CRÉATION D'UN CHATPDF

Maram Assili
Ecole supérieure de la statistique et de l'analyse d'information

Introduction

Chez PwC, on travaille avec divers clients, chacun fournissant un PDF décrivant ses besoins , souvent des données financières complexes. Lire et analyser manuellement ces documents est long et laborieux. Ainsi, on a développé ChatPDF, un chatbot permettant aux utilisateurs de poser des questions et d'obtenir des réponses instantanées basées sur les informations des PDF, rendant le traitement des données plus rapide et efficace.

Méthodologie

Afin de créer ChatPDF, on a adopté la méthodologie **RAG (Retrieval Augmented Generation)**, qui repose sur trois blocs essentiels :

1. **Bloc de Récupération** : Ce bloc permet d'extraire les passages pertinents du PDF en fonction de la question posée par l'utilisateur.
2. **Bloc de Génération** : À partir des extraits récupérés, un modèle de génération de texte produit une réponse naturelle et fluide, adaptée au contexte.
3. **Bloc d'Augmentation** : Ce bloc permet d'intégrer les informations récupérées dans le contexte pour enrichir le modèle de génération.

Cette approche nous permet de traiter efficacement les PDFs tout en fournissant des réponses claires et rapides aux utilisateurs.
Voici une représentation de ce concept **RAG** :

Fig. 1: Vue d'ensemble de l'approche RAG

L'approche RAG se compose de trois blocs principaux. Le **Bloc (A,B,C,D)** concerne la Préparation des PDF, où nous extrayons et nettoyons le texte, puis l'encoder en vecteurs. Le **Bloc (1,2)** se concentre sur la Préparation des requêtes, où nous prétraitons et encodons les questions des utilisateurs. Enfin, le **Bloc (3,4,5)** s'occupe de la Génération des réponses, où nous recherchons les informations pertinentes et générons des réponses claires. Cette structure optimise le traitement des données et améliore la qualité des réponses.

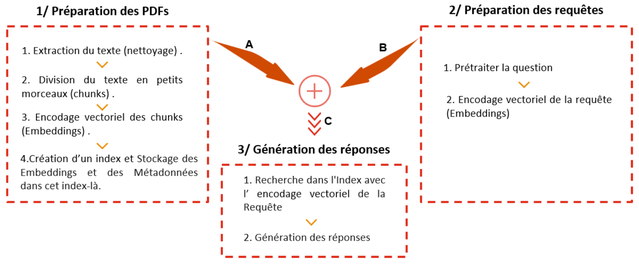


Fig. 2: Le concept

Outils

Pour le développement de ChatPDF, on a utilisé plusieurs outils et bibliothèques à différentes étapes du processus, comme détaillé ci-dessous :
1. Extraction et Nettoyage des Données :
- **pdfplumber** : permet d'extraire le texte des fichiers PDF tout en conservant leur mise en forme.
- **spaCy** : est utilisé pour le traitement du langage naturel: le nettoyage et l'analyse du texte (tokenisation, lemmatisation, NER...).
2. Chunking :
- **LangChain** : Ce framework permet de diviser le texte extrait en petits morceaux (ou **chunks**) .
3. Encodage Vectoriel (Embeddings) :
- **FinBERT** : est une variante de BERT fine-tunée spécifiquement sur des documents financiers. FinBERT utilise le concept des transformeurs pour créer des représentations vectorielles des textes, ce qui permet de mieux capturer les nuances et les relations dans les données financières.

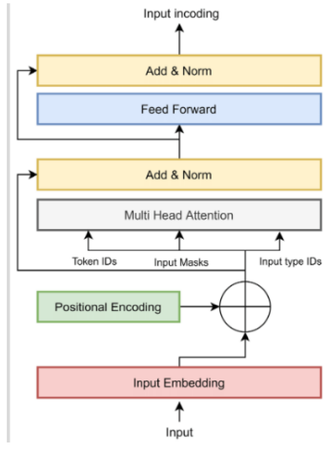
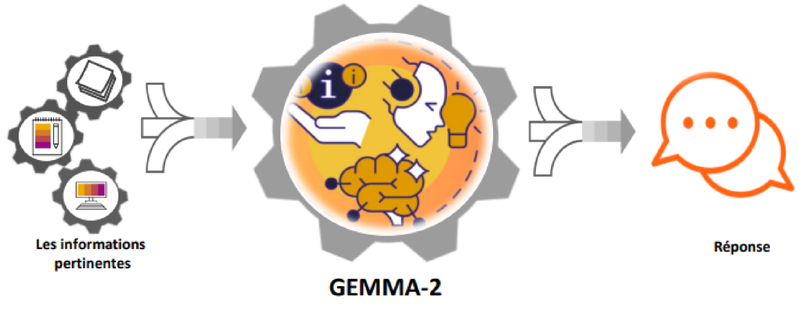


Fig. 3: Architecture du Finbert

4. Indexation et Requêtes :
- **FAISS (Facebook AI Similarity Search)** : permet de créer des index de manière efficace et de rechercher la similarité entre les embeddings. Cela facilite la récupération rapide des informations pertinentes en réponse aux requêtes des utilisateurs.
5. Génération des Réponses :
- **Gemma 2** : est un modèle de langage développé par **Google**, capable de reformuler les informations extraites de manière fluide, améliorant ainsi l'expérience utilisateur en fournissant des réponses en **anglais** précises et contextuelles.
- **Helsinki-NLP/opus-mt-en-fr** : Si la requête est en **français**, la réponse générée par le modèle GEMMA sera utilisée comme entrée pour ce modèle de **traduction**, permettant ainsi de fournir une réponse dans la langue de la requête tout en conservant le sens et le contexte.



Métriques

Le **ROUGE-1** score est utilisé pour évaluer les réponses générées par le modèle de génération de texte en comparant les résultats avec des textes de référence c'est-à-dire en comparant les n-grammes du texte généré avec ceux du texte de référence. Cei est déterminé par:

- **Rappel** = $\frac{\text{Nombre de mots communs}}{\text{Nombre total de mots dans le texte de référence}}$
- **Précision** = $\frac{\text{Nombre de mots communs}}{\text{Nombre total de mots générés}}$
- **F-mesure** est une combinaison des scores de précision et de rappel pour obtenir une mesure globale de la qualité.

```
Metrics:
          Metric     Score
0  ROUGE-1 Precision  0.811111
1     ROUGE-1 Recall  0.474026
2  ROUGE-1 F-measure  0.598361
Total Time: 14.22 seconds
Search Time: 0.00 seconds
Generation Time: 14.06 seconds
ROUGE Calculation Time: 0.02 seconds
```

Fig. 5: Métrique

⇒ La nature de chaque question et les informations extraites pour répondre à la requête agissent sur les résultats.

Résultats

On a recouru à **Gradio**, une bibliothèque open-source en Python, pour concevoir une interface utilisateur interactive dans le cadre de ce projet.

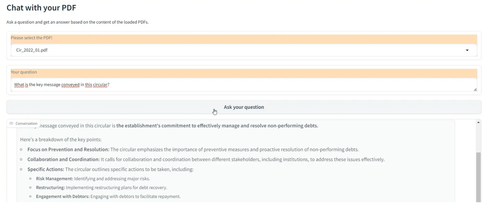


Fig. 6: Produit du travail

Conclusion

Cet outil optimise le temps de lecture des PDF en automatisant l'extraction et la compréhension des informations clés. Grâce à des modèles avancés, il facilite la navigation dans de grands volumes de texte et fournit des réponses pertinentes.

perspectives

L'intégration de l'**OCR(Optical character recognition)** qui permet de traiter des PDFs scannés.

Références

- Naik, Krish. Complete Road Map To Prepare NLP - Follow This Video - You Will Be Able to Crack Any DS Interviews. YouTube, 2020
- Béranger. NLP : Extraire des caractéristiques pour utiliser des algorithmes ML. Medium, 6 Oct. 2020.
- Pinecone. "FAISS Tutorial." Pinecone.