

# Web Mining

## Laboratoire 1

---

*Edward Ransome (laboratoire réalisé par Antoine Rochat)*

---

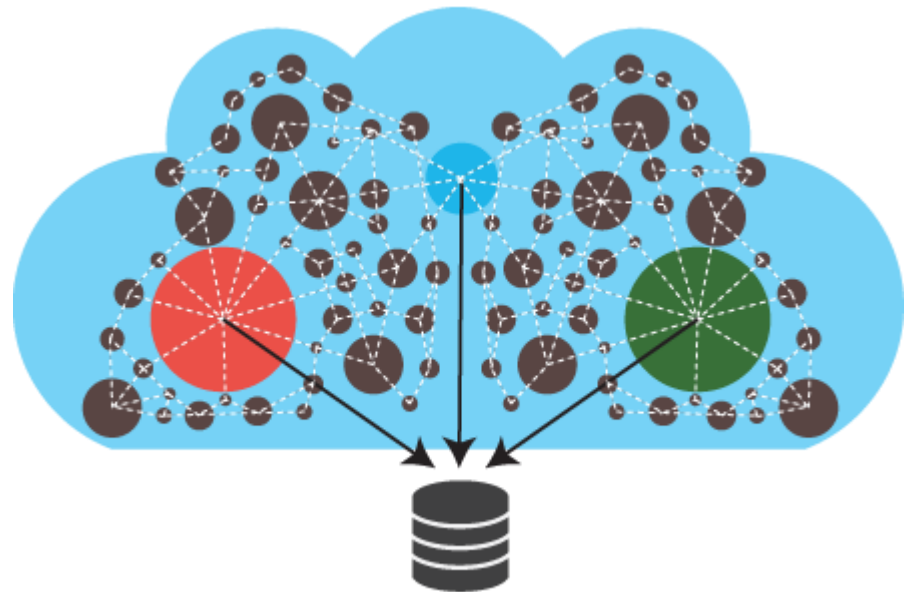
*Institut des Technologies de l'Information et de la Communication (IICT)*

---

# Crawling, indexation et recherche de pages Web

Les points étudiés:

- *Crawling* de pages web
  - Construction d'un index avec *ElasticSearch*
  - Implémentation d'une fonctionnalité de recherche
  - Deux questions théoriques
- 
- A rendre sur *Moodle* avant le **dimanche 19.03.2023 à 23h59**
  - Groupes de 2 ou 3 personnes



# Partie 1

## Crawler

- Réalisation d'un logiciel en *Python* utilisant la librairie *Scrapy* pour réaliser un crawler
    - On ne visitera que des formats de fichiers textuels (et avec un certain intérêt)
    - On se limitera à un seul domaine
    - Attention de ne pas se faire «blacklister»
  - On souhaite récupérer et indexer spécifiquement des informations sur une page
    - Champs spécifiques
    - Données structurées
- > Selectors de Scrapy

```
import scrapy

class QuotesSpider(scrapy.Spider):
    name = "quotes"

    start_urls = [
        'https://quotes.toscrape.com/page/1/',
        'https://quotes.toscrape.com/page/2/',
    ]

    def parse(self, response):
        print(response.css('title::text').get())
```

## *Partie 2*

# Indexation

- Utilisation d'*ElasticSearch* pour l'indexation des champs
  - Plateforme logicielle pour l'indexation et la recherche de données
  - Basée sur la bibliothèque *Apache Lucene*
  - Utilisable à l'aide d'une API basée sur HTTP ou de librairie plus haut niveau
  - Associé à Kibana pour la visualisation des données
- Il vous faudra configurer précisément les champs (ainsi que leur type) que l'on souhaite indexer



## Partie 3

# Recherche

- Le but de cette partie est d'utiliser l'index que l'on vient de créer pour effectuer des recherches
- Un deuxième programme *Python* mettant en œuvre cette fonctionnalité de recherche
- Cette recherche devra privilégier certains champs par rapport à d'autres et afficher le score des documents retournés

Haute École d'ingénierie et de gestion du canton de Vaud  
16.27323

<https://fr.wikipedia.org/wiki/HEIG-VD>

Haute École de gestion du canton de Vaud  
15.942445

<https://fr.wikipedia.org/wiki/HEG-VD>

École d'ingénieurs du canton de Vaud  
15.564051

<https://fr.wikipedia.org/wiki/EIVD>

Haute École spécialisée de Suisse occidentale  
15.480998

<https://fr.wikipedia.org/wiki/HES-SO>

## Partie 4

### Questions théoriques

## Deux sujets à développer

- Indexation multilingue
- Recherche floue  
*Fuzzy query*

