Web Sensor

PROJET DE SYNTHÈSE – MASTER 1 IISC
SOUTENANCE FINALE
17/06/2020

TUTEURS TECHNIQUES:

- DIMITRIS KOTZINOS
- WASSIM SWAILEH

ENCADRANT DE GESTION DE PROJET:

- TIANXIAO LIU

MEMBRES DE L'ÉQUIPE:

- MATHIEU VOISIN
- LYDIA KHELFANE
- GABRIEL CHEVALLIER
- MARTIN GUILBERT-LEJEUNE

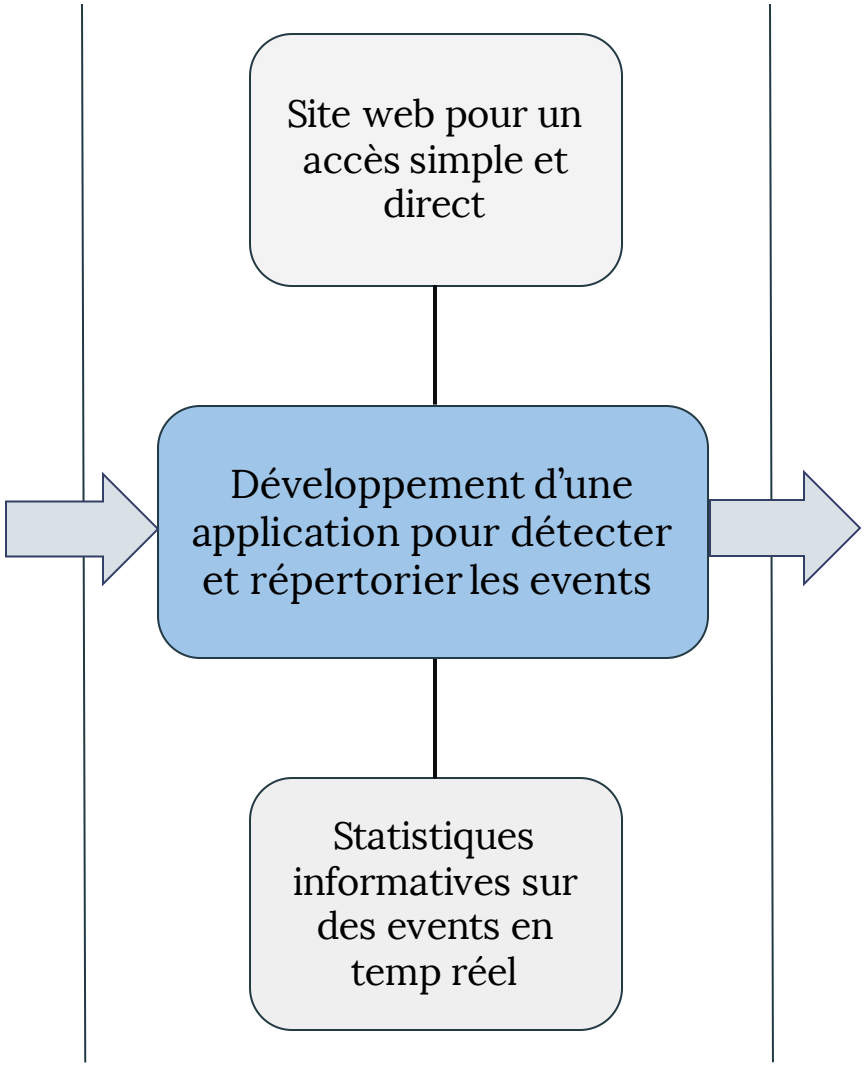
Sommaire

- Introduction
- Architecture technique
- Choix des bases de données
- Natural Language Processing
- Vectorisation des données
- Algorithmes de clustering
- Prédiction d'events
- Affichages des events
- Gestion de projet
- Conclusion et perspectives

Introduction: Objectif du projet

Twitter :

- Grand nombre de tweets
- Beaucoup d'informations
- Pourcentage de tweet "utile" faible
- Tendances parfois peu intéressantes
- ...



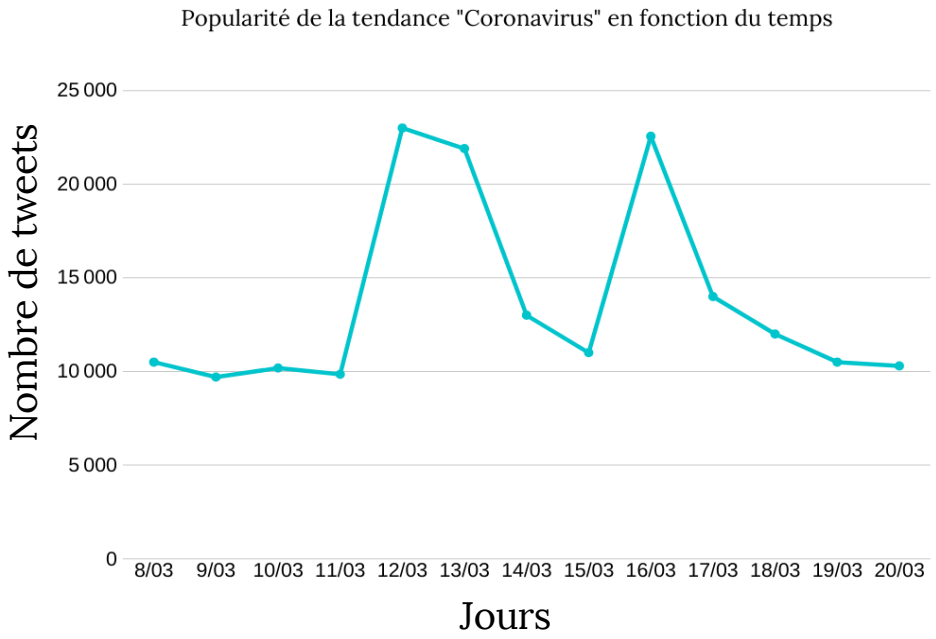
Notre application :

- Recherches avancées d'événements
- Visualisations statistiques
- Comparaison de popularité
- Consultation d'events populaires
- Prédiction de popularité
- Temp réel
- ...

Mise en scénario

Où en est le coronavirus en France?

➢ Regardons les statistiques sur l'application

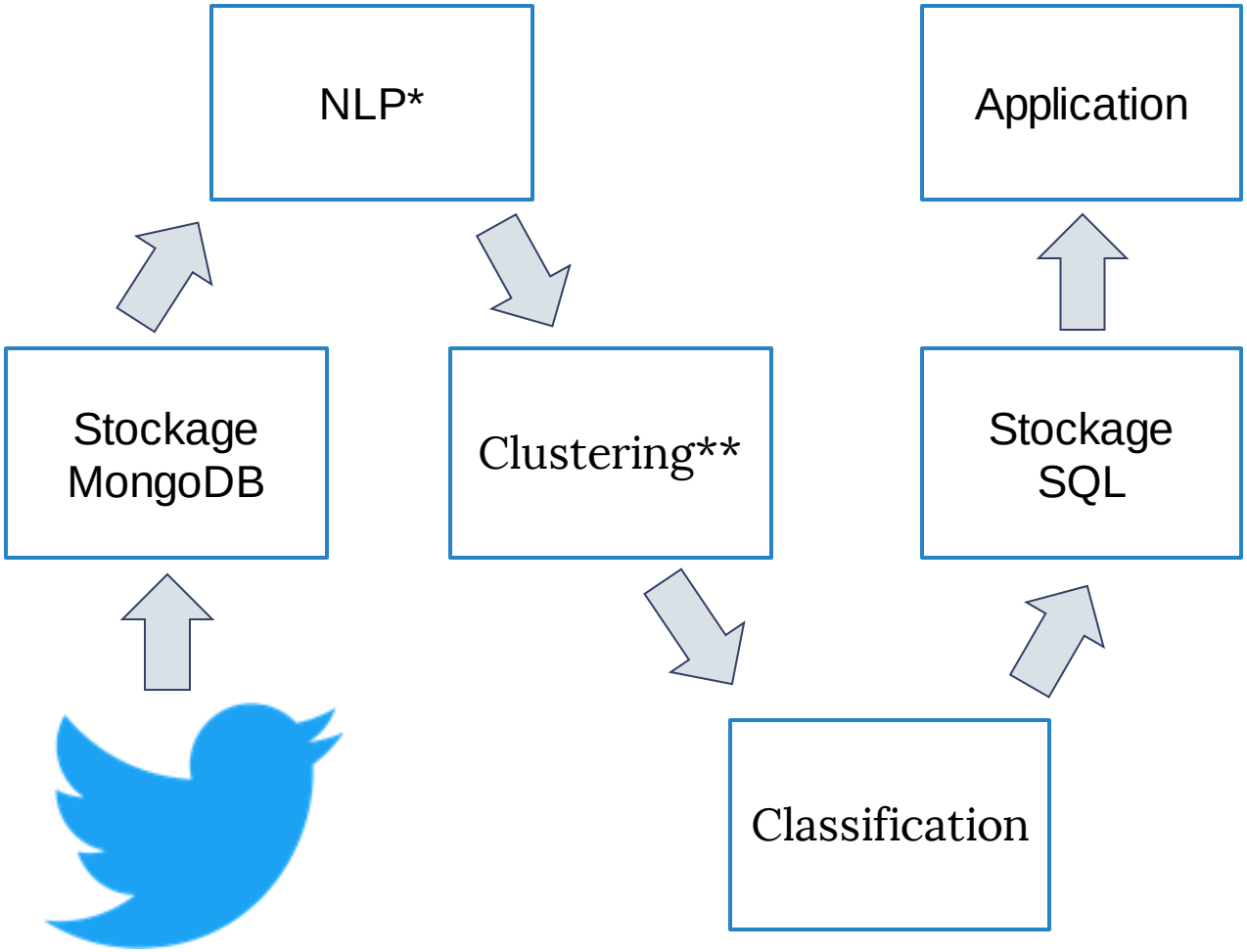


Ah! Il s'est passé quelque chose le 12 et le 16 Mars… et peut-être aussi le 13 !

Effectivement, le 12 et le 16 il y avait des discours du président, et le 13 du premier ministre…

En tout cas on a pas fini d'en entendre parler… Ah le président s'est exprimé sur le sujet sur twitter !

Architecture technique : Étapes de traitement

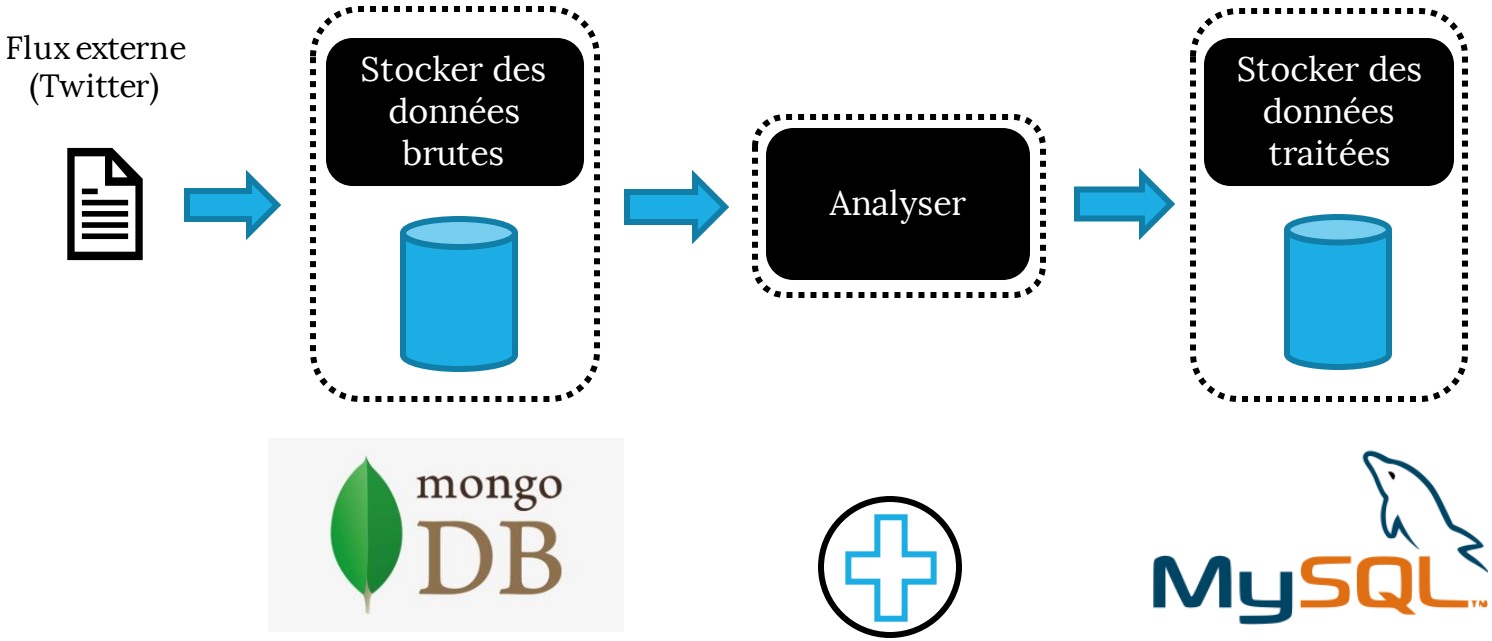


*NLP : Natural Language Processing

- Traitement automatique du langage naturel

**Clustering : Regroupement des données similaires sous un même ensemble

Choix des bases de données



Stockage brut des données :

- MongoDB permet de stocker tout type de données et de les analyser en temps réel
- Permet de requêter twitter et obtenir des tweets sous un format JSON.
- Le modèle de document JSON de MongoDB facilite le stockage

Stockage des données traitées :

- Permet l'exploitation du contenu structuré répartie dans différentes tables
- Accès plus pratique aux données par des requêtes SQL
- La gestion d'un grand nombre de données sans avoir de problèmes de performances

Natural Language Processing (NLP)

"Il faut rester confiné à cause du méchant coronavirus. :'("

Retrait Majuscules	"il faut rester confiné à cause du méchant coronavirus. :'("
Retrait Stop Words	"faut rester confiné cause méchant coronavirus. :'("
Lemmatization	"falloir rester confiner cause méchant coronavirus. :'("
Retrait Accents	"falloir rester confiner cause mechant coronavirus. :'("
Retrait Caract. Spéciaux	"falloir rester confiner cause mechant coronavirus"

Objectifs :

- Avoir une liste de mots
- Tous les verbes à l'infinitif
- Pas de caractères spéciaux "gênants"

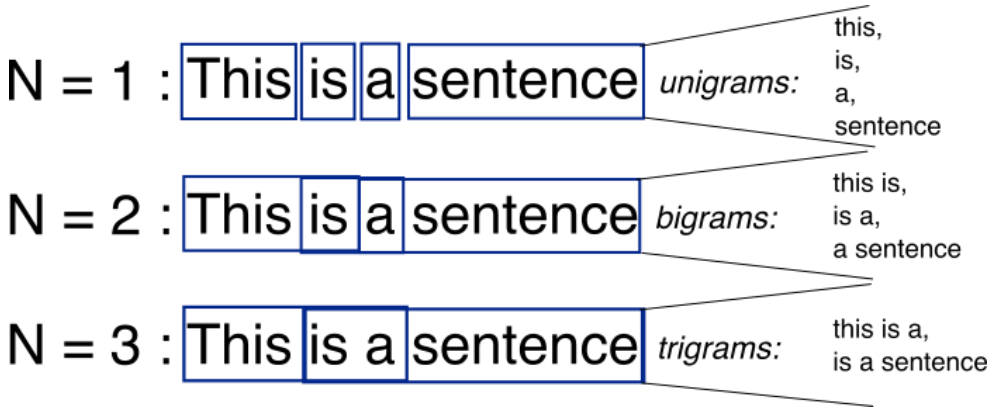
✓ On obtient donc un texte normalisé permettant un traitement unifié

Vectorisation des données traitées

2 choses à faire: **Définir un vocabulaire** et **calculer le TFIDF**

Division d'une section en sous éléments de n termes

Exemple:
L'expression "Prince Charles" devient-elle:
"Prince", "Charles"
ou
"Prince Charles" ?



$$TFIDF_{t,d,D} = TF_{t,d} \times IDF_{t,D}$$

Importance d'un terme t dans un document d

Fréquence d'un terme t dans un document d

Importance du terme t dans l'ensemble des documents D

$$TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}}$$

$$IDF(t,D) = \log \frac{n_D}{df(t)} + 1$$

Exemple de Vectorisation

Phrase 1 : “Le Prince Charles a le coronavirus !”
Phrase 2 : “Le Prince Charles a un cheval blanc.”
Phrase 3 : “Le coronavirus.”

Phrase 1 : [prince charles],[coronavirus]
Phrase 2 : [prince charles],[cheval], [blanc]
Phrase 3 : [coronavirus]

Calcul du TF:

	prince charles	coronavirus	cheval	blanc
Phrase 1	0.5	0.5	0	0
Phrase 2	0.33	0	0.33	0.33
Phrase 3	0	1	0	0

Calcul de IDF:

	IDF(t)
prince charles	1.18
coronavirus	1.18
cheval	1.48
blanc	1.48

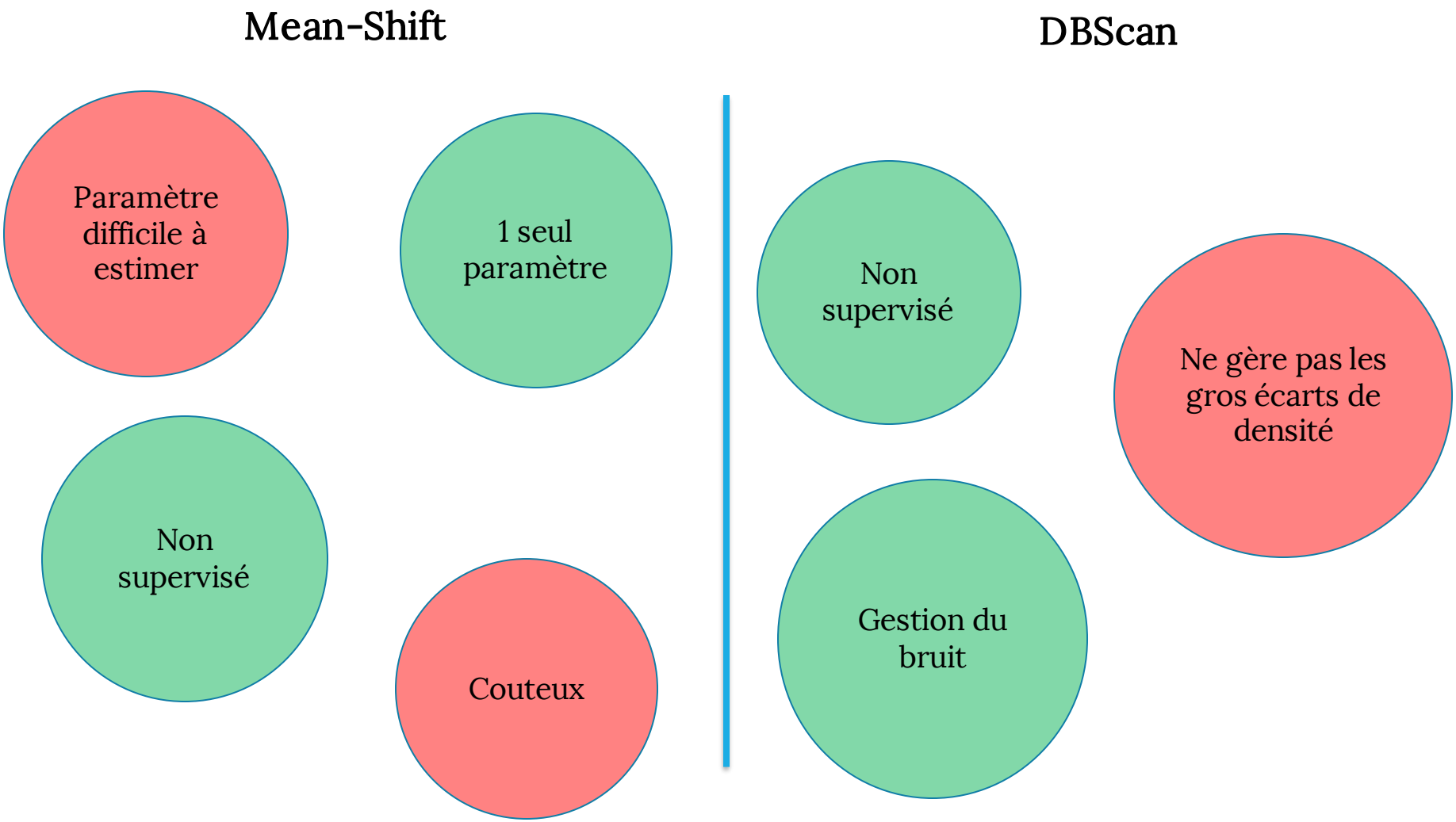
Vectorisation des phrases avec le TF-IDF :

	prince charles	coronavirus	cheval	blanc
Phrase 1	0.59 (0.71)	0.59 (0.71)	0	0
Phrase 2	0.39 (0.49)	0	0.49 (0.61)	0.49 (0.61)
Phrase 3	0	1.18 (1)	0	0

Normalisation des termes :

$$v = \frac{0.59}{\sqrt{0.59^2+0.59^2}} = 0.71$$

Clustering : Algorithmes étudiés



Clustering : Algorithme Mean-Shift

Méthode basée sur le concept KDE :

- Calcul de densité
- Regroupement par densité

Pour toutes les valeurs (données) :
Tant que valeur != centre de masse
On calcule le centre de masse m(x)
On déplace la valeur

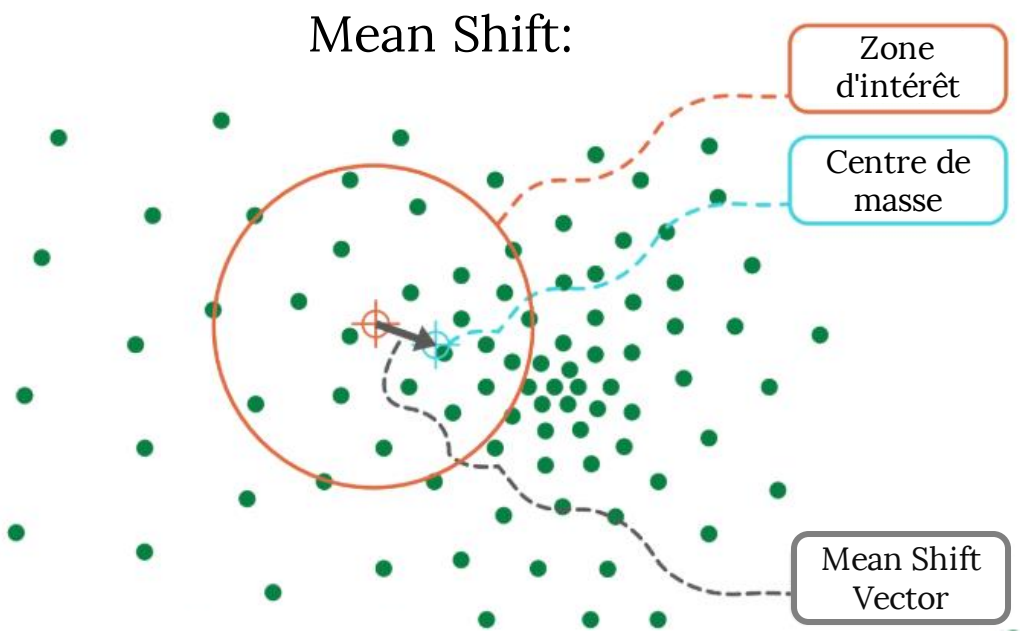
$$m(x) = \frac{\Sigma_{xi \in N(x)} K(x_i - x) x_i}{\Sigma_{xi \in N(x)} K(x_i - x)}$$

m(x) est le centre de masse

K(x) est la fonction de densité

X est la donnée en cours de traitement

N(x) est le voisinage (région d'intérêt)



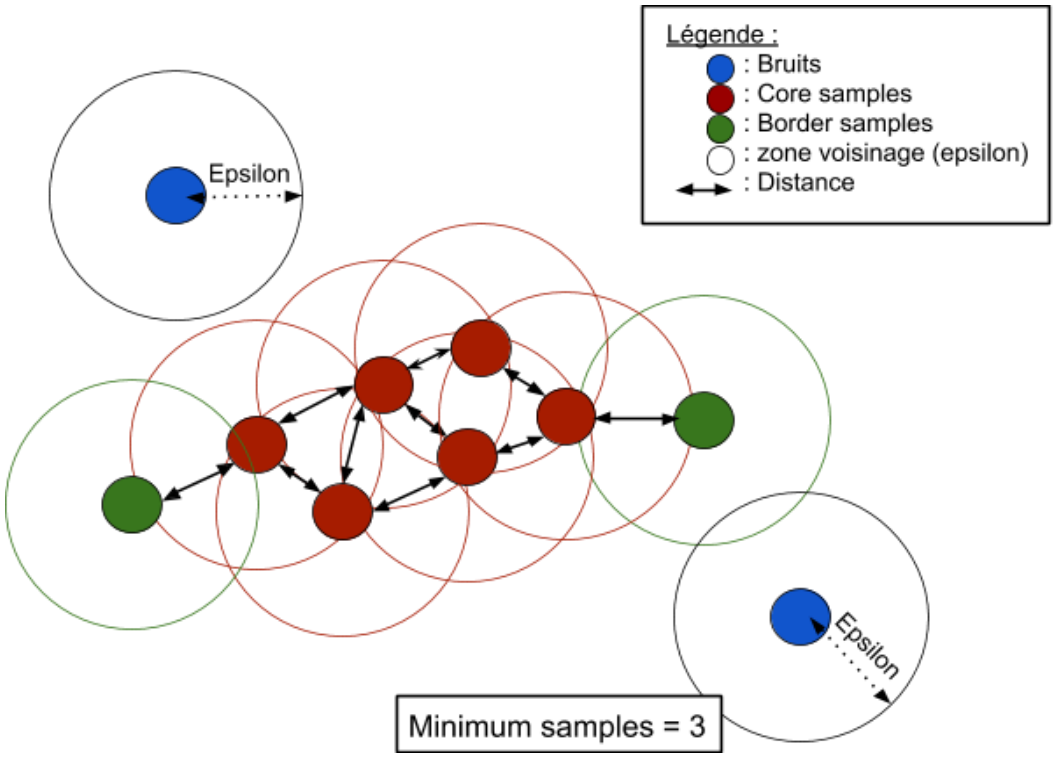
Clustering : Algorithme DBSCAN

Différents paramètres:
- ε : La distance observé autour de chaque donnée
- minSample : Le nombre d'éléments minimum dans un ε-voisinage pour définir un core sample

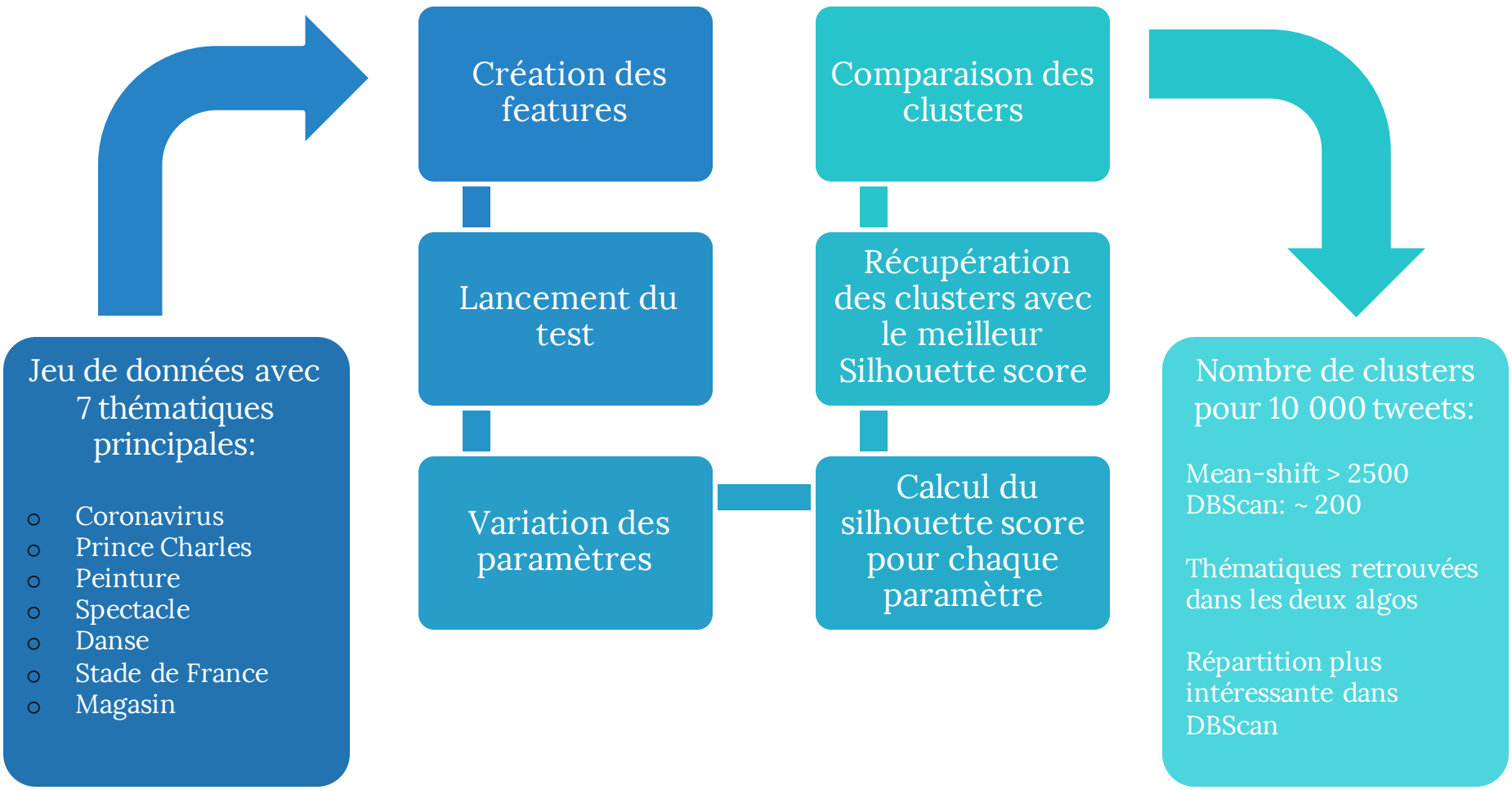
Core sample:
Sample ayant au moins minSample autres samples dans son ε-voisinage

Border sample:
Nombre de core samples dans l'ε-voisinage ∈ [1;minSample[

Bruit:
Nombre de core samples dans l'ε-voisinage = 0



Tests de comparaisons



Prédictions : Modèle ARIMA

Trois paramètres :
- p : nombre de terme auto-regressif.
- d : nombre de différenciation nécessaire.
- q : nombre de terme moyenne-glissante.

Avec **Y** : La série temporelle
α : coefficient initial
B : coeffcient de décalage estimé par le modèle
ε : l'erreur d'estimation
Φ : coefficients de moyenne mobile du modèle

Modèle auto-régressive AR :

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

Modèle de moyenne mobile MA :

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Modèle ARIMA :

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

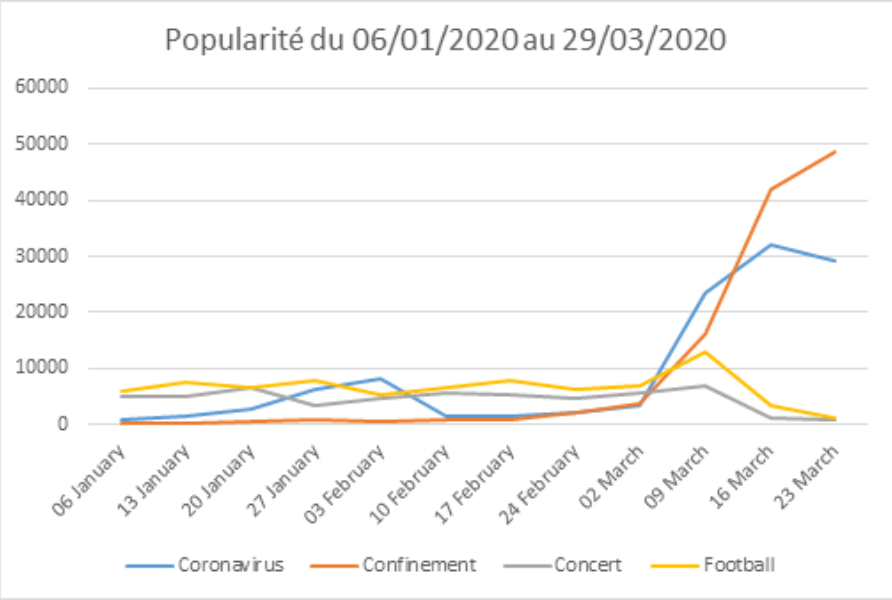
Affichages des events

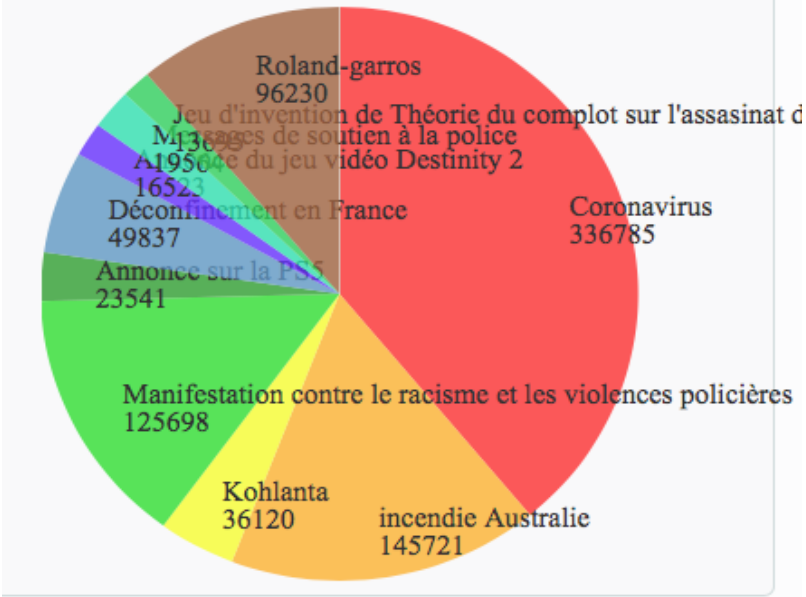
- Les langages et bibliothèques utilisés pour l'application Web :

Chart.js

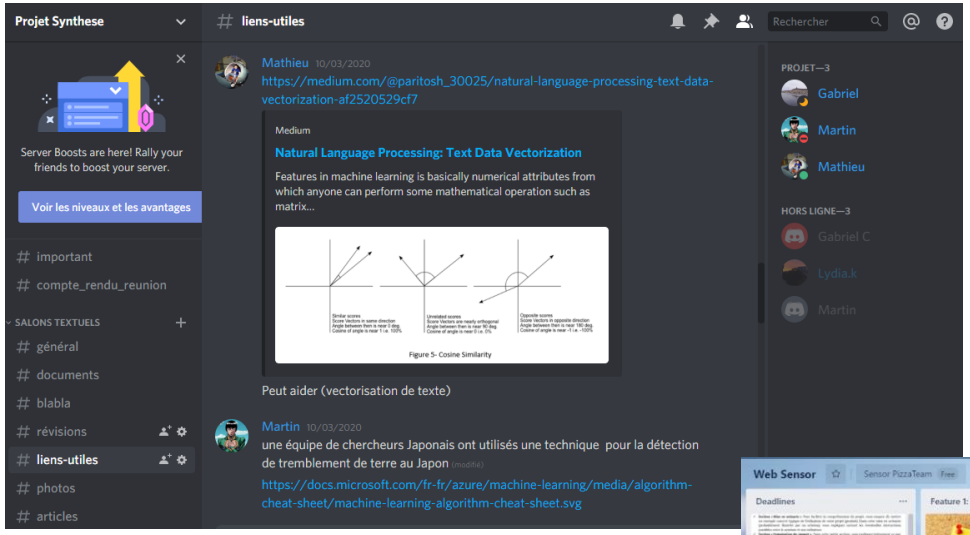
Data-Driven Document

- Visualisations comparatives de la popularité des events :



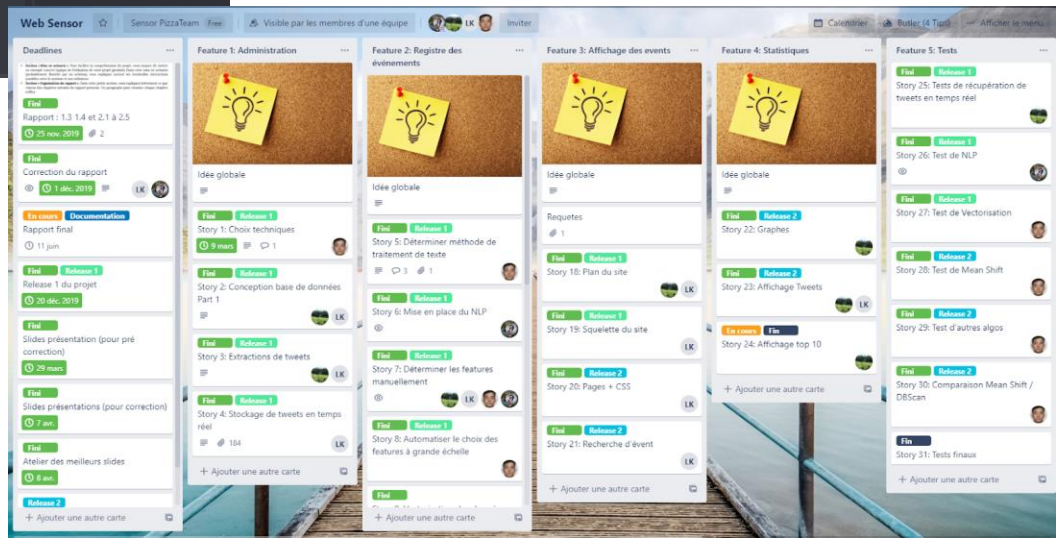


Gestion de projet: Outils utilisés



Discord:
Outil de communication
Indispensable pendant le confinement
Messages conservés et triées par salons

Trello:
Outil de gestion de projet
Tableau divisé en 5 features
Plus de 30 stories réalisées



Github

Gestion de projet: Releases et répartition

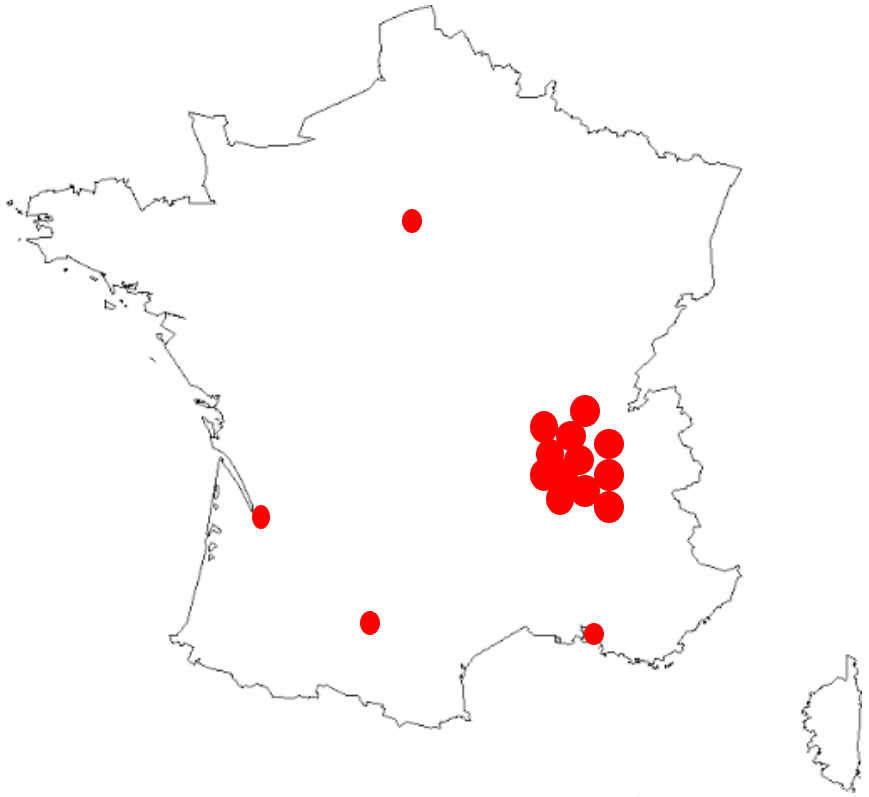
Date de release	Fonctionnalités	Mathieu	Martin	Gabriel	Lydia
09/03/2020	Mise en place des outils de gestion	X			
	Mise en place des BDs			X	X
	Récupération des données			X	
	Mise en place du NLP	X	X		
	Tests NLP	X	X		
	Squelette de site web				X
24/04/2020	Algorithme de clustering	X	X		
	Premiers affichages statistiques			X	X
	Tests de clustering		X		
	Slides + documentation du projet	X	X	X	X
19/06/2020	Clustering sur du temps réel	X	X		
	Prédiction d'évent populaire		X		
	Stockage et requêtes SQL			X	X
	Fin du site web			X	X
	Fin de documentation	X	X	X	X

Conclusion et perspectives

Découverte de nouvelles notions:
- NLP
- Clustering
- Prédiction

Différentes applications de ces notions:
- Traduction automatique
- Partitionnement de données
- Meteo

Perspectives d'amélioration:
- Traitement en temps réel
- Représentation géographique des tweets



Event: Tremblement de terre à Lyon

Merci de votre attention
Questions?