

# Web Scraping

CPE 393: Text Analytics

***Dr. Sansiri Tarnpradab***

*Department of Computer Engineering  
King Mongkut's University of Technology Thonburi*

*Intro*

*Pattern  
Matching*

*Text  
Visualization*

*Web Scraping*



*Text  
Preparation*

*Text Feature  
Representation*

*Text  
Classification*

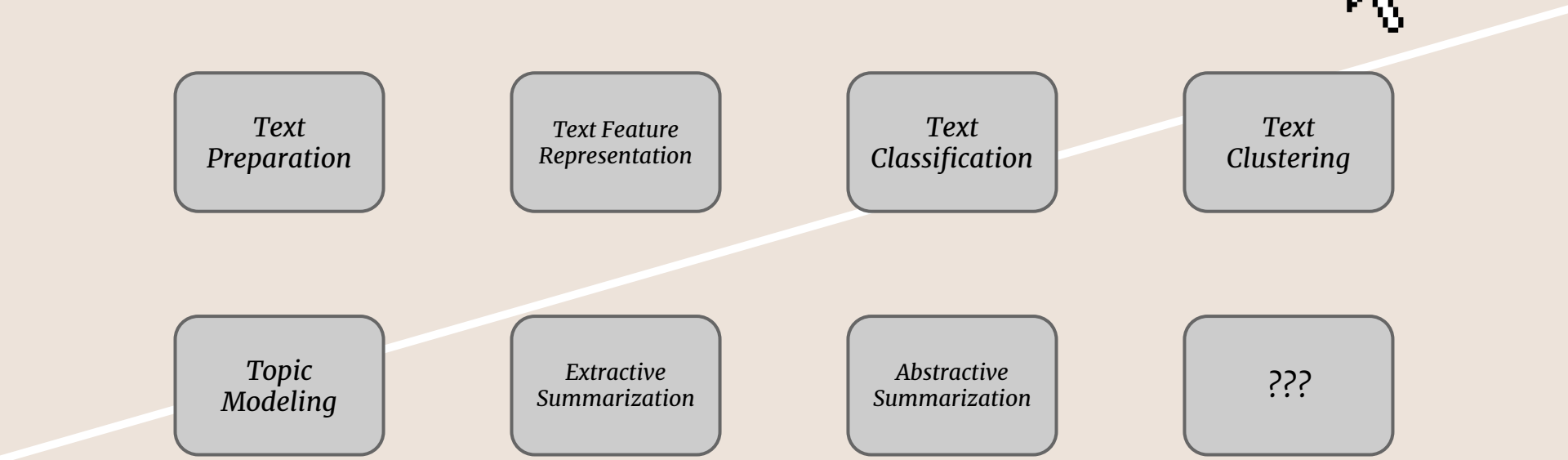
*Text  
Clustering*

*Topic  
Modeling*

*Extractive  
Summarization*

*Abstractive  
Summarization*

*???*

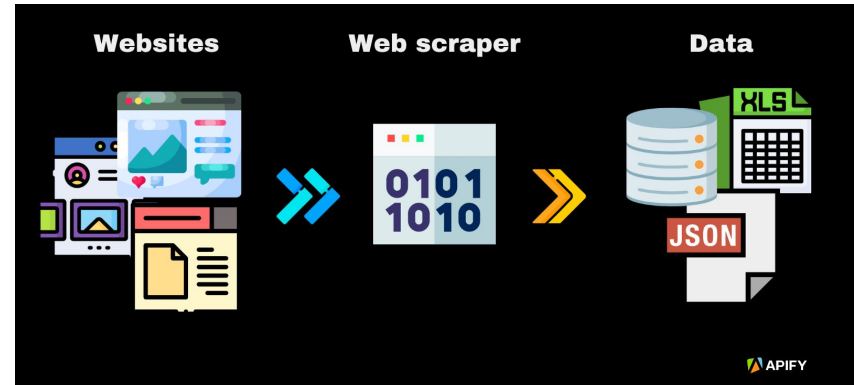


# Outline

- Introducing web scraping
- Importance
- API
- Overview of web scraping
- Web scraping vs Web crawling
- Tools & technologies
- Challenges

# What is Web Scraping?

- A process to automatically extract data from websites
- Aka:
  - Web data extraction
  - Web harvesting
- Unstructured data (mostly)



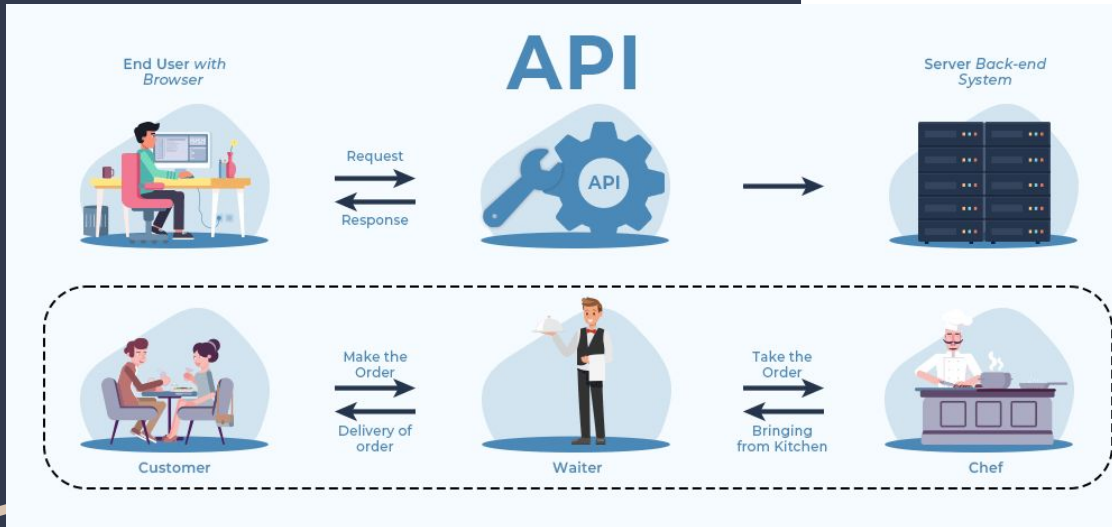
Ref: <https://blog.apify.com/what-is-web-scraping/>

# *Importance of Web Scraping*

- Data Collection
- Provides a way to access a wealth of data from websites that may not offer APIs or structured datasets
- Benefits:
  - Market research
  - Price monitoring
  - Content aggregation
  - Academic research
  - etc...

# API

- Application Programmable Interface
- A set of rules, protocols, and tools
- Allows software applications to communicate with each other



Common APIs:

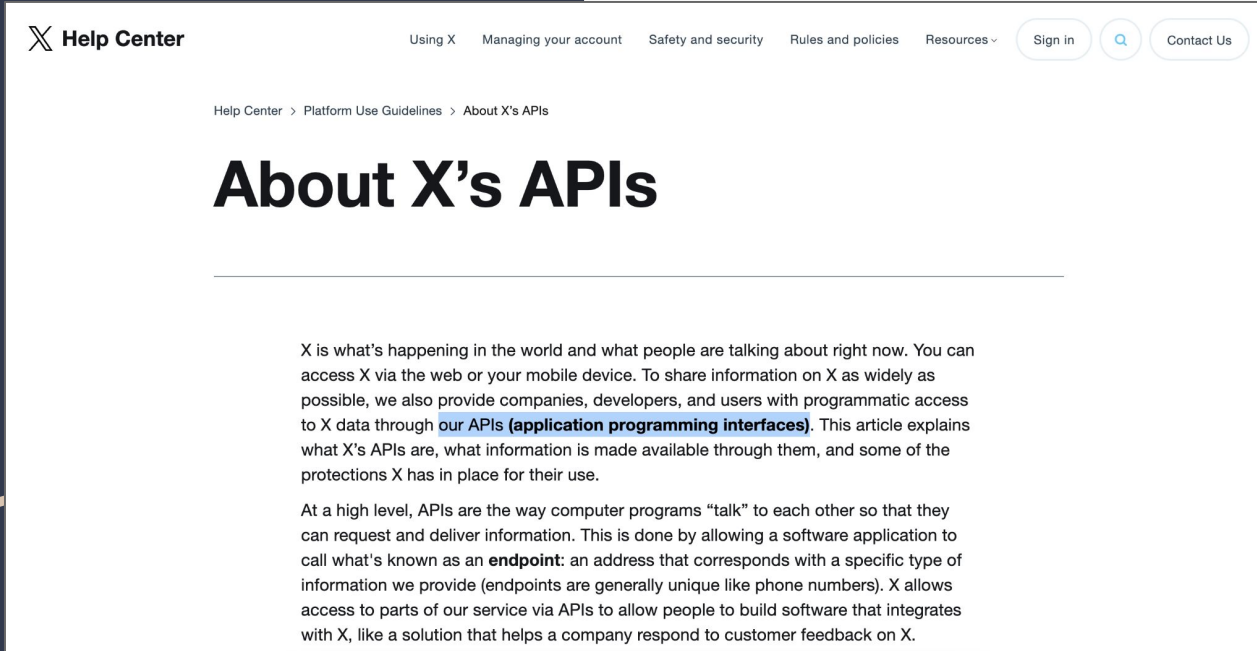
- Web APIs
- RESTful APIs
- SOAP APIs
- Public APIs
- Private APIs

Ref: <https://www.geeksforgeeks.org/what-is-an-api/>

# API & Text Analytics

To extract textual data from social media

For example...



The screenshot shows the X Help Center interface. At the top, there's a navigation bar with the X logo and 'Help Center' on the left, and links for 'Using X', 'Managing your account', 'Safety and security', 'Rules and policies', 'Resources', 'Sign in', a search icon, and 'Contact Us' on the right. Below the navigation bar, a breadcrumb trail reads 'Help Center > Platform Use Guidelines > About X's APIs'. The main heading is 'About X's APIs' in a large, bold font. Below the heading, there's a horizontal line. The text explains that X is what's happening in the world and what people are talking about right now. It states that users can access X via the web or their mobile device. To share information on X as widely as possible, X also provides companies, developers, and users with programmatic access to X data through their APIs (application programming interfaces). This article explains what X's APIs are, what information is made available through them, and some of the protections X has in place for their use. It then explains that at a high level, APIs are the way computer programs "talk" to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an endpoint: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers). X allows access to parts of our service via APIs to allow people to build software that integrates with X, like a solution that helps a company respond to customer feedback on X.

X Help Center

Using X Managing your account Safety and security Rules and policies Resources Sign in Search Contact Us

Help Center > Platform Use Guidelines > About X's APIs

## About X's APIs

X is what's happening in the world and what people are talking about right now. You can access X via the web or your mobile device. To share information on X as widely as possible, we also provide companies, developers, and users with programmatic access to X data through [our APIs \(application programming interfaces\)](#). This article explains what X's APIs are, what information is made available through them, and some of the protections X has in place for their use.

At a high level, APIs are the way computer programs "talk" to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an **endpoint**: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers). X allows access to parts of our service via APIs to allow people to build software that integrates with X, like a solution that helps a company respond to customer feedback on X.

Ref: <https://help.twitter.com/en/rules-and-policies/x-api>

# Overview of Web Scraping

Fetching

Parsing

Parsing

Preprocessing

Storing

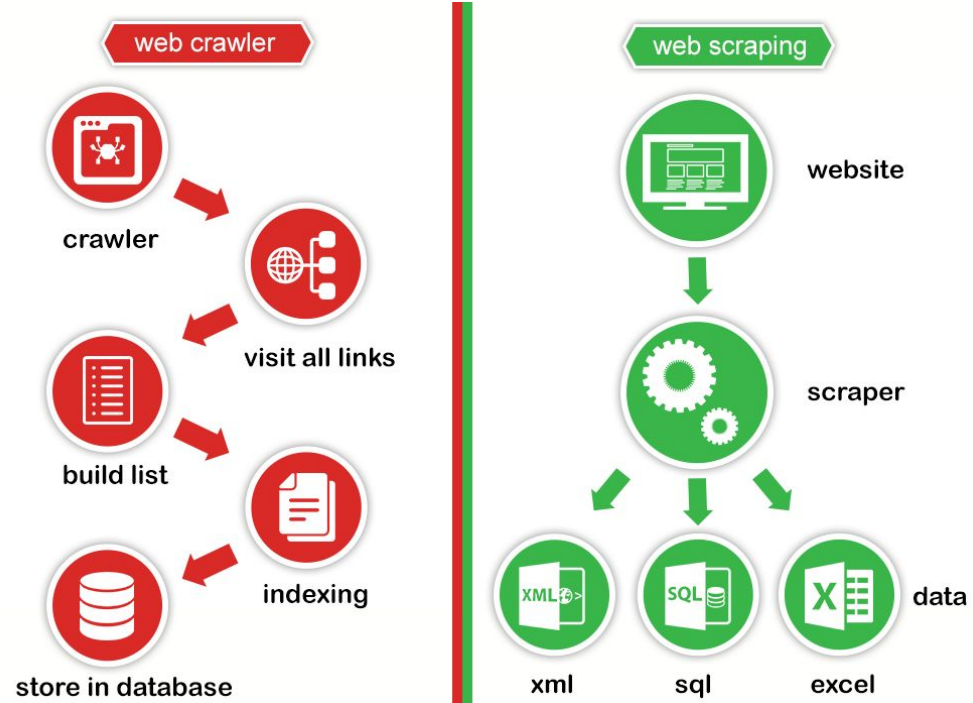
1. **Send HTTP requests** to the website's server
2. The server **sends back the requested web page** (typically in HTML format)
3. **Parse the content** by making use of the structure e.g. tags and attributes
4. **Extract data** from the parsed content
5. **Clean** the data
6. **Preprocess** the data
7. **Store** the data



# Web Scraping

vs

# Web Crawling



# Web Scraping

vs

# Web Crawling

(Cont.)

## Web Scraping

- **What:** Extracting data/content from webpages
- **How:** May require interaction with webforms
- **Purpose:** To collect data of interest for various purposes
- **Product:** Extracted content e.g. texts, images, tables, etc.

## Web Crawling

- **What:** Browsing and indexing web pages
- **How:** Navigating through web pages by following hyperlinks
- **Purpose:** index web pages
- **Product:** Metadata e.g. URLs, page titles, etc.

# *Tools & Tech for Web Scraping*

BeautifulSoup

Scrapy

Requests

Selenium

# *Challenges in Web Scraping*

## Technical

- Dynamic content
- Updates of website

## Data

- Quality
- Amount

## Legal/ Ethics

- Legal and ethics concerns
- Anti-scraping mechanism

# Conclusion

- Introducing web scraping
- Importance
- API
- Overview of web scraping
- Web scraping vs Web crawling
- Tools & technologies
- Challenges