Projet inter-promotions 2021

Berger-Levrault

Contexte du projet

éditeur de logiciels international et multisectoriel

Projet en collaboration avec la Direction Recherche et Innovation de Berger-Levrault

Différents domaines d'expertise

- Gestion administrative
- Citoyens, Familles et Élus
- Étudiants et Équipe pédagogique
- GMAO et Asset Management
- Patients, Résidents et Bénéficiaires

Objectif principal du projet

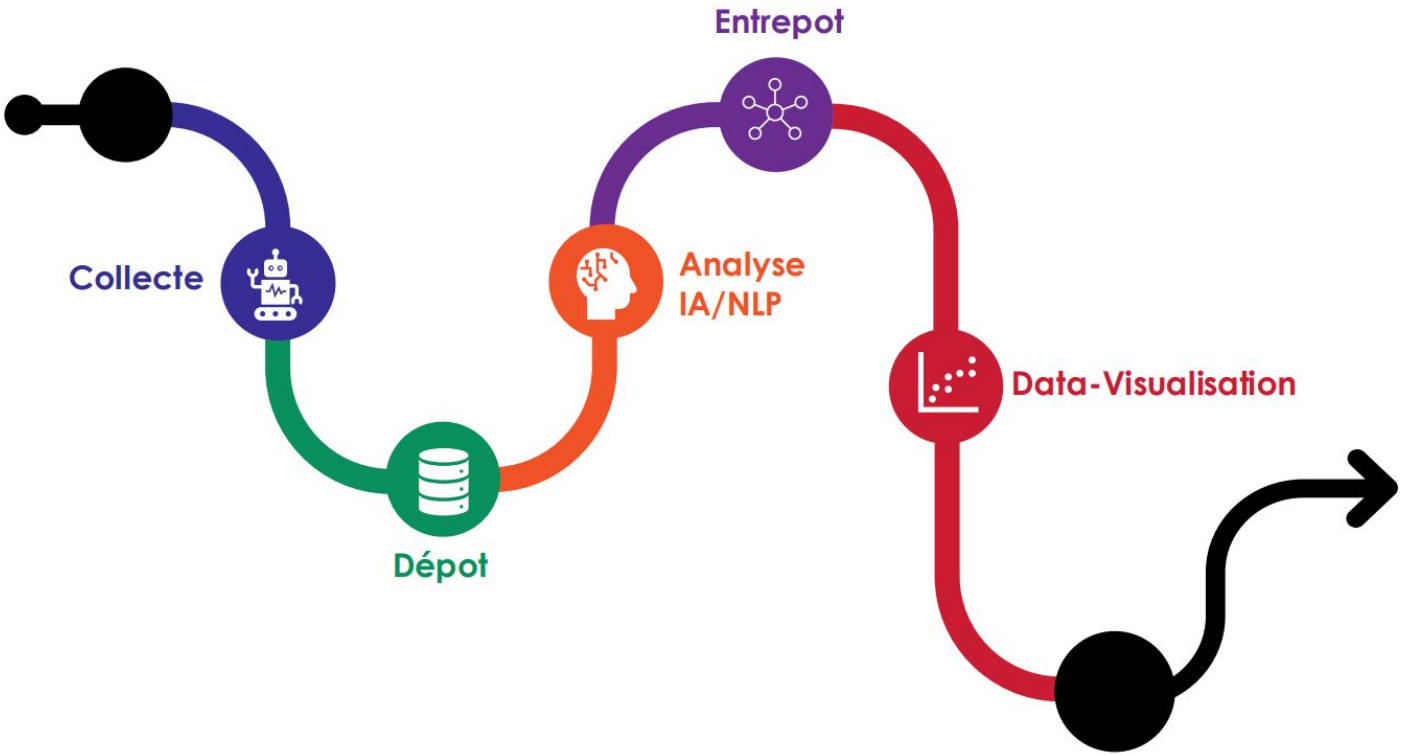
Construire un observatoire des innovations autour des métiers de la gestion

Créer un outils de veille technologique, le Google News de Berger-Levrault

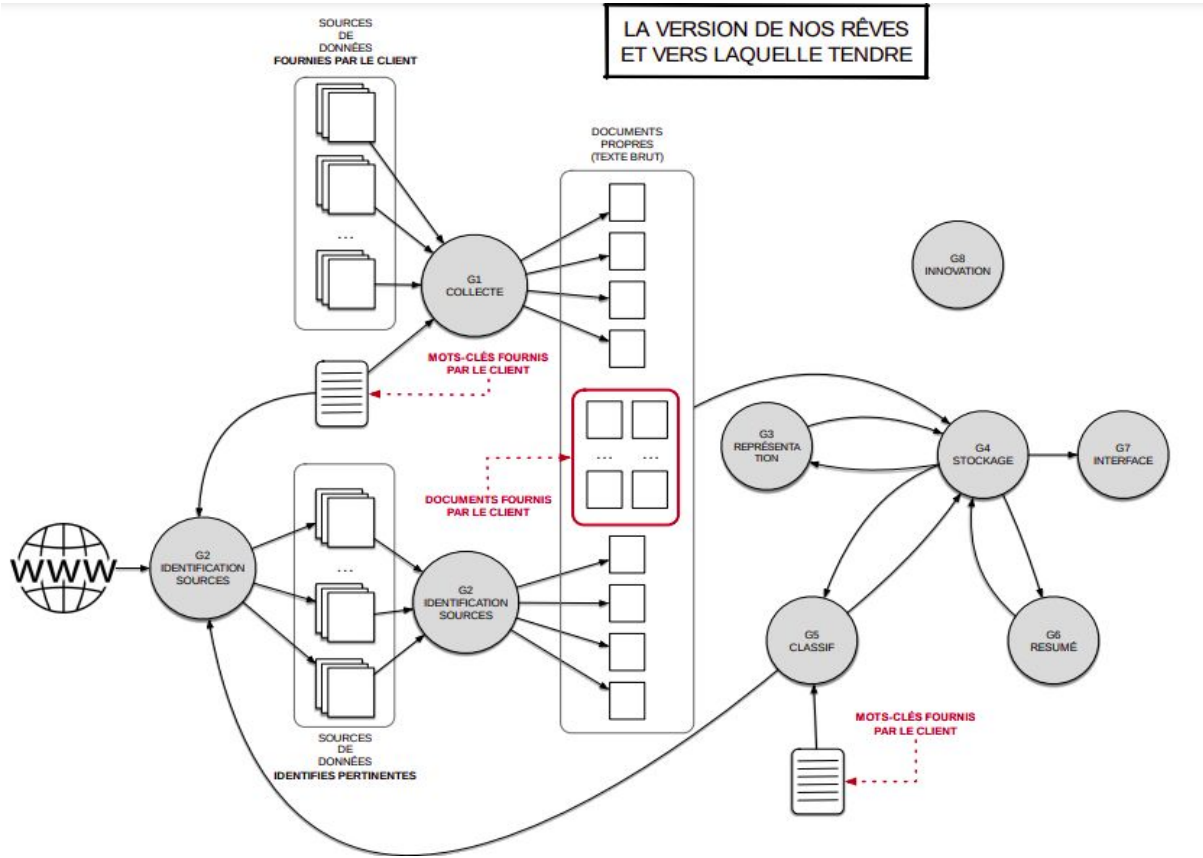
À partir de :

- Récolte de données textuelles
- Apprentissage divers sur ces données
- Visualisation de résumés et de statistiques sur le corpus

Principe de fonctionnement du projet

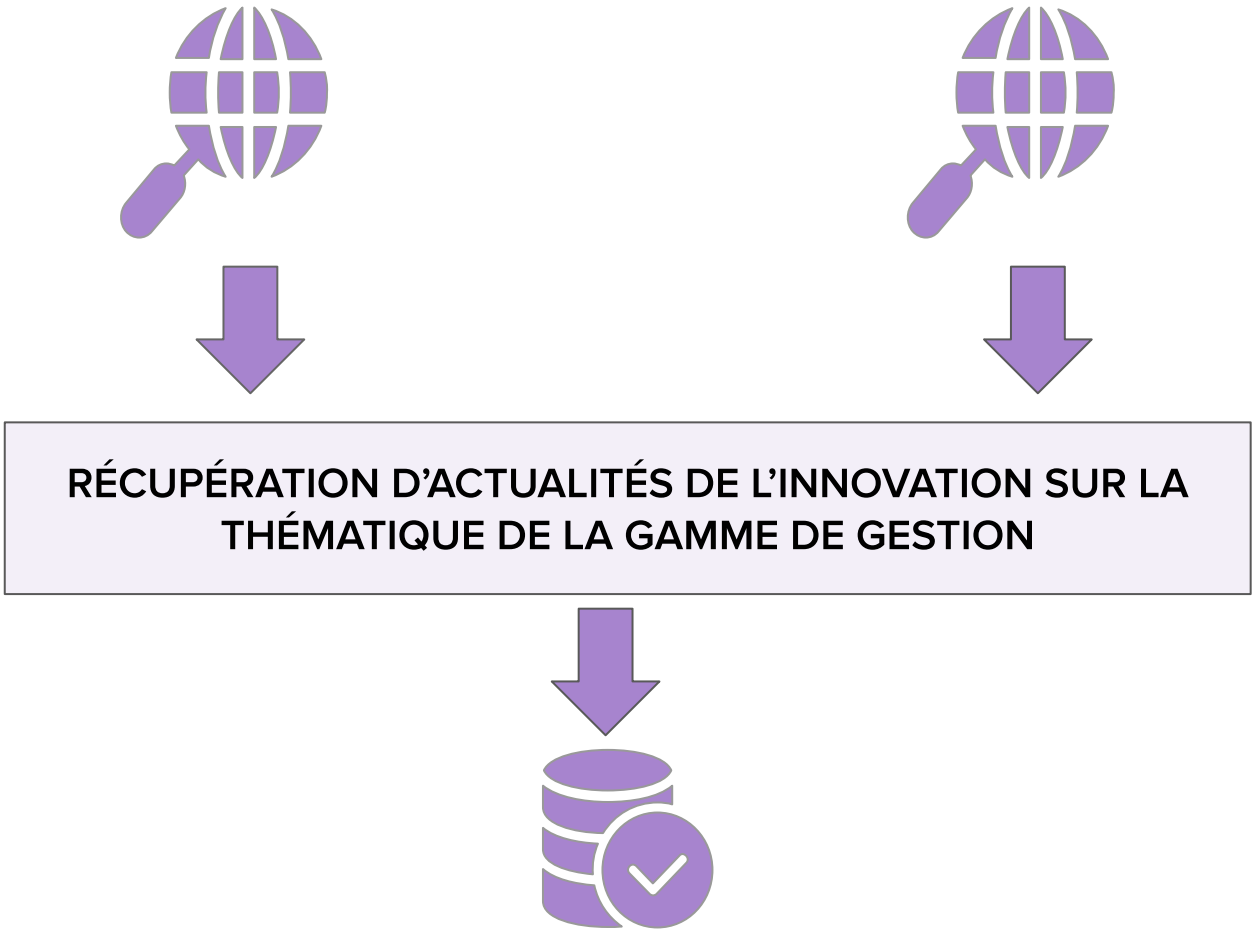


Architecture



Mais d’où proviennent les données ?

LE SCRAPING : RÉCOLTE DE DONNÉES SUR LE WEB



LE SCRAPING : LES INFORMATIONS RETENUES

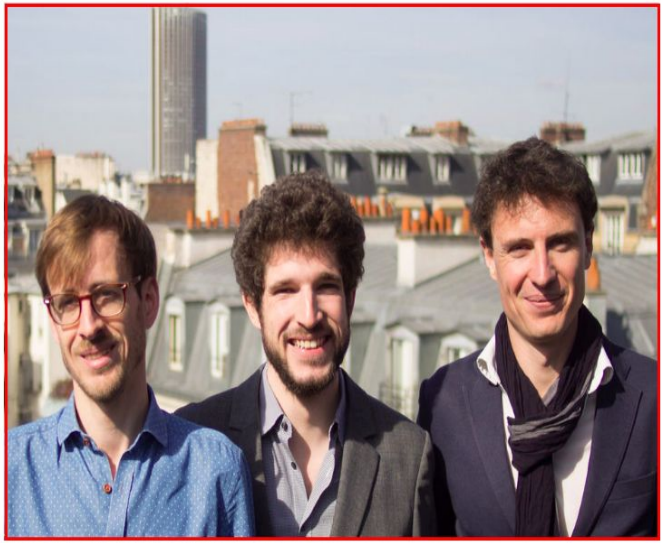
ETS Global rachète la start-up française Pipplet et sa solution d'évaluation des langues

Alors que l'évaluation linguistique dans le cadre des recrutements est en croissance, Pipplet, start-up parisienne créée en 2015, a séduit la filiale européenne d'ETS Global. Une opération qui va notamment lui faciliter l'accès au marché international.

AUDE CHARDENON | PUBLIÉ LE 16 SEPTEMBRE 2020 À 11H20

DIGITAL RH, FORMATION, COMMUNICATION

TWITTER FACEBOOK LINKEDIN FLIPBOARD EMAIL



De gauche à droite : Matthieu Herman, Adrien Wartel et Baptiste Derongs. © Pipplet

A LIRE AUSSI

Lokalise lève 6 millions de dollars pour sa solution de traduction automatisée[...]

27 000 articles de Wikipédia ont été rédigés dans une langue... qui n'existe pas

La filiale européenne d'ETS Global a officialisé mardi 15 septembre l'acquisition de la start-up parisienne Pipplet. Le montant du rachat de la société française, âgée de 5 ans, n'est pas communiqué.

Ce rapprochement permet à l'entreprise globale, basée à Princeton et créatrice des tests TOEIC et TOEFL, de compléter son portefeuille de produits à destination des professionnels et se renforcer dans l'évaluation en ligne des compétences. Elle a surtout pour objectif de proposer une offre répondant aux besoins des entreprises, des institutions académiques et des particuliers, "*partout dans le monde avec une gamme d'évaluations en langues disponible en distanciel et en présentiel*", expliquent les partenaires.

UNE ÉVALUATION SOUS FORME DE MISE EN SITUATION

Créée en 2015 par Baptiste Derongs, Matthieu Herman et Adrien Wartel, Pipplet a lancé une solution permettant à des recruteurs de valider la capacité de leurs candidats à travailler dans une langue étrangère. "*Nous sommes partis du constat que les compétences exigées dans une entreprise sont très différentes de que l'on apprend à l'école*", expose Matthieu Herman, CEO de la start-up.

Dans la pratique, Pipplet est une plateforme de mise en situation. Elle propose des tests en 30 minutes pour évaluer les compétences des candidats en langues étrangères via des sessions enregistrées. Celles-ci sont ensuite envoyées à des examinateurs, puis évaluées selon la grille CECRL, le cadre européen commun de référence pour les langues. Les entreprises utilisent ensuite ces tests pour le recrutement, la mobilité interne, les audits linguistiques, et la certification des compétences.

ACCÉLÉRER LES PROCESSUS DE RECRUTEMENT

Pipplet propose des tests standardisés pour le secteur académique, des tests d'évaluation orale pour des secteurs spécifiques et une expérience digitalisée pour les candidats. Elle s'intègre dans les logiciels RH du marché via une distribution entièrement en ligne et propose également sa solution aux organismes de formation. La start-up basée à Paris se rémunère sous forme de crédit de tests, dont le prix dépend du volume et de la formule de test souhaités. "*Certains tests sont spécifiques en fonction de l'industrie et des métiers*", poursuit l'entrepreneur.

La plateforme vise à accélérer les processus de recrutement et cible autant les start-up en phase de croissance à l'international que les grands groupes de conseil. "*Pipplet s'est établie comme un standard dans les processus de recrutement, notamment dans certains secteurs où le multilinguisme et l'oral sont importants, comme le service client*", poursuit-il.

DES CLIENTS COMME RENAULT, SHOPIFY, ELECTRONIC ARTS OU ENCORE ACCENTURE

Cette opération va aider Pipplet à poursuivre son développement indépendant avec des moyens financiers renforcés et lui donner un accès au réseau de partenaires d'ETS Global dans 80 pays à travers l'Europe, l'Asie, le Moyen Orient, l'Afrique et l'Amérique du Sud. La solution sera aussi intégrée au catalogue d'ETS Global en complément de son offre existante.

Pipplet compte parmi ses clients Renault, Shopify, Electronic Arts (EA) ou encore Accenture. Elle précise que "*plus de 30 000 candidats ont ainsi été testés dans 28 langues, de l'anglais au vietnamien*". Elle emploie dix collaborateurs et travaille avec une centaine d'organisations.

AUDE CHARDENON
@CharderonA

Mais aussi :

- Nom de la source
- URL de la source
- URL de l'article
- Le contenu de l'article avec balises HTML
- Langue de publication de l'article

LE SCRAPING : DONNÉES FOURNIES

50 SITES INTERNET

- Sites gouvernementaux
- Sites d'entreprises
- Sites de médias ...

Sites sous abonnement

Pas d'articles publiés en 2020

ÉPURATION

Scraping interdit

Sites non mis à jour

31 SOURCES EXPLOITÉES

Les Echos

AMF
ASSOCIATION DES MAIRES DE FRANCE ET DES PRÉSIDENTS D'INTERCOMMUNALITÉ

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Vie publique
Au cœur du débat public

LA TRIBUNE

sopra hr
SOFTWARE

LE SCRAPING : DONNÉES FOURNIES

50 SITES INTERNET

- Sites gouvernementaux
- Sites d'entreprises
- Sites de médias ...

Sites sous abonnement

Pas d'articles publiés en 2020

ÉPURATION

Scraping interdit

Sites non mis à jour

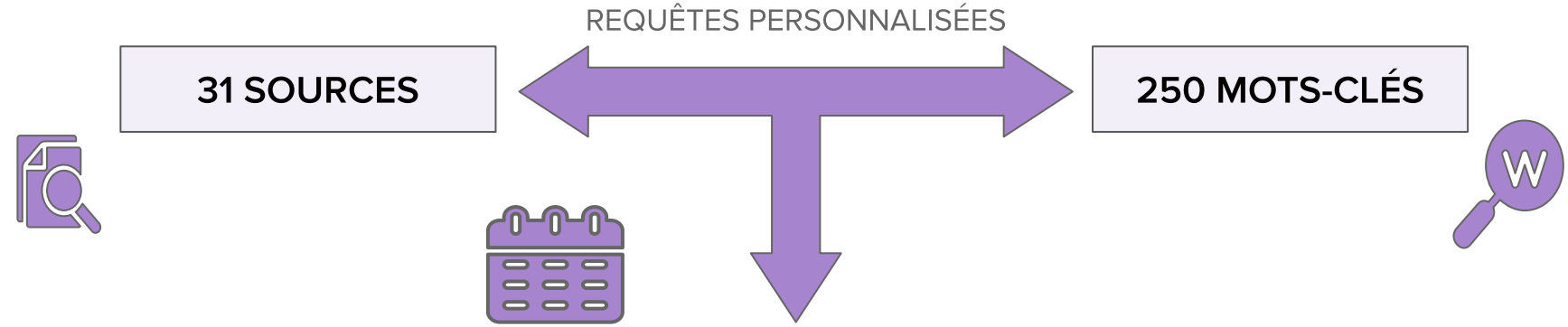
31 SOURCES EXPLOITÉES

LISTES DE 250 MOTS-CLÉS

LEXIQUE GESTION

LEXIQUE INNOVATION

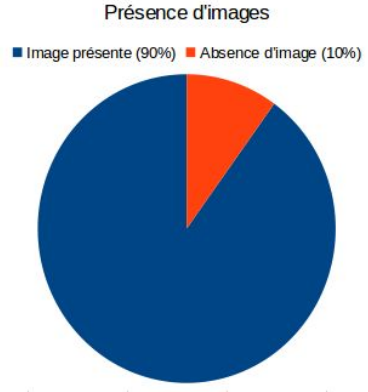
LE SCRAPING : LE RÉSULTAT

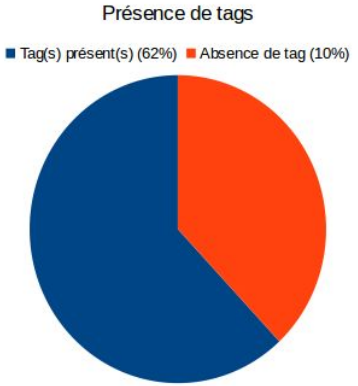


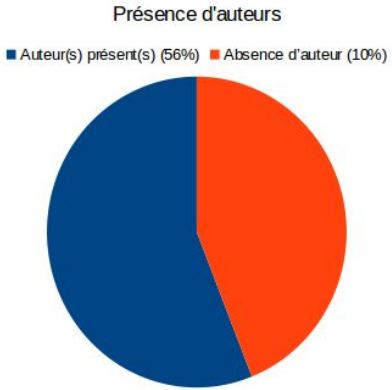
1864 ARTICLES RÉCUPÉRÉS

et insérés en BD

LE SCRAPING : STATISTIQUES BRÈVES







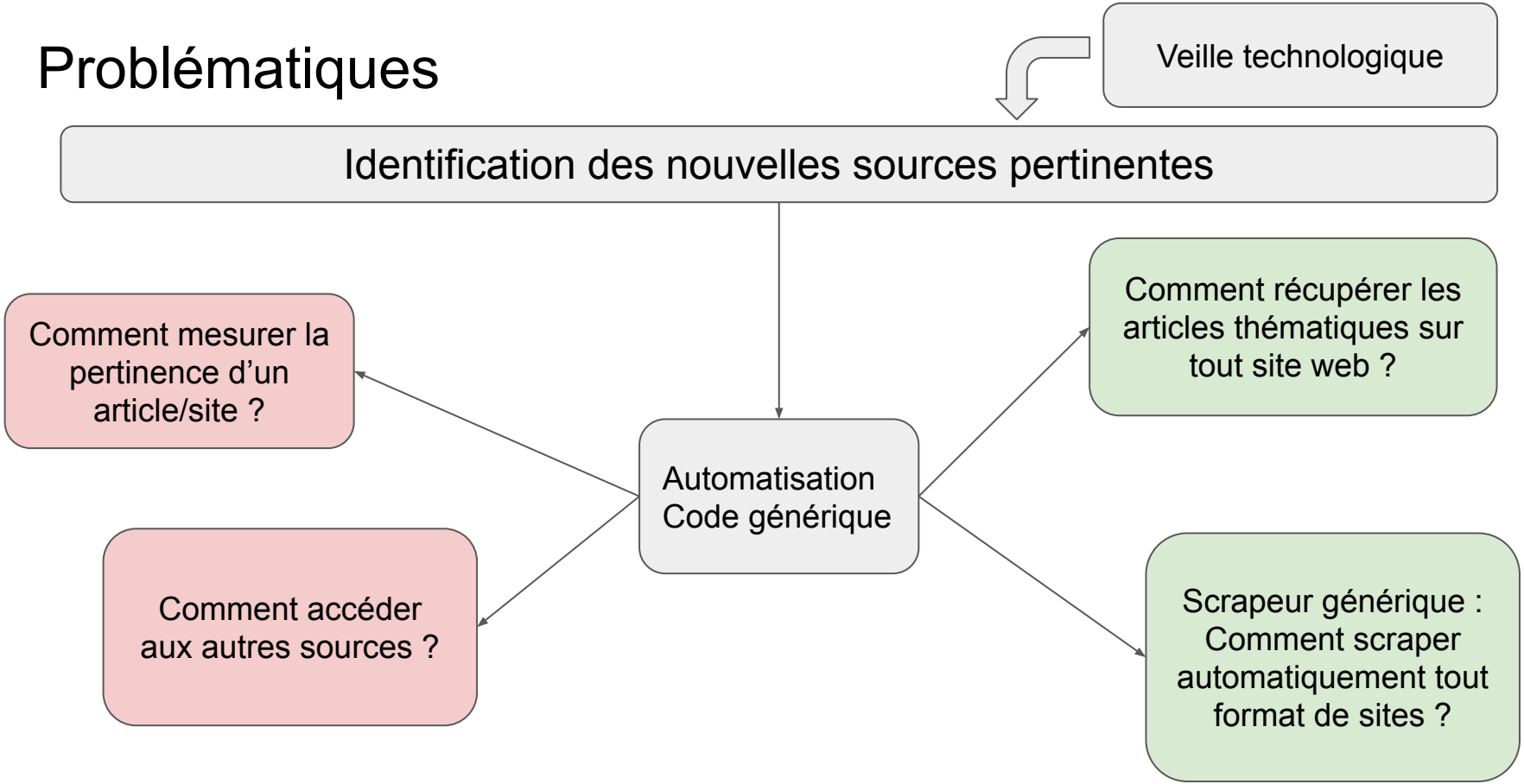
Nuage de mots sur les tags des articles

LE SCRAPING : CONCLUSION

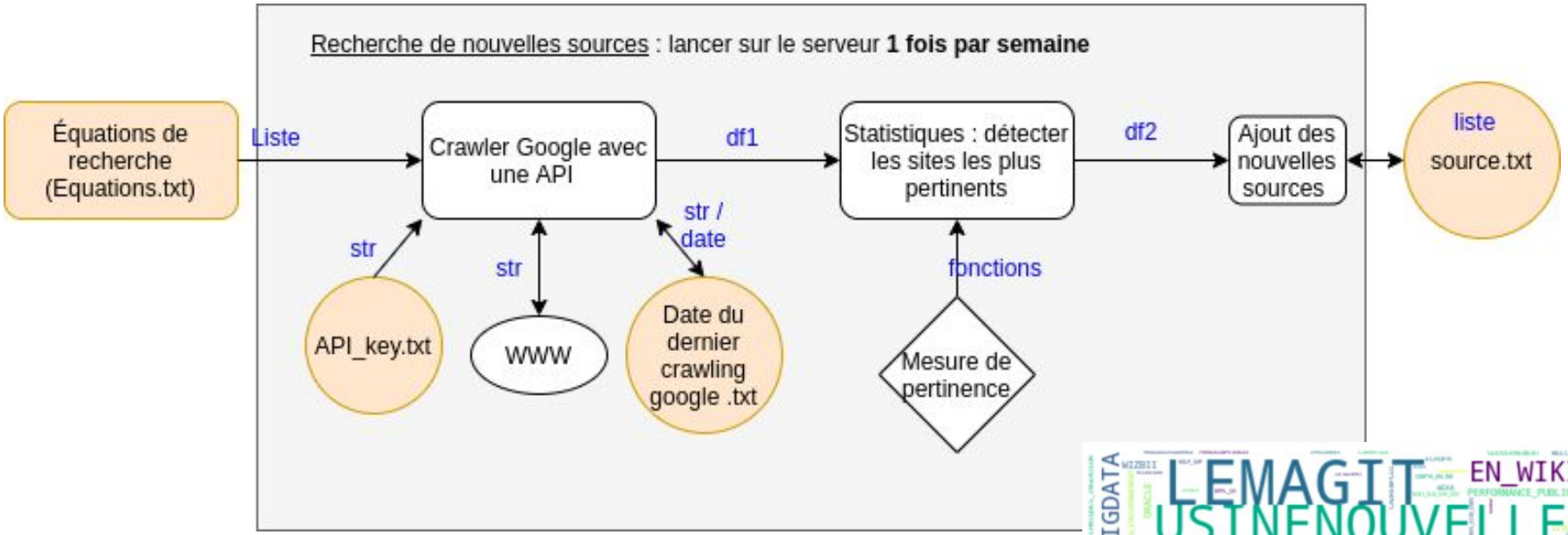
- ➢ Une majorité des sources scrapées
- ➢ Fonction de scraping commune à toutes les sources
- ➢ Facilité de modification de l'algorithme

- ➢ Choix des mots-clés à améliorer → G8 : Innovation
- ➢ + de sources à exploiter → G2 : Scraping

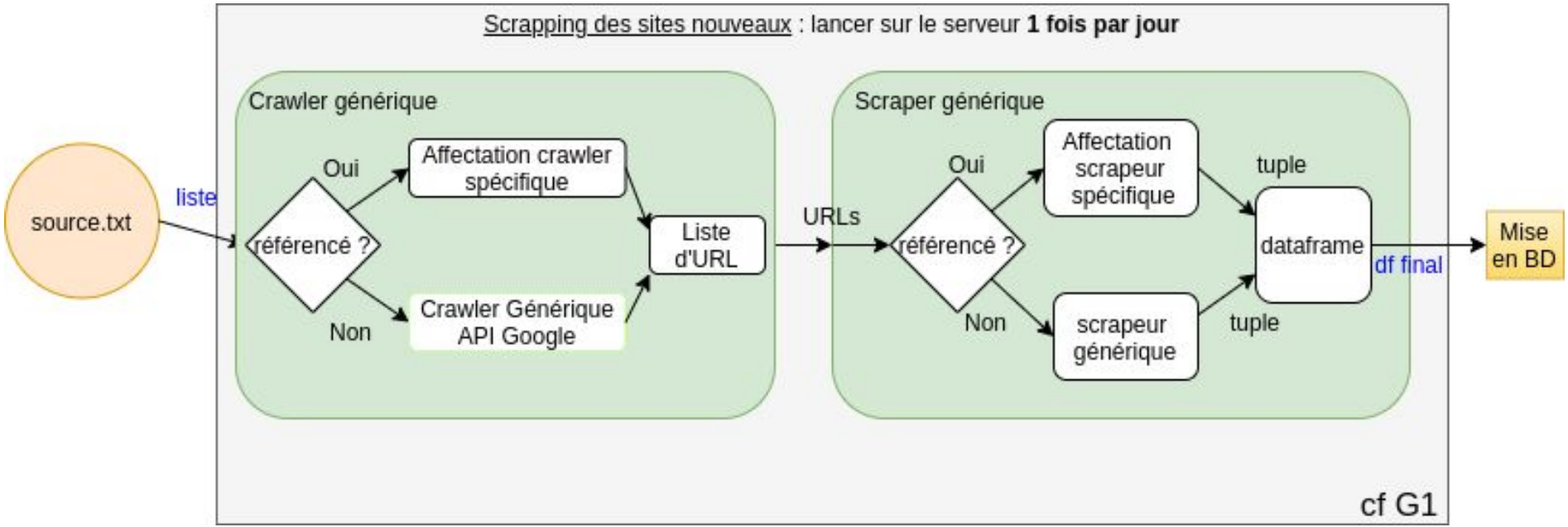
Problématiques



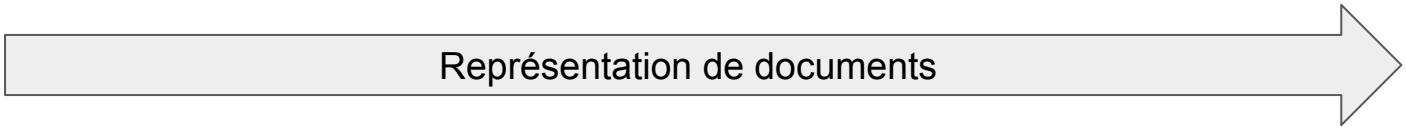
Identification des nouvelles sources pertinentes



Scrapping des sources pertinentes



Ce qui été prévu	Ce qui a été fait
	Essayer de définir les tendances des mots clefs
Utiliser une API Google pour identifier de nouvelles sources	Utiliser une API Google pour identifier de nouvelles sources
Faire des mesures de pertinences	Définir des mesures de pertinences des sites
	Initier un crawler générique pour chercher des articles thématiques sur un site particulier + crawler spécifique
Initier la construction d'un scrapeur générique	Initier la construction d'un scrapeur générique + scrapeur spécifique



G3 - Représentation de données

1. Récupération des données dans la BD

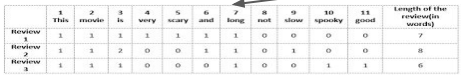
Non réalisé

2. Élimination des doublons et nettoyage des données

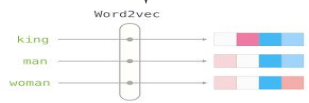
réalisé

réalisé

3. Représenter les documents



	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6



BOW, TF IDF

Word2Vec, Doc2Vec, FastText, GloVe

Pos Tagging, NER

4. Insertion des représentations dans la BD

Non réalisé

Travail réalisé pour les représentations

1. Nettoyage des données en fonction des représentations et des demandes des Groupes 5/6/7

2. Entraînement des différents modèles à partir de nos données / Recherche de modèles pré-entraînés sur du français

3. Réalisation des représentations

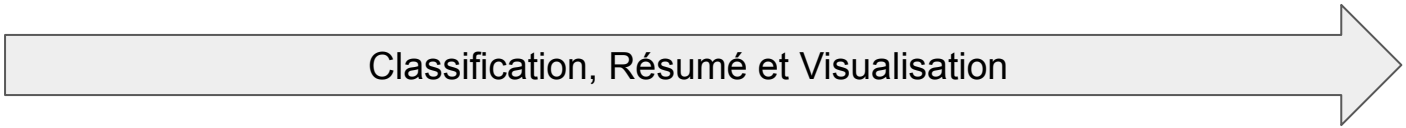
- TF IDF / BOW 1,2,3-Gram sur le contenu des articles et les titres
- Word2Vec / FastText / GloVe pondérés ou non avec l'IDF
- Doc2Vec entraîné sur notre corpus
- POS Tagging / NER à partir de différents modèles

Synthèse du travail

- La majeure partie du travail demandé a été réalisé dans le temps imparti
- Équipe efficace et dynamique
- Bonne entente/ambiance entre les membres du groupe

Piste d'amélioration :

- Réaliser une traduction français → anglais sur notre corpus pour utiliser les librairies en anglais sur Python



G5 - Classification

3 Objectifs :

- Détecter les documents liés à la thématique d'Innovation (et la thématique Gamme de Gestion)
- Détecter les documents avec un nouveau contenu
- Classifier les documents par thèmes

Détection de documents liés à la thématique d'innovation

Méthode en deux étapes :

- Semi-supervisé
- Supervisé

Sur **7490** articles :

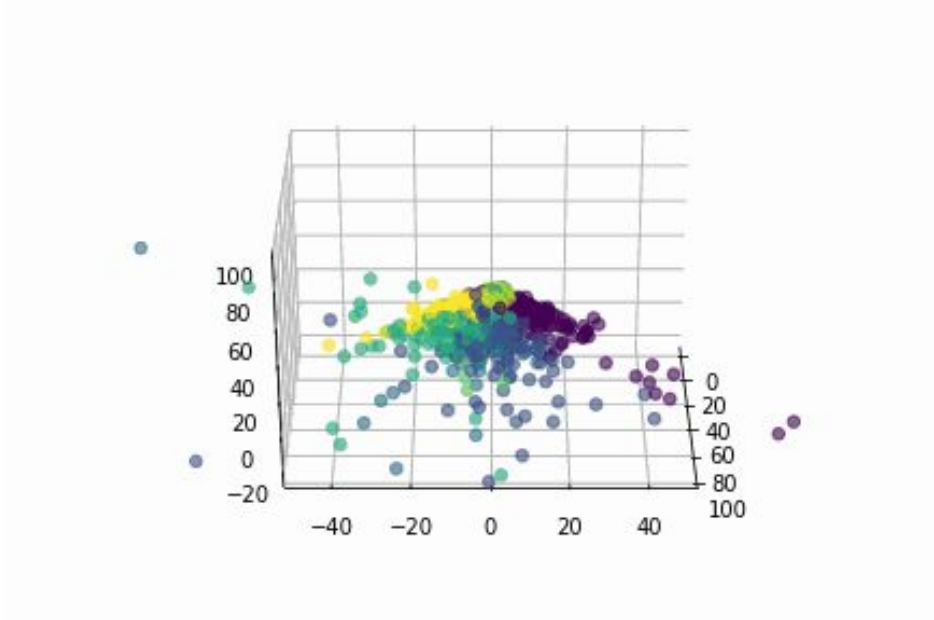
+ de **2800** liés à la thématique d'**innovation**
+ de **4500** liés à la thématique **gamme de gestion**
+ de **1700** liés aux **2 thématiques**

Classification des documents par thèmes

Utilisation d’un algorithme de clustering

Sur **1742** articles :

- **Entre 100 et 500 articles** par thèmes
- Réparties en **9 thèmes**



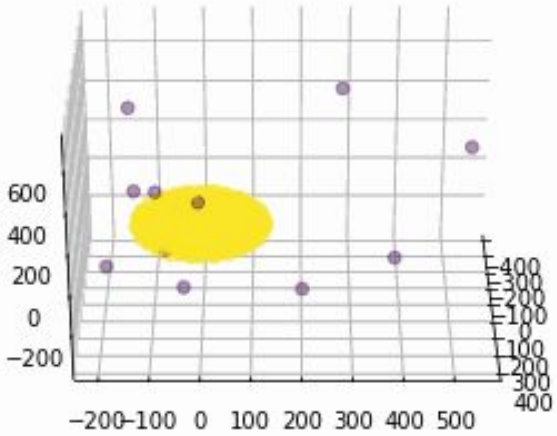
Détection de documents avec un nouveau contenu

Objectif détecter les documents sémantiquement anormaux :

- Algorithme de détection d'anomalies

Sur **1742** articles :

- **16** classés comme **nouveaux**



G6 - Résumé

2 Objectifs initiaux:

- Résumé : application de méthodes de résumé automatique sur un seul document

- Stories : construction d'un résumé multi-documents pour chaque story.
  - représenter tous les aspects de la story
  - minimiser la redondance

Résumé

Extractif

- Lead_3
- TextRank
- SummaRuNNer
- BertSumExt

Abstractif

- Seq2Seq
- Pointer-Generator
- Bottom-Up
- BertSumAbs
- BART

Résumé

Méthodes approfondies

Méthodes testées

Méthodes abandonnées

Extractif

- **Lead_3 (v0)**
- **TextRank (v1 & v2)**
- SummaRuNNer
- BertSumExt

Abstractif

- Seq2Seq
- Pointer-Generator
- Bottom-Up
- BertSumAbs
- **BART (v2)**

Résumé

Extractif (TextRank v2)	Titre	Abstractif (Bart v2)
Dans une première partie, le rapport de la Cour des comptes traite de l'évolution des risques professionnels et des limites de leur connaissance et de leur prise en compte dans les systèmes de protection sociale et d'indemnisation des accidents du travail et des maladies professionnelles (maladie, maternité, accidents du travail, maladie professionnelle).	'La gestion du risque accidents du travail et maladies professionnelles'	Depuis quelques années d'ailleurs, certains professionnels RH s'efforcent de mesurer chaque année le taux d'engagement au sein de leur entreprise : un calcul généralement fondé sur la combinaison des résultats d'une enquête de satisfaction interne et d'indicateurs sociaux (taux d'absentéisme, taux de rétention) et de performance (chiffre d'affaires, etc.) 'Si les actions pour favoriser l'engagement en entreprise relèvent forcément d'une approche en partie individualisée, quelques bonnes pratiques s'imposent.
Cette délégation lui confie l'exécution d'une mission de service public, lui confère un monopole dans l'organisation et la régulation des compétitions officielles, ainsi que des prérogatives de puissance publique, qui lui permettent, notamment, de constituer et gérer les équipes nationales de son sport.	'Sylvain Grisot : « Plus qu'un changement des règles d'urbanisme, c'est un changement de système qui est nécessaire »'	« Chaque aménagement correspond à un usage, on fait une ville qui est finalement touchée par une forme d'obsolescence programmée » Pour avoir travaillé comme maître d'ouvrage sur la structure publique, j'ai constaté que les projets étaient généralement mal pensés. Ce travail doit aussi être réalisé sur le périurbain économique sur lequel

Résumé

Extractif (TextRank v2)

deuxième partie ?

Dans une première partie, le rapport de la Cour des comptes traite de l'évolution des risques professionnels et des limites de leur connaissance et de leur prise en compte dans les systèmes de protection sociale et d'indemnisation des accidents du travail et des maladies professionnelles (maladie, maternité, accidents du travail, maladie professionnelle).

Titre

'La gestion du risque accidents du travail et maladies professionnelles'

'Sylvain Grisot : « Plus qu'un changement des règles d'urbanisme, c'est un changement de système qui est nécessaire »'

Quelle est-elle ?

Cette délégation lui confie l'exécution d'une mission de service public, lui confère un monopole dans l'organisation et la régulation des compétitions officielles, ainsi que des prérogatives de puissance publique, qui lui permettent, notamment, de constituer et gérer les équipes nationales de son sport.

Abstractif (Bart v2)

d'ailleurs ?

Depuis quelques années d'ailleurs, certains professionnels RH s'efforcent de mesurer chaque année le taux d'engagement au sein de leur entreprise : un calcul généralement fondé sur la combinaison des résultats d'une enquête de satisfaction interne et d'indicateurs sociaux (taux d'absentéisme, taux de rétention) et de performance (chiffre d'affaires, etc.) 'Si les actions pour favoriser l'engagement en entreprise relèvent forcément d'une approche en partie individualisée, quelques bonnes pratiques s'imposent.

« Chaque aménagement correspond à un usage, on fait une ville qui est finalement touchée par une forme d'obsolescence programmée » Pour avoir travaillé comme maître d'ouvrage sur la structure publique, j'ai constaté que les projets étaient généralement mal pensés. Ce travail doit aussi être réalisé sur le périurbain économique sur lequel

Arrêt en milieu de phrase

Résumé

Extractif (TextRank v2)

\+

- Résultats cohérents dans l'ensemble
- Simplicité de mise en place

\-

- termes anaphoriques
- longueur "aléatoire"

Abstractif (Bart v2)

\+

- paraphrase le texte initial
- génère de nouvelles phrases
- inclut de nouveaux mots

\-

- résumé de taille 50 un peu courts pour exploiter tout le potentiel

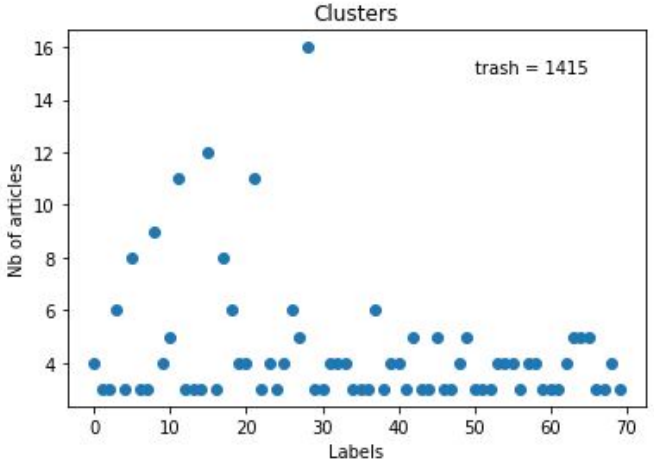
Stories (Agrégat de documents)

- Besoin initial :
  - résumé multi-document sur les thèmes du groupe 5
- Constat:
  - thèmes trop génériques et trop chargés pour notre problème
- Solution:
  - génération de clusters dynamiques denses sur les documents “récents”
  - regroupement par titres (à la google news)

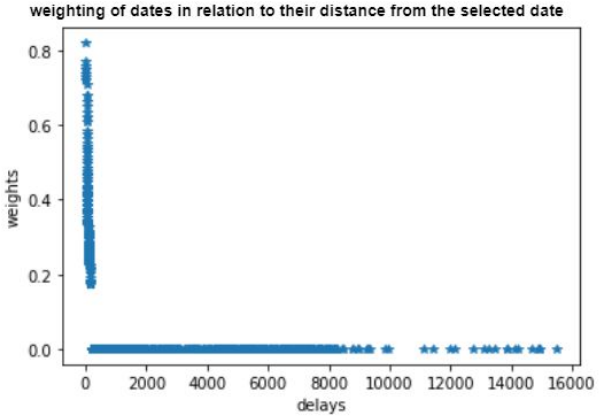
Stories (Agrégat de documents)

Méthode de clustering utilisée:

- DBSCAN, pour avoir des résultats très denses



Repondération des clusters en fonction de la date de publication des articles



L'eau, enjeu majeur de la ville de demain
Source : Ville intelligente Mag
2020-10-28T00:00:00

Urban Chronicles : un media collaboratif pour donner les clés de la ville de demain
Source : Ville intelligente Mag
2020-10-20T00:00:00

3 questions à ... Juliette Auricoste, Directrice du programme Petites Villes de demain
Source : Association Petites Villes France
2020-10-15T00:00:00

Lancement du programme Petites villes de demain
Source : Ministère Cohésion Territoires
2020-10-01T00:00:00

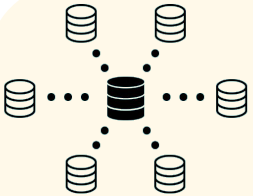
G8 - Objectifs du Groupe Innovation

Développement d’un moteur de recherche IA

Intégration des retours utilisateurs pour améliorer l’interface globale du projet

Moteur de recherche IA

User Preferences



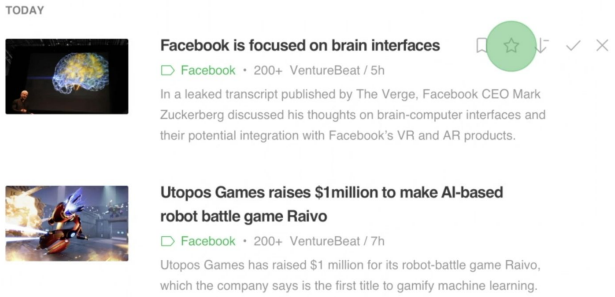
Indexation

LTR Pipeline

Ranking Model

Query

User Feedback



Intégration du feedback utilisateur

Likes & Favorites

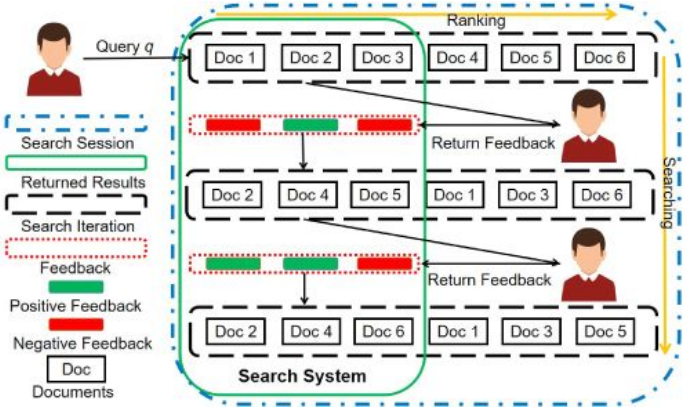
Click History

Requests

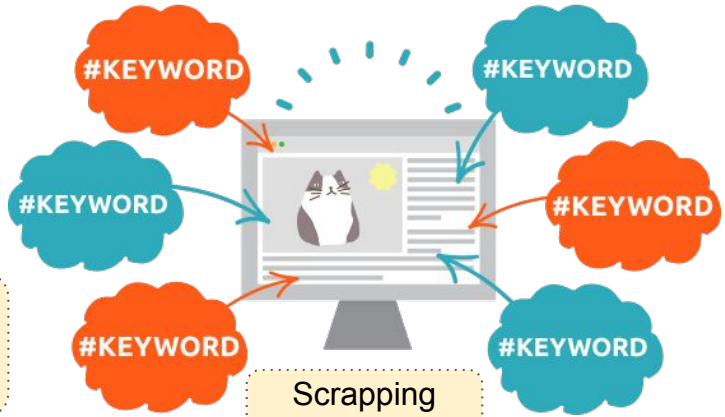
Comments

Abstractive Summarization

Stronger Text Representation



More relevant results



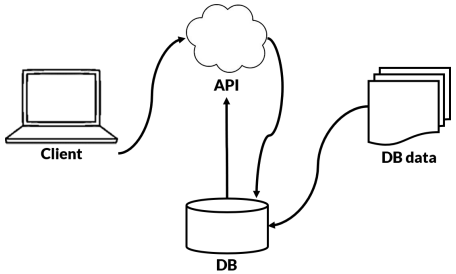
Scrapping Optimization

https://github.com/G8-Innovation/ProjetIP2021

Objectifs du groupe Stockage

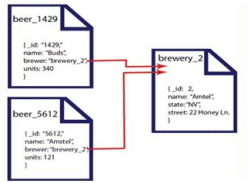
- Choisir un SGBD adapté et proposer une modélisation
- Implémenter une base de données fonctionnelle

- Version idéale : faire le lien entre les groupes → donner accès aux données, en lecture et écriture



Choix du SGBD

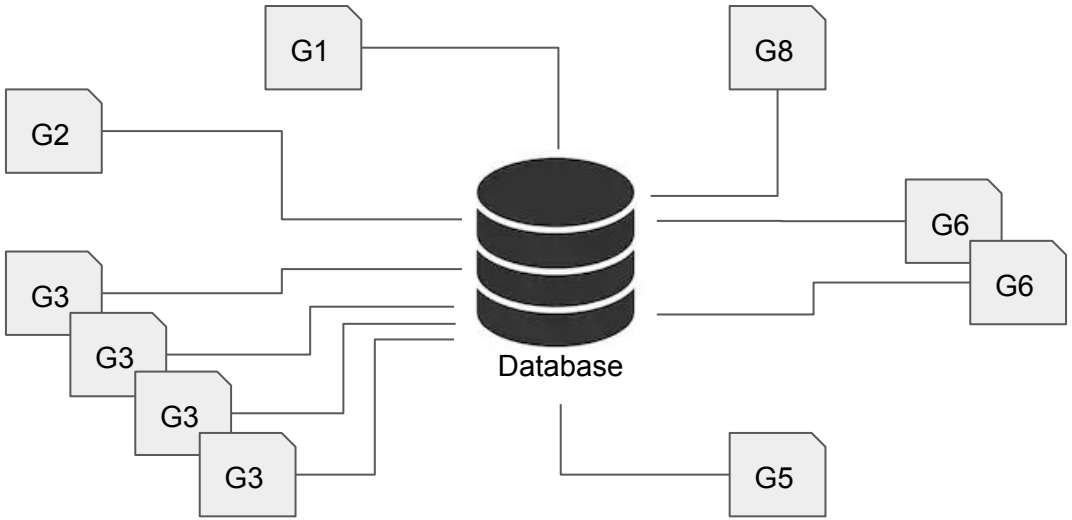
- Relationnel
- NoSQL
  - Modèle clé-valeur
  - Modèle orienté colonne
  - **Modèle orienté document**
  - Modèle orienté graphe



mongoDB

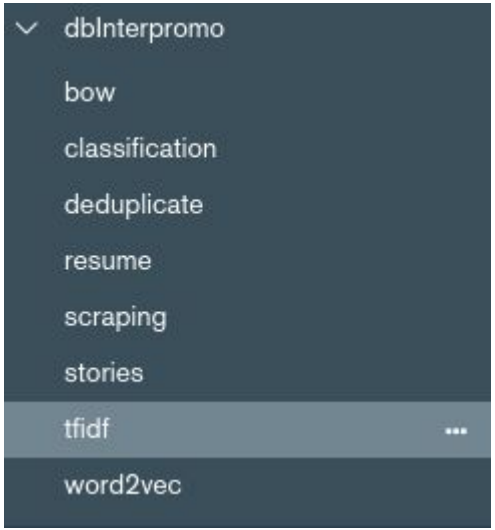
Modélisation

- Modélisation par groupe
  - Temps
  - Dépendance entre groupes

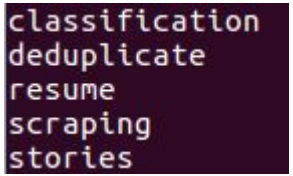


Implémentation

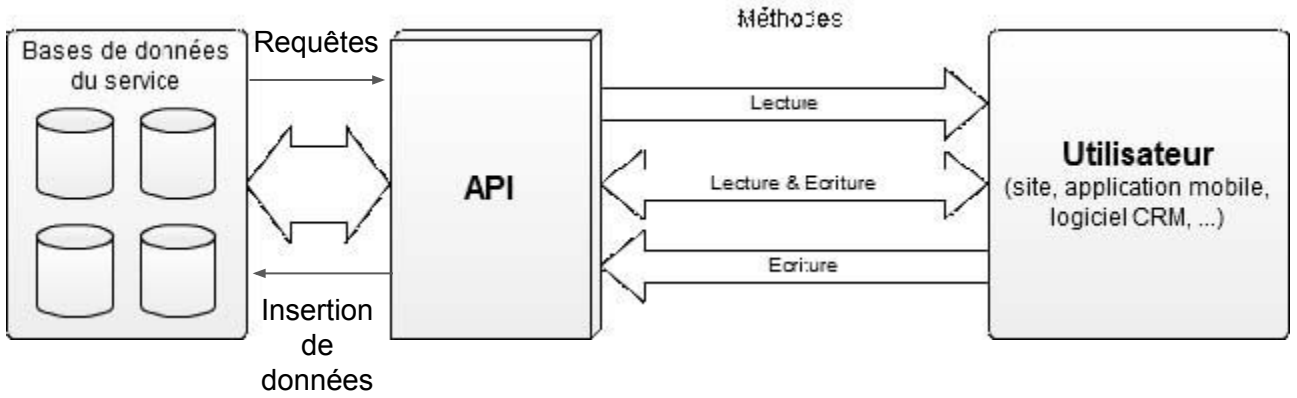
Base implémentée en local



Base implémentée sur le serveur

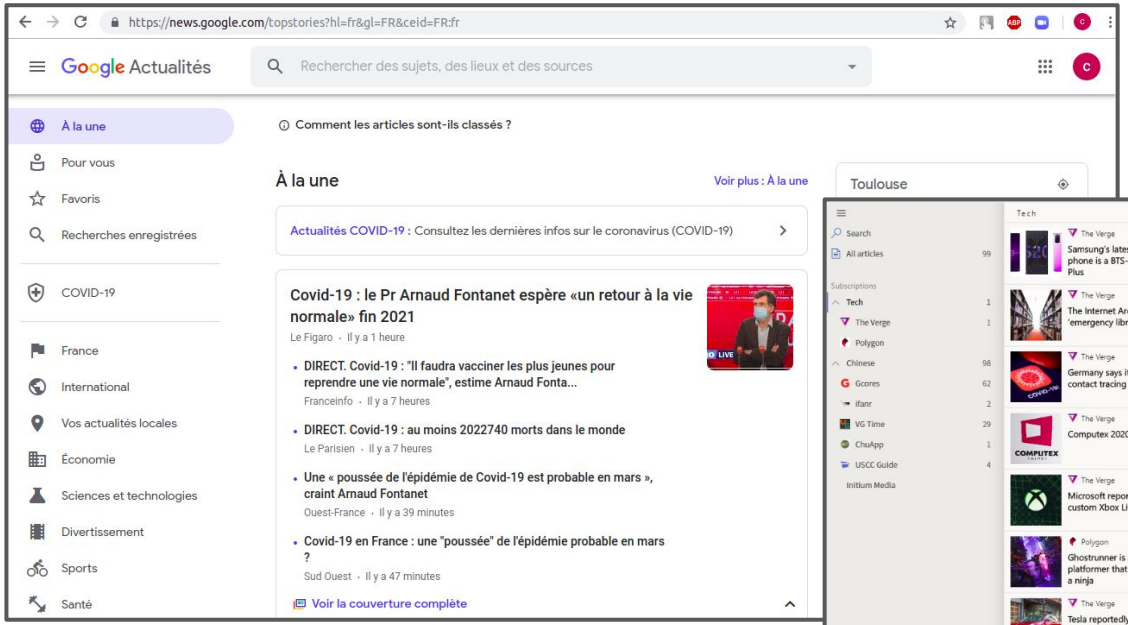


API et déploiement

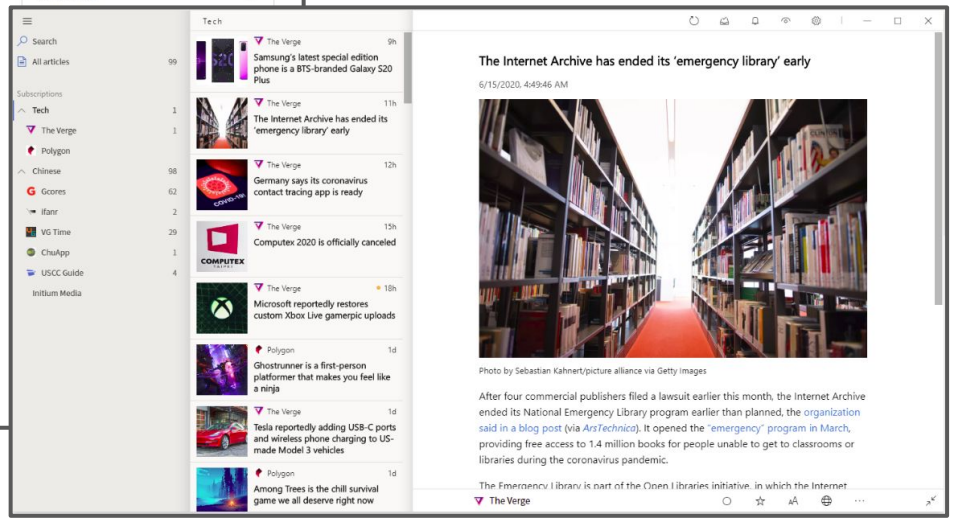


Objectifs du groupe Visualisation

Exemple de site de veille/news



Fluent reader



Organisation du travail

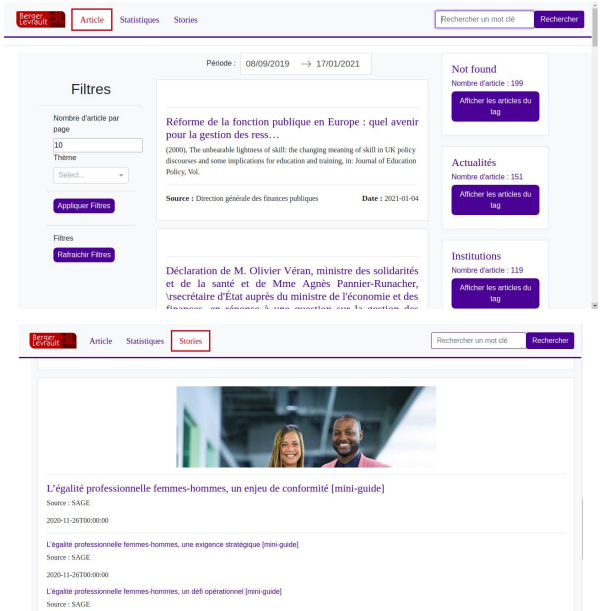
Groupe statistique

Groupe squelette

maquette

Groupe CSS

Apparence



Conclusion

Points négatifs

- Problèmes de serveur
- Biais dans les données
- Distanciel

Points positifs

- Communication intra et inter groupes
- Chaque groupe est apte à fournir des résultats
- Participation de tous, bonne implication générale

Merci pour votre attention !

Des questions ?