Mémoire de PFE

Formation FIL

IMT Atlantique

Amélioration des performances de détection automatique d'anomalies dans des logs applicatifs

Oscar Gloaguen

Direction Générale des Finances Publiques

Table des matières

Remerciements

Je souhaite remercier la Direction Générale des Finances Publiques et tout particulièrement Olivier Blanc, pour m'avoir permis d'effectuer mon alternance dans cette équipe et pour m'avoir accompagné tout au long jusqu'au PFE.

Je tiens a remercier personnellement Robin Gries, qui a su être un vrai coéquipier tout au long de ce projet malgré sa complexité.

Je voudrais aussi remercier IMT Atlantique ainsi que ses professeurs, et surtout mon tuteur pédagogique Thomas Ledoux qui s'est mis a ma disposition pour le PFE.

Résumé

Les logs sont une source importante de données détaillant le fonctionnement interne d'une application, mais ne sont pourtant que rarement utilisés a leur plein potentiel. Dans ce mémoire, je vais détailler le processus d'évolution d'un outil d'apprentissage automatique qui utilise les logs pour détecter et même tenter de prévoir des anomalies logicielles. Le projet s'apparentant plus a un projet de recherche, la démarche scientifique sera détaillée, ainsi que les caractéristiques techniques et le déroulement de l'implémentation. L'organisation du projet avec un stagiaire et moi-même sera développée, ainsi que ses impacts humains et économiques.

Abstract

Logs are an important source of data when it comes to the internal workings of software, but they are rarely used to their full potential. In this memoir, I will explain the evolution of a machine learning tool which uses logs to detect and even attempt to predict software anomalies. The project being similar to a research project, the scientific protocol will be detailed, as well as the technical characteristics and the course of the implementation. Project management with an intern and myself will be developed, as well as the human and economic impacts of the project.

Mots-clés traitement automatique du langage (TAL), apprentissage automatique, apprentissage profond, analyse de logs applicatifs

Introduction

1-2 pages

Le projet `analyse-logs` a été développé par deux précédents apprentis de la Direction Générale des Finances Publiques (DGFiP), Rémi puis Léa. Cependant, les performances des algorithmes implémentés n'étant pas satisfaisantes, c'est ici que nait le sujet de ce PFE. Ce mémoire détaille le processus d'évolution de ce projet dans l'objectif d'amélioration des performances.

J'ai travaillé sur cette problématique en binôme avec Robin, un stagiaire en dernière année de master. Ce document touchera aussi sur l'organisation du sujet entre nous ainsi que les bénéfices et difficultés a travailler en équipe.

Ce sujet de PFE s'intègre dans la stratégie d'innovation du SI de la DGFiP. L'objectif est de montrer l'efficacité de ces outils, pour mettre en valeur l'innovation et pousser leur utilisation au sein des bureaux. Dans ce cadre, une structure proche de celle d'un projet de recherche a été suivie. Pour cela, nous avons d'abord composé et étudié un état de l'art des algorithmes de détection d'anomalies existants. Ils utilisent pour la plupart des méthodes d'apprentissage automatique, avec des algorithmes classiques ou de l'apprentissage profond.
[1]

Chapitre 1

La DGFiP, administration en évolution constante

La DGFiP est une administration française née de la fusion de la Direction Direction Générale des Impôts (DGI) et de la Direction Générale des Comptes Publics (DGCP) (DGCP) en 2008. Elle hérite alors des missions des deux entités, en faisant un service public très étendu, en charge notamment de la collecte des impôts et taxes et de la législation fiscale.

Cette administration possède une hiérarchie forte séparée en 8 services, qui définit leur domaine de travail, ainsi que différentes directions (voir organigramme 1 en annexes). Une majorité des services sont des services métiers, directement en lien avec les missions de la DGFiP, mais 3 de ces services sont des services support, ou transverses. Ces derniers sont le service des Ressources Humaines, le service de Stratégie, Pilotage et Budget, et le Service des systèmes d'Information (SI). Malgré une interaction indirecte avec le domaine métier des dinances publiques, ils répondent aux besoins de la DGFiP en permettant le bon fonctionnement des autres services, ou même en améliorant leur performance.

Comme toute grande structure aujourd'hui, la DGFiP possède un besoin très fort en technologies de l'information, qui est rempli par le SI. Ce service est indispensable, car de nombreuses missions de la DGFiP reposent sur des programmes (e.g., calcul et déclarations d'impôts et des taxes) permettant de traiter de larges quantités de données en un temps restreint. Les premières versions de ces applications datent des années 80, et ont pour la plupart évolué et sont restées utilisées jusque aujourd'hui. L'informatique est donc au centre de cette administration, autant pour les agents en interne que pour les utilisateurs externes.

Le SI est séparé en de nombreux bureaux (voir annexe 2). Il comprend lui même des bureaux transverses qui facilitent le bon fonctionnement des autres bureaux. Le reste des bureaux est regroupé sous la Direction des projets numériques (DPN), et sont chargés du pilotage, du développement et de la maintenance d'applications d'un domaine précis.

Cette nouvelle hiérarchie date de 2021, où une réorganisation a eu lieu. En effet, la

DPN n'existait pas avant cela, et les bureaux étaient regroupés sous deux directions, "étude et développement" et "production". Le SI comporte aujourd'hui plus de bureaux, qui sont donc plus spécialisés, avec par exemple des bureaux en charge d'une unique mission importante.

Le Bureau du SI des professionnels (BSI-3) est un bureau de la DPN chargé de la fiscalité des professionnels. Il a a sa charge une dizaine d'applications qu'il spécifie, développe et maintient. L'une d'entre elles est ME-ca-ni-sa-tion Des Opé-ra-tions Comp-tables (MEDOC) qui est une application d'encaissement d'impôts et de gestion de comptabilité de l'État. Le BSI-3 possède une mission particulière de modernisation de cette application, tâche très complexe étant donné son échelle et sa complexité.

Chapitre 2

L'existant : le projet `analyse-logs`

2.1 Contexte du projet

2.2 Blocs logiciels

DeepLog

Autres algorithmes

Blockly

Chapitre 3

Le processus scientifique au cœur du projet

3.1 État de l'art de l'analyse de logs

3.2 Analyse des jeux de données disponibles

3.3 Organisation des tâches

Chapitre 4

Développement du nouveau projet

Chapitre 5

Projection de l’impact du projet

Conclusion

Bibliographie

[1] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog : Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1285–1298, New York, NY, USA, 2017. Association for Computing Machinery.

Glossaire administratif

BSI-3 Bureau du SI des professionnels : Bureau dépendant de la DPN, chargé du développement et de la maintenance des applications du domaine de la fiscalité des professionnels.

DGCP Direction Générale des Comptes Publics (DGCP) : Aussi appelée «trésor public», ancienne administration chargée de la gestion des comptes de l'État et du recouvrement des impôts.

DGFiP Direction Générale des Finances Publiques : Service public de l'État rattaché au ministère des finances, chargé des missions de gestion publique et de fiscalité.

DGI Direction Générale des Impôts : Ancienne administration chargée de la liquidation des impôts.

DPN Direction des projets numériques : Direction au sein du SI regroupant les différents bureaux en charge de la direction et de la réalisation de projets dans le numérique.

MEDOC ME-ca-ni-sa-tion Des Opé-ra-tions Comp-tables : Application traitant l'encaissements des taxes (telles que la TVA) ainsi que la génération d'écritures comptables pour l'État.

SI Service des systèmes d'Information : Service support de la DGFiP chargé de la gestion informatique et du développement d'applications.

transverse Utilisé fréquemment a la DGFiP comme synonyme de transversal, dans le sens “recoupant plusieurs disciplines ou secteurs”.

Glossaire technique

apprentissage automatique Ou *machine learning* en anglais, ensemble de méthodes visant a développer des algorithmes généraux basés sur l'apprentissage de données, s'opposant a un algorithme explicite classique.

apprentissage profond Branche de l'apprentissage automatique se basant sur des réseaux de neurones artificiels, comportant plusieurs couches de traitement de données permettant d'extraire des caractéristiques complexes.

log De l'anglais log (journal), sortie d'une application (souvent un fichier) représentant le chemin d'exécution d'une application.

TAL traitement automatique du langage : Ensemble de méthodes visant à analyser et traiter le langage naturel.

Annexes

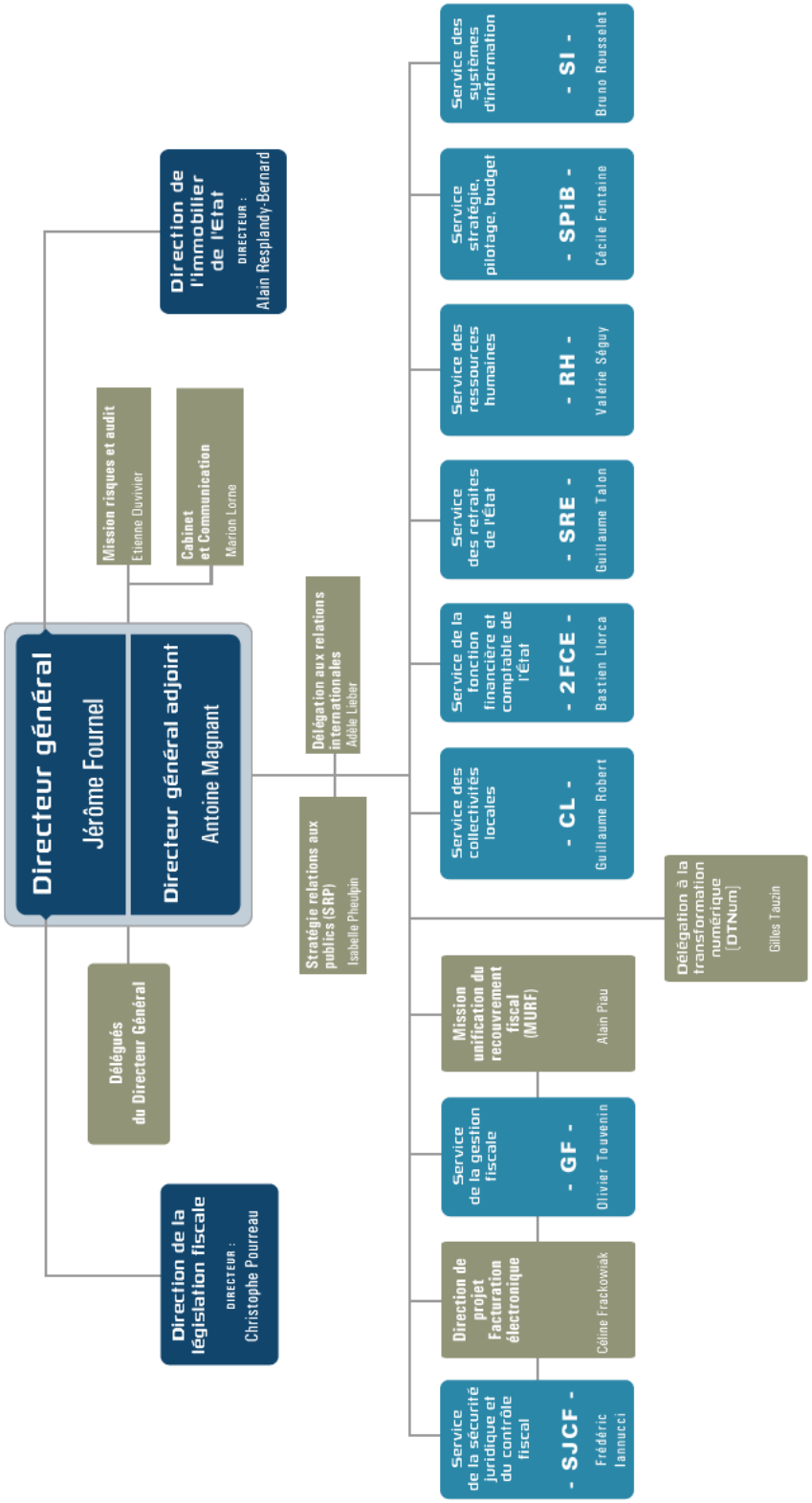

FIGURE 1 – Organigramme des services de la DGFiP

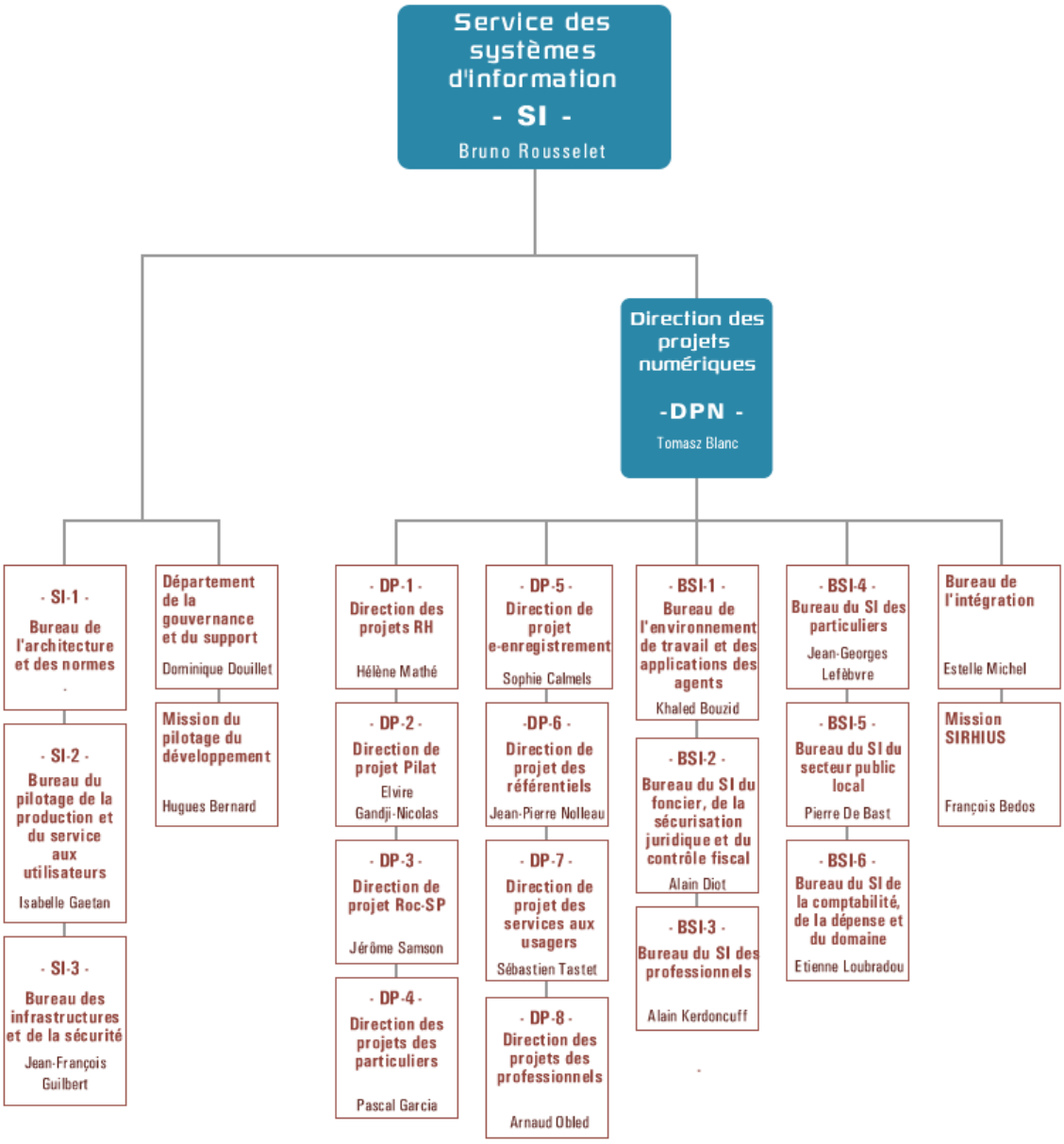

FIGURE 2 – Organigramme des bureaux du SI