script or a program, in order to transform it to allow its use in another context.



## Scraping + Parsing



# Live extraction : Web Scraping & Loading Text

## What is HTML ?

# What is HTML(Hyper Text Markup Language) ?

HTML is the markup language designed to **represent web pages**. **HTML allows you to structure semantically and logically** and to **format the content of pages**, to include multimedia resources (images, videos, ...), input forms and computer programs. It is often used in conjunction with the JavaScript programming language and cascading style sheets (CSS)

 WIKI

WIKIPÉDIA

#### **Portails thématiques**

Contact

## Contributor

[Aide](#)

## **Modifications récentes**

#### **Plane um den**

Páginas libres

Sainte-Croix pagines libres  
Télécharger au format

Unacademy

Informations sur la page  
Réseau WhatsApp

[Clear entire page](#)

#### **Winnipeg**

CONTINUOUS  
WIRING

www.oxfordjournals.org

[Créer un livre](#)

Unchargeable costs  
PDF

## languages

## Benzene

Wiki Loves Monuments : photographiez un monument historique, aidez Wikipédia et gagnez !

En apprendre plus

## Hypertext Markup Language

(Rédigé depuis [Hmtl])

L'**Hypertext Markup Language**, généralement abrégé **HTML**, est le langage de balisage conçu pour représenter les pages web. C'est un langage permettant d'écrire de l'hypertexte, d'un son nom, HTML, permet également de structurer sémantiquement et logiquement et de mettre en forme le contenu des pages, d'instruire des ressources multimédias dont des images, des formulaires de saisie et des programmes informatiques. Il permet de créer des documents interopérables avec des équipements très variés de lecture conforme au standard **Standard Generalized Markup Language** de vers. 1 et est aussi utilisé conjointement avec le langage de programmation JavaScript et des feuilles de style en cascade (CSS). HTML a inspiré la **Structured Generalized Markup Language** (SGML). Il n'a pas un format écrit.

**Sommaire** [masquer]

- 1 Démonstrations
- 2 Évolution du langage
  - 2.1 1989-1992 - Origine
  - 2.2 1993 - Apparition du Mosaic
  - 2.3 1994 - Sortie du Netscape Navigator
  - 2.4 1995-1996 - World Wide Web
  - 2.5 1997 - HTML 3.2 et 4.0
  - 2.6 2000 - XHTML
  - 2.7 De 2002 à nos jours - HTML 5 et abandon du XHTML 2
  - 2.8 Evolution de HTML : sans numéro de version ?
- 3 Description de HTML
  - 3.1 Syntaxe de HTML
  - 3.2 Structure des documents HTML
  - 3.3 Balise et attribut
  - 3.4 Attributs de HTML
  - 3.5 Jeux de caractères
  - 3.5.1 Technique d'échappement
- 4 Interopérabilité de HTML
- 5 Notes et références
- 6 Voir aussi
- 7 Articles connexes

**HTML**  
**Hypertext Markup Language**

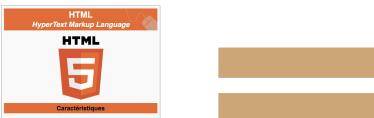
**HTML**

**5**

**Caractéristiques**

Type MIME	text/html
Développé par	World Wide Web Consortium & WHATWG
Version initiale	1.0
Type de format	Langage de balisage
Basé sur	Standard Generalized Markup Language (SGML)
Origine de la norme	XHTML 1.0
ISO	ISO/IEC 15445
Spécifications	Format ouvert
Sites web	<a href="http://www.w3.org/html/">www.w3.org/html/</a> <a href="https://github.com/whatwg/html/pull/1">https://github.com/whatwg/html/pull/1</a>

modifier + inclure le code + modifier Historique



# Structure of a HTML Page

```
<!DOCTYPE html PUBLIC "-//IETF//DTD HTML 2.0//EN">
<html>
  <head>
    <title>
      Example of HTML
    </title>
  </head>
  <body>
    This is a sentence with <a href="page.html">hyperlink</a>.
    <p>
      This is a paragraph where there is no hyperlink.
    </p>
  </body>
</html>
```

## HTML

Head

Title

Texte

Body

Texte

a

Texte

p

Texte

# Live extraction : Web Scraping & Loading Text

## How to retrieve an HTML page from website ?

# How to retrieve an HTML page from website ?

## With browser tools

The screenshot shows a web browser displaying the Babelio homepage ([babelio.com/livrespopulaires\\_debut.php?p=1](http://babelio.com/livrespopulaires_debut.php?p=1)). The page features a navigation bar with links like 'Rejoignez Babelio pour découvrir vos prochaines lectures', 'CONNEXION AVEC facebook', and 'Inscription Classique'. Below the navigation is the Babelio logo and a search bar. The main content area displays a grid of book covers under the heading 'LIVRES LES PLUS POPULAIRES'. To the right of the page, the browser's developer tools are open, specifically the 'Elements' tab. This tab shows the raw HTML code for the page, including sections for 'Découvrir', 'Accueil', 'Mes livres', and 'Ajouter des livres'. The 'Elements' tab also includes a 'Styles' panel where CSS rules for various elements like 'body' and 'div' are listed, and a 'Properties' panel showing specific style values like 'background-color: #e6f2ff; background-image: url(/img/bg\_top\_center.jpg); background-position: top center; background-repeat: no-repeat; border-bottom: 1px solid #ccc; font-family: 'Open Sans', sans-serif; font-size: 14px; height: 100%; width: 100%; color: #333; margin: 0; padding: 0;'. The developer tools also show a visual representation of the page's layout with colored boxes indicating element boundaries.

# How to retrieve an HTML page from website ?

## With command line interface

```
curl -X GET https://www.babelio.com/livrespopulaires_debut.php
```

# How to retrieve an HTML page from website ?

## With command line interface

```
[21:19:01] maxime:~ $ curl -X GET https://www.babelio.com/livrespopulaires_debut.php
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" "https://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd"><html version="XHTML+RDFa 1.0" xmlns="https://www.w3.org/1999/xhtml" xml:lang="fr"><head><meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" /><meta http-equiv="Content-language" content="fr-FR" /><meta http-equiv="cache-control" content="no-cache"><meta http-equiv="pragma" content="no-cache" /><meta http-equiv="expires" content="-1"><link rel="SHORTCUT ICON" href="/faviconbabelio_.ico" /><meta charset="iso-8859-1"><meta http-equiv="X-UA-Compatible" content="IE=edge"><meta name="viewport" content="user-scalable=no, width=device-width, initial-scale=1, maximum-scale=1"><meta property="og:type" content="website"></meta><title>Babelio - Découvrez des livres, critiques, extraits, résumés</title><meta name="title" content="Babelio - Découvrez des livres, critiques, extraits, résumés" /><meta name="description" content="Le site où les passionnés de lecture partagent et échangent autour de leurs lectures" /><meta name="keywords" content="babelio, livre, livres en ligne, bibliothèque en ligne, critiques livres, classer livres, logiciel gestion bibliothèque, livre occasion, livre photo, livre enfant, livre ancien, vente livre, livre scolaire, littérature, littérature, bandes dessinées, bande dessinée, bd, contes, recette de cuisine, dictionnaire, dictionnaire anglais français, dictionnaire français, dictionnaire des synonymes, dictionnaire anglais, mangas, mangas x, jeunesse, policier, roman policier, polar, machefer, roman, harry potter, asterix, tintin, star wars, point de croix, philosophie, atlas, art, prix goncourt, science fiction, poésie, livre poche, pleiade, tourisme, histoire érotique, histoire, lecture" /><link rel="stylesheet" type="text/css" href="/css_cache/17,18,20,21_45.css" media="all"/><link rel="canonical" href="https://www.babelio.com/livrespopulaires_debut.php" /><meta property="og:url" content="https://www.babelio.com/livrespopulaires_debut.php"><script type='text/javascript'>
var yieldlove_site_id = "babelio.com_deconnecte";
</script><script type='text/javascript' src='//cdn-a.yieldlove.com/yieldlove-bidder.js?babelio.com_deconnecte'></script><script type="text/javascript">
    var habillage_state = 0;
    var googletag = googletag || {};
    googletag.cmd = googletag.cmd || [];
    (function() {
        var gads = document.createElement("script");
        gads.async = true;
        gads.type = "text/javascript";
        var useSSL = "https:" === document.location.protocol;
        gads.src = (useSSL ? "https:" : "http:") + "//www.googletagservices.com/tag/js/gpt.js";
        var node = document.getElementsByTagName("script")[0];
        node.parentNode.insertBefore(gads, node);
    })();
</script><script type='text/javascript'>
googletag.cmd.push(function() {
```

# How to retrieve an HTML page from website ?

With python

```
import requests
```

```
req = requests.get("https://www.babelio.com/livrespopulaires_debut.php")
```

```
req.text
```

# How to retrieve an HTML page from website ?

# With python

# Live extraction : Web Scraping & Loading Text

## How to parse, extract and load HTML content ?

# How to parse, extract and load HTML content ?

Python Tools to parse and extract

Beautiful Soup

LXML

REGEX

# Live extraction : API

## What is an API ?

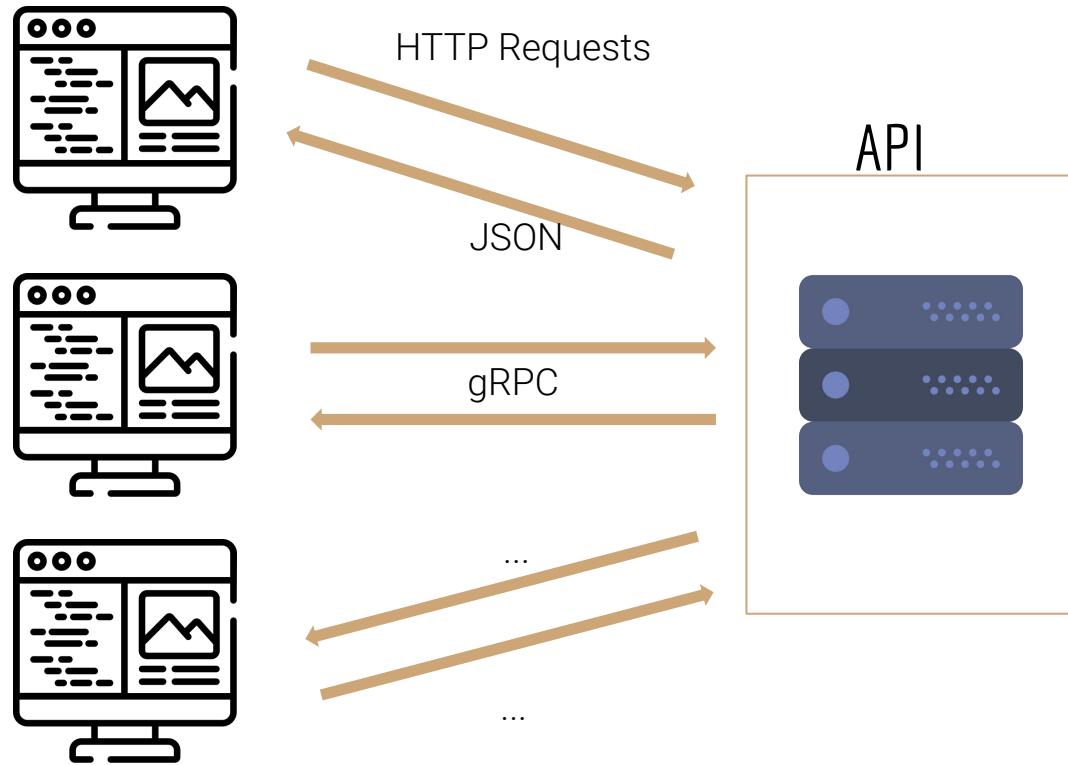
# What is an API (Application Programming Interface) ?

## Definition

An Application Programming Interface (API) can be summed up as a **computer solution that allows applications to communicate with each other and exchange services or data with each other**. It is actually a set of functions that facilitate, through a programming language, access to the services of an application.

# What is an API (Application Programming Interface) ?

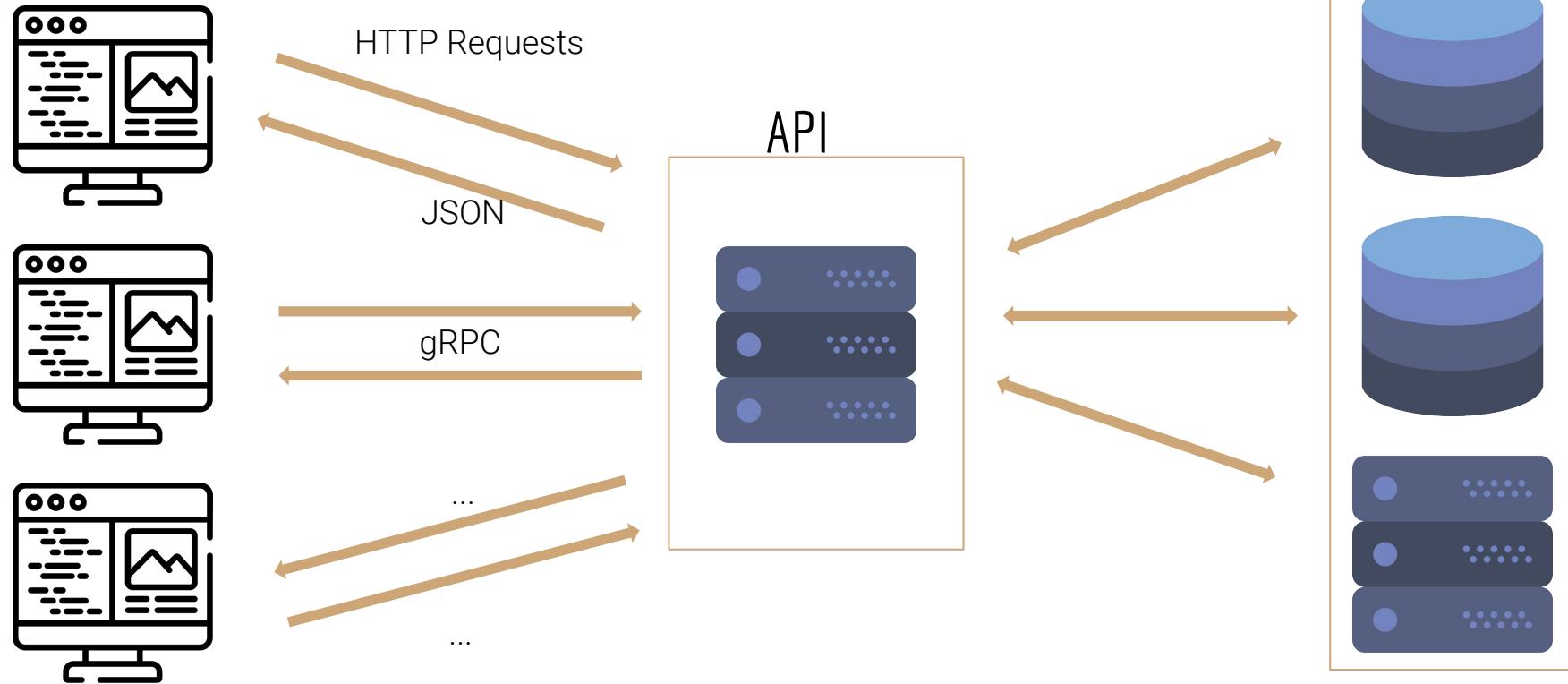
## How does it really works ?



ETL - Extract

# What is an API (Application Programming Interface) ?

## How does it really works ?



# Live extraction : API

## How to interact with an API ?

# How to interact with an API ?

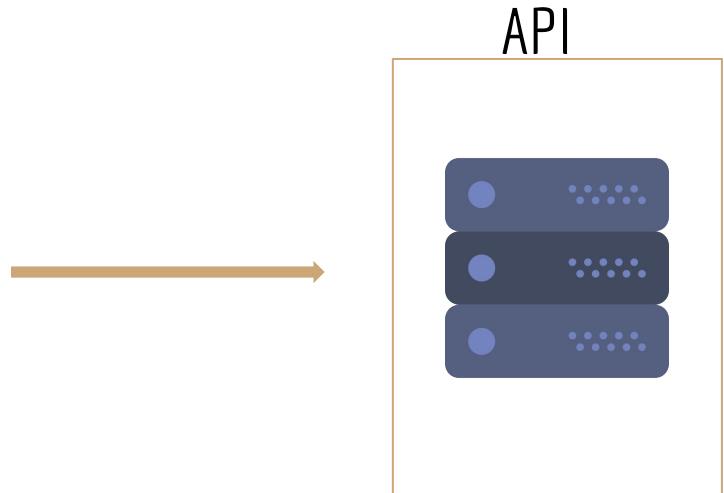


**GET** (Query Data)  
get all books

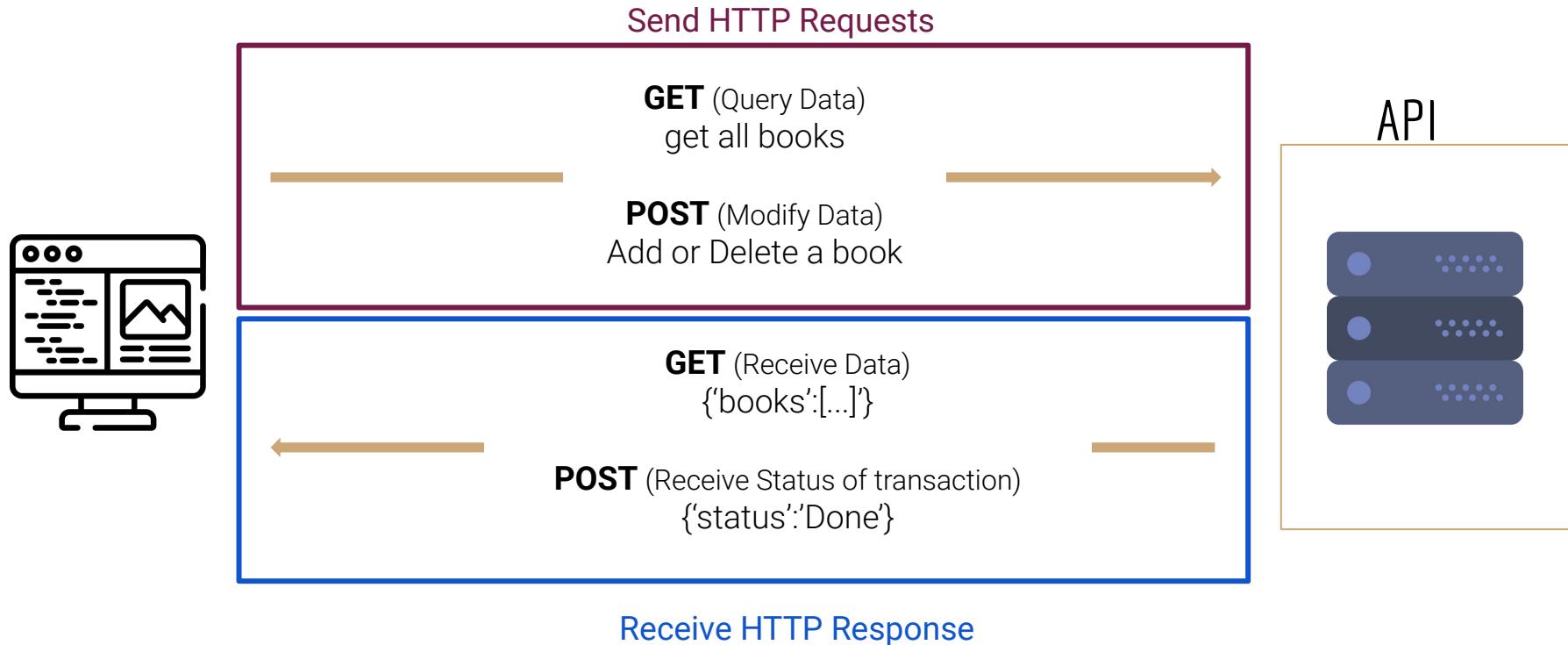
**PUT** (Modify Data)  
Insert a new book

**POST** (Modify Data)  
Update or Replace a  
book

**DELETE** (Modify Data)  
Delete a book



# How to interact with an API ?



# Live extraction : API

## Tools to interact with an API

# Tools to interact with an API

## With a software



Postman

My Workspace ▾ [Invite](#) Sign In

POST http://loc... GET http://loc... GET https://fr... + ... No Environment

GET https://fr.wikipedia.org/w/api.php?explaintext=True&action=query&format=json&exintro=True&prop=extracts|categories|titles=Victor Hugo

Send Save

Params Authorization Headers (7) Body Pre-request Script Tests Cookies Code Comments (0)

Query Params

KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/> explaintext	True	
<input checked="" type="checkbox"/> action	query	
<input checked="" type="checkbox"/> format	json	
<input checked="" type="checkbox"/> exintro	True	
<input checked="" type="checkbox"/> prop	extracts categories	
<input checked="" type="checkbox"/> titles	Victor Hugo	

Body Cookies (3) Headers (23) Test Results

Pretty Raw Preview JSON

```

1 {
2   "continue": {
3     "clcontinue": "3135|Article_à_référence_nécessaire",
4     "continue": "|{|extracts"
5   },
6   "query": {
7     "pages": [
8       "3135",
9       {
10         "pageid": 3135,
11         "ns": 0,
12         "title": "Victor Hugo",
13         "extract": "Victor Hugo est un poète, dramaturge, écrivain, romancier et dessinateur romantique français, né le 26 février 1802 (7 ventôse an X selon le calendrier républicain encore en vigueur) à Besançon et mort le 22 mai 1885 à Paris. Il est considéré comme l'un des plus importants écrivains de langue française, il est aussi une personnalité politique et un intellectuel engagé qui a eu un rôle idéologique majeur et occupe une place marquante dans l'histoire des lettres françaises au XIXe siècle, dans des genres et des domaines d'une remarquable variété. \n\n Théâtre, Victor Hugo se manifeste comme un des chefs de file du Romantisme français lorsqu'il expose sa théorie du drame romantique dans les préfaces qui introduisent Cromwell en 1827, puis Hernani en 1830 qui sont de véritables manifestes, puis par ses autres œuvres dramatiques : Ruy Blas en 1838, mais aussi Lutrèce Borgia et Le Roi s'amuse.\n\n Victor Hugo est aussi un poète lyrique avec des recueils comme Odes et Ballades (1826), Les Feuilles d'automne (1831) ou Les Contemplations (1856), mais il est aussi poète engagé contre Napoléon III dans Les Châtiments (1853) ou encore poète épique avec La légende des siècles (1859 et 1877). Ses romans rencontrent également un grand succès populaire, avec notamment Notre-Dame de Paris (1831), et plus encore avec Les Misérables (1862). Son œuvre multiple comprend aussi des discours politiques à la Chambre des pairs, à l'Assemblée constituante et à l'Assemblée législative, notamment sur la peine de mort, l'école ou l'Europe, des récits de voyages (Le Rhin, 1842, ou Choses vues posthumes, 1887 et 1890)... une correspondance abondante... ainsi que de nombreux croquis et dessins à la plume et au lavis.\n\n Victor Hugo a fortement influencé la littérature mondiale et reste l'un des auteurs les plus lus et les plus traduits de tous les temps."}
14     ]
15   }
16 }
```

Status: 200 OK Time: 149ms Size: 2.63 KB Save Response

Bootcamp

# Tools to interact with an API

With command line interface

```
curl -X GET 'https://fr.wikipedia.org/w/api.php?explaintext=True&action=query&format=json&exintro=True&prop=extracts|categories&titles=Victor%20Hugo'
```

# Tools to interact with an API

## With command line interface

```
[11:48:07] maxime:~ $ curl -X GET 'https://fr.wikipedia.org/w/api.php?explaintext=True&action=query&format=json&exintro=True&prop=extracts|categories&titles=Victor%20Hugo'
>{"continue":{"clcontinue":"3135!Article_\u00e0_\u00e9_\u00e9f\u00e9_\u00e9rence_\n\u00e9_\u00e9cessaire","continue":"!extracts"},"query":{"pages":{"3135":{"pageid":3135,"ns":0,"title":"Victor Hugo","extract":"Victor Hugo est un po\u00e8te, dramaturge, \u00e9crivain, romancier et dessinateur romantique fran\u00e7ais, n\u00e0 le 26 f\u00e9vrier 1802 (7 vent\u00e9mber selon le calendrier r\u00e9publicain encore en vigueur) Besan\u00e7on et mort le 22 mai 1885 Paris. Il est consid\u00e9r\u00e9 comme l'un des plus importants \u00e9crivains de langue fran\u00e7aise. Il est aussi une personnalit\u00e9 politique et un intellectuel engag\u00e9 qui a eu un r\u00e9el influence majeur et occupe une place marquante dans l'histoire des lettres fran\u00e7aises au XIXe si\u00e8cle, dans des genres et des domaines d'exception remarquable vari\u00e9t\u00e9. Au th\u00e9âtre, Victor Hugo se manifeste comme un des chefs de file du Romantisme fran\u00e7ais lorsqu'il expose sa th\u00e9orie du drame romantique dans les pr\u00e9faces qui introduisent Cromwell en 1827, puis Hernani en 1830 qui sont de v\u00e9ritables manifestes, puis par ses autres \u00e9uvres dramatiques : Ruy Blas en 1838, mais aussi Lucifer\u00e8ce Borgia et Le Roi s'amuse. Victor Hugo est aussi un po\u00e8te lyrique avec des recueils comme Odes et Ballades (1826), Les Feuilles d'automne (1831) ou Les Contemplations (1856), mais il est aussi po\u00e8te engag\u00e9 contre Napol\u00e9on III dans Les Ch\u00e2timents (1853) ou encore po\u00e8te \u00e9pique avec La L\u00e9gende des si\u00e8cles (1859 et 1877). Ses romans rencontrent \u00e9galement un grand succ\u00e8s populaire, avec notamment Notre-Dame de Paris (1831), et plus encore avec Les Mis\u00e9rables (1862). Son \u00e9uvre multiple comprend aussi des discours politiques la Chambre des pairs, l'Assembl\u00e9e constituante et l'Assembl\u00e9e legislative, notamment sur la peine de mort, l'engagement de l'Europe, des routes de voyages (Le Rhin, 1842, ou Choses vues, posthumes, 1887 et 1890), une correspondance abondante, ainsi que de nombreux croquis et dessins la plume et au crayon. Victor Hugo a fortement contribu\u00e9 au renouvellement de la po\u00e8sie et du th\u00e9âtre. Il a \u00e9t\u00e9 admir\u00e9 par ses contemporains et l'est encore, mais il a aussi \u00e9t\u00e9 contest\u00e9 par certains auteurs modernes. Il a permis de nombreuses g\u00e9n\u00e9rations de d\u00e9velopper une r\u00e9flexion sur l'engagement de l'homme dans la vie politique et sociale gr\u00e2ce ses multiples prises de position, choisissant de s'exiler pour vivre Guernesey pendant les vingt ans du Second Empire. Ses choix, l'homme la fois moraux et politiques, durant la deuxi\u00e8me partie de sa vie, et son \u00e9uvre hors du commun ont fait de lui un personnage embl\u00e9matique, que la Troisi\u00e8me R\u00e9publique a honor\u00e9 par des fun\u00e9rales nationales, qui ont accompagn\u00e9 le transfert de sa d\u00e9position au Panth\u00e9on de Paris le 1er juin 1885, dix jours apr\u00e8s sa mort."}, "categories": [{"ns":14, "title": "Cat\u00e9gorie:Acad\u00e9mie des Jeux floraux"}, {"ns":14, "title": "Cat\u00e9gorie:Adversaire de la peine de mort"}, {"ns":14, "title": "Cat\u00e9gorie:Anticl\u00e9rical"}, {"ns":14, "title": "Cat\u00e9gorie:Article contenant un appel \u00e0 traduction en anglais"}, {"ns":14, "title": "Cat\u00e9gorie:Article contenant un lien mort"}, {"ns":14, "title": "Cat\u00e9gorie:Article de Wikip\u00e9dia avec notice d'autorit\u00e9"}, {"ns":14, "title": "Cat\u00e9gorie:Article de qualit\u00e9 en polonais"}, {"ns":14, "title": "Cat\u00e9gorie:Article de qualit\u00e9 en serbo-croate"}, {"ns":14, "title": "Cat\u00e9gorie:Article pouvant contenir un travail in\u00e9dit"}, {"ns":14, "title": "Cat\u00e9gorie:Article \u00e0 pr\u00e9cision n\u00e9cessaire"}]}]}
```

# Tools to interact with an API

## With python

```
import requests
```

```
params = {  
    'action': "query",  
    'titles': "Victor Hugo",  
    'format': "json",  
    'prop': 'extracts|categories',  
    'explaintext': True,  
    'exintro': True  
}
```

```
req = requests.get(url="https://fr.wikipedia.org/w/api.php", params=params)
```

```
req.json()
```

# Tools to interact with an API

## With python

```
In [5]: import requests

In [6]: params = {
...     'action': "query",
...     'titles': "Victor Hugo",
...     'format': "json",
...     'prop': 'extracts|categories',
...     'explaintext': True,
...     'exintro': True
... }

In [7]: req = requests.get(url="https://fr.wikipedia.org/w/api.php", params=params)

In [8]: req.json()
Out[8]:
{'continue': {'clcontinue': '3135!Article_à_référence_nécessaire',
  'continue': '||extracts',
  'query': {'pages': {'3135': {'categories': [{"ns': 14,
    'title': 'Catégorie:Académie des Jeux floraux'},
   {'ns': 14, 'title': 'Catégorie:Adversaire de la peine de mort'},
   {'ns': 14, 'title': 'Catégorie:Anticlérical'},
   {'ns': 14,
    'title': 'Catégorie:Article contenant un appel à traduction en anglais'},
   {'ns': 14, 'title': 'Catégorie:Article contenant un lien mort'},
   {'ns': 14,
    'title': 'Catégorie:Article de Wikipédia avec notice d'autorité'},
   {'ns': 14, 'title': 'Catégorie:Article de qualité en polonais'},
   {'ns': 14, 'title': 'Catégorie:Article de qualité en serbo-croate'},
   {'ns': 14,
    'title': 'Catégorie:Article pouvant contenir un travail inédit'},
   {'ns': 14, 'title': 'Catégorie:Article à précision nécessaire'}],
  'extract': "Victor Hugo est un poète, dramaturge, écrivain, romancier et dessinateur romantique français, né le 26 février 1802 (7 ventôse an X selon le calendrier républicain encore en vigueur) à Besançon et mort le 22 mai 1885 à Paris. Il est considéré comme l'un des plus importants écrivains de langue française. Il est aussi une personnalité politique et un intellectuel engagé qui a eu un rôle idéologique majeur et occupe une place marquante dans l'histoire des lettres françaises au XIXe siècle, dans des genres et des domaines d'une remarquable variété. Au théâtre, Victor Hugo se manifeste comme un des chefs de file du Romantisme français lorsqu'il expose sa théorie du drame romantique dans les préfaces qui introduisent Cromwell en 1827, puis Hernani en 1830 qui sont de véritables manifestes, puis par ses autres œuvres dramatiques : Ruy Blas en 1838, mais aussi Lucrece Borgia et Le Roi s'amuse. Victor Hugo est aussi un poète lyrique avec des recueils comme Odes et Ballades (1826), Les Feuilles d'automne (1831) ou Les Contemplations (1856), mais il est aussi poète engagé contre Napoléon III dans Les Châtiments (1853) ou encore poète épique"}]
```