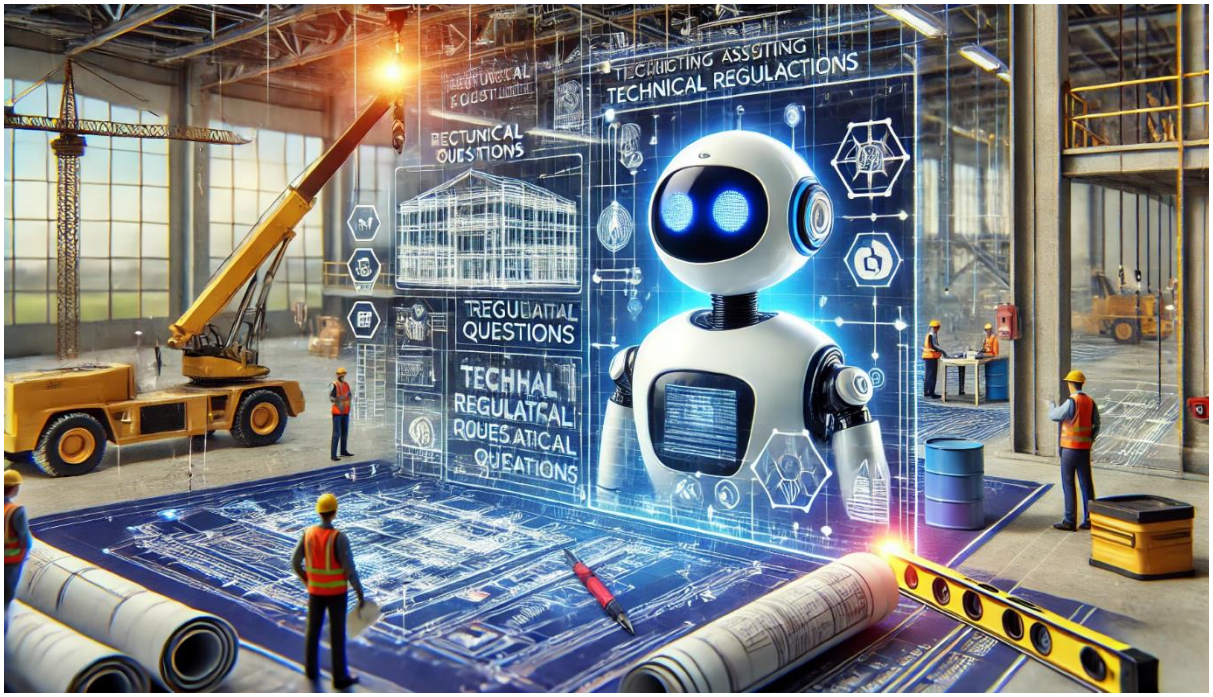
DataScientest • com

Projet NLP
Chatbot réglementation technique



Charafeddine MECHRI
Marine MERLE
David MICHEL
NLP-Sept24

Table des matières

1. Contexte et Objectifs

Dans de nombreux secteurs, la complexité des réglementations et des normes en vigueur représente un défi majeur pour les professionnels. C'est en particulier le cas dans le secteur du bâtiment pour les conducteurs de travaux sur le chantier. Ces derniers consacrent une part importante de leur temps à rechercher dans une multitude de sources les réponses précises à leurs questions. Cette situation peut donc s'avérer très chronophage.

Dans ce contexte, la mise en place d'un outil IA, capable de répondre rapidement et précisément aux questions réglementaires, s'impose comme une solution novatrice. Le projet de création d'un chatbot répondant aux questions sur la réglementation dans le bâtiment s'inscrit ainsi dans cette démarche. En automatisant la recherche d'informations pertinentes, ce projet vise à améliorer l'efficacité des conducteurs de travaux tout en leur permettant de se concentrer davantage sur leurs tâches opérationnelles. Ce projet constitue également une avancée majeure pour le secteur, en alliant technologie et expertise métier.

Un des membres du projet travaillant chez Eiffage Construction, un des majors du BTP, ce projet fil rouge Datascientest est donc une application directe de la formation reçue en entreprise.

2. Extraction et traitement des données

Les bâtiments peuvent être classés en plusieurs catégories réglementaires en fonction de leur usage. Parmi les principaux types, on distingue :

- Les **établissements recevant des travailleurs (ERT)**, qui comprennent les bâtiments où s'exercent des activités professionnelles nécessitant des dispositions relatives à la sécurité et à la santé des travailleurs.
- Les **établissements recevant du public (ERP)**, tels que les écoles, hôpitaux ou centres commerciaux, qui accueillent un public extérieur et imposent des normes spécifiques de sécurité et d'accessibilité.
- Les **habitations**, regroupant les bâtiments résidentiels, notamment les immeubles collectifs et maisons individuelles.

Chacune de ces catégories est soumise à des réglementations spécifiques :

- Les **ERT** sont principalement régies par le **Code du travail**, qui impose des normes de sécurité, d'aération, d'éclairage et de prévention des risques pour les travailleurs.
- Les **ERP** suivent des arrêtés et des **instructions techniques** détaillées, ainsi que des règles de sécurité incendie et d'accessibilité pour les personnes handicapées.

- Les **habitations** relèvent des exigences du **Code de l'habitation et de la construction**, qui couvrent des aspects comme l'isolation thermique, la salubrité et la conformité technique.

Ces réglementations se retrouvent dans différents types de documents, tels que :

- Les **textes législatifs et réglementaires** (lois, décrets, arrêtés) accessibles sur les plateformes comme **Légifrance**.
- Les **Documents Techniques Unifiés (DTU)**, qui contiennent les règles de l'art pour la construction et sont publiés par des organismes comme l'AFNOR.
- Les **normes techniques** et **guides professionnels**, disponibles sur des sites spécialisés tels que COBAZ ou AFNOR.

Pour ce projet, nous avons choisi de concentrer nos efforts sur les textes réglementaires disponibles sur Légifrance, une plateforme officielle et exhaustive permettant l'accès aux textes législatifs et réglementaires consolidés par la direction de l'information légale et administrative en France. Légifrance constitue donc une source de données libre et fiable, parfaitement adaptée pour une première phase du projet.

La collecte des données sur Légifrance a été réalisée via du web scraping fait avec Beautifulsoup.

Nous avons ensuite suivi les étapes suivantes :

a) Détermination des textes liés à la conception et à l'exécution des ERT, ERP et habitations.
   Nous n'avons en effet pas souhaité récupérer les données non nécessaires au projet.
b) Analyse de la structure des pages et identification des sections pertinentes.
   Bien que tous présents sur Legifrance, les Codes ont des structures différentes

Tableau 1 : structure des Codes (données)

Source	Arborescence
ERT	Partie > livre > titre > chapitre > section (parfois) > articles
ERP	Partie > livre > titre > section > sous-section > articles
Habitations	Règlement > livre > chapitre > section (parfois) > article

c) Analyse des codes html des pages
d) Scrapping et récupération des données sous format structurés de chacune des pages

Cette approche nous a alors permis de constituer une première base de données (base_de_donnees.xlsx). Des pistes pour élargir les sources d'information, comme l'intégration des DTU et des normes AFNOR, sont également envisagées pour les étapes futures du projet.

NOTA – Une API est également disponible sur le site du gouvernement mais nous avons préféré scrapper la page afin de mettre en application les modules vus lors de la formation.

Tableau 2 : Capture d'un extrait du Code du Travail et structure HTML de la page

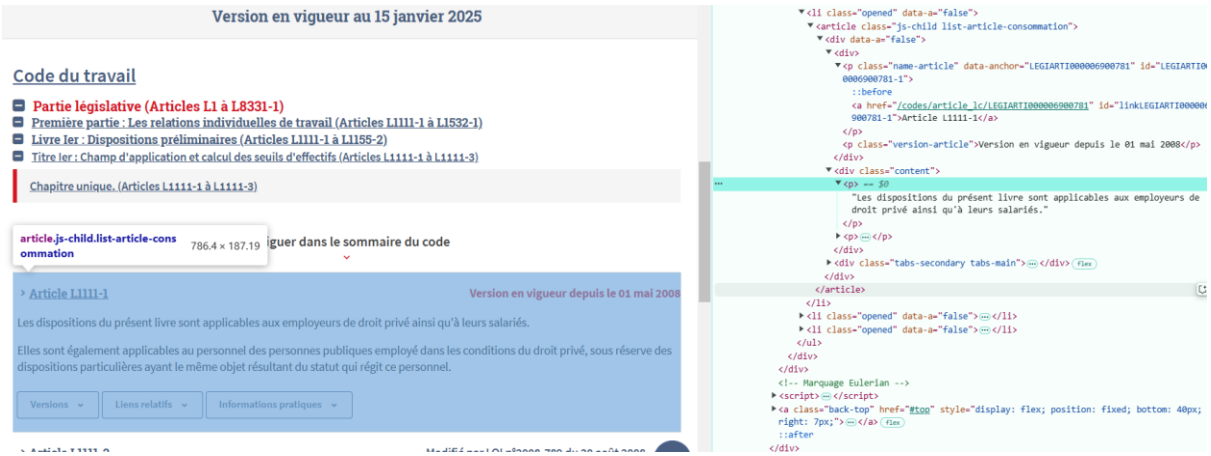


Tableau 3 : Capture d'un extrait de dataframe, structuré, des données scrappées

	Num_Titre	Titre	Num_Chapitre	Chapitre	Num_Section	Section	Numéro Article	Type décret	Mise en application	Contenu
0	Titre Ier	Obligations du maître d'ouvrage pour la concep...	Chapitre Ier	Principes généraux	Section 1	Champ d'application et définitions	R4211-1	Création	Décret n°2008-244 du 7 mars 2008 - art. (V)	Les dispositions du présent titre déterminent,...
1	Titre Ier	Obligations du maître d'ouvrage pour la concep...	Chapitre Ier	Principes généraux	Section 1	Champ d'application et définitions	R4211-2	Création	Décret n°2008-244 du 7 mars 2008 - art. (V)	Pour l'application du présent titre, on entend...
2	Titre Ier	Obligations du maître d'ouvrage pour la concep...	Chapitre Ier	Principes généraux	Section 2	Dossier de maintenance	R4211-3	Modifié par	Décret n°2011-1461 Décret n°2010-1017\r du 30...	Le maître d'ouvrage élabore et transmet aux ut...
3	Titre Ier	Obligations du maître d'ouvrage pour la concep...	Chapitre Ier	Principes généraux	Section 2	Dossier de maintenance	R4211-4	Création	Décret n°2008-244 du 7 mars 2008 - art. (V)	Le dossier de maintenance des lieux de travail...
4	Titre Ier	Obligations du maître d'ouvrage pour la concep...	Chapitre Ier	Principes généraux	Section 2	Dossier de maintenance	R4211-5	Création	Décret n°2008-244 du 7 mars 2008 - art. (V)	Le dossier de maintenance des lieux de travail...

3. Visualisation des données

La visualisation des données est une étape cruciale dans tout projet de NLP. Elle permet de mieux comprendre la structure et les caractéristiques des données, facilitant ainsi la prise de décisions sur les traitements à appliquer et les modèles à déployer. Une bonne compréhension des données est essentielle pour identifier les éventuels biais, les lacunes ou les structures spécifiques qui pourraient influencer les résultats du projet. Dans ce cadre, nous avons réalisé plusieurs visualisations pour explorer et analyser les données recueillies.

Les visualisations faites sont les suivantes :

Nombre de documents issus de chacune des sources

Cette visualisation a permis de mesurer la contribution relative des différentes sources (ERT, ERP, habitations, autres catégories) à notre corpus. Cela nous a donné une vue d'ensemble sur la répartition des documents et nous a permis d'identifier les éventuelles disparités.

Tableau 4 : Analyse du nombre de documents par source de données

```
Nombre de documents par source :
 Code de la construction et de l'habitation                                          2628
Règlement de sécurité contre les risques d'incendie et de panique. Etablissements recevant du public    1161
Etablissement recevant des travailleurs                                              299
Réglementation sécurité et incendie. Etablissement recevant du public                 44
Name: Source, dtype: int64
```

Longueur des phrases de chacun des contenus (articles)

L'analyse des longueurs de phrases nous a aidé à comprendre la complexité linguistique des textes. Les réglementations se caractérisent souvent par des phrases longues et complexes, ce qui peut poser des défis pour les modèles de NLP. Cette visualisation nous a permis d'adapter nos stratégies de traitement, par exemple en décidant de segmenter les phrases particulièrement longues pour une meilleure compréhension.

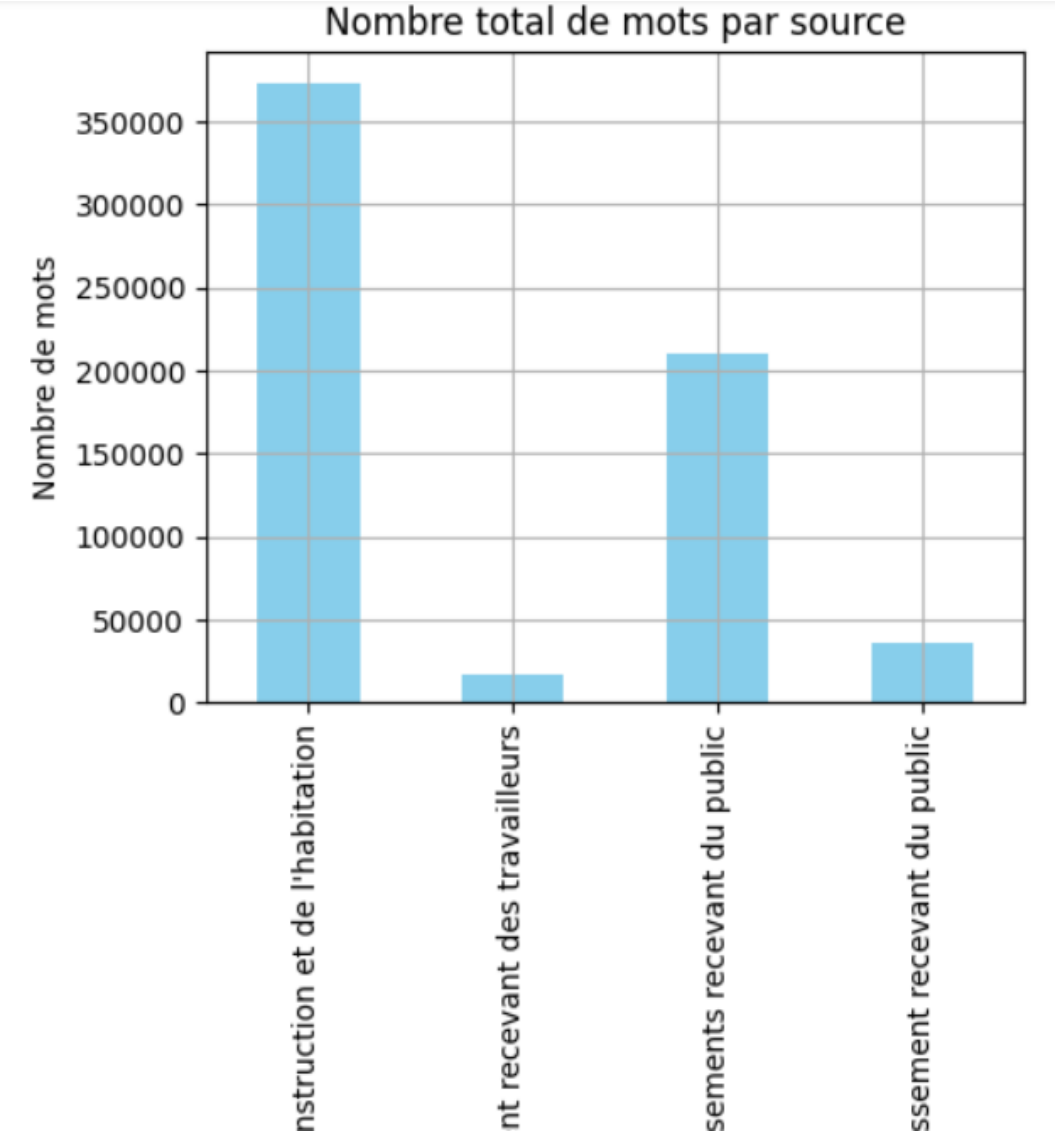
Tableau 5 : Analyse de la longueur moyenne du texte par sources de données

```
Longueur moyenne du texte par source :
 Source
Code de la construction et de l'habitation                                          848.645738
Etablissement recevant des travailleurs                                             347.304348
Règlement de sécurité contre les risques d'incendie et de panique. Etablissements recevant du public   1072.582257
Réglementation sécurité et incendie. Etablissement recevant du public              4556.750000
Name: text_length, dtype: float64
```

Nombre total de mots par source

En visualisant le volume de mots provenant de chaque source, nous avons pu identifier les différences de granularité entre les textes. Les textes réglementaires provenant de certaines sources contenaient parfois des informations plus concises ou, au contraire, des détails redondants. Cette analyse a renforcé notre compréhension de la densité d'information dans les documents.

Tableau 6 : Analyse du nombre de mots des différentes source de notre corpus



Wordclouds pour chaque catégorie de texte

Un wordcloud est une représentation visuelle des mots les plus fréquents dans un texte, leur taille étant proportionnellement à leur fréquence. Cette visualisation est particulièrement utile pour identifier les thèmes principaux et les mots-clés dominants. Nous avons créé des nuages de mots pour les textes de chaque catégorie (ERT, ERP, habitations et autres) afin de repérer les termes récurrents propres à chaque type de réglementation. Ces insights ont servi à mieux comprendre le vocabulaire spécifique à chaque domaine et à orienter les choix des intégrations pour le modèle.

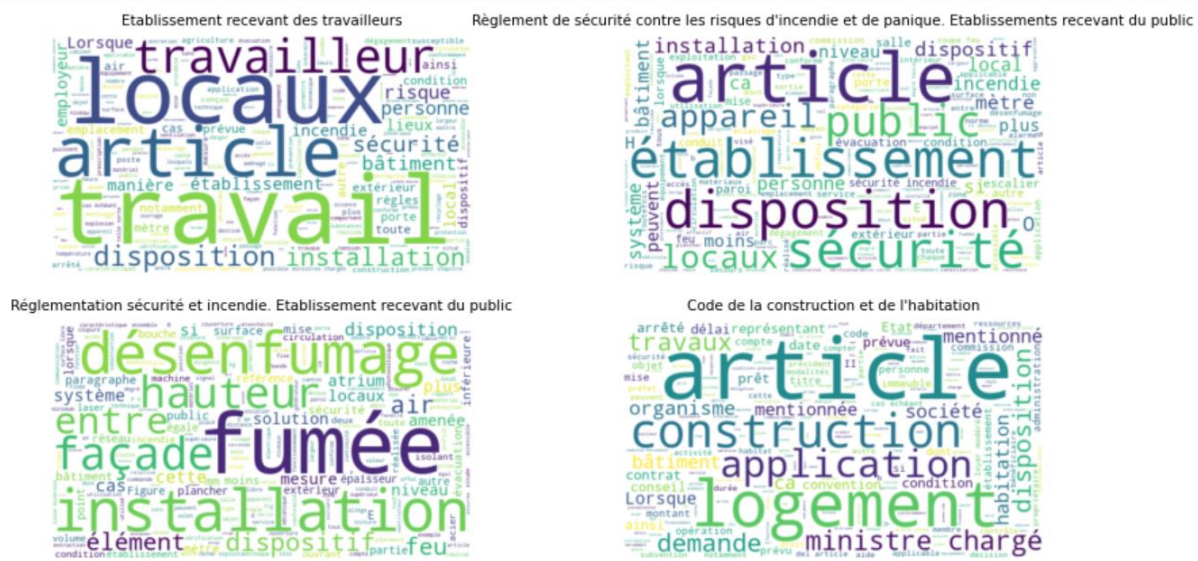


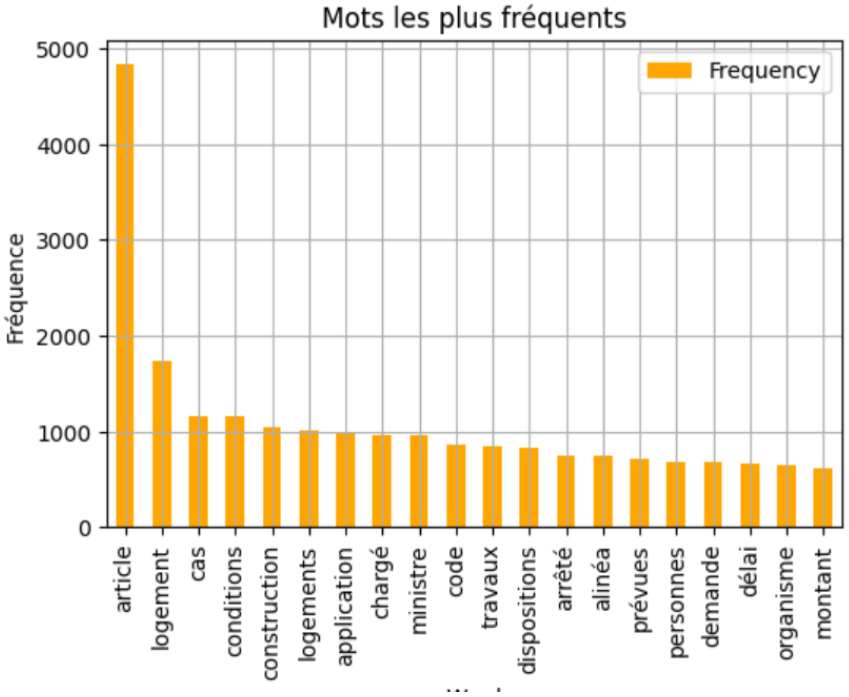
Tableau 7 : Wordcloud du contenu des différentes sources de données

Analyse de la fréquence des mots des différents articles

Une analyse approfondie de la fréquence des mots a été réalisée pour chaque source. Cela nous a permis d'identifier les termes les plus représentatifs et ceux qui pourraient nécessiter un traitement particulier, comme les noms propres ou les termes juridiques spécifiques. Nous avons également pu repérer des mots vides récurrents et décider s'ils devaient être exclus ou conservés selon leur importance contextuelle.

Toutes les analyses effectuées n'ont pas nécessairement été directement pertinentes pour la suite du projet. Cependant, elles présentent notre démarche exploratoire, visant à comprendre nos données et à maximiser la pertinence et l'efficacité des modèles de NLP.

Tableau 8 : Analyse de la fréquence des mots de notre corpus documentaire

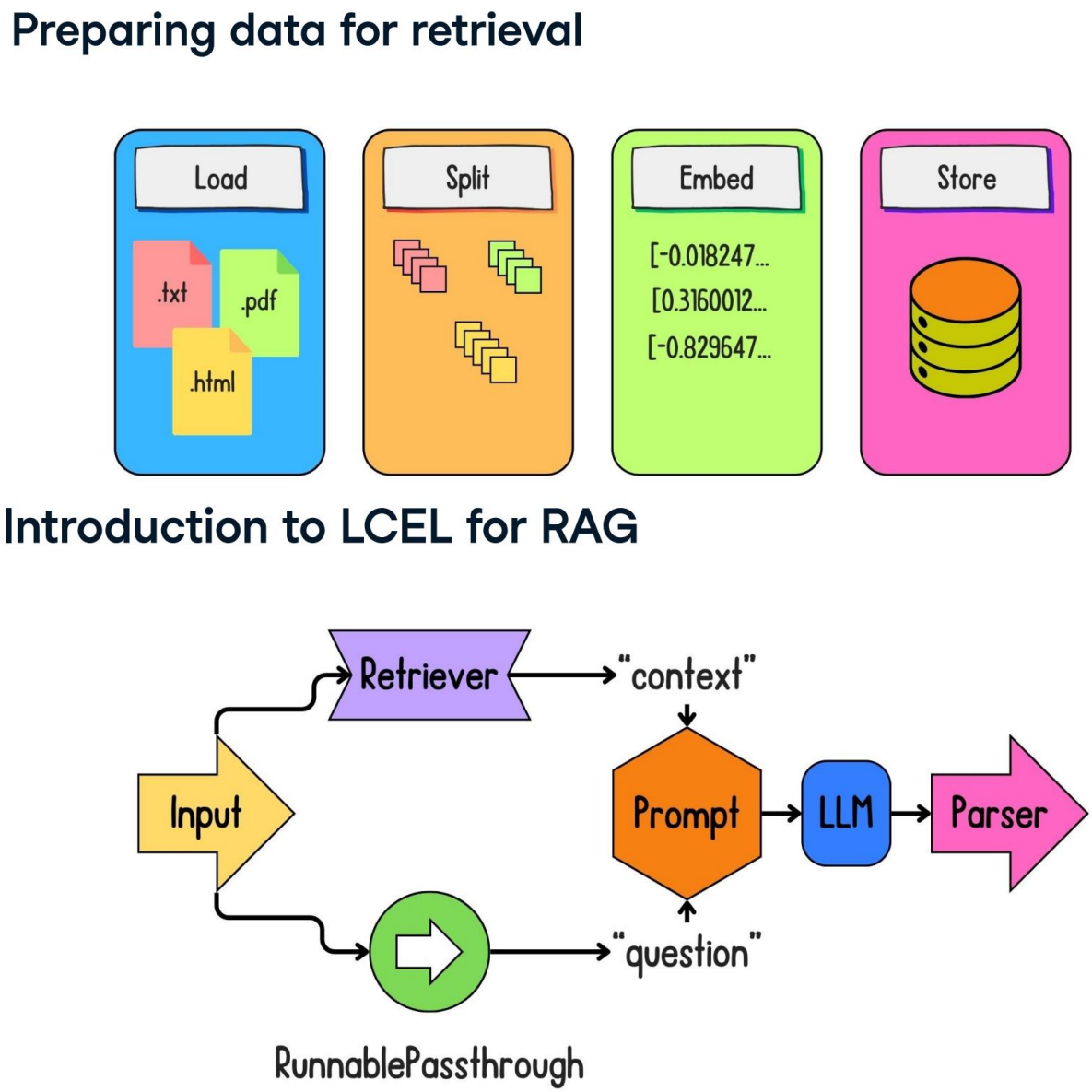


4. Implémentation de RAG avec embedding standard

L'implémentation d'un système RAG (Retrieval-Augmented Generation) avec des embeddings standards a constitué une étape clé pour permettre à notre chatbot de répondre précisément aux questions réglementaires. Cette approche repose sur l'intégration de techniques de recherche d'informations pertinentes et de génération de réponses contextuelles.

Le schéma suivant présente les différentes étapes à suivre pour la mise en place d'un système RAG – étapes que nous avons suivi et que nous allons expliquer :

Tableau 9 : Les étapes du RAG (source : Datacamp)



Découpage des contenus

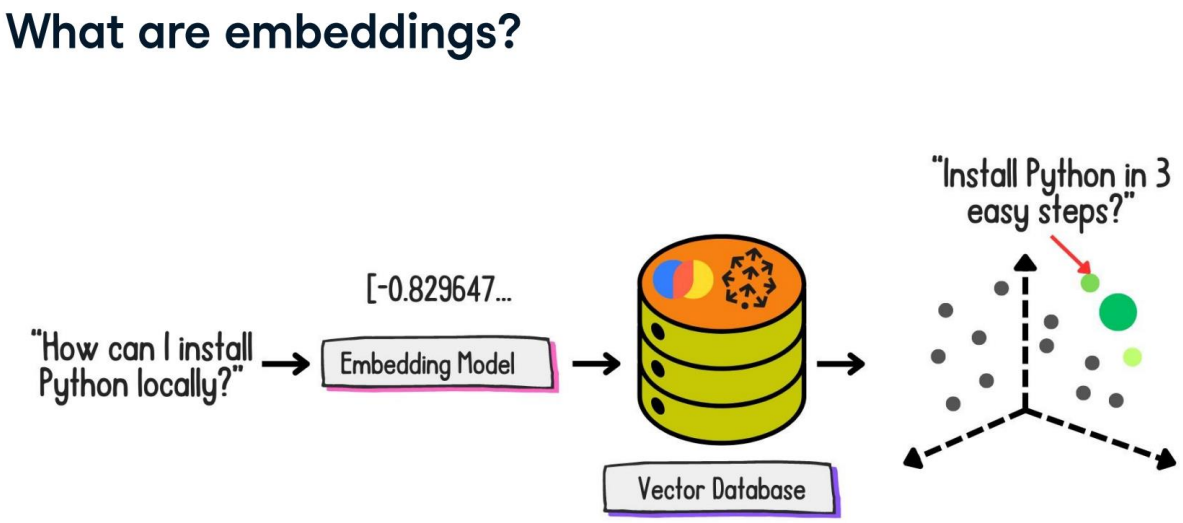
Pour structurer les données textuelles, nous avons utilisé l'outil RecursiveTextSplitter pour découper les documents en "chunks", de taille raisonnable. Cette étape est essentielle car les modèles comme GPT-4, que nous avons utilisé par la suite, ont une limite sur la longueur des textes qu'ils peuvent traiter.
Le découpage a été effectué en respectant les structures naturelles des textes, comme les paragraphes ou les sections, afin de minimiser la perte de contexte.

Transformation des chunks en embeddings

Une fois les chunks créés, nous les avons transformés en vecteurs numériques à l'aide du modèle SentenceTransformer("all-MiniLM-L6-v2").
Les embeddings sont des représentations numériques des textes qui capturent leur signification sémantique et permettent de mesurer la similarité entre des textes en

calculant la distance entre leurs vecteurs.
Ce modèle d'encodage a été choisi pour sa légèreté et ses performances optimisées pour des tâches de recherche d'information.

Tableau 10 : Schéma expliquant le fonctionnement des embeddings



Création d'une base de données vectorielles avec Chroma

Nous avons utilisé Chroma pour stocker et gérer les embeddings. Chroma est une base de données vectorielles spécialisée qui permet, entre autres, d'indexer efficacement les intégrations et d'interroger les données pour trouver les vecteurs les plus similaires à une requête de l'utilisateur.
Associer des métadonnées aux vecteurs, comme la source du texte, permet d'enrichir les résultats.
Nous avons créé un client persistent afin de nous assurer que la base de données vectorielles reste accessible même après l'arrêt du projet.

Chargement des documents dans une collection

L'ensemble des chunks a été chargé dans une collection Chroma**.** Chaque document a été associé à ses métadonnées afin de permettre une récupération contextuelle plus précise.

Test de la base de données avec une requête

Pour vérifier le bon fonctionnement de la base de données vectorielles, nous avons effectué une requête en interrogeant la collection. L'objectif était de retrouver les 5 documents les plus pertinents**,** ainsi que leurs métadonnées. Ce test nous a permis de valider :

- La qualité des embeddings et leur capacité à capturer les relations sémantiques.
- La capacité de Chroma à répondre rapidement et précisément aux requêtes.

Création d'un vectorstore et d'un retriever

Le vectorstore est un composant qui stocke les embeddings et offre des mécanismes pour les interroger. Il constitue le cœur de la recherche dans notre système RAG.
Le retriever agit comme une interface pour extraire les documents les plus pertinents à partir du vectorstore. Il joue un rôle clé en combinant les requêtes utilisateur avec les informations stockées pour fournir les résultats les plus proches contextuellement.

Création d'une chaîne RAG

Pour permettre une interaction fluide avec le chatbot, nous avons configuré une chaîne ConversationalRetrievalChain**.** Cette chaîne combine plusieurs éléments :

- Un LLM (Large Language Model) basé sur OpenAI GPT-4o. Le LLM génère des réponses naturelles et contextuelles en utilisant les documents récupérés par le retriever.
- Une mémoire ConversationBufferWindowMemory : cette mémoire permet au chatbot de conserver le contexte des derniers échanges avec l'utilisateur, améliorant ainsi la continuité des réponses dans une conversation.

5. Implémentation de RAG avec embedding finetuné

Nous avons ensuite essayé de finetuner l'espace d'embedding afin de mettre en avant les similarités spécifiques et capturer des nuances métier. Dans ce but, nous avons chargé plusieurs types de modèles de la famille Bert orientés français ou multilingues et nous avons évalué les résultats avant et après finetuning. La procédure s'appuie sur les 3 étapes suivantes :

- **Chargement des données et des modèles pré-entrainés :**

La première étape consiste à charger et prétraiter les données fusionnées dans le fichier .csv, incluant des textes, les contextes associés et la source d'extraction. Les contextes sont définis comme une liste d'intitulés des éléments hiérarchiques de l'extraction [Source, Livre, Chapitre, Section, Sous-section, Paragraphe]. Voici un exemple :

```
['Etablissement recevant des travailleurs', "Obligations du maître d'ouvrage pour la conception des lie
ux de travail  (Articles R4211-1 à R4217-2)", 'Sécurité des lieux de travail (Articles R4214-1 à R4214-
28)', 'Caractéristiques des bâtiments (Articles R4214-1 à R4214-8)', 'Article R4214-4', 'Créé par Décre
t n°2008-244 du 7 mars 2008 - art', '(V)', '']
Les surfaces des planchers, des murs et des plafonds sont conçues de manière à pouvoir être nettoyées o
u ravalées en vue d'obtenir des conditions d'hygiène appropriées.
```

Les données sont encodées à l'aide de modèles de langage pré-entrainés comme Flaubert, Camembert, Bert et DistilBert qui sont des modèles spécialisés / et qui prennent en charge la langue française. Chaque texte et contexte sont transformés en embeddings via ces modèles. Ce processus permet de préparer les données pour l'étape suivante : la recherche de similarités entre les requêtes de l'utilisateur et la base de savoir du modèle. Une fois les embeddings générés, on évalue les similarités à l'aide de la cosine similarity, afin de déterminer les réponses les plus en lien avec la requête utilisateur. Un exemple de requête et de recherche d'éléments proches est donné ici :

```
User query: Je dois faire des modification dans un établissement recevant le public. Par qui l'autorisation est elle délivrée

----------------------------------------------------------------------------------------------------------------------
Cosine Similarity: 0.5724

Source plus proche:Code de la construction et de l'habitation

 Contexte: Code de la construction et de l'habitation.x. Article R641-19.x. Dispositions permettant de faire face à des difficultés particulières de loge
ment..x. Mesures tendant à favoriser la construction d'habitations..x. Sociétés coopératives de construction..x. Agrément des contrôleurs techniques.x. D
iagnostic portant sur les déchets issus de rénovations et de démolitions.x. Modifié par Décret n°99-340 du 29 avril 1999 - art. 1 () JORF 5 mai 1999.x.

Texte: Le délai supplémentaire prévu àl article L.641-1, alinéa 4, ne peut être accordé au bénéficiaire de l attribution d office lorsque le propriétaire
notifie qu il entre dans une des catégories prévues àl article L.641-2.
```

Les embeddings du modèle pré-entrainés donne une ressemblance maximale de 57,24 %. Le texte proposé en réponse à la requête peut en effet être jugé relativement proche.

- **Finetuning du modèle**

Afin d'améliorer la précision des réponses, un finetuning du modèle est effectué sur un jeu de données spécialement conçu pour l'apprentissage contrastif. Ce jeu de données est composé de paires de textes avec des labels continus sur l'intervalle -1, 1 indiquant leur similarité.
La base de données est construite avec des paires dont la ressemblance est calculée en fonction du nombre d'éléments communs de la hiérarchie de contextes. Si 2 contenus proviennent de la même source, même livre, même chapitre mais section, sous-section et paragraphe différents, ils auront un ratio de 3/6. Si les paragraphes proviennent de sources différentes, le ratio des éléments communs est 0. Ce ratio est ensuite transformé en coefficient de ressemblance entre -1 et 1 en utilisant une fonction analytique:

$$ressemblence = 1 - 2\,e^{-4{,}2 * ratio}. Ex: (0\% \leftrightarrow 1), (15\% \leftrightarrow\ \ 0.3), (30\% \leftrightarrow 0.5)..$$

Le dataset ainsi obtenu est composé de 11833 éléments :

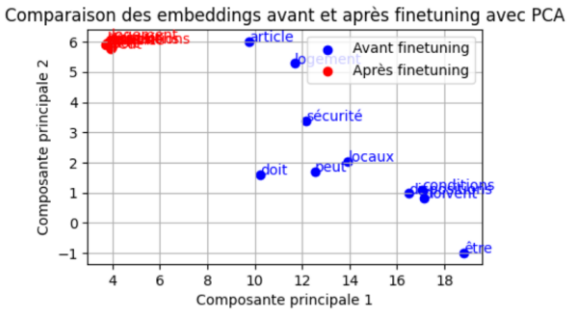
- 3833 associations positives entre le contexte et le contenu de chaque ligne
- 8000 associations basées sur le nombre de mots communs dans la hiérarchie.

```
Dataset composé de: 3833 de relations positives (1) entre contexte et contenu
Dataset composé de: 8000 de relations entre -1 et 1 entre contenus différents
```

Un modèle contrastif est ensuite entraîné, utilisant des réseaux neuronaux pour apprendre à différencier les paires similaires et non similaires. En plus de la couche de projection, une seconde couche a été prévue pour changer la dimension de l'embedding. Pour l'entrainement actuel, la dimension de l'embedding a été conservée (768). Après l'entraînement, une analyse est réalisée pour évaluer la qualité des embeddings produits par le modèle finetuné. Une première comparaison avec la même requête des éléments les proches montre que les éléments fournis ne sont pas particulièrement plus pertinents mais ils présentent des scores de similarité plus grands : 62,7%

```
User query: Je dois faire des modification dans un établissement recevant le public. Par qui l'autorisation est elle délivrée


----------------------------------------------------------------------------------------------------------------------------
Cosine Similarity: 0.6270

Source plus proche:Code de la construction et de l'habitation

 Contexte: Code de la construction et de l'habitation.x. Article R171-5.x. Construction, entretien et rénovation des bâtiments.x. PERFORMANCE ÉNERGÉTIQUE
ET ENVIRONNEMENTALE.x. RÈGLES GÉNÉRALES DE PERFORMANCE ÉNERGÉTIQUE ET ENVIRONNEMENTALE.x. Agrément des contrôleurs techniques.x. Diagnostic portant sur l
es déchets issus de rénovations et de démolitions.x. Création Décret n°2021-872 du 30 juin 2021 - art..x.

Texte: Un arrêté conjoint des ministres chargés de la construction et de l énergie précise les modalités d application de la présente section.
```

Une deuxième comparaison est faite à travers une représentation 2D de la position de certains mots dans l'espace d'embedding finetuné réduit à ces deux composantes principales (PCA).
L'action du finetuning est trop forte.



En vue de leur utilisation dans un RAG par la suite, les données, modèle et config sont sauvegardés.

- **Mise en place du RAG**

Pour intégrer une recherche efficace dans le chatbot, une architecture de RAG est mise en place. Afin d'enrichir les résultats fournis au générateur, plusieurs retrievers ont été mis en place. Faiss est la première méthode de recherche de similarité utilisée. Elle utilise la représentation vectorielle finetunée pour déterminer les éléments les plus proches. Une méthode complémentaire est ElasticSearch qui fait de la recherche textuelle et qui peut donc classer ou enrichir les propositions fournies par Faiss. Malheureusement cette méthode n'a pas fonctionné pour incompatibilité de librairie. L'étape suivante consiste à utiliser le modèle AutoModelForSeq2SeqLM (T5-base) pour générer une réponse à partir des documents récupérés. Ce modèle est capable de traiter la question et d'extraire l'information nécessaire pour générer une réponse concise et précise, en combinant la recherche et la génération de texte pour fournir une réponse plus adaptée au contexte.
Malheureusement les réponses sont totalement aléatoires.

6. Poursuite du projet

Dans la continuité du projet, nous avons exploré la possibilité d'enrichir la base de données en intégrant d'autres types de données, notamment les Documents Techniques Unifiés (DTU) au format PDF. Les DTU constituent une source essentielle pour le secteur du bâtiment car ils détaillent les bonnes pratiques et les règles de l'art en matière de construction. Cependant, leur intégration dans une base de données vectorielles a présenté des défis spécifiques.

Exploration des outils de chargement de PDF

Pour extraire le contenu des fichiers PDF, nous avons exploré plusieurs outils de parsing adaptés aux documents complexes tels que pdfreader, pdfplumber**,** pdfminer**.** Après des tests comparatifs, nous avons retenu pdfreader qui est une solution simple d'utilisation et permet d'extraire efficacement le texte brut des DTU.

Nettoyage des données extraites

L'extraction des textes depuis les PDF nécessite une phase importante de nettoyage des données**.** Les documents DTU contiennent typiquement des en-têtes, des pieds de page qui doivent être supprimés pour ne pas nuire à la cohérence des textes.

Conservation de l'arborescence du texte grâce aux regex

Un des principaux enjeux de cette tâche a été de conserver l'arborescence du texte, essentiel pour préserver le contexte et les relations hiérarchiques des informations. Les DTU sont souvent organisées en sections et sous-sections numérotées (par exemple, 1, 1.1, 1.1.1), et il était crucial de maintenir cette structure.
Pour cela, nous avons développé des expressions régulières (regex) permettant d'identifier automatiquement les titres de section et de sous-section, de reconstruire une arborescence fidèle au texte initial.

Résultats et perspectives

Bien que nous n'ayons pas encore finalisé cette piste, les premières expérimentations ont montré que cette méthode permet de récupérer efficacement le texte tout en conservant sa structure. Cependant, des défis subsistent, notamment liés à

- la variabilité des formats des DTU
- la présence de tableaux
- la présence d'image ou figures

Compte tenue l'importance des DTU dans le bâtiment, cette piste reste malgré tout à explorer et sera à poursuivre après le projet.

Tableau 11 : Capture d'une page de DTU

6.3 Méthodes de montage

6.3.1 Opérations préliminaires à la pose

L'entreprise doit vérifier que le niveau NGF (Niveau général de la France) a bien été matérialisé par le Maître de l'Ouvrage.

6.3.2 Axes ou points de départ, répartition des découpes en rives

Les axes de départ sont fixés par le plan de calepinage.

Ces axes perpendiculaires doivent être situés de façon à ce que les dalles découpées en rives aient toujours une dimension supérieure à 100 mm pour des raisons de stabilité (voir figure 1).

NOTE Le marquage au sol des emplacements des vérins n'est dû par l'installateur que dans le cas où il est clairement exigé dans les DPM.

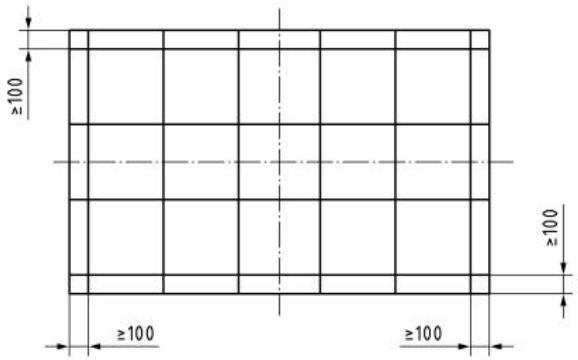


Figure 1 — Exemple de répartition des découpes en rives

6.3.3 Pose proprement dite

Ce travail ne débute jamais en rives. Il peut être mené suivant différentes méthodes qui peuvent être

Nous souhaiterions également explorer les pistes suivantes :

Optimisation de l'embedding finetuné

- Moins agressif sur le finetuning: Low-Rank Adaptation of Large Language Models (LoRA)
- Tester d'autres retrievers
- Exploiter la structure hiérarchique des contenus pour filtrer avec Langchain
- Tester d'autres générateurs (bloom, LLaMA)

7. Conclusion

Ce projet a été une occasion précieuse de mettre en œuvre l'ensemble des nouvelles compétences acquises au cours de la formation, en suivant un processus complet, de l'extraction des données brutes à leur transformation en embeddings, jusqu'à l'implémentation d''un système de RAG.
Chaque étape, depuis la collecte et le nettoyage des données jusqu'à la conception d'un chatbot intelligent, a permis de consolider les concepts appris en les appliquant sur un cas pratique.
Ce projet a été particulièrement stimulant en raison de son application directe au secteur du bâtiment.
Au-delà des compétences techniques, ce projet a renforcé notre compréhension des enjeux liés à la structuration et à l'exploitation des données textuelles dans des domaines spécialisés. Il ouvre également la voie à des améliorations futures, comme l'intégration de sources supplémentaires, telles que les DTU et les normes, pour enrichir les capacités du chatbot.
En conclusion, ce projet a non seulement permis de développer une solution innovante, mais aussi de mesurer concrètement l'apport de cette nouvelle technologie dans un contexte professionnel.

Sources

Récolte de données

(1) API gouv - https://www.data.gouv.fr/fr/dataservices/legifrance/
(2) CNIL - https://www.cnil.fr/fr/focus-interet-legitime-collecte-par-moissonnage
(3) Code du travail - https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072050/LEGISCTA000018488606/#LEGISCTA000018532586
(4) Code de la construction et de l'habitation - https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074096/LEGISCTA000006112857/#LEGISCTA000006112857
(5) Arrêté du 25 juin 1980 applicables aux ERP - https://www.legifrance.gouv.fr/loda/id/LEGITEXT000020303557

RAG

(6) Datacamp - Génération augmentée de RAG avec Langchain
(7) Datacamp - application with LLM