Elasticsearch, Kibana
-
Indexation, recherche et visualisation

Victor Ballu - victor.ballu@clever-cloud.com

clever cloud

IHQST

elico

Équipe de recherche de Lyon en sciences de l'information et de la communication

Victor Ballu - victor.ballu@clever-cloud.com

clever cloud

- Cloud managé
- Gestion des backups
- blue green déploiement
- Auto scaling
- Anticipation des besoins

Victor Ballu - victor.ballu@clever-cloud.com

Institut d'Histoire et de Philosophie des Sciences et des Techniques

- Philosophie de la physique, connaissance scientifique, unité des sciences
- Logique et langage, philosophie des mathématiques
- Philosophie de la biologie et de la médecine
- Décision, rationalité, interaction
- Histoire de la philosophie des sciences

Équipe de recherche de Lyon en sciences de l'information et de la communication

- Identités, langages et pratiques médiatiques
- Bibliothèques numériques, documents numériques et médiations
- Data, Big Data, et Open Data
- Cultures écrites, cultures numériques
- Savoirs informationnels et scientifiques : élaboration, circulation, appropriation

Etude Typologique du concept d'interprétabilité des mécanismes mis en oeuvre en Intelligence artificielle

- Explicabilité
- Interprétabilité
- Emergences
- Réseaux de neurones

BNP PARIBAS
CORPORATE & INSTITUTIONAL BANKING

- Détection de comportements anormaux
- Indexation et recherche d'information
- Recommandation
- NLP
- Détection de topics

clever cloud

- Traitement du signal
- Forecasting
- Donnée times series
- NLP
- Identification de topic
- Détection de fraudes
- Anticipation des besoins
- ...

Victor Ballu - victor.ballu@clever-cloud.com

Sommaire

- Intro
- Présentation générale Elastic Stack
- Notions importantes en bases de données et data science
- Information retrieval et indexation
- Cas pratiques


elasticsearch


kibana

- Elasticsearch / Kibana/ Logstash / Beats / APM...
- Créé en 2004 par Shay Banon. Première version publique février 2010
- Basé sur le projet Lucene
- Modèle économique open core

- Moteur de recherche et d'analyse
- Moteur d'indexation
- NoSql data
  - données textuelles, numériques, géographiques
- Outil distribué
- Basé sur le moteur de recherche open-source Apache Lucene

- Plateforme d'analyse et de visualisation
- Outil de management de la suite Elastic

Notions

- Distribution
  - Primary shard= Répartition de la ressource (de l'index)
  - Replicas = Répartition de la charge des requêtes (copies des shards primaires)
- NoSQL / SQL
- API (Application programming interface) / REST (Representational state transfer)
- Systèmes transactionnels
  - ACID properties :
    - Atomicity : Tout est impacté ou rien
    - Consistency : La transaction est correcte
    - Isolation : Résilient aux opérations concurrentes
    - Durability : Le changement d'état est stocké et disponible d'une manière aussi qualitative qu'une vieille donnée
- CAP Théorème : Un système distribué ne peut pas garantir à la fois :
  - Consistency (Consistance)
  - Availability (Disponibilité)
  - Partition tolerance (Tolérance aux erreurs de transferts)

Rapport à Elasticsearch

Elasticsearch est une système

- Distribué
- NoSQL
- Utilisant les protocole de communication par API REST

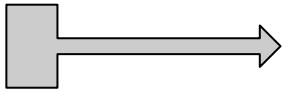
CEPENDANT, Elasticsearch n'est pas un systèmes transactionnels:

- Les propriétés ACID ne sont pas respectés, et ne peuvent pas l'être selon le CAP theorem

> *C'est pourquoi Elasticsearch n'est pas utilisé en base de donnée, mais en base d'indexation*

Recherche d'information

- Information retrieval
  - Deux facettes
    - Indexation
    - requêtes



On retrouve ces deux logiques dans Elasticsearch

Sans index, un moteur de recherche devrait scanner tous les docs pour trouver une correspondance

Victor Ballu - victor.ballu@clever-cloud.com

Recherche d'information

- Indexation

Objectif : être capable de retrouver rapidement un document

REMARQUE

Le *question answering* est similaire, mais tente en plus de construire une réponse

Recherche d'information

- index

  - DOC 1 : "*le vol spatial prend son essor à la fin de la Seconde Guerre mondiale* "
  - DOC 2 : "*La Seconde Guerre mondiale, ou Deuxième Guerre mondiale, est un conflit armé à l'échelle planétaire qui dure du 1er septembre 1939 au 2 septembre 1945.*"
  - DOC 3 : "*Un planétaire est un ensemble mécanique mobile, figurant le Système solaire (le Soleil et ses planètes) en tout ou partie.*"

MOT	guerre	planétaire	conflit
DOC 1	X		
DOC 2	X	X	X
DOC 3		X	

Recherche d'information

- index inversé

  - DOC 1 : *"le vol spatial prend son essor à la fin de la Seconde Guerre mondiale "*
  - DOC 2 : *"La Seconde Guerre mondiale, ou Deuxième Guerre mondiale, est un conflit armé à l'échelle planétaire qui dure du 1er septembre 1939 au 2 septembre 1945."*
  - DOC 3 : *"Un planétaire est un ensemble mécanique mobile, figurant le Système solaire (le Soleil et ses planètes) en tout ou partie."*

MOT	DOC 1	DOC 2	DOC 3
guerre	X	X	
Planétaire		X	X
Conflit		X	

Recherche d’information

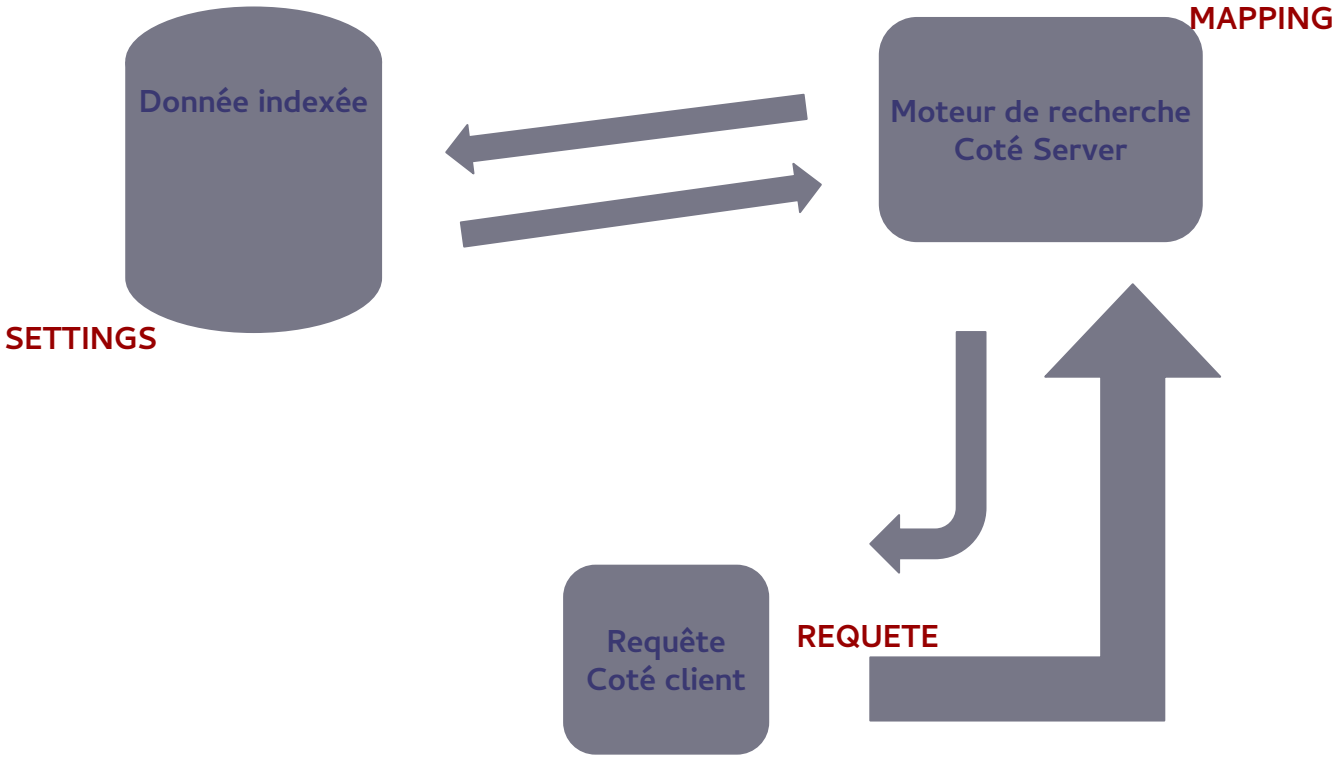
- Requêtes

Objectif : Retrouver tout ou partie d’un document selon certains critères

Recherche d'information

- **Lemmatisation : Ramener les mots à leur forme condensée de base**
  - Ex : *grandeur*, *grandes*, *grands* renvoient à *grand*
- **Stemming : ne garder que la racine des mots**
  - ex : *chercher* devient *cherch*
  - Attention erreur possible : *université* et *univers* ont la même racine: *univers*
- **Stop Words**

Recherche d'information



Recherche d'information

Pour effectuer une recherche d'information, il faut donc une représentation COMPARABLE des documents et des requêtes.

Recherche d'information

➔ La comparaison

Plusieurs manière de mesurer la similarité entre deux représentations. Généralement basé sur la comparaison vectorielle, et notamment la cos similarité.

- Longueur d'un document est :
  - $|d| = (X_1^2 + X_2^2 + \dots + X_n^2)^{1/2}$
- Term Frequency : Nombre d'occurrences de mots par documents
- TF-IDF : term frequency-inverse document frequency
  - $d_y = (W_{X1} + W_{X2} + \dots + W_{Xn})_y$
  - $tf_{x,y}$: fréquence de x dans doc y
  - N : Nombre total de doc
  - df_x : Nombre de doc contenant x

$$\cos(\theta) = \frac{\mathbf{x}_1 . \mathbf{x}_2}{|\mathbf{x}_1| \, |\mathbf{x}_2|}$$

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

Recherche d'information

- Précision (*precision*) : nbre de résultat valide retournés /résultats retournés total

Point de vu de l'algorithme

- Rappel (*recall*) : nbre résultat valide retournés / résultats attendus
- F1 Score : (précision * rappel) / (précision + rappel)

Point de vu de omniscient

TD

- RESSOURCES
  - CODE https://github.com/vballu/cnam-lessons
  - ELASTICSEARCH DOC :
    - MAPPING: https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping.html
    - SETTING:
      - https://www.elastic.co/guide/en/elasticsearch/reference/current/indices-update-settings.html
      - https://www.elastic.co/guide/en/elasticsearch/reference/current/indices-create-index.html
    - QUERY:
      - https://www.elastic.co/guide/en/elasticsearch/reference/current/search-search.html
      - https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html
      - https://www.elastic.co/guide/en/elasticsearch/reference/current/search-request-body.html
  - Kibana DOC :
    - KQL : https://www.elastic.co/guide/en/kibana/current/kuery-query.html
    - Discover: https://www.elastic.co/guide/en/kibana/current/discover.html
    - Visualize: https://www.elastic.co/guide/en/kibana/current/visualize.html
    - Dashboard: https://www.elastic.co/guide/en/kibana/7.6/dashboard.html
    - dev tools - Console : https://www.elastic.co/guide/en/kibana/current/console-kibana.html

TD

- OBJECTIF
  - Prendre en main les concepts d'elasticsearch et kibana
  - Comprendre les différentes notions (mapping, setting,query). Kibana VS Elastic
  - Comprendre la structuration et manipulation de données
  - Appréhender les enjeux d'un outil de recherche
  - Valoriser des données à travers un dashboard
- CONCRÈTEMENT:
  - Améliorer la performance du moteur de recherche (voir ressources CODE)
  - Faire une beau Dashboard

Clever Cloud Paris
137 rue vieille du temple 75003 Paris

Clever Cloud Nantes
3 rue de l'allier 44000 Nantes
02 85 52 07 69

https://www.clever-cloud.com

CONTACT

victor.ballu@clever-cloud.com