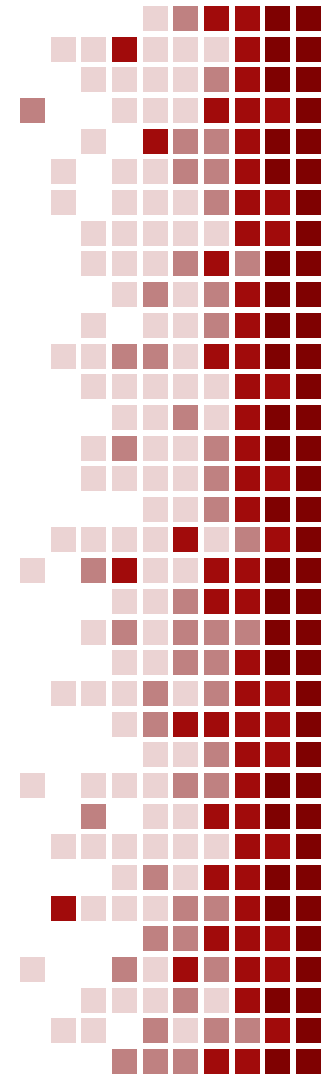


# Web Digital Footprints & Data Privacy

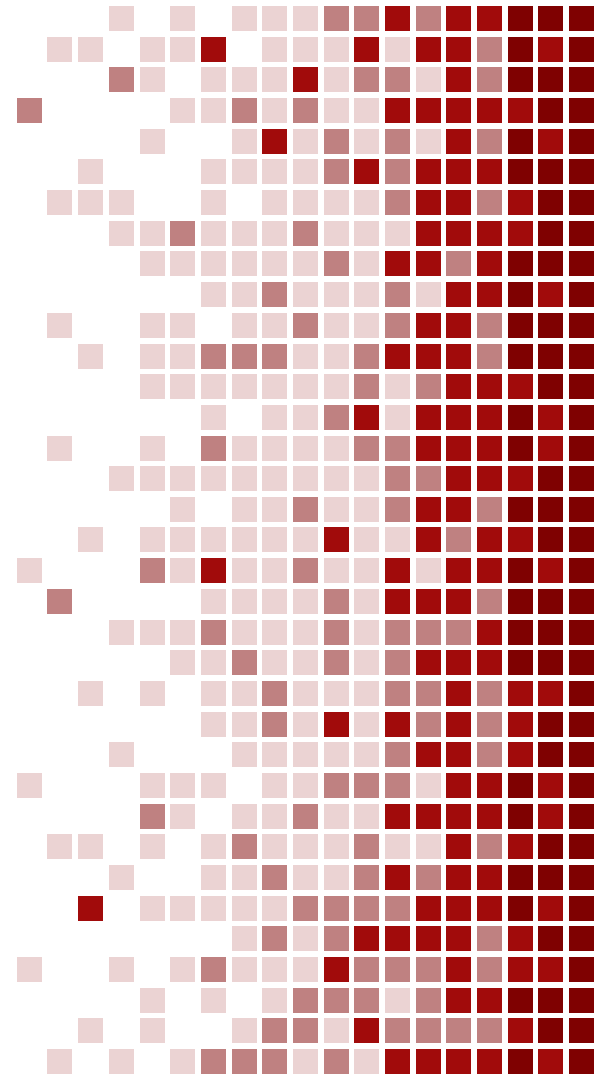
Kewin Dousse  
2018

# PLAN

1. Contexte
2. Objectifs
3. Etat de l'art
4. Démonstration
5. Architecture
6. Résultats
7. Conclusion



# 1. Contexte



# CONTEXTE

## Situation

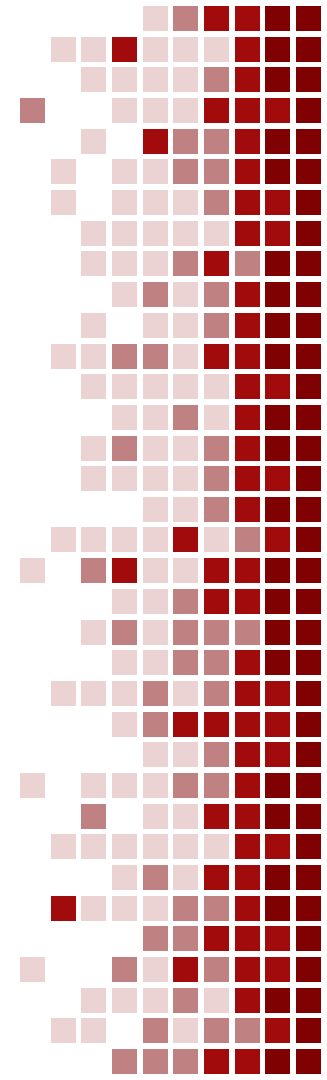
- Révélations d'une étude de Michal Kosinski en 2017
- Facebook likes 👍  
→ profil personnel
- Profiling de plus en plus utilisé :  
Publicités,  
Campagne politique

## Motivation

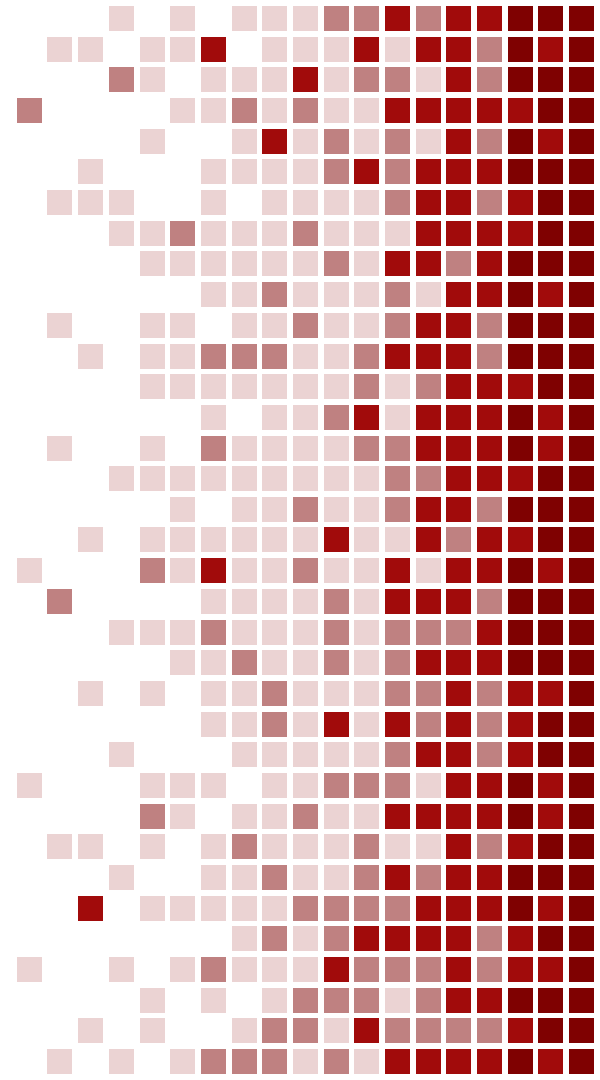
Volonté de **sensibiliser** le public au profiling

## Rôles

- Nastaran Fatemi (Responsable)
- Félicien Fleury (Mendant)
- SDIPI




## 2. Objectifs



“ *Le but de ce projet est de proposer un outil de visualisation pour sensibiliser le public à la question du profiling sur internet.* ”

# OBJECTIFS : Forme

- Extension pour **Google Chrome** 
- Développement **open-source**  
[github.com/sdipi](https://github.com/sdipi)

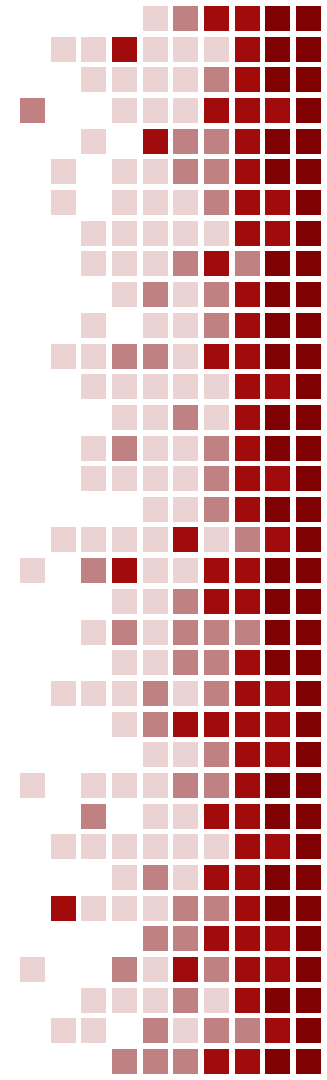
# OBJECTIFS : Fonctionnalités

## Profiling

- **Demande** d'informations de base
- **Récolte** des données de navigation
- **Création** d'un profil
- **Evaluation** de la pertinence

## Tracking

- **Identification** des trackers
- **Navigation** parmi les données récoltées



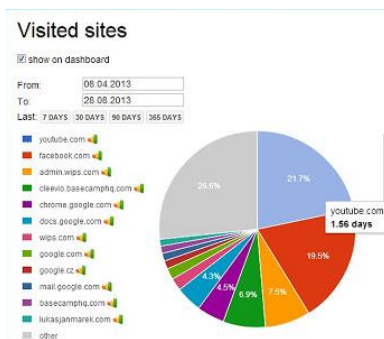


# 3. Etat de l'art

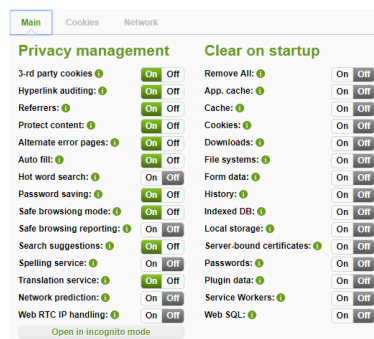


# ETAT DE L'ART : Extensions

- Statistiques de navigation
- Masquage de données personnelles
- Détection de trackers



timeStats



Privacy manager

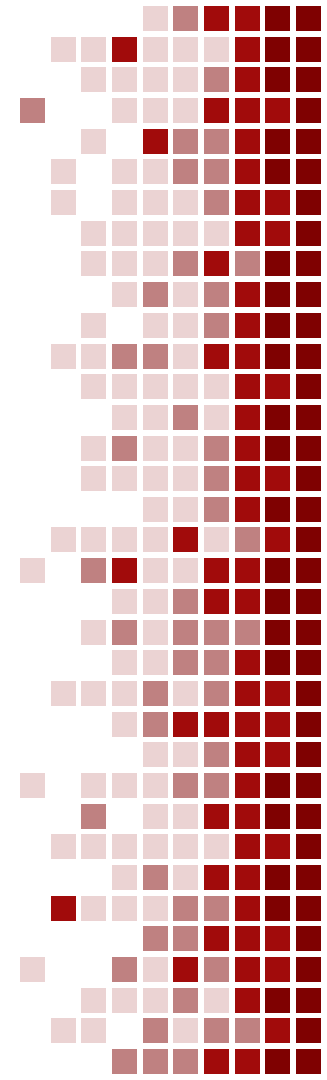
# ETAT DE L'ART : Analyse de texte

## Keyword extraction

- Analyse d'un corpus de documents
- Sélection de mots importants

## Topic modeling

- Analyse d'un corpus de documents
- Reconnaissance/génération de thèmes sous-jacents



# ETAT DE L'ART : Keyword extraction

## TF-IDF

Mot important :

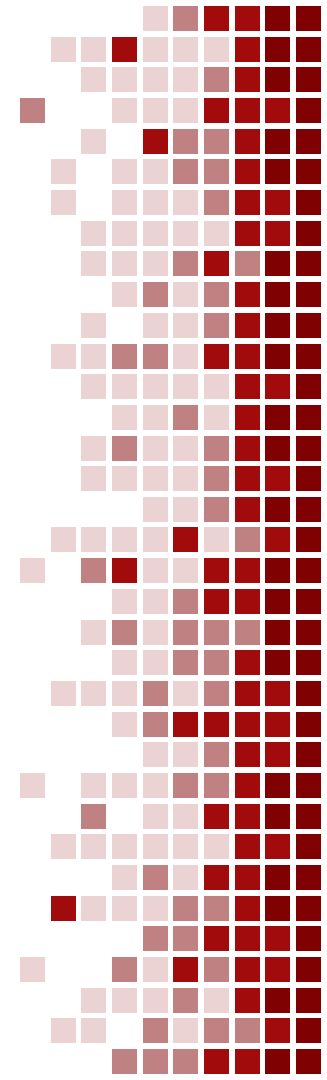
- Beaucoup d'occurrences
- Présence dans peu de documents
- Score par mot par document

## RAKE

- Séparation en groupes de mots
- Nombre d'occurrences des mots et groupes

## TextRank

- Séparation en phrases
- Identifie les phrases semblables
- Inspiré de PageRank



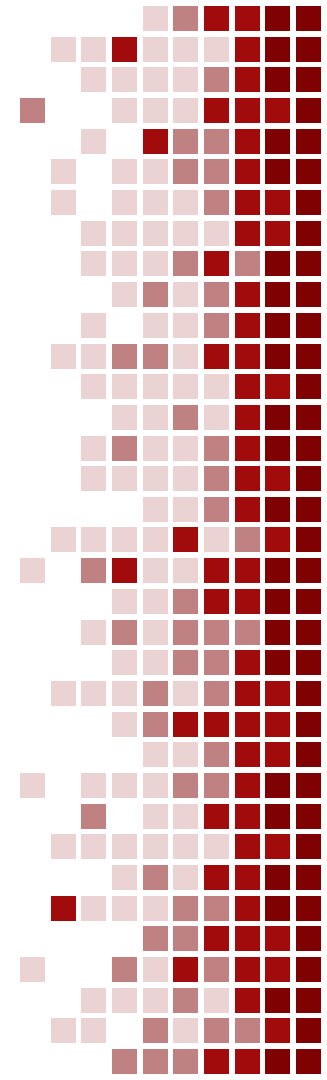
# ETAT DE L'ART : Topic Modeling

**But :** Déterminer des *thèmes* sous-jacents aux pages web, et regrouper les mots qui les composent.

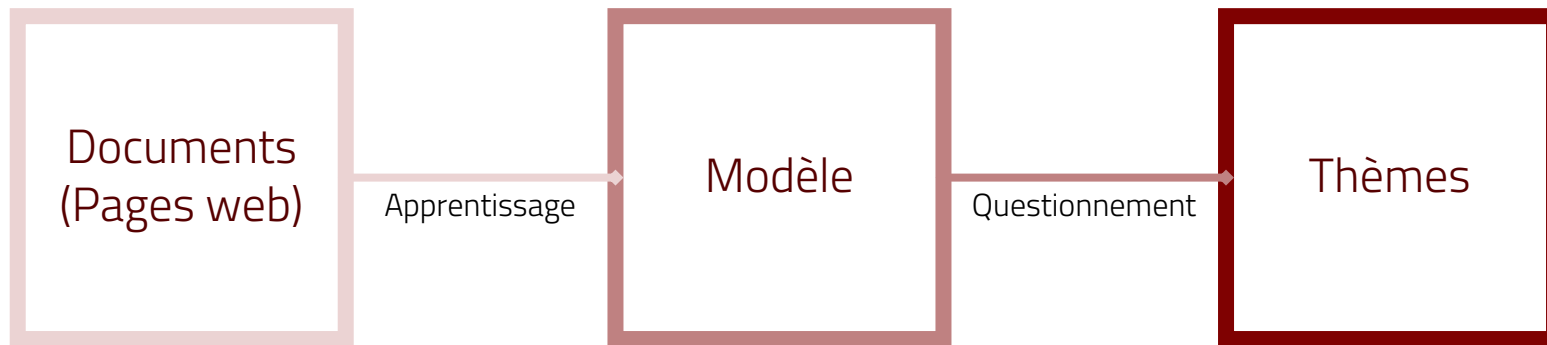
**Méthode :** Création de modèles *probabilistiques* en analysant le contenu des pages web

**Algorithmes :**

- LSA
- pLSA
- **LDA**

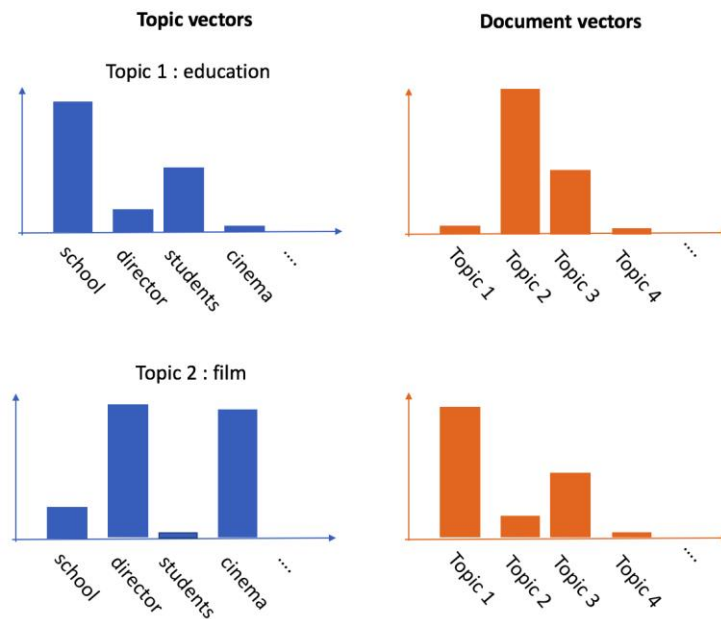


# ETAT DE L'ART : Topic Modeling



# ETAT DE L'ART : LDA

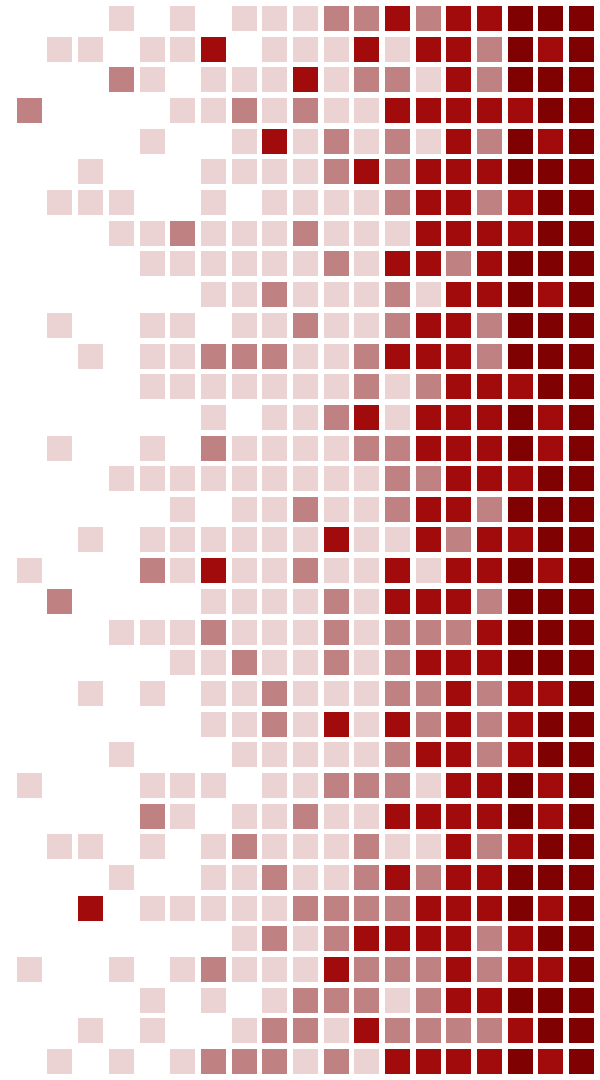
Exemple de modèle :



Source : DataCamp

4.

Démonstration

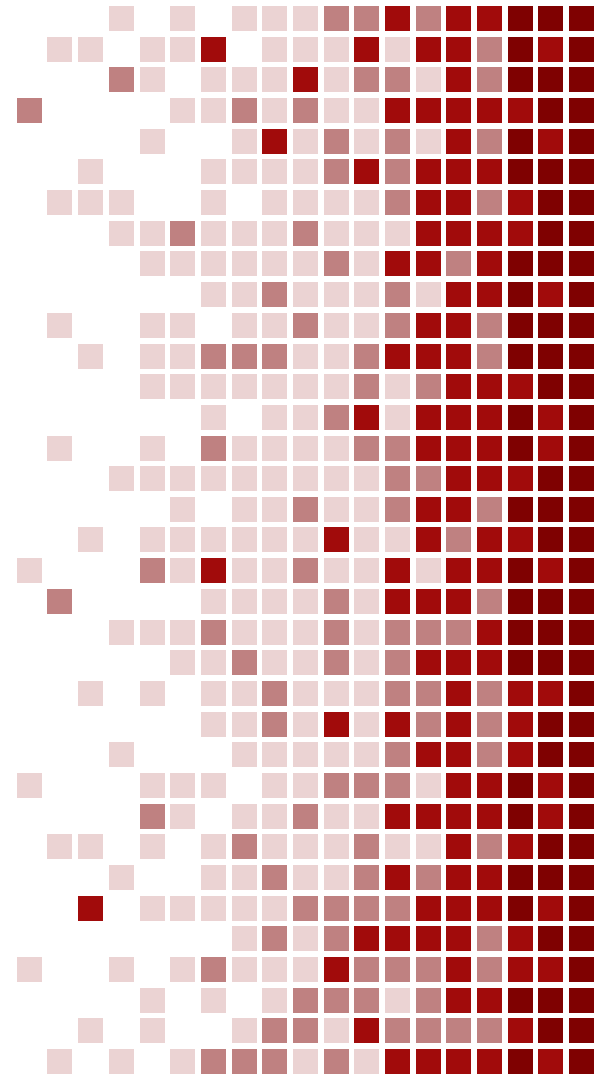




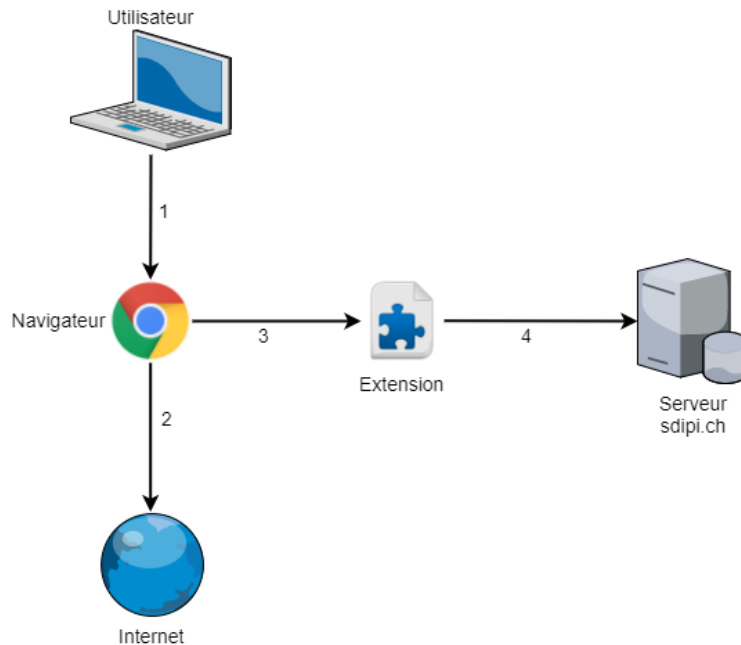
# DEMO



# 5. Architecture

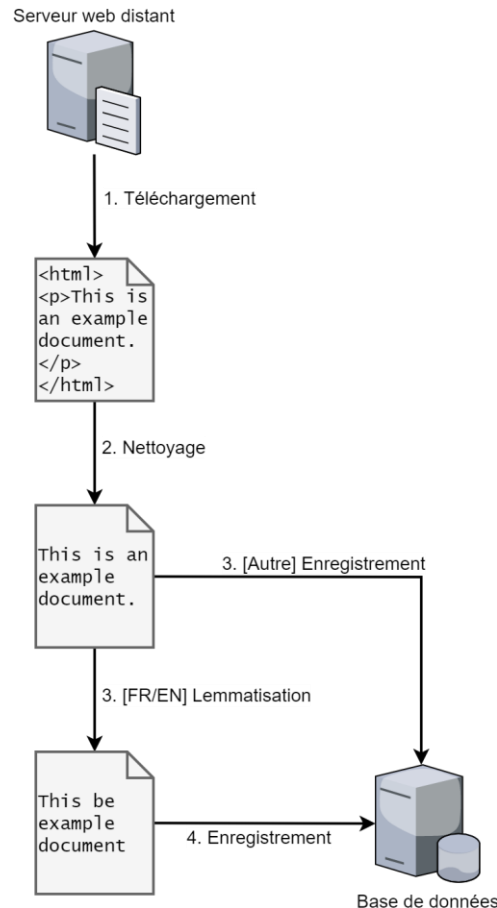


# ARCHITECTURE : Vue d'ensemble



# ARCHITECTURE : Récolte

1. Notre serveur télécharge la page HTML.
2. Simulation dans un navigateur Chrome. On ne garde en sortie que le texte.
3. Regroupement des différentes formes d'un mot [EN/FR].
4. Enregistrement du texte final.





# ARCHITECTURE : Wordcloud

## 1. Offline

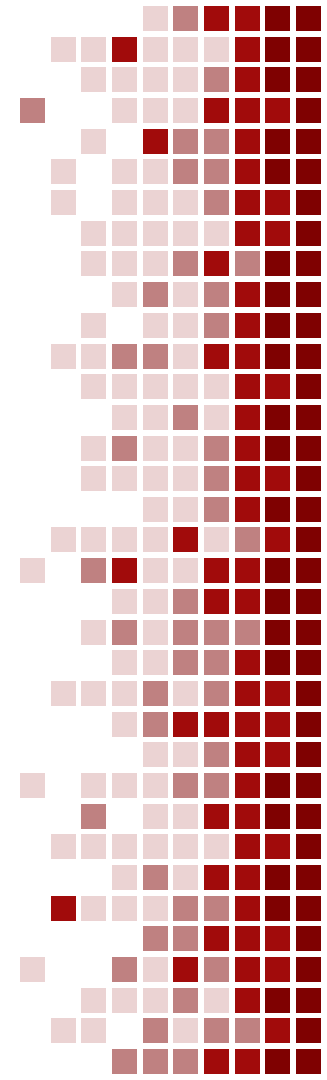
- Calcul du score TF-IDF de chaque mot
- Mémorisation des mots les plus importants par page

## 2. Serveur

- Somme du temps de visualisation de chaque page pour l'utilisateur

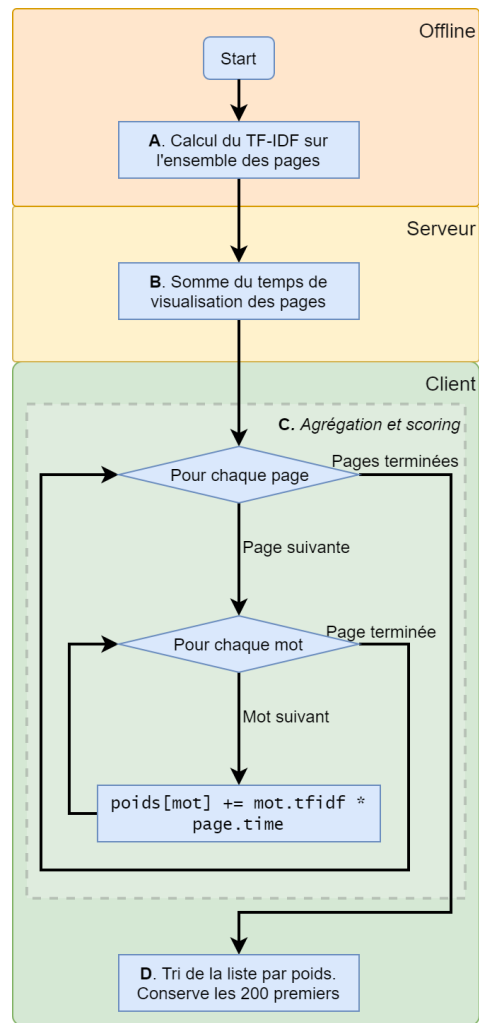
## 3. Client

- Calcul du score d'importance final pour chaque mot
- Affichage des meilleurs mots



# ARCHITECTURE : Wordcloud

- **Offline** : Effectué avant le démarrage du serveur
- **Serveur** : Effectué lorsque demandé
- **Client** : Effectué sur le navigateur



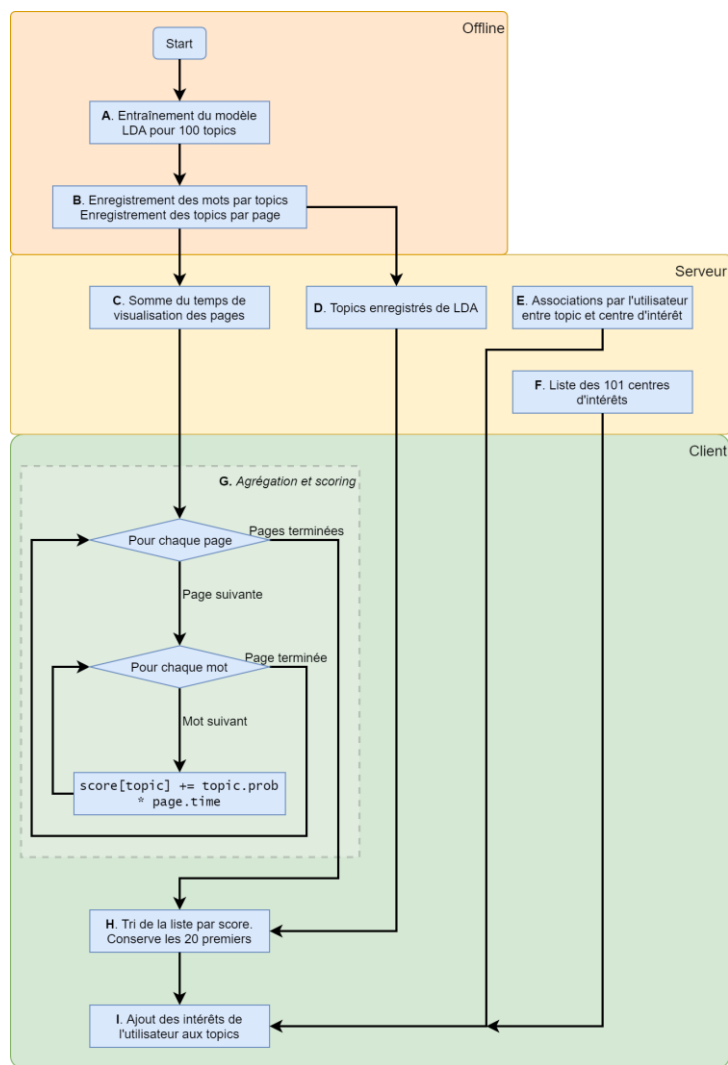
# ARCHITECTURE : Topics

## Topics List

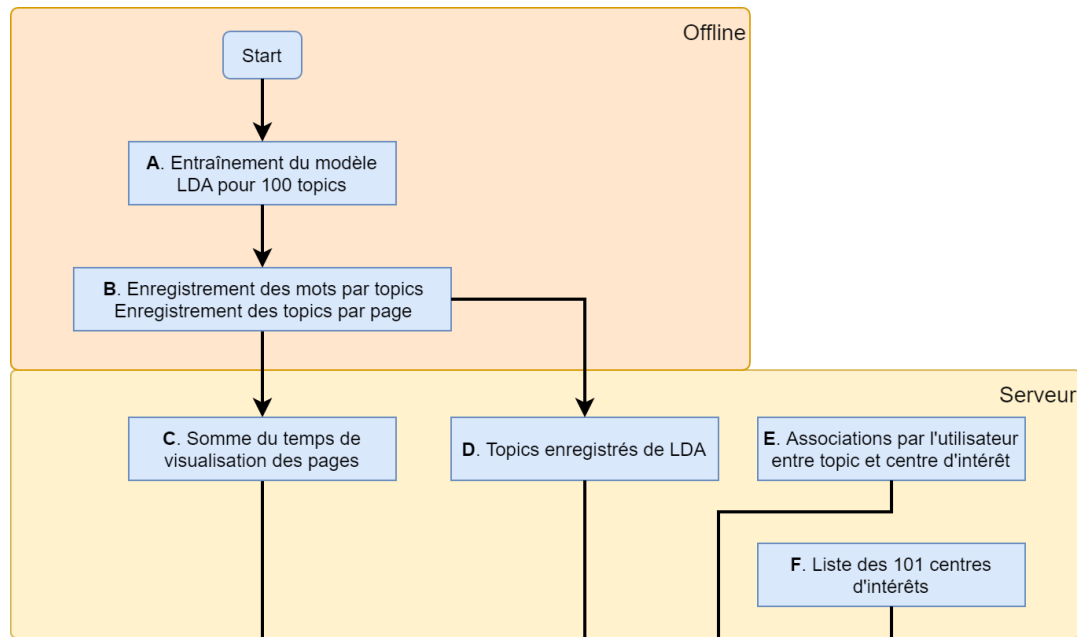
#	Words	Related interest	Estimated interest
1	comment reddit post	Social Media Enthusiast	100% <div><div></div></div>
2	share facebook link	Social Media Enthusiast	53% <div><div></div></div>
3	account sign email		45% <div><div></div></div>
4	rate earn win	Gamer	43% <div><div></div></div>
5	amp log src		40% <div><div></div></div>
6	anime girl manga	Social Media Enthusiast	37% <div><div></div></div>
7	chrome api web	Technophile	37% <div><div></div></div>
8	example vector product		36% <div><div></div></div>
9	like get just		25% <div><div></div></div>
10	leagueoflegends champion http	Hardcore Gamer	23% <div><div></div></div>
11	play n64 subreddit	Gamer	19% <div><div></div></div>
12	self topic word	Technophile	19% <div><div></div></div>



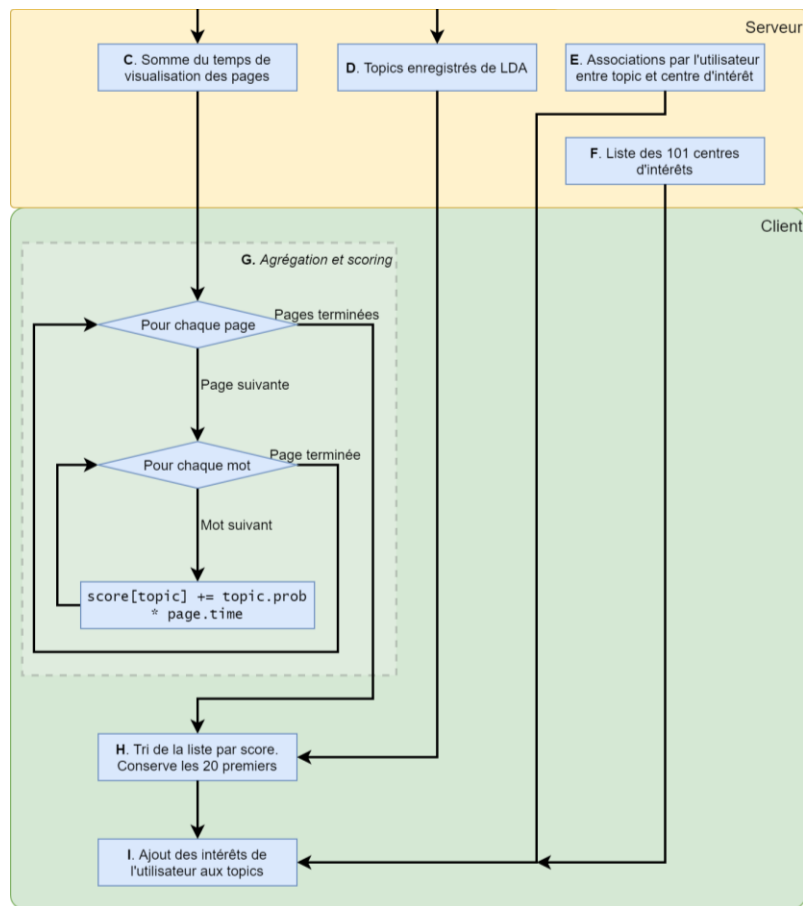
# ARCHITECTURE : Topics



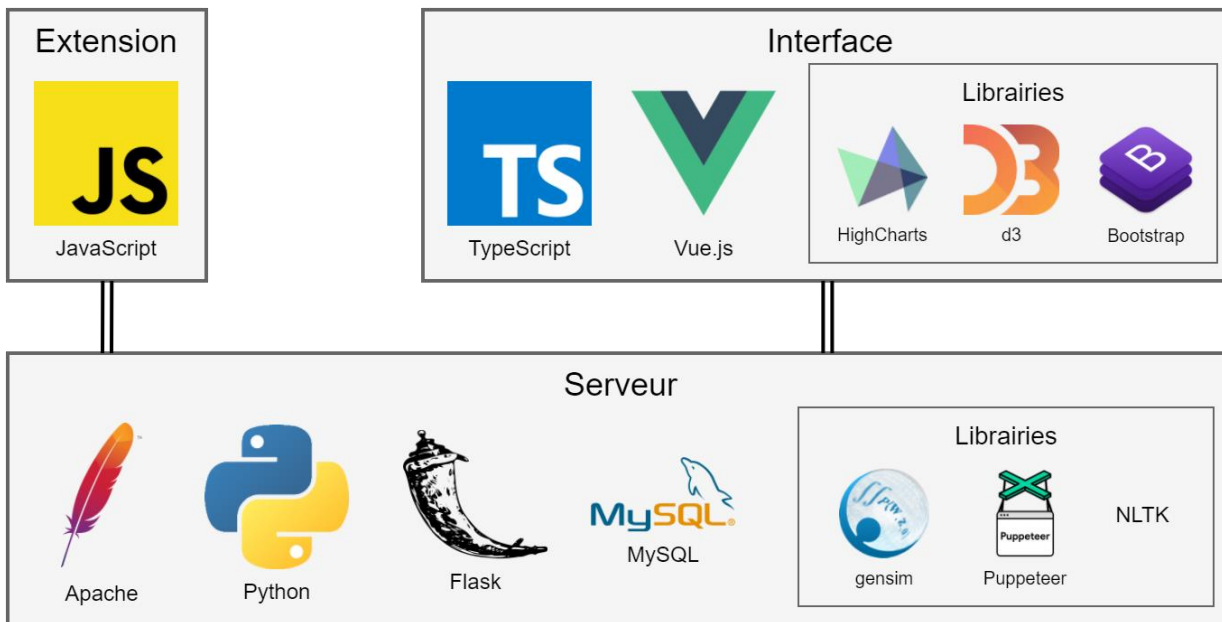
# ARCHITECTURE : Topics



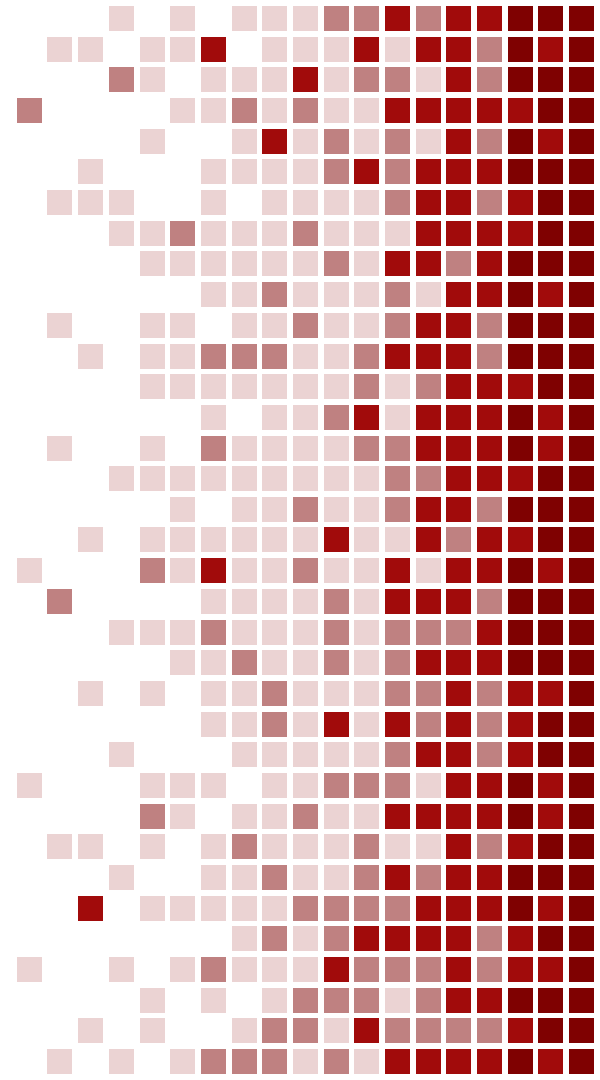
# ARCHITECTURE : Topics



# ARCHITECTURE : Stack technologique

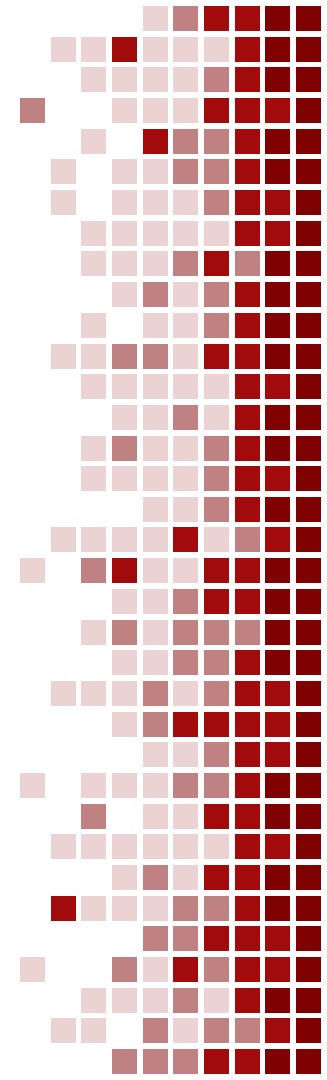


# 6. Résultats



# RESULTATS : Evaluation

- Extension disponible sur le Chrome Web Store : Début janvier 2018
- Récolte de données → début février 2018 (1 mois)
- Evaluation de la performance de :
  - TF-IDF
  - LDA
  - Topics suggérés
  - Visualisations



# RESULTATS : TF-IDF

- **Utilisation :** Affichage de mots clés par site web

#	Domain	Keywords
1	<a href="http://www.reddit.com">www.reddit.com</a>	<i>reddit submit comment</i>
2	<a href="http://df.sdipi.ch">df.sdipi.ch</a>	<i>phpmyadmin past welcome</i>
3	<a href="http://wdf.sdipi.ch">wdf.sdipi.ch</a>	<i>footprints digital extension</i>
4	<a href="http://www.draw.io">www.draw.io</a>	<i>gmdl eng proc</i>

- Satisfaisant lorsque conditions réunies
- Non-sens dans les autres cas

# RESULTATS : Topic Modeling

- **Utilisation :** Génération de thèmes, regroupant des mots

#	Words	Related interest
1	comment reddit post	<input type="text" value="Social Media Enthusiast"/>
2	share facebook link	<input type="text" value="Social Media Enthusiast"/>
3	example vector product	<input type="text" value="Technophile"/>
4	amp log src	<input type="text"/>

- 2/3 des topics sont «sensés»



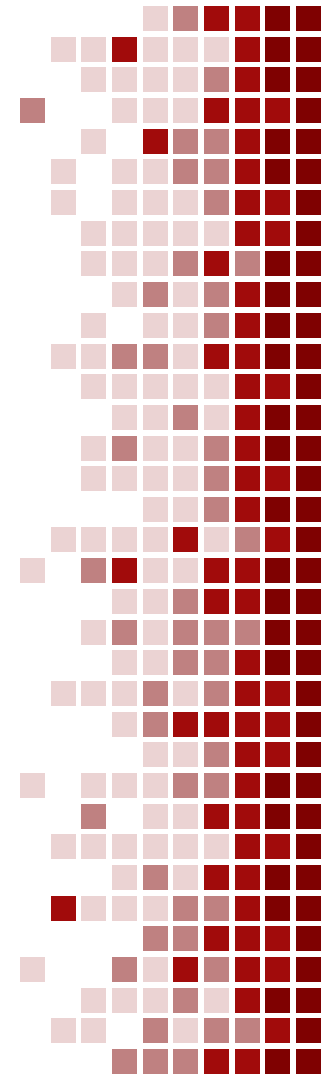
6 users

7'183 URLs

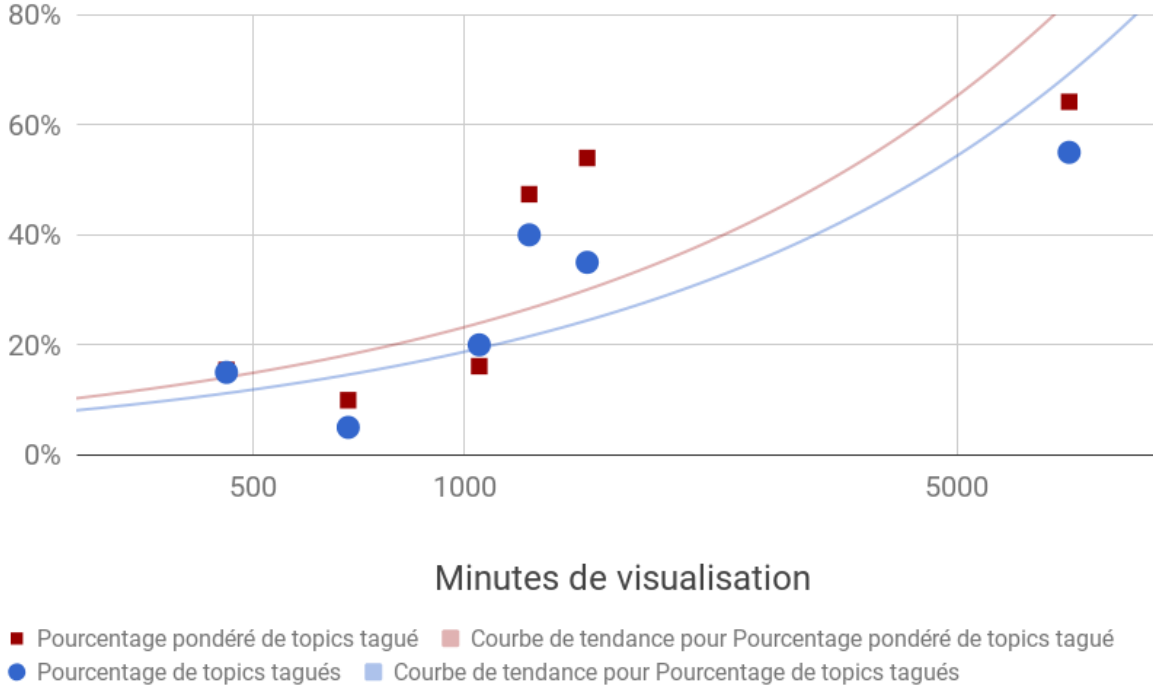
Pages distinctes

4.5 Go

Taille de la DB



## Taux d'association des topics proposés

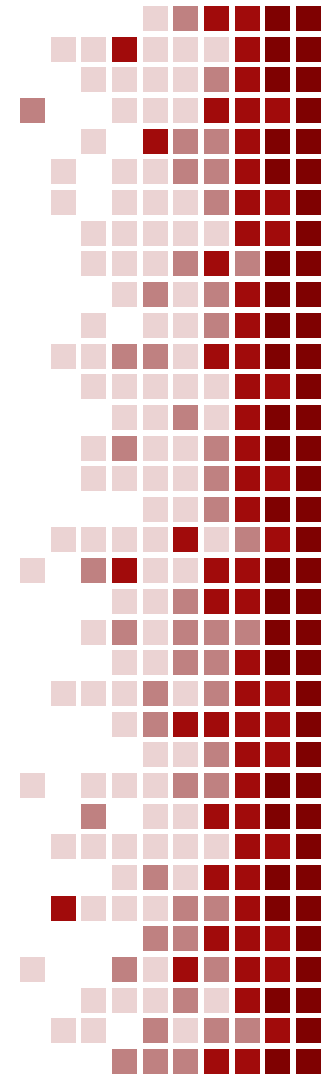


34 / 120

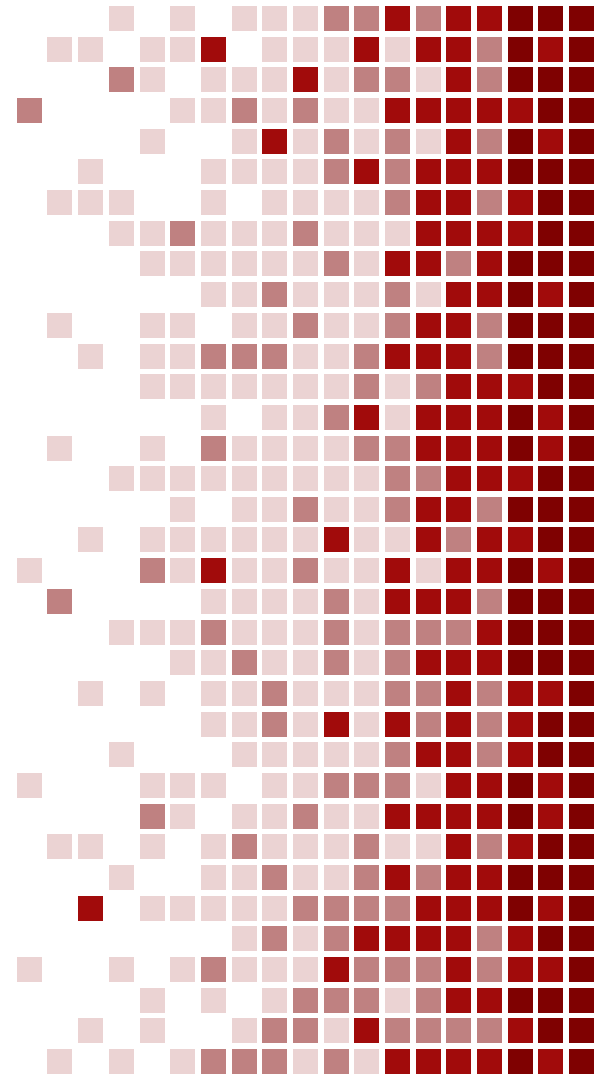
Intérêts identifiés

34.5 %

Correspondance pondérée



# 7. Conclusion



# CONCLUSION : Réalisations

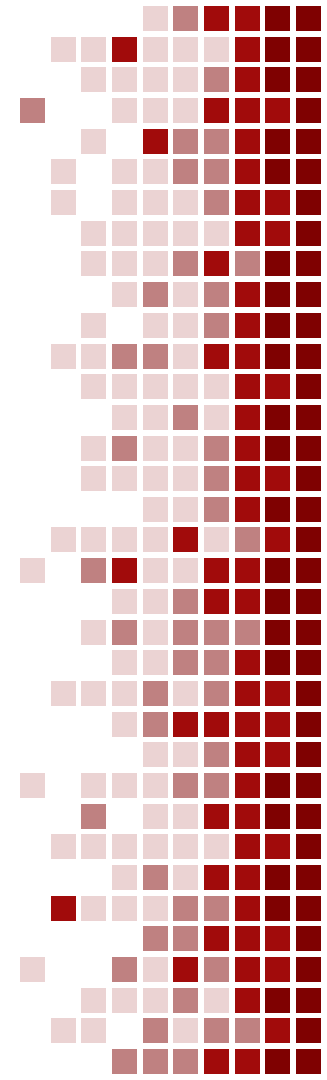
- Etat de l'art de techniques de tracking, profiling
- Outil fonctionnel de **récolte** de données de navigation, puis **génération** et **visualisation** de profils
- **Analyse** des données récoltées, révélant le **potentiel de détection** de données personnelles

# CONCLUSION : Tâches

Tâche	Bilan	Commentaire
Etat de l'art : Tracking		
Etat de l'art : Analyse de texte	-	Suffisant, aurait pu être plus complet et plus poussé
Conception de la solution	+	A évolué au fil de plusieurs itérations
Implémentation de la solution		Application fonctionnelle et stable
Evaluation de la solution et résultats		Interprétation subjective inévitable

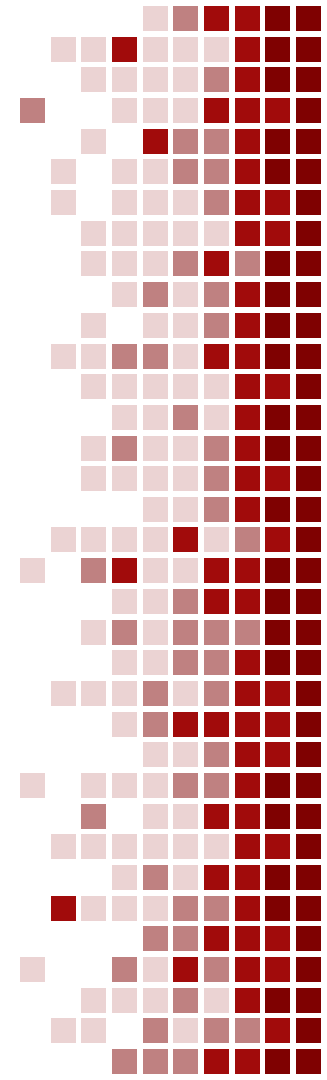
# CONCLUSION : Travaux futurs

- **Plus d'utilisateurs** : Meilleure évaluation
- Meilleur **cleaning** de données
- Utilisation de différents algorithmes
- Capture de plus de données



# CONCLUSION : Personnel

- Utilisation de technologies du Web
- Recherche
- Nécessité d'analyser plusieurs aspects
- Produit final fonctionnel



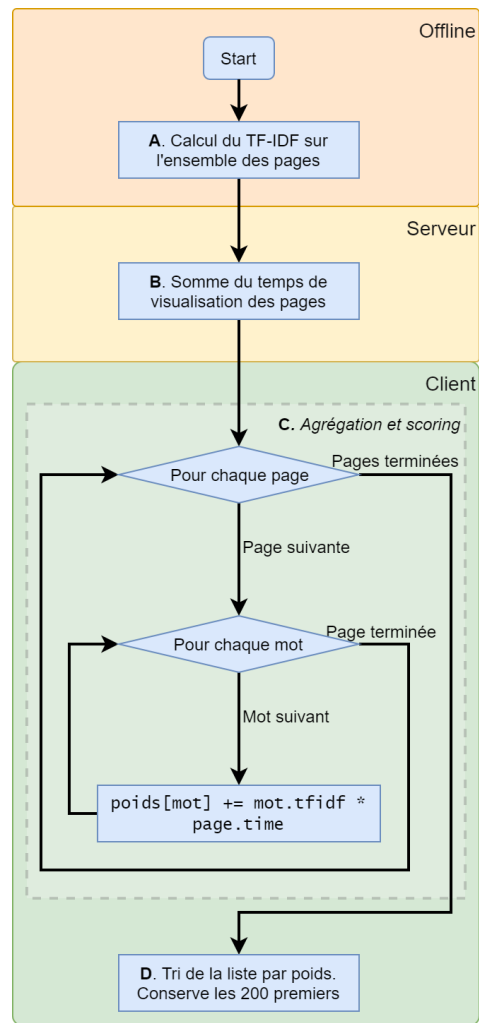


# MERCI !

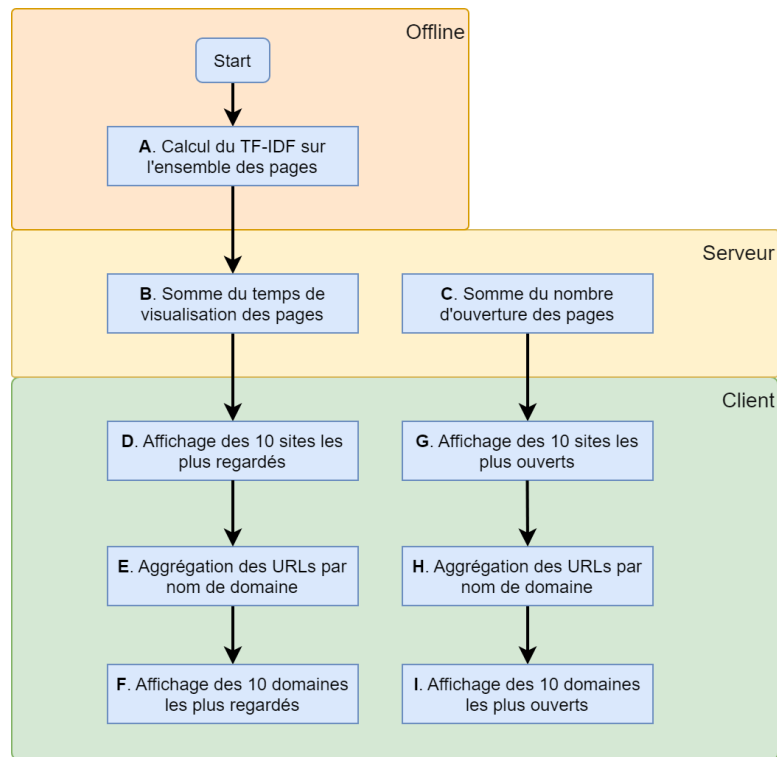
## Questions ?

# ARCHITECTURE : Wordcloud

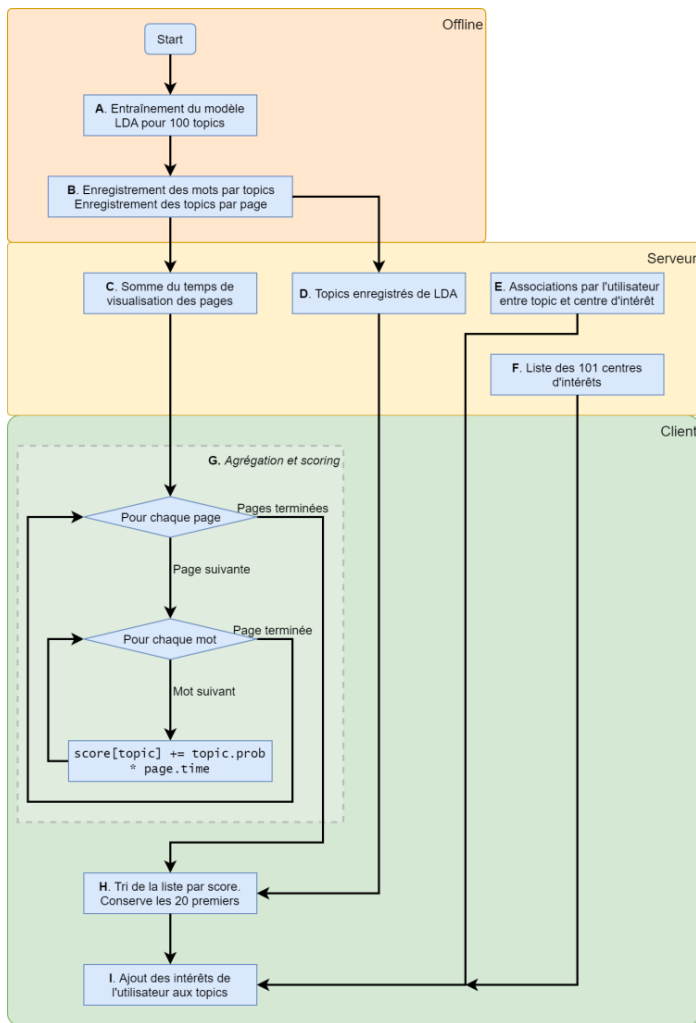
- **Offline** : Effectué avant le démarrage du serveur
- **Serveur** : Effectué lorsque demandé
- **Client** : Effectué sur le navigateur



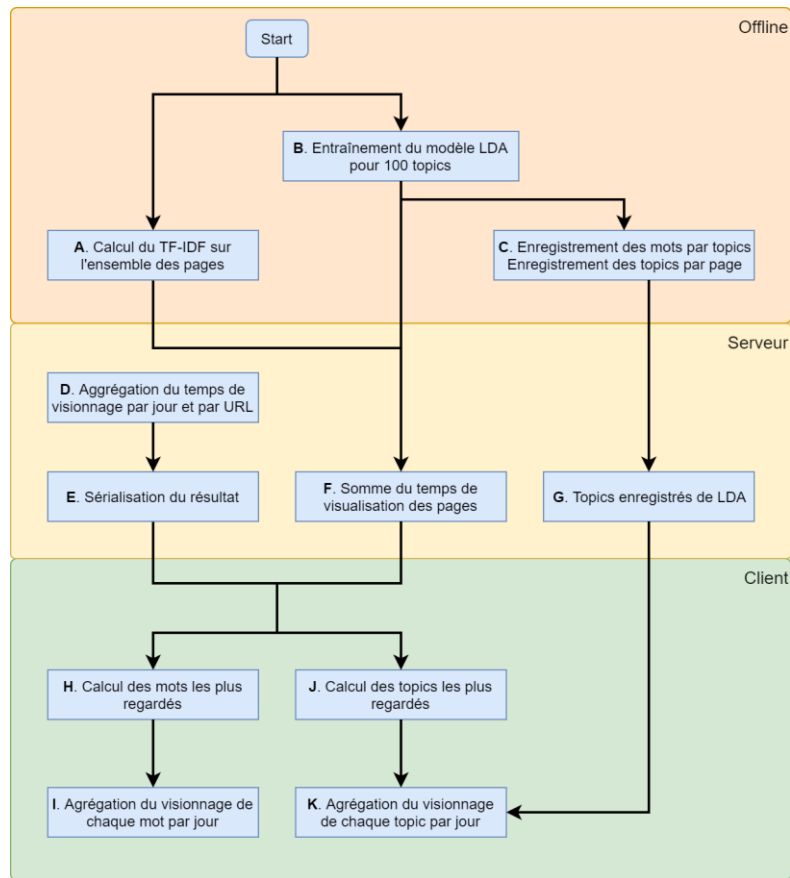
# ARCHITECTURE : Most watched



# ARCHITECTURE : Topics

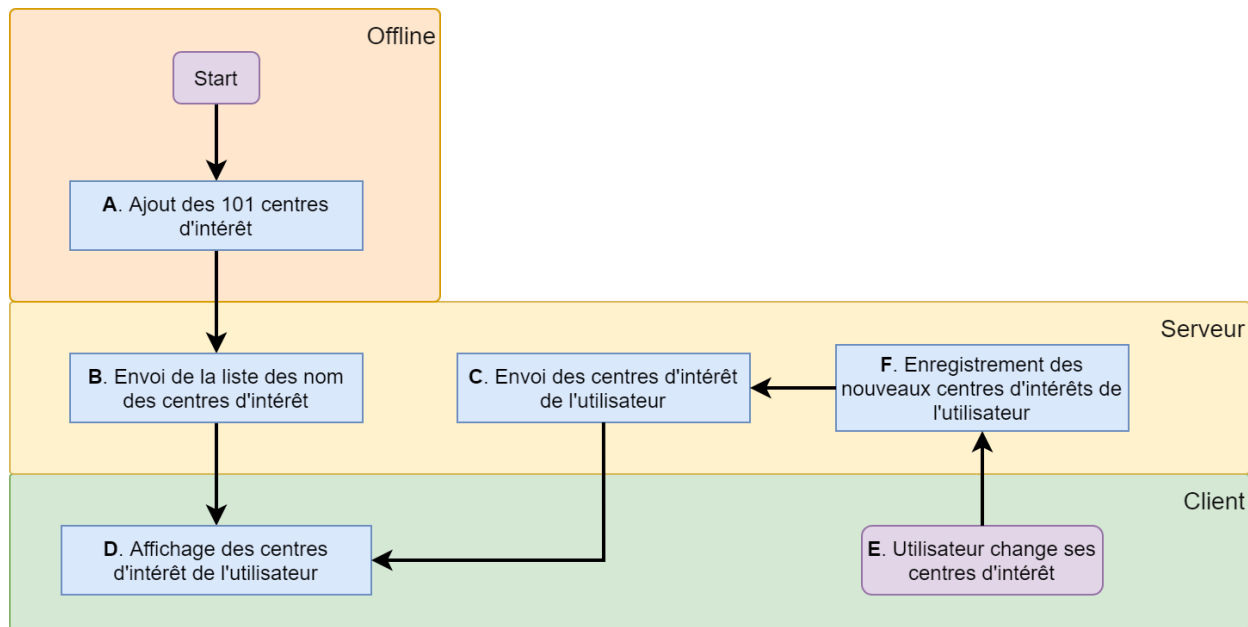


# ARCHITECTURE : History



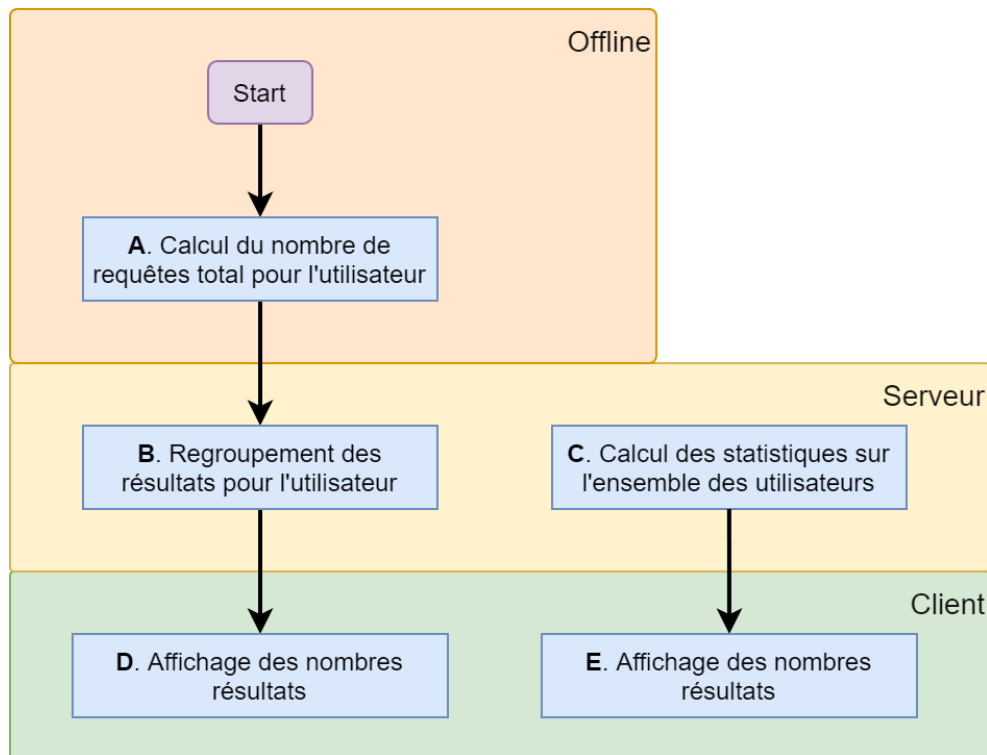
# ARCHITECTURE :

## Settings

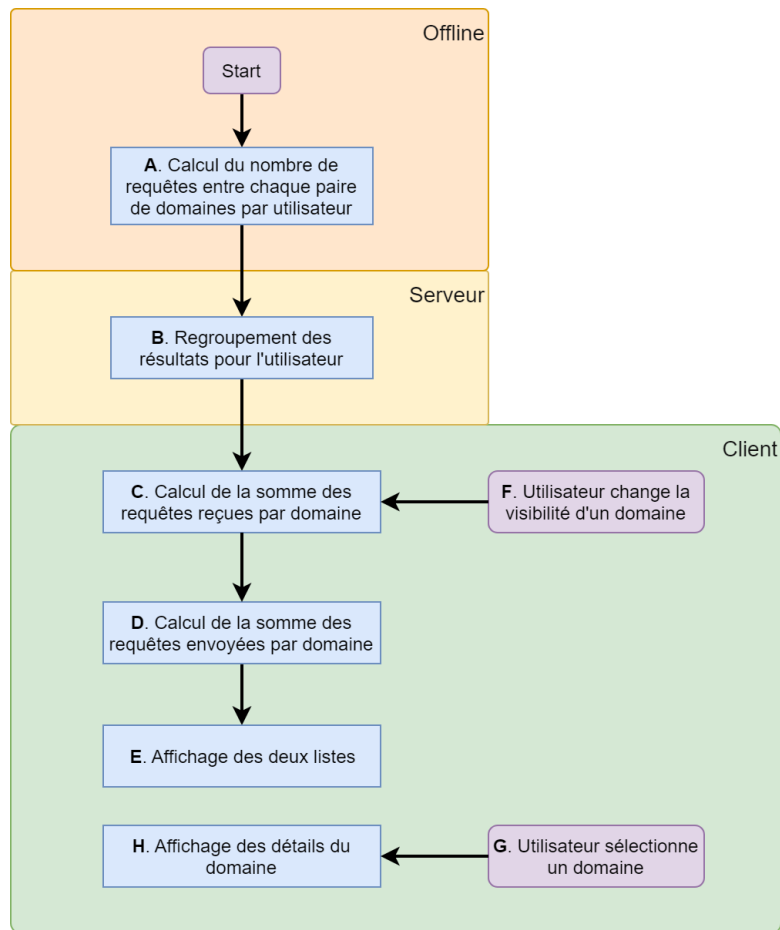


# ARCHITECTURE :

## Stats

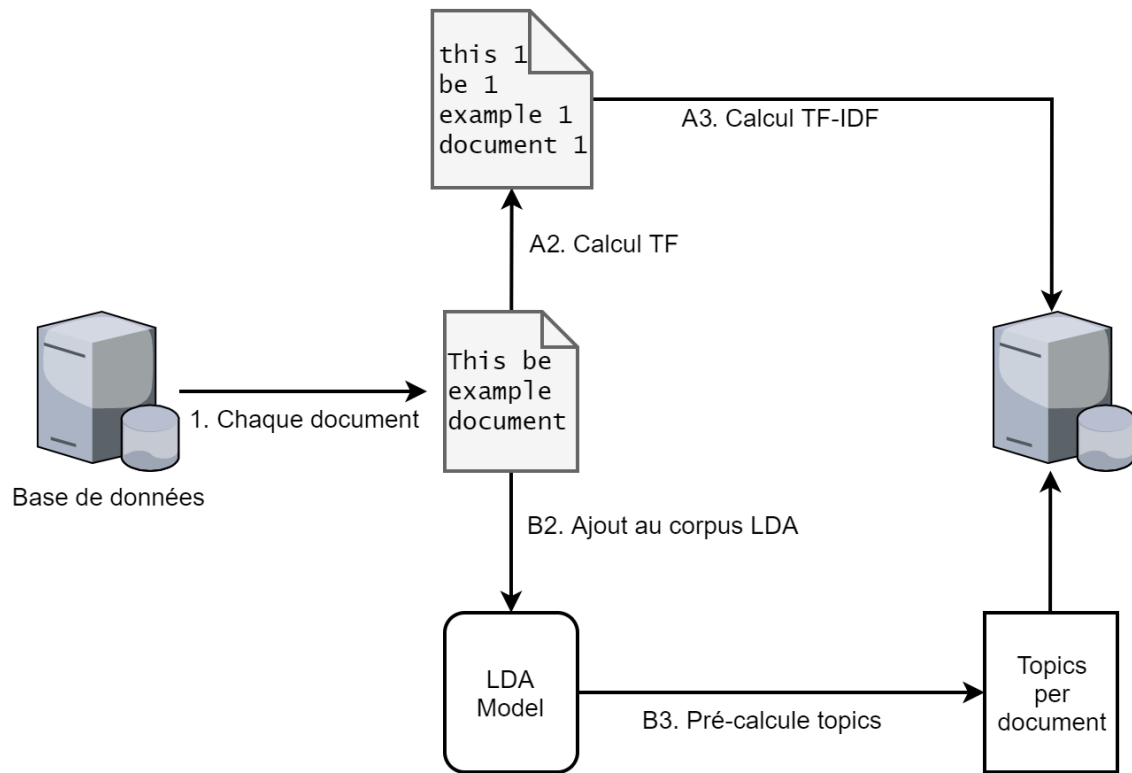


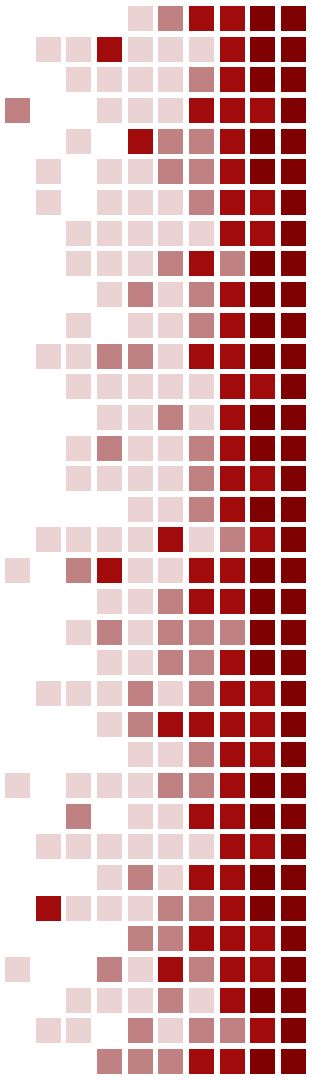
# ARCHITECTURE : Trackers





# Traitement offline





# Règlement Général sur la Protection des Données

- Entre en vigueur le 25 mai 2018
- Droits pour la personne fournissant des données
  - Le droit à l'information
  - Le droit d'accès
  - Le droit de rectification
  - Le droit d'effacement
  - Le droit à la limitation du traitement
  - ...

# Données envoyées aux Trackers

