Data Science en R/Python au Centre de Services des Données

NOM Prénom :
FRANK Shir

Branche :
Informatique et Systèmes d'Information - VDC

Responsable Pédagogique :
MATTA Nada

Semestre :
Printemps 2023

Résumé (150 mots)

Mon stage s'est déroulé au sein de l'association Gaea21 en tant que Lead développeuse, dans le département IT.
Ma mission consistait dans le développement Full-Stack du site web du projet "Répertoire Vert", à l'aide des frameworks Symfony et ReactJS. En tant que chef de projet, je devais également gérer l'équipe et avoir une vision globale des tâches.
La gestion du projet se faisait selon la méthode Agile, avec différentes phases répétées cycliquement :
Recueil des besoins du responsable de l'association, Mise en place d'un planning et répartition des tâches dans l'équipe, Mise en oeuvre, Réunion hebdomadaire avec le responsable et l'équipe et Déploiement de la version du site en cours.
Le but est de livrer une version du site web fonctionnelle permettant le référencement de tous les produits et/ou services professionnels verts proposés dans une région définie, destinée pour le moment aux entreprises.

Entreprise : Deutsche Bundesbank

Lieu : ***Mainzer Landstraße 16, 60325 Francfort-sur-le-Main, Allemagne***

Responsable : Stefan BENDER

Mots clés (cf Thésaurus) :

- **Mise en place, mise en oeuvre**
- **Services des organismes financiers**
- **Informatique**
- **Analyse des données - Logiciels**

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

DEUTSCHE
BUNDESBANK
EUROSYSTEM

Remerciements

Je tiens tout d'abord à remercier Yvan, le responsable de l'association, ainsi que l'équipe RH de Gaea21 pour m'avoir recrutée au sein de l'association, et en particulier Charlotte Boll, ma coach RH, pour sa gentillesse, sa bonne humeur et son perfectionnisme.

Merci aussi à Alexandre Nguyen et Jonathan Capitao, mes tuteurs, pour leur aide quotidienne dans les tâches que je devais réaliser.

Sans oublier mon équipe de développeurs du Répertoire Vert, avec qui la cohésion était présente et qui étaient motivés pour atteindre nos objectifs, et mes autres collègues toujours là pour aider.

Sommaire

Introduction

La Deutsche Bundesbank est la banque centrale de la République fédérale d'Allemagne. Son rôle principal est de maintenir la stabilité des prix et la valeur de l'euro. Elle est également chargée de superviser les banques et les institutions financières en Allemagne, d'assurer la sécurité et l'efficacité des paiements et des opérations financières, de gérer les réserves de change du pays, et de participer à la formulation et à la mise en œuvre de la politique monétaire européenne en tant que membre de l'Eurosystème.
Elle est donc importante au niveau national, mais également international :
- Elle est membre de la Banque des règlements internationaux (BRI) et participe activement aux forums internationaux tels que le G20 et le Fonds monétaire international (FMI).
- Au niveau européen, elle est un membre important de la Banque centrale européenne (BCE) et joue un rôle clé dans la mise en œuvre de la politique monétaire de la zone euro.

La banque compte 30 filiales, comptant la centrale et les antennes régionales. Ces dernières mettent en œuvre les missions de la banque centrale allemande au niveau régional et sont responsables de la coordination avec les acteurs économiques locaux et les autorités régionales pour favoriser le développement économique régional.

Mon stage s'est déroulé à la Bundesbank Zentrale à Francfort-sur-le-Main, le siège principal de la banque. Il est le centre décisionnel et opérationnel de la banque et coordonne les activités des filiales régionales.

Je travaille dans le département DSZ (Datenservicezentrum) faisant partie du département Statistique, et dirigé par Stefan Bender. Son rôle est de fournir un accès aux données statistiques produites par la Bundesbank, ainsi qu'à d'autres sources de données pertinentes pour la politique monétaire et la supervision bancaire.
Plus spécifiquement, la DSZ offre les services suivants :

- Collecte et traitement des données provenant de différentes sources (banques centrales, agences gouvernementales et organisations internationales)
- Diffusion des données sous forme de tableaux, graphiques, rapports et bases de données interactives, accessibles aux décideurs politiques, chercheurs et analystes.

Il est composé de plusieurs sous-unités :

- DSZ 10 et 20 pour la recherche qui utilisent les données collectées par les autres sous-unités
- DSZ 30 chargé des tâches de traitement de données avec Machine-Learning, Data Engineering et la mise en place de la plateforme de données interne
- DSZ 40 pour la collecte de données
- DSZ 1 (Sustainable Finance Data Hub) travaille en étroite collaboration avec la DSZ 40

Plus précisément dans la DSZ, je fais donc partie de l'équipe du Sustainable Finance Data Hub (SFDH), une unité spécialisée qui collecte et enrichit les données liées au climat pour soutenir la prise de décision optimale en matière de changement climatique. Il fournit un point d'accès central aux données, répond aux questions méthodologiques et prend part à des projets de génération de données innovants.

I Sujet et place dans le service

A Sujet

A.1 Sujet défini avant mon arrivée

Avant le début de mon stage, il était prévu que je travaille sur un projet innovant visant à développer une preuve de concept pour extraire des informations numériques sur les émissions de carbone à partir de rapports de durabilité en pdf. Pour ce faire, nous devions utiliser des technique de Natural Language Processing en python.

Le soutien aux tâches en cours afin de connaître toutes les fonctions clés de l'équipe faisait également partie intégrante du stage. Cela aurait pu inclure la documentation des données, la liaison et la comparaison des ensembles de données pour la gestion de la qualité des données et la participation aux groupes de travail internes et externes.

A.2 Sujet réel

Mes tâches en réalité étaient plus variées et sur divers projets :

- Crawling de résultats de recherche Google pour récupérer les rapports de durabilité des entreprises, et extraction d'informations des PDF
- Optimisation de code Python pour la génération et comparaison de gros jeux de données
- Rapprochement de jeux de données provenant de différentes sources avec R (analyses)
- Génération automatisée d'un rapport à partir de résultats stockés dans un fichier csv
- Projet de recherche NLP pour trouver quels datasets et méthodologies sont utilisé(e)s dans quels papiers de recherche

B Fonction occupée dans le service

Ma fonction dans le SFDH, et plus généralement dans la DSZ était celle de stagiaire Data Scientist, aux côtés d'un autre stagiaire occupant exactement la même fonction. Comme détaillé dans le sujet, des tâches diverses en rapport avec les données m'ont été assignées sur de nombreux projets.

Le point commun entre toutes mes tâches était qu'à partir d'un gros volume de données, je devais en créer de nouvelles qui soient utilisables pour le but souhaité et compréhensibles par les différents corps de métier présents à la Bundesbank. Autrement dit, il s'agissait de tâches de valorisation des données. Pour cela, je passais par diverses méthodes de traitement de données à l'aide de Python ou R. J'ai parfois effectué le travail de collecte de données, notamment pour le projet GAIA (crawling).

Les projets ne se restreignaient pas au SFDH, ils concernaient aussi (et pour la plupart) d'autres services de la DSZ, notamment ceux dans la recherche : Certaines tâches m'ont alors été attribuées non pas par mon tuteur, mais par d'autres collègues d'autres services.

II Explication des projets

A Projet Gaia

A.1 Contexte

Le projet GAIA vise à simplifier l'utilisation des rapports de durabilité des entreprises grâce à une plateforme centralisée et à faciliter la recherche en réduisant les efforts manuels. Il cherche également à créer une référence fiable pour les informations liées au climat.

Ce projet a été lancé par la Deutsche Bundesbank en réponse à la prise de conscience croissante de l'importance de la finance durable. En effet, des risques financiers en découlent, et il faut donc en avoir conscience et savoir les gérer.

La motivation derrière le projet GAIA est de fournir aux investisseurs et aux décideurs des informations fiables et cohérentes sur la durabilité des investissements. Il s'agit de donner aux investisseurs les moyens de prendre des décisions d'investissement plus éclairées et de mieux comprendre les risques et les opportunités associés aux investissements durables.

La Bundesbank a également pour objectif de promouvoir la stabilité financière en intégrant les risques liés au changement climatique dans ses analyses et en encourageant les investisseurs à intégrer ces risques dans leur propre prise de décision.
L'équipe du projet est composée de membres de la Bundesbank bien sûr, mais se fait surtout en collaboration avec d'autres banques internationales (Banque d'Espagne, Banque Centrale Européenne, etc...).

Concrètement, le projet consiste tout d'abord à récupérer les rapports de durabilité des entreprises, disponibles publiquement en ligne. Ensuite, le texte et les tableaux / graphiques sont extraits, et les données d'empreinte carbone des entreprises (Scope 1, 2, 3) sont extraites à l'aide de Natural Language Processing (NLP).

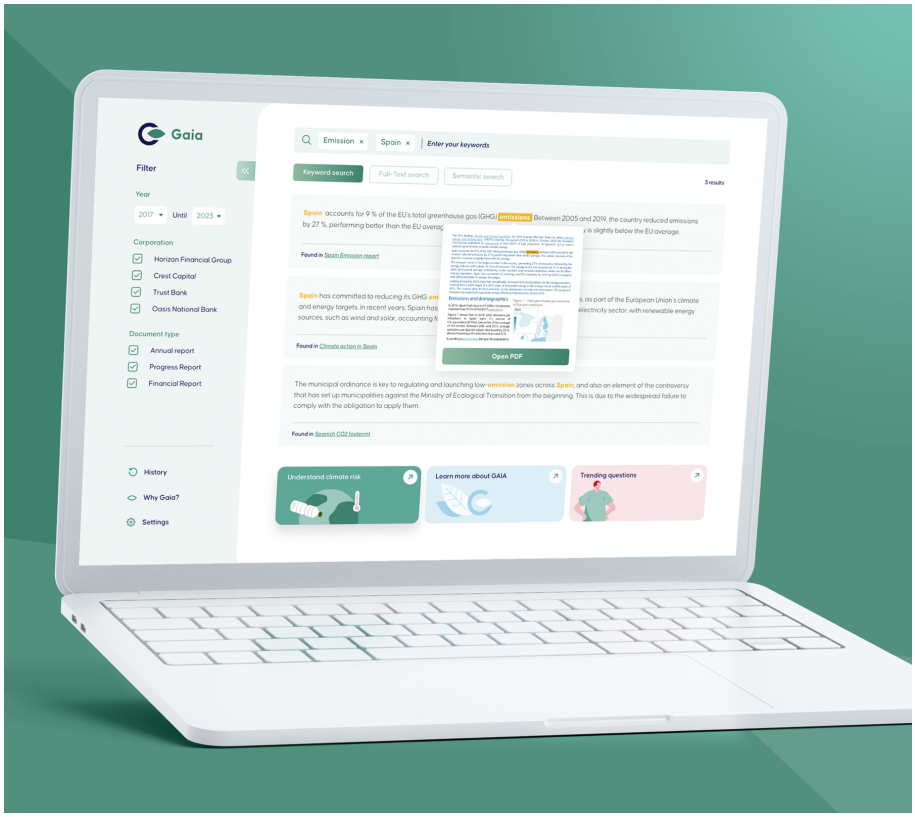

FIGURE 1 – Mock-up de la plateforme Gaia

A.2 Antécédents

Mises à part les spécifications et objectifs de la plateforme qui étaient dejà clairs, rien n'avait été concrètement commencé avant mon arrivée.

A.3 Objectif visé

L'objectif est de présenter une preuve de concept du projet. Mon principal objectif était donc de développer un crawler qui parcourerait les résultats de recherche Google pour les rapports de durabilité des entreprises majeures dans le monde. Les PDFs trouvés doivent ensuite être téléchargés. Parallèlement, les PDFs/liens trouvés doivent être classés, selon s'ils sont justes ou faux pour une requête donnée. Une fois que cela serait fait, je devrais présenter mes résultats à l'équipe du projet.

B Projet ESCB Exchange

B.1 Contexte

ESCB signifie Eurosystem/European System of Central Banks. Cela se réfère au système européen de banques centrales. L'Eurosystem est composé de la Banque centrale européenne (BCE) et des banques centrales nationales des pays de la zone euro.

L'ESCB Exchange est un forum permettant aux banques centrales de collaborer autour de données climatiques communes. Ce forum favorise l'échange de connaissances, de code et d'applications pour améliorer la valeur analytique des données climatiques utilisées. Chaque banque centrale désigne jusqu'à 2 experts en données climatiques pour participer au forum. Les rencontres du forum ont lieu chaque mois, et des groupes volontaires et agiles pilotent les sujets d'intérêt commun. L'initiative encourage l'échange sur le nettoyage des données, le partage de code et les leçons apprises concernant les données climatiques granulaires (des entreprises).

L'un des sujets abordés lors des forums est notamment le rapprochement de données sur les émissions carbones, censées être identiques mais provenant de différentes sources. En effet, les banques collectent ces données et les utilisent pour des recherches. Il est donc primordial de choisir la meilleure source de données climatiques. Ma mission portait exactement là-dessus.

B.2 Antécédents

Concernant le rapprochement de données provenant de différentes sources, un code d'analyses de données en R ainsi qu'une présentation présentant la comparaison avaient déjà été faits. La comparaison concernait les sources ISS et Carbon 4 Finance, tous deux fournisseurs de données.

B.3 Objectif visé

Ma mission consistait en la même analyse que celle faite précédemment, sauf qu'au lieu de faire une analyse bilatérale, je devais en faire une trilatérale en ajoutant aux sources précédentes celle de TRUCOST. Le Powerpoint associé devait également être fait en aval.

FIGURE 2 – Logos des sources de données à rapprocher

C Projet CSDB

C.1 Contexte

CSDB signifie "Central Securities Database". Il s'agit d'une base de données centrale des titres en Allemagne, utilisée pour la surveillance des marchés financiers, le suivi des transactions de titres et la facilitation des processus de règlement-livraison. Au sein de la DSZ, les chercheurs utilisent cette base de données.

Les datasets sont sous la forme de fichiers CSV et il y en a 1 pour chaque mois de chaque année depuis 2009. Tous les mois environ, une nouvelle version d'un ou plusieurs de ces datasets peut parfois être générée, avec des nouvelles colonnes, nouvelles valeurs, etc... Autrement dit, une nouvelle version du dataset de 2022-04 (différente de l'original) peut être créée en Avril 2023.

Pour chaque dataset, 4 variantes doivent être générées : En effet, des normes existent, et chaque groupe de chercheurs possède des droits de lecture différents sur les colonnes du dataset. Par conséquent, chaque variante possède le même contenu dans ses colonnes, mais les colonnes présentes sont différentes. Et cela doit être fait dès qu'une nouvelle version d'un dataset existe. Or, chaque dataset fait plusieurs Go, et actuellement, cette génération des 4 variantes ainsi que la comparaison de chaque nouveau dataset avec sa version originale prend 1 semaine.

Quand la procédure doit être répétée tous les mois, cette solution n'est pas viable telle quelle et doit être optimisée.

C.2 Antécédents

La situation à mon arrivée était la suivante :

Les nouvelles et les anciennes versions du "même" ensemble de données (par exemple les données 2009-04) étaient comparées après la génération des 4 ensembles de données. Cela engendre beaucoup d'effort et de temps de calcul évitable.

Les fichiers csv ont été compressés afin de tenter une baisse du temps, ce qui a effectivement beaucoup accéléré le temps de lecture des csv. Cependant le temps de compression semblait fortement ralentir le processus.

La comparaison des datasets comportait trop d'informations (moyenne, médiane de chaque colonne des datasets, et bien d'autres valeurs statistiques), ce qui utilisait très certainement du temps supplémentaire inutile, quand seulement un très bref résultat était attendu.

C.3 Objectif visé

Afin de pouvoir générer les 4 versions de datasets uniquement sur les versions modifiées des jeux de données de la même période, la comparaison doit être réalisée en amont. De plus,

la comparaison doit être optimisée et n'indiquer à la fin que si les datasets sont identiques ou non, et sinon montrer les différences.

D Projet NFiG

D.1 Contexte

NFiG signifie "pour usage interne seulement" (Nur für internen Gebrauch) et concerne les données des chercheurs du département. Ces derniers utilisent Stata, un logiciel de statistiques pour générer des graphiques ou effectuer différentes analyses. Lorsque ces graphiques doivent être publiés, ils doivent être adaptés de manière à ne plus être confidentiels.

Par conséquent, chaque graphique doit être retrouvé dans le code qui l'a généré, ce qui implique de connaître le fichier stata et la ligne à laquelle l'output est généré.

Le but est donc de retrouver automatiquement où est créé chaque graphe dans le code, ou de manière plus générale, de faciliter leur "traçage". Or, les exportations peuvent être générées par des boucles, ce qui rend la recherche des noms de fichiers difficile.

D.2 Antécédents

Un code python avait déjà été écrit pour tenter d'automatiser cette recherche, mais n'a pas vraiment réussi. Des rapports Excel on été faits à la main afin de recenser quel graphique ou autre fichier output a été généré dans quel fichier de code / fichier log et à quelle ligne.

D.3 Objectif visé

La tâche qui m'a été donnée

E Projet de recherche NLP

E.1 Contexte

E.2 Antécédents

E.3 Objectif visé

III Déroulement du travail

A La méthodologie

Durant toute la durée du stage, il y avait un suivi régulier, permettant une gestion de projet selon la méthode agile.

Ainsi, cela fonctionnait par cycles répétés :
Tous les jeudis, une réunion avec l'ensemble du département IT et le client (Yvan, le responsable de l'association) avait lieu afin de recueillir ses besoins pour le site.
Si nécessaire, un poker planning était réalisé pour répartir les tâches efficacement selon les difficultés et le temps prévu.

Tous les jours, une réunion de suivi avec le tuteur et l'équipe permettait de partager notre avancée et obtenir de l'aide sur des points de blocage.

Tous les mardis, une réunion avec le département Design permettait de mettre au clair certains points ou de faire des commandes, car le développement d'un site est étroitement lié au design.

Une fois toutes les 2 semaines, une réunion transversale pour le projet Répertoire Vert avait lieu.
Pour finir, tous les lundi et jeudi, une mise en ligne du site était nécessaire pour garder la version en ligne toujours à jour et présentable au client.
D'un point de vue technique, il fallait mettre à jour le site sur un serveur Linux distant à l'aide de Git, un logiciel de gestion de versions que j'ai particulièrement utilisé durant ce stage.

Justement, il était important de faire des **commits et push** au moins 1 fois par jour, pour garder un contrôle des versions du site. Chaque membre de l'équipe possédait une branche à son nom. Plusieurs fois par semaine, un merge des toutes les branches sur la branche de développement était réalisé, avant de faire un merge sur master de la branche de Développement. En effet, la branche master est celle utilisée pour la mise en production, elle doit donc être propre, fonctionnelle et régulièrement mise à jour.

D'autre part, il ne fallait pas négliger le remplissage les outils de suivi, à savoir :
- **le journal de bord** pour les heures et les tâches réalisées par jour/semaine/mois
- **le tableau des tâches** pour organiser les tâches à faire/en cours/faites, et répertorier le temps mis pour chacune d'entre elles.

(voir Annexes)

Des points début, mi-stage et fin de stage étaient réalisés avec le tuteur et la coach RH pour vérifier le remplissage des outils et voir si tout allait bien.

B Application de la méthode et Résultats

B.1 Projet Gaia

B.2 Projet ESCB Exchange

B.3 Projet CSDB

B.4 Projet NFIG

B.5 Projet de recherche NLP

C Planning général suivi

Conclusion

Durant ce stage chez Gaea21, j'aurai donc réussi avec mon équipe à terminer les principales fonctionnalités de la version 0 du site Répertoire Vert.

Ainsi, une entreprise peut à présent :

- S'inscrire et se connecter
- Visualiser son profil et modifier ses informations
- Désactiver son compte
- Ajouter des produits et/ou services
- Visualiser ses produits/services, les supprimer et les modifier
- Visualiser ses statistiques, comme le nombre de clics sur ses produits, sans bug
- Parcourir les différentes catégories et sous-catégories, et voir la liste des entreprises appartenant à chaque sous-catégorie.
- Visualiser les profils des autres entreprises ainsi que leurs produits/services

Mes compétences se sont nettement améliorées en Git, que je sais maintenant utiliser en ligne de commande. Je sais désormais mettre un site en production, et coder avec Symfony et ReactJS.

J'ai pu apprendre à gérer une équipe et un projet avec des deadlines serrées, en utilisant des outils de gestion.

Mes collègues étaient pour la plupart plus âgés que moi, mais j'étais la plus ancienne sur le projet (à cause d'un turn-over important). De ce fait, je connaissais assez le projet pour pouvoir les accompagner. Mais la différence d'âge ne se ressentait pas, travailler avec eux était très fluide. Il y avait toujours un climat d'entraide et beaucoup de communication dans l'équipe.

Les réunions journalières avec mon tuteur et l'équipe me permettait de suivre leur avancée, et additionnellement si besoin, nous nous appelions avec un ou plusieurs membres, par exemple s'ils avaient besoin d'aide ou que quelque chose devait être rectifié. Nous communiquions beaucoup par Skype. Aussi, je mettais en commun notre travail plusieurs fois par semaine, ce qui me permettait de bien suivre la progression du projet.

L'outil de gestion que j'utilisais pour planifier était un tableau Google Sheet présentant les tâches, leur difficulté, et leur durée estimée que je remplissais avec mon équipe.

Avoir une certaine responsabilité était très formateur et m'a donné une idée du métier de Lead Programmer, qui me plairait d'ailleurs beaucoup. Cependant, je souhaiterais m'orienter dans le domaine de l'IA et du Big Data, c'est pourquoi je choisirai la spécialité Valorisation des Connaissances pendant mon cursus.

Bibliographie

[1] Vladimir Agafonkin. *Leaflet - a JavaScript library for interactive maps*. [en ligne]. Mis à jour le 4 septembre 2020 [Consulté le 14 décembre 2021]. URL : https://leafletjs.com/.

[2] Massimiliano Arione. « KnpLabs/KnpPaginatorBundle : SEO friendly Symfony paginator to sort and paginate ». In : *Github*. [en ligne]. Mis à jour le 2 décembre 2021 [Consulté le 17 décembre 2021]. URL : https://github.com/KnpLabs/KnpPaginatorBundle.

[3] Jeff Atwood et Joël Spolsky. *Stack Overflow - Where Developers Learn, Share, and Build Careers*. [en ligne]. Mis à jour le 30 décembre 2021 [Consulté le 17 décembre 2021]. URL : https://stackoverflow.com/.

[4] Ghaida Bouchâala. « Transform your excel data into a relational database : Symfony | Medium ». In : *Ghaida Bouchâala – Medium*. [en ligne]. Mis à jour le 5 juillet 2020 [Consulté le 22 novembre 2021]. URL : https://ghaidabouchala.medium.com/import-excel-data-in-the-database-symfony-back-end-e14efea51cd2.

[5] Doctrine. *Welcome to Doctrine 2 ORM's documentation ! - Doctrine Object Relational Mapper (ORM)*. [en ligne]. Mis à jour le 21 décembre 2021 [Consulté le 17 décembre 2021]. URL : https://www.doctrine-project.org/projects/doctrine-orm/en/2.10/index.html.

[6] Facebook. *React – Une bibliothèque JavaScript pour créer des interfaces utilisateurs*. [en ligne]. Mis à jour le 22 mars 2021 [Consulté le 8 novembre 2021]. URL : https://fr.reactjs.org/.

[7] Flickity. *Flickity · Touch, responsive, flickable carousels*. [en ligne]. Mis à jour le 19 décembre 2021 [Consulté le 2 décembre 2021]. URL : https://flickity.metafizzy.co/.

[8] Grafikart. *Tutoriels et Formations vidéos sur le développement web*. [en ligne]. Mis à jour le 28 décembre 2021 [Consulté le 18 août 2021]. Disponible sur : URL : https://grafikart.fr/.

[9] Theo Lampert. « react-flickity-component - npm ». In : *npm*. [en ligne]. Mis à jour le 30 août 2021 [Consulté le 8 novembre 2021]. URL : https://www.npmjs.com/package/react-flickity-component.

[10] Symfony SAS. *Symfony, High Performance PHP Framework for Web Development*. [en ligne]. Mis à jour le 29 décembre 2021 [Consulté le 17 décembre 2021]. URL : https://symfony.com/doc/current/index.html.

[11] SymfonyCasts. *SymfonyCasts - PHP and Symfony Video Tutorial Screencasts*. [en ligne]. Mis à jour le 23 décembre 2021 [Consulté le 23 août 2021]. URL : https://symfonycasts.com/.

[12] Alexander Weissman. *Getting Started | Select2 - The jQuery replacement for select boxes*. [en ligne]. Mis à jour le 3 mai 2020 [Consulté le 28 septembre 2021]. URL : https://select2.org/.

Table des figures