**Présentation : Générateur de notes RESF**

**Introduction**

Bonjour à toutes et à tous, on a le plaisir de vous présenter aujourd’hui le projet de Générateur de notes RESF, réalisé dans le cadre de la certification IA. Cette présentation va s’articuler autour des enjeux métiers, de la méthodologie IA mise en place, des résultats obtenus et des perspectives pour le service FIPU de la Banque de France.

**I. Enjeux métiers**

1. Présentation du Service DCPM / FIPU

La DCPM est en charge de la réalisation des prévisions économiques de court et moyen terme de la Banque de France. Elle regroupe 3 services, qui travaillent en étroite collaboration :

* DIACONJ
* SEMAP
* FIPU

Le processus de prévision est un processus itératif, c’est-à-dire qu’une prévision s’appuie sur les résultats de chaque service.

Le service FIPU en particulier regroupe 5 à 6 économistes chargés de la réalisation des prévisions de finances publiques pour la Banque de France. En matière de finances publiques, en raison de l’indépendance entre la politique monétaire et la politique budgétaire, la Banque de France ne dispose pas d’informations privilégiées et dépend donc entièrement des documents publiés par d’autres institutions, comme le Trésor, la Cour des comptes, l’Insee ou la Commission européenne.

2. Difficultés rencontrées

A partir de ces documents budgétaires, chaque année, les économistes doivent produire quatre prévisions, dont deux intermédiaires et deux complètes, en s’appuyant sur près de 2 000 pages de rapports.

Les documents à analyser incluent le PLF, le RESF, le programme de stabilité, les prévisions de la Commission, les rapports de la Cour des comptes, les notes méthodologiques de l’Insee et des articles de presse.

La longueur des documents, l’éparpillement des informations utiles dans de nombreuses pages de rapport, l’extraction de données des documents et la réalisation de graphiques et tableaux limitent le temps consacré à l’analyse économique de fond.

3. Opportunité de l’IA

L’automatisation de la lecture, de l’analyse, de la rédaction de notes et de l’élaboration de tableaux grâce à l’IA permettrait de gagner en efficacité et de recentrer les experts sur l’analyse de fond :

* Les OCR (Optical Character Recognition, ou Reconnaissance Optique de Caractères en français) est une technologie qui permet de convertir automatiquement une image contenant du texte (comme un document scanné, une photo ou un PDF d’image) en un texte lisible et modifiable par un ordinateur. L’OCR analyse l’image pour détecter les formes des lettres, des chiffres et des symboles, des images, des tableaux, puis reconstitue le contenu textuel sous forme de données numériques, comme un document texte. De tels outils sont particulièrement intéressants pour détecter les graphiques, tableaux et structures des documents. Pour les documents administratifs et budgétaires, structurés en chapitres, ils s’avèrent extrêmement performants.
* Les LLM sont des outils qui peuvent s’avérer très utiles pour comprendre les demandes utilisateurs et extraire l’information pertinente du texte. Les LLM ne peuvent prendre en entrée, que les données texte, et s’appuient donc sur les sorties des outils OCR.

**II. Type de document testé : le RESF**

Le Rapport Économique, Social et Financier (RESF) est associé à chaque Projet de Loi de Finances (PLF) et publié chaque année par le Trésor. C’est l’un des documents les plus importants pour l’exercice de prévisions de finances publiques. Il comporte environ 225 pages, dont une trentaine consacrée à des tableaux divers.

Ce rapport détaille la stratégie économique de la France, les perspectives macroéconomiques, l’évolution des finances publiques et des dossiers thématiques d’actualité, comme la crise Covid ou la guerre en Ukraine.

Il est en comptabilité nationale, à la base des sciences économiques, tandis que les autres documents sont en comptabilité générale, voire budgétaire. Il s’agit donc du document le plus facilement interprétable, et donc, celui qui nécessite le moins d’expertise de finances publiques. C’est pourquoi nous avons décidé de concentrer nos efforts dans le traitement automatique de ce document, car il n’y a quasiment pas de retraitement d’experts des chiffres et tableaux présents dans ce document.

Par ailleurs, il est souvent intéressant de se référer aux RESF des années antérieures, les prévisions du gouvernement s’étalant sur plus d’un an, afin de comprendre les changements opérés par le gouvernement (mesures nouvelles, suppression ou amendement de mesures existantes), depuis l’exercice budgétaire précédent. Un outil permettant de retrouver rapidement les informations pertinentes des années antérieures ajouterait de la profondeur à l’analyse économique.

**III. Méthodologie et outils IA**

Un des leitmotivs de ce projet, c’est de pouvoir être utilisé par le service concerné dès aujourd’hui. Nous avons fait en sorte que tous les modèles utilisés dans ce projet soient open source, fonctionnant en API, c’est-à-dire ne nécessitant pas d’ordinateurs ayant de puissantes cartes graphiques. Nous avons donc fait le choix de nous tourner vers les algorithmes de Mistral offrant une bonne performance, sans coût, avec une limite d’appels assez haute pour être exploitable par tout agent ayant accès au code, et disposant d’un ordinateur banque.

1. Organisation des dossiers

Le projet s’appuie sur une organisation structurée avec un dossier paramètre centralisant les paramètres de fonctionnement, un dossier *input* pour les fichiers d’entrée (RESF\_année.pdf) et un dossier *output* pour les résultats (notes générées, tableaux extraits, conversion texte, embeddings).

2. Conversion PDF vers texte

La conversion des PDF en texte exploitable par un LLM est réalisée via l’OCR Mistral (mistral-ocr-latest), accessible par API et compatible avec des ordinateurs sans GPU. Cette technologie permet de reconnaître non seulement le texte mais aussi les tableaux et images présents dans le document.

Afin d’être certains que l’outil fonctionne bien, nous l’avons testé sur l’ensemble des RESF, et avons vérifier que :

* Toutes les informations soient bien retranscrites dans la sortie de l’OCR. Nous n’avons remarqué aucune perte d’informations entre le document et la sortie OCR
* Que l’ensemble des tableaux soient correctement identifiés. Les tableaux sont délimités par le signe « | » dans la sortie OCR, ce qui nous permet ensuite facilement de les isoler. On vérifie ensuite que leurs chiffres, en-tête (pas de décalage lignes/colonnes), soient bien conformes à ce qui est inscrit dans le document. Globalement, les résultats de cet outil open source sont époustouflants. Seul un ou deux tableaux par documents étaient problématiques (absence de la ligne d’en-tête notamment). Le reste des tableaux était parfaitement identifié et extrait.
* Que l’ensemble des images / graphiques contenus dans le document sont également extraits.

L’outil n’est donc pas infaillible, mais :

* Le temps de contrôle est substantiellement moins long que l’extraction manuelle qui était habituellement faite,
* Et les erreurs sont moins courantes que lorsque les sorties sont faites par un humain.

3. Mise à profit de la sortie OCR

Nous avons produit plusieurs programmes à partir de la sortie OCR :

* Un programme d’enregistrement automatique de toutes les images du document (graphiques / images / logo) :
  + Le lien des images est contenu dans la sortie OCR (sortie texte). Il faut identifier ce lien et ensuite les télécharger en se connectant au serveur sur lesquelles elles sont enregistrées.
  + Les images sont extraites et enregistrées dans un sous-dossier avec comme nom leur titre dans le document, ce qui peut être intéressant pour la réalisation de présentation. Les images sont plus rapidement identifiables et trouvables par les analystes.
  + Il n’y a plus besoin de rechercher à travers tout le document l’image souhaitée, ce qui est un gain de temps non négligeable quand il s’agit de rapports de plusieurs centaines de pages, comprenant chacun une cinquantaine d’images / graphiques.
  + Limites identifiées :
    - Absence de documentation au moment de la création du programme. Cela a pris donc beaucoup de temps, en écumant les fonctions python de la librairie, pour comprendre comment récupérer les images de la sortie OCR. Les images sont stockées sur un serveur, et non pas dans le document texte de sortie OCR. Il faut donc les appeler depuis le serveur, pour les enregistrer localement.
    - Toutes les images sont enregistrées y compris les logos, ce qui est d’un moindre intérêt. Impossibilité de distinguer les images pures, des graphiques, infographies et logos
    - 50 % des images avec un titre valide lors de l’utilisation du programme au départ. Dans une version de notre code, nous avons tenté d’améliorer cette statistique en aidant le programme à mieux identifier les titres des images : exemple « Graphique X ». Légère amélioration : un peu moins des 2/3 des images avaient un titre valable ensuite. On pourra à l’avenir améliorer cette fonction d’identification.
    - Dernière limite : les images ne sont pas enregistrées localement au départ, mais sur un serveur, ce qui peut poser problème en cas d’utilisation sur des documents confidentiels.
* Un programme d’extraction des tableaux et d’enregistrement automatique sous format Excel, avec pour nom, le titre du tableau dans le document, facilitant leur réutilisation par les experts.
  + Les tableaux sont identifiés un à un et sont enregistrés sous format texte et Excel.
  + Limites :
    - Création d’une base de données unifiée très difficile :
      * Tous les tableaux sont identifiés, y compris ceux ne comprenant pas de séries temporelles (séries utiles à la prévision macroéconomique). Nous avons essayé d’imposer un filtre sur les tableaux, en rejetant les tableaux ne contenant pas de chiffres, mais dans la plupart des tableaux ne contenant pas de séries temporelles, on trouve des années, des numéros de bas de page, ce qui complique l’utilisation de ce genre de filtres.
      * Les séries temporelles n’ont pas toutes la même période temporelle, ce qui complique la possibilité de créer facilement une base de données. Il faudrait éventuellement utiliser R, un autre langage de programmation, qui dispose de meilleurs outils que Python, pour joindre des tableaux ensemble.
      * Par conséquent, la création d’une base de données unifiée n’est pas impossible, mais très difficile et chronophage
    - Difficulté dans l’identification des titres des tableaux : plus de 50 % des tableaux n’ont pas le bon titre. Nous avons là encore essayé d’aider le modèle à mieux identifier les titres à l’aide de mots clés, comme (« tableau X ») ; mais cela n’a pas permis d’améliorer ce chiffre. Cela tient à la structure du texte dans le document autour des tableaux. En général, le titre se trouve dans le tableau, et non pas autour comme pour les images, ce qui complique l’identification.
* Un outil d’identification des chapitres du document :
  + Cela permet de séparer le document en thématiques
  + L’identification des chapitres s’appuie sur la table des matières. On repère les titres, et les numéros de page de ces chapitres. On extrait ensuite le texte entre ces numéros de pages avec le titre du chapitre comme nom de fichier
  + Ce code n’a pas été utile pour le RESF, ce document comprenant moins de 200 pages, mais dans le cas de documents plus volumineux, ce genre de programme pourrait être utile pour la généralisation à des documents plus volumineux du programme que nous vous présentons ci-après

4. Génération de notes d’analyse

Le texte est découpé en segments égaux en nombre de caractères avec un chevauchement fixe, puis chaque segment est vectorisé via l’embedding MistralAI en fonction de leur signification sémantique. Pour connaître les paramètres optimaux, nous avons fait plusieurs essais à partir d’un programme de questions – réponses. Nous avons ensuite établi un score en fonction du nombre de bonnes réponses (10 questions) et regardons la vitesse d’exécution en fonction des paramètres choisis. Voici la liste des 10 questions que nous avons sélectionnées.

|  |
| --- |
| Quelle est la prévision de croissance du PIB pour l’année 2022 dans le rapport ? |
| Quelles hypothèses macroéconomiques sous-tendent la trajectoire budgétaire (taux d’intérêt, prix du pétrole, inflation, etc |
| Quelle est l’évolution attendue de l’inflation en 2022 selon le rapport ? |
| Quels sont les principaux moteurs de la croissance identifiés pour 2022 ? |
| Comment le rapport évalue-t-il les risques pesant sur la reprise économique post-Covid ? |
| Quel est le niveau prévu du déficit public en 2022, en pourcentage du PIB ? |
| Quelle est la trajectoire de la dette publique prévue dans le rapport ? |
| Comment évoluent les dépenses de l’État en 2022 par rapport à 2021 ? |
| Quelles sont les principales mesures nouvelles de politique budgétaire présentées ? |
| Quelle est la stratégie du gouvernement pour ramener les finances publiques vers les critères de Maastricht ? |
| Quelles priorités économiques sont affichées par le gouvernement pour 2022 ? |
| Le rapport prévoit-il des mesures pour soutenir l’investissement public ou privé ? |
| Comment le rapport justifie-t-il la fin ou la prolongation des dispositifs de soutien exceptionnels (type PGE, activité partielle, etc |
| Quels éléments de la politique de l’emploi ou de la formation sont détaillés dans le rapport ? |
| Quelles sont les projections d’emploi et de chômage pour 2022 ? |
| Le rapport présente-t-il une évaluation de la pauvreté ou des inégalités en 2022 ? |
| Quelle place est accordée à la transition écologique dans le cadrage budgétaire ? |
| Quels dispositifs sociaux ou fiscaux sont ajustés ou créés pour soutenir le pouvoir d’achat ? |
| Comment la France se positionne-t-elle par rapport à ses partenaires européens en matière de croissance, de dette et de déficit ? |
| Quel est le déficit public en 2021 ? |

On remarque que les modèles avec de faibles chevauchements donnent de meilleures performances aux questions réponses. Les modèles avec des chunk de tailles 1000 et 1500 donnent également de meilleurs résultats. Voici le classement des modèles par score. On remarque que le meilleur modèle est le modèle avec une taille de chunk de 1000 caractères et un chevauchement de 50. On pourrait encore affiner la recherche de ces paramètres dans le cas d’une mise en production, en faisant les essais sur plusieurs documents et en agrégeant les scores.

Une fois les paramètres choisis, on peut se lancer dans l’élaboration des notes d’analyses. À partir d’un prompt bien construit, d’exemples de notes d’experts et d’instructions spécifiques, le système génère une note d’analyse finale enregistrée au format Word.

IV. Résultats obtenus

Les RESF 2020, 2021 et 2022 ainsi que les notes d’analyse des experts ont servi de base d’apprentissage. Pour l’année test 2025, le système a généré automatiquement une note d’analyse au format Word et les tableaux extraits au format Excel. Le temps d’exécution sur une machine Onyxia est d’environ 11 minutes, ce qui représente un gain de temps significatif pour les économistes, leur permettant de se concentrer sur l’analyse qualitative et la prise de décision. L’élaboration de ce genre de notes par des économistes expérimentés est d’environ une demie journée en moyenne, entre la lecture, la création des tableaux et la rédaction de la note.

Ces notes ne constituent pas un produit fini en soi, car elles devraient aussi contenir des comparaisons aux prévisions internes Banque de France. Or, pour des raisons de confidentialité, nous n’avons pu fournir au modèle ces prévisions. Cependant, les compléter requiert moins de temps que de créer la note ex-nihilo.

V. Outils subsidiaires et perspectives

Nous avons continué le projet, afin de montrer toutes les potentialités de l’utilisation de ces outils. Nous avons créé un agent FIPU. Cet agent FIPU permet la réalisation automatique et rapide de graphiques à partir de demandes utilisateurs. En effet, il est courant que pour des analyses ou des présentations, nous ayons besoin de réaliser des graphiques à partir de séries temporelles. Ce genre d’outil permet de créer ces graphiques sans code et de manière simple et efficace.

Cet outil fonctionne à partir d’un LLM qui extrait les informations pertinentes des demandes utilisateurs. Ces informations servent ensuite comme paramètres dans des fonctions préalablement construites. Par exemple, création d’un histogramme, ou de graphiques de courbes, extraction d’une série temporelle spécifique, etc…

Voici ce que donne l’outil.

Il est possible de lui ajouter d’autres fonctions à l’avenir pour étendre ces capacités d’actions, comme par exemple rédiger et insérer un graphique dans un mail.

A l’heure actuelle, cet outil fonctionne en local sur nos ordinateurs, qui ont de bonnes cartes graphiques. On pourrait envisager de créer ce genre d’outils sur des machines internes de la Banque.

Conclusion

Le générateur de notes RESF répond à un besoin métier fort : automatiser la lecture et l’analyse de rapports volumineux, tout en fiabilisant et accélérant la production de notes d’expertise. L’IA permet de recentrer les experts sur la valeur ajoutée de l’analyse, tout en garantissant la traçabilité et la reproductibilité des analyses. Les prochaines étapes consistent à généraliser l’outil, l’adapter à d’autres types de documents et renforcer l’intégration de l’IA dans les processus métiers.