INGÉNIEUR EN MATHÉMATIQUES APPLIQUÉES ET MODÉLISATION SPÉCIALITÉ GESTION MULTIMÉDIAS DES DONNÉES MASSIVES

RAPPORT DE SECOND SEMESTRE D'APPRENTISSAGE

SUJET
Valorisation des méthodes de Data Science pour la Caisse Nationale des Allocations Familiales (CNAF)

Apprenti: Nicolas TURPAULT (MAM5 - GMD)

Maître d'apprentissage : Clémentine NEMO / Luc CHAMPALLE

Tuteur Enseignant : Lionel FILLATRE

Entreprise : ATOS

20/02/2017 - 22/09/2017

Résumé

Ce rapport rapporte la seconde période d'apprentissage et fait suite à un précédent rapport discutant la première période.
Lors de ce rapport, on discutera des idées ainsi que du travail mis en place pour améliorer l'outil de gestion des incidents de la CNAF. Ceci passe tout d'abord par un travail d'étude des données, suivi de présentations de plusieurs idées avant l'acceptation et la réalisation de celles-ci. La réalisation finale est un algorithme permettant l'aide à la décision afin de gagner du temps à la réalisation des solutions aux problèmes. La solution utilisant des réseaux de neurones sera justifiée en la comparant avec des approches plus classiques.

Remerciements

Premièrement, j'aimerais remercier l'entreprise ATOS de Sophia Antipolis de m'avoir accueilli lors de cet apprentissage. Merci à tous les collaborateurs qui ont participé au bon déroulement de cet apprentissage. Tout le monde a été très accueillant, prêt à aider et cela m'a permis de découvrir de façon très positive la vie au sein d'une multinationale comme Atos.

Un remerciement particulier à tous les membres de l'équipe travaillant sur le portail partenaire de la CNAF qui m'ont intégré de façon très rapide au sein de leur équipe.

De plus, je remercie Gabriel KEPEKLIAN (Directeur en Recherche et Développement sur le site d'Atos à Bezons) et Olivier CORBY (enseignant chercheur à l'INRIA) de m'avoir aidé techniquement lors de la première partie de mon alternance.

Je tiens à remercier Kévin ROYERE pour sa disponibilité et son écoute lors de mon alternance. Il m'a permis de m'intégrer aux équipes, de comprendre le bon fonctionnement de l'entreprise et de découvrir un nouveau domaine m'apportant diverses compétences lors de la première partie de mon apprentissage.

Un remerciement particulier à Alain CIVEL, directeur technique du site, qui m'a donné quelques travaux d'analyses de données simples ainsi que la mise en place de réunion avec le client afin de présenter mes compétences et travaux. Il a su me faire confiance et me présenter au client pour déboucher à la mise en place de ce sujet de deuxième partie d'alternance.

L'investissement et l'aide de la CNAF ont été très importants pour la réalisation de mon travail. Je tiens particulièrement à remercier Laurence SALAGNAT, Christophe BACILIERE et André DEWRDT pour leur aide, leur confiance et leur soutien lors de mon projet. Je tiens aussi à remercier Stéphane DONNE et Pierre COLLINET de la direction statistique de la CNAF qui ont été mes interlocuteurs techniques et toujours de bons conseils.

Je tiens aussi à souligner la disponibilité de Luc CHAMPALLE (directeur des projets CNAF à ATOS) qui a été d'une écoute et de conseils importants à la réalisation de cette alternance. Je le remercie d'avoir assuré la place de tuteur en cette fin d'alternance.

Finalement, je tiens à remercier Clémentine NEMO, ma tutrice de cette alternance. Elle m'a été d'une aide précieuse au bon déroulement de cet apprentissage par sa disponibilité et ses critiques positives et négatives à l'égard de mon travail. Je tiens aussi à souligner ses précieux conseils lors de mon alternance ainsi que pour ma poursuite professionnelle. Je la remercie pour sa disponibilité et la confiance qu'elle a su m'accorder. J'aurai souhaité aboutir mon alternance avec elle, et je la félicite pour cet heureux événement.

Plan du rapport

Introduction

Atos Sophia est au cœur d'une multinationale, elle offre des services informatiques aux entreprises.
Dans le cadre d'un marché public avec la CNAF, ATOS prend en charge la maintenance des applications du domaine « Gestion de la relation allocataires et échanges partenaires ». Plus de 50 personnes travaillent à Atos sur ce domaine.

Atos est prestataire de service pour la CNAF depuis de nombreuses années. Afin de garder la relation commerciale et face à la concurrence, il est vital d'être force de proposition. C'est dans ce cadre que mon alternance a été proposée, afin d'apporter un axe « data science » qui n'a jamais été proposé auparavant.

Les premiers mois de mon apprentissage ont été découpés en deux tâches simultanées :

- Aide à l'analyse fonctionnelle pour la CNAF.
  Ces travaux m'ont permis d'acquérir diverses compétences qui ne sont pas toutes dans mon domaine d'étude. Ceci va de compétences fonctionnelles à des compétences en web sémantique que j'ai étudié cette année à Polytech. Ces travaux ne sont pas développés dans ce rapport.

- Valorisation des méthodes de data science à Atos, dans le contexte CNAF.

La deuxième partie de mon apprentissage sera totalement consacrée à cette deuxième tâche.

Valorisation des méthodes de data science à Atos, dans le contexte CNAF :

La data science n'est pas encore un domaine de compétence développé par l'agence Atos de Sophia-Antipolis. Dans ce contexte mon apprentissage est l'occasion pour Atos de fédérer les collaborateurs à la data science et de démarcher les clients sur de nouveaux axes de réflexion.

Dans le contexte du marché avec la CNAF, le premier sujet de travail portait sur la valorisation des données de l'open-data CNAF. Les 5 premiers mois m'ont permis d'aborder ce sujet mais aussi de faire d'autres propositions plus intéressantes pour Atos et la CNAF (architecte, commerciaux, …).
Lors de la première partie d'apprentissage présentée en détail lors d'un premier rapport, j'ai présenté divers travaux à la CNAF :

- Une compréhension de l'Open Data utilisant des visualisations diverses.
- Une proposition d'amélioration de l'Open Data grâce au web sémantique
- Une analyse des tweets mentionnant la CNAF ou les CAF (caisse d'allocations familiales)

Lors de la présentation de mes travaux à la CNAF, l'analyse de tweets, et donc de textes à susciter leur curiosité. Ils n'ont pas souhaité continuer à travailler sur l'Open Data étant

donné qu'une équipe de la CNAF travaille déjà à son amélioration. Au vu de mes compétences, ils m'ont proposé de faire un travail de recherche sur les données de leur outil de gestion des incidents (ces données sont essentiellement du texte) sur lequel ils ont actuellement une surcharge de travail et des améliorations semblent possibles.

Ce rapport présente la démarche de recherche sur ces données, des propositions faites pour l'amélioration du logiciel ainsi que le travail réalisé.
Pour mener à bien ce travail, des travaux de réflexion, de recherche, de présentation et d'implémentation ont dû être réalisés. L'ensemble de ces parties sera développée en s'appuyant sur la démarche suivie lors de mon apprentissage.

Notons que ce travail a été mené techniquement en autonomie. Bien que mon encadrante ne soit pas sur le même site que moi, elle était très disponible pour répondre à mes interrogations de même que mon second tuteur d'apprentissage qui est le responsable du marché avec la CNAF.

Description du travail proposé

Le travail proposé par la CNAF est une recherche de problématiques susceptibles d'être améliorée par la Data Science. Pour ce faire, ils m'ont fourni une extraction des données de leur logiciel afin que je puisse collecter des problématiques qui permettraient de réduire la surcharge de travail.
J'ai été libre sur ce travail grâce à une confiance d'ATOS et de la CNAF. Cela m'a permis de passer du temps à la recherche, la proposition de nouvelles idées et une liberté dans mes choix sans avoir une pression de rentabilité de ce travail.

Présentation de l'outil de gestion des incidents

L'outil de gestion des incidents permet aux collaborateurs utilisant des applications de la CNAF de rapporter un problème, demander une évolution et faire des demandes techniques ou fonctionnelles.

L'extraction de données qui m'a été fournie représente l'ensemble des tickets qui ont été ouverts ces dernières années. Ceci représente plus de 300 000 tickets[1].

Mise en place du sujet

Les travaux de réflexion et de recherche sont parties intégrantes de mon apprentissage car ils m'ont permis de proposer et décrire le travail que j'allais réaliser par la suite.
Le sujet étant assez vague, le premier travail à réaliser est de s'imprégner des données et de les comprendre.
Suite à de nombreux échanges, de visualisations et de compréhension du processus, j'ai identifié deux problèmes qui pourraient être amélioré avec de l'analyse de données.

[1] Un ticket est présenté dans la section « données »

Premièrement, le nombre de tickets similaires est assez important (10%), cependant ceux-ci sont parfois identifiés tardivement dans le processus. Cela entraine une redondance d'activité jusqu'à l'identification de la similarité des tickets. Il y a parfois 4 ou 5 tickets qui rapportent le même problème. Le but de mon étude de similarité est de prédire si un ticket entrant est similaire d'un ticket de la base de données.

Le deuxième problème identifié est l'attribution d'un ticket à une équipe. En effet, le processus de redirection des équipes fonctionne comme ceci :

- Une équipe support reçoit le ticket, elle essaye de résoudre le problème ou redirige le ticket vers l'équipe qu'elle identifie capable de résoudre le problème : appelons la équipe A.
- L'équipe A reçoit le ticket, soit elle résout problème, soit elle redirige le ticket vers l'équipe support. Celle-ci pourra ensuite identifier une nouvelle équipe plus pertinente. Dernière option, l'équipe A redirige elle-même le ticket vers une nouvelle équipe.
- On recommence l'étape précédente, jusqu'à ce qu'une équipe résolve le problème

Lors de ce processus, on a pu remarquer avec les équipes de la CNAF qu'un ticket qui transite par plusieurs équipes met beaucoup plus de temps à être résolu à cause d'un temps de transition parfois long entre les équipes. Le but de mon analyse va donc être d'essayer de prédire l'équipe susceptible de résoudre le problème pour éviter le transit important des tickets.

Il est important de noter que les deux analyses que je propose sont des aides à la décision. Celles-ci sont destinées à l'équipe support. L'algorithme prédit une similarité ou une équipe et ensuite l'équipe support est libre de suivre cette prédiction ou non.

Problématiques scientifiques

Une fois les problématiques identifiées, je les ai présentés à diverses personnes de la CNAF. Les problématiques et le travail à faire ont été validés de façon fonctionnelle. Des échanges techniques avec des personnes des équipes statistiques de la CNAF ont aussi permis d'identifier des problématiques scientifiques. En effet, un ticket contient des données textuelles et des données catégorielles. Cependant, la direction statistique de la CNAF a vu un intérêt important dans l'analyse de textes car celle-ci pourrait être utilisée dans d'autres projets. C'est pourquoi, lors de mon travail, j'ai proposé de faire une analyse utilisant seulement le texte, puis une autre utilisant le texte et les catégories.

Description du travail réalisé

Toutes les solutions qui sont proposées dans cette section sont des méthodes d'apprentissage supervisé, c'est-à-dire qu'il faut une base de données d'apprentissage avec des résultats déjà connus pour pouvoir entrainer l'algorithme. Puis c'est dans un second temps que l'on va prédire des résultats pour de nouvelles données.

Données

La base de données d'apprentissage utilisée est l'extraction de données de tickets représentant plus de 300 000 tickets.

Ticket :

Un ticket est ouvert par un utilisateur qui rencontre un problème, demande une évolution, etc. Celui-ci renseigne divers informations comme le type de demande (problème, évolution) ou la priorité.
Il y a 17 données que l'utilisateur remplit grâce à une liste de propositions, ces données sont appelés variable catégorielles. Il doit ensuite rajouter une description textuelle de son problème.
Le ticket est ensuite complété par les réponses techniques et fonctionnelles des équipes résolvant le ticket.
De plus, un ticket contient des informations fournies automatiquement : identifiants, dates.

Un premier travail de nettoyage de données fut nécessaire. En effet, dans l'extraction de tickets, je n'ai gardé que certaines informations :

- Une colonne identifiant de manière unique le ticket
- 9 colonnes sur les 17 données catégorielles que l'utilisateur peut remplir en créant un ticket. 7 colonnes étaient soit trop souvent nulles, ou non informatives car elles avaient toujours la même valeur.
- Le ticket similaire s'il y en a un de fournit (c'est la dernière donnée catégorielle).
- La description que l'utilisateur fournit

Il faut noter que les données sont assez bruitées. En effet, un travail de nettoyage a été nécessaire à cause des données nulles ou de textes contenant beaucoup d'abréviations et de fautes typographiques. Pour ce faire, j'ai supprimé la ponctuation, les chiffres, les adresses web puis j'ai effectué un travail de lemmatisation. Cette méthode est une méthode qui permet de remplacer un mot par son mot « du dictionnaire » associé. Une des difficultés rencontrée lors de cette étape est de trouver un algorithme de lemmatisation efficace pour le français. Beaucoup d'algorithmes existent pour l'anglais, mais peu sont vraiment efficaces pour le français.

Prédiction d'équipe :

Dans l'extraction fournit au départ, la donnée « équipe » qui a résolue le problème n'était pas fournie. Une nouvelle base de données avec l'identifiant du ticket, l'ensemble des équipes ainsi que les actions réalisées m'a été fournie. C'est depuis cette table que j'ai pu identifier l'équipe qui a résolue le problème pour l'ajouter à mes données.
La totalité des données a été utilisée pour la prédiction d'équipe, ce qui représente 220 000 tickets après suppression des données inutilisables (identifiant manquant, pas d'équipe ayant résolu le ticket identifié, ticket similaire)

IDT	COL1	…	COLX	EQUIPE
INC0268059	Prod	…	Tech	011EALLOC
INC0135715	Rec	…	FONC	031EPILOTAGE

Tableau 1 Données utilisées pour la prédiction d'équipe

Prédiction de similarité

Les données utilisées pour prédire la similarité ont dû être modifiées et n'ont pas pu être toutes utilisées. En effet, il y a 10% de données similaires. Compte tenu de la différence avec les données qui n'ont pas de tickets similaires, je n'ai pas pu utiliser toutes les données qui n'ont pas de tickets similaires afin d'éviter que mes algorithmes ne prédisent que des tickets non similaires. En effet, si je donne 90% de données non similaires et que l'algorithme ne prédit que des données non similaires, mon pourcentage de réussite sera de 90% sans avoir prédit un seul ticket similaire.
J'ai donc utilisé les 10% de données qui sont liés entre eux étant similaires et 30% de données qui ne sont pas similaires. Le nombre de données utilisées est d'environ 130 000 données, ce qui représente 65 000 lignes dans le tableau ci-dessous.

IDT1	IDT2	COLASimilar	COLBSimilar	COL…Similar	Similar
INC0268059	INC0273574	True	False	…	True
INC0135715	INC00607982	False	False	…	False

Tableau 2 Données utilisées pour la prédiction de similarité

Méthodes de validation des résultats

Une étape importante lorsque l'on fait de l'analyse de données est la façon de valider ses résultats. En effet, on a un ensemble de données d'apprentissage qui nous permet d'entrainer notre algorithme. Le but étant d'avoir un taux de réussite important sur cet ensemble d'apprentissage, on entraine notre algorithme de façon à optimiser les résultats. Cependant, l'application a pour but de prédire une équipe pour des données qu'il n'a jamais rencontré précédemment, pour lequel on ne l'a donc pas optimisé. On parle ici de capacité de généralisation de l'algorithme, c'est-à-dire de généraliser sur l'ensemble des donnés de ce type en ayant vu qu'une partie. Si un algorithme ne généralise pas bien, on peut avoir un très fort taux de réussite pour la base d'apprentissage, mais un taux de réussite très faible pour les données réelles. Pour éviter cette erreur, j'ai utilisé 85% des données d'entrainement présentées dans le paragraphe précédent pour entrainer mon algorithme et j'ai gardé 15% de données pour tester mes résultats. C'est-à-dire que l'algorithme va pouvoir s'adapter avec 85% des données. Puis je prédirai ensuite les équipes des 15% restant que l'algorithme n'aura jamais vu. En comparant les prédictions avec les données réelles pour ces 15% de données que l'algorithme n'a jamais vu, j'obtiendrai un score de référence qui tient compte de la généralisation. Lors de la présentation des résultats, vous trouverez donc *train accuracy* qui est le pourcentage de réussite sur les données d'entrainement et *test accuracy* qui est le pourcentage de réussite sur des données inconnues que l'on appelle données de test.

Prédiction d'équipe

L'équipe à prédire est l'équipe qui a résolue le problème. Il y a un peu plus de 700 équipes différentes. Ceci rend la prédiction vraiment difficile. En effet, la meilleure prédiction de hasard possible est de prédire toujours l'équipe la plus représentée dans la base d'apprentissage tout le temps.

Utilisant le texte seulement

Pour prédire l'équipe à partir du texte seulement, j'ai présenté une méthode utilisant l'apprentissage profond (aussi appelé *deep learning*), c'est-à-dire des réseaux de neurones qui sont utilisés en recherche pour ce type de problème (texte) même si elles sont plus connues pour le son ou l'image.
Cependant, la direction statistique ne connaissant pas cette méthode mais plutôt une approche plus traditionnelle, j'ai décidé de comparer l'approche traditionnelle avec ma méthode utilisant le *deep learning*.

Approche traditionnelle

L'approche traditionnelle est une méthode statistique qui utilise le nombre de mots ou caractères associés afin d'attribuer un score puis de classifier le texte à partir de ces évaluations.
Cette méthode est présenter dans le rapport de recherche associé[2].

Méthode proposée

La méthode proposée s'inspire de la recherche et est adaptée à mon problème. Ayant travaillé en autonomie, j'ai favorisé des méthodes avec lesquelles j'avais déjà travaillé dans les possibilités que la recherche m'offrait avec des résultats similaires. Des libertés ont aussi été prises et des comparaisons de méthodes ont été faites. Un rapport plus adapté à la recherche a été fait et peut être fourni sur demande pour montrer la totalité des avancées que j'ai faites, des comparaisons de méthodes, des résultats et implémentations réalisées. Ce rapport va donc présenter de façon sommaire la méthode qui a donné les meilleurs résultats mais ne donnera aucun détail technique. Les principales idées seront présentées car les idées et l'implémentation représentent un temps important de mon apprentissage.

J'ai proposé cette approche car les résultats de recherche surpassent souvent les méthodes d'analyse de textes classiques. Cependant, de nombreuses autres méthodes utilisant le *deep learning* existent.

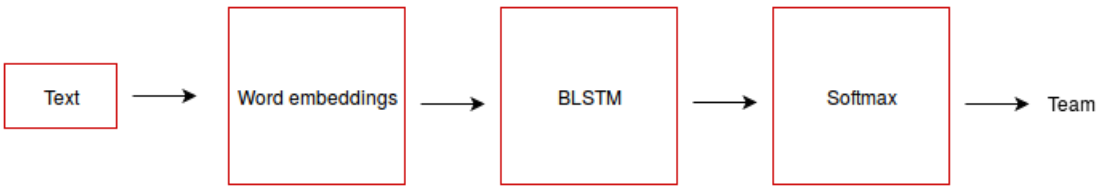

Figure 1 Meilleur modèle permettant la prédiction d'équipe

Dans la figure ci-dessus où les étapes seront explicitées, il y a deux réseaux de neurones. Je ne les développerai pas en détail car mon rapport de recherche le fait. Cependant, le principe de base d'un réseau de neurones, est de relier des neurones qui ont une structure simple en informatique. Ce ne sont pas de vrais neurones, même si cela s'est inspiré de la biologie, il reste encore de grandes différences. Pour mieux comprendre une idée basique des réseaux de neurones, voici un exemple de neurone simple utilisé dans certaines architectures simples de réseaux de neurones :

[2] Rapport fourni sur demande à l'adresse suivante : turpaultn@gmail.com

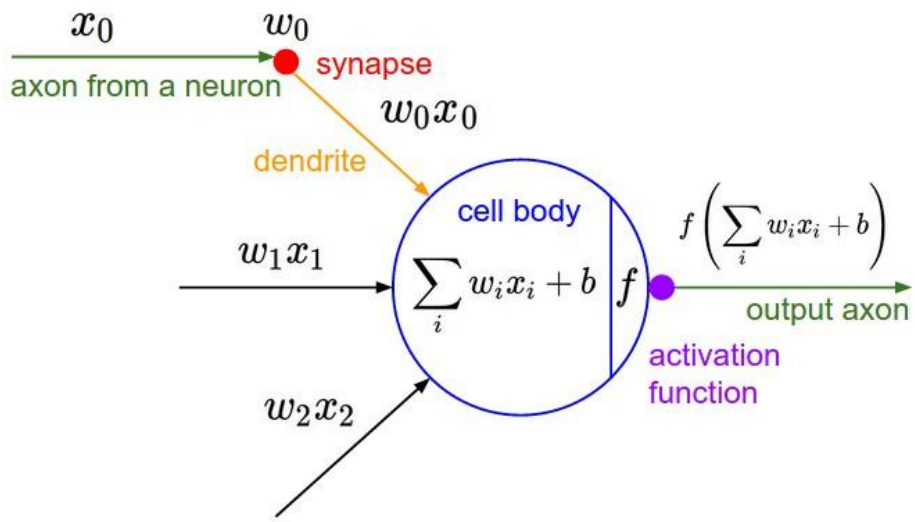

Figure 2 Exemple de neurone informatique

Cette figure explique qu'un neurone représente la somme des entrées multipliée par ce que l'on appelle un poids (*w* dans le neurone au-dessus) puis à ce résultat, on applique une fonction. Dans ce neurone, c'est le poids *w* que l'on essaye d'optimiser afin que la sortie soit celle attendue pour l'ensemble des données qu'on lui fournira en entrée. Lorsque le problème est très simple, un neurone suffit car une valeur du poids w donne toutes les sorties correspondant aux entrées.
Ceci est l'idée principale. Il y a ensuite des détails techniques et mathématiques qui s'ajoutent pour faire des réseaux entiers et plus complexes.

Dans la Figure 1, la méthode de *Word Embeddings* permet de convertir les mots en vecteur de nombre réels. Ceci se fait grâce à un réseau de neurones. Ici on utilise une méthode qui s'appelle Word2vec[21]. C'est une étape qui me semblait indispensable à l'application de l'algorithme suivant. Cette étape est déjà codée, il m'a fallu utiliser la librairie et choisir les paramètres adaptés à mon problème.

BLSTM[16] signifie Bidirectional Long Short Term Memory. C'est un réseau de neurones que l'on appelle récurrents car il permet d'apprendre des dépendances longues au court du temps. J'ai utilisé cet algorithme car c'est un algorithme que je connaissais. En effet, lors de mon stage de Juin à Septembre 2016, j'ai utilisé un *LSTM*[7] qui est quasiment le même algorithme que le BLSTM mais ce *LSTM* était appliqué au son. Il y a différents types de problèmes identifiés lorsque l'on applique ce type d'algorithme. Les résultats peuvent être mauvais si on ne fait pas attention à certains pièges qui rendent l'algorithme inefficace. Or grâce à mon stage j'ai été confronté à un grand nombre de ces problèmes d'apprentissage que l'on rencontre avec cet algorithme et on a pu les résoudre avec mon tuteur. C'est une des raisons qui m'a fait choisir cet algorithme en plus de papiers de recherche avec de très bons résultats utilisant cet algorithme.

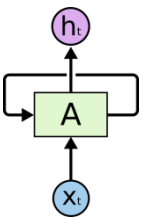

Figure 3 Exemple de neurone récurrent

Je connaissais le *LSTM* mais pas le *BLSTM*, la différence qu'il y a entre ces deux algorithmes est une spécificité technique. En effet, grâce au *LSTM* et un ensemble des réseaux de neurones récurrents, on peut apprendre des dépendances dans le temps. Le *LSTM* est connu pour identifier de longues dépendances. Prenons l'exemple d'une phrase comme :

« Le chat de ma voisine vient de manger une souris »

Le principe de l'algorithme est de lire les mots un à un et de garder en mémoire les mots qu'il a vu pour être capable de garder une dépendance. Dans la phrase ci-dessus par exemple, l'algorithme peut réussir à trouver des dépendances entre « souris » et « chat ». Le *BLSTM* à la différence d'un *LSTM* ne lit pas la phrase dans un seul sens, mais dans les deux sens. Il va donc recevoir les mots un à un en commençant par le début puis recevoir l'ensemble des mots en commençant par la fin. Il a été montré que dans plusieurs cas, ceci donne de meilleurs résultats. C'est le cas pour mon problème ayant effectué la comparaison. La comparaison n'était pas prévue initialement mais m'a permis de justifier scientifiquement l'ensemble de mes choix.

Dans la Figure 1, je présente ma méthode où j'utilise un réseau de neurones pour la partie *word embeddings*, puis j'utilise les résultats de celui-ci comme entrées pour la brique d'après qui est le *BLSTM*, un autre réseau de neurones. Ces deux briques peuvent être rassemblées en un seul réseau de neurones spécifique si on le souhaite, une comparaison avec cette approche a aussi été faite dans mon rapport de recherche.

La dernière brique de la Figure 1 s'appelle *Softmax*, cette méthode permet de passer de valeurs réelles à une probabilité entre 0 et 1 pour chacune des équipes. On choisit ensuite l'équipe ayant la plus grande probabilité.

La dernière chose que j'ai appliquée pour ce modèle qui m'a permis d'avoir de meilleurs résultats est une méthode que l'on appelle *finetuning*. En effet, après une analyse des équipes qui étaient représentées et de leur distribution, j'ai constaté que la distribution des équipes est modifiée au cours du temps. C'est-à-dire que l'équipe la plus représentée en 2012 n'est pas la même que la plus représentée en 2013 et ainsi de suite. Pour ce faire et améliorer notre prédiction, l'idée est de spécialiser un peu notre algorithme. On ré-entraine donc notre algorithme sur les derniers 20% de la base d'apprentissage. Cela permet d'avoir une distribution de prédiction d'équipe plus proche que celle qu'il y a dans la base de test. Cependant, on parle bien ici de réentrainement, on garde donc l'ensemble de la base d'apprentissage pour entrainer une première fois l'algorithme qui voit donc chaque donnée d'entrée plusieurs fois (~20 fois), puis on redonne en entrée que quelques fois (5) une petite partie de la base d'apprentissage pour le spécialiser.
Cette étape ne faisait pas partie de mes plans initiaux mais elle m'a permis de gagner un pourcentage de réussite important.

Résultats

Modèle	*Train accuracy (%)*	*Test accuracy (%)*
Hasard suivant la distribution	*1.84*	*1.21*
Prédiction de l'équipe la plus représentée	7.34	4.54
Approche traditionnelle	65.61	23.26
Approche proposée avant *finetuning*	51.02	32.07
Approche proposée après *finetuning*	67.24	47.12

On voit dans le tableau ci-dessus que la tâche est vraiment difficile au vu des résultats obtenus avec le hasard ou l'équipe la plus représentée. Je n'ai pas mis le hasard pur selon les 700 équipes qui a des résultats bien inférieurs à 1% (de l'ordre de 0.14%).
On voit aussi que l'approche traditionnelle connu par la direction statistique de la CNAF ne classifie pas aussi bien les textes que l'approche que j'ai proposée utilisant des réseaux de neurones. De plus, la spécialisation de l'algorithme en utilisant le *finetuning* fut une étape vraiment importante de mon analyse permettant un gain de 15% de prédiction.

Utilisant le texte et les variables catégorielles

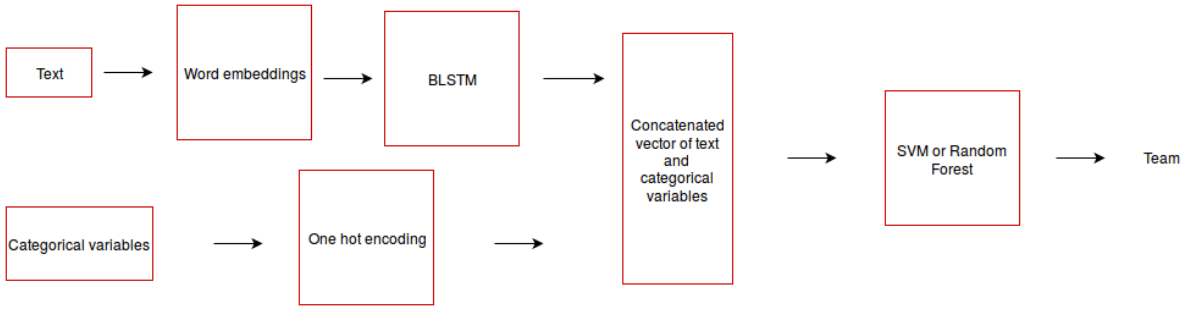

Figure 4 Architecture utilisée pour prédire l'équipe utilisant textes et variables catégorielles

La figure ci-dessus présente l'architecture utilisée pour faire cette prédiction. La partie supérieure utilisant le texte a déjà été présentée.
One hot encoding signifie de remplacer une variable catégorielle par un vecteur de 0 et de 1. Par exemple, une variable qui prend comme valeurs (chat, chien, oiseau) peut être représenté par un vecteur où :
(1, 0, 0) représente chat
(0, 1, 0) représente chien
(0, 0, 1) représente oiseau
Je regroupe toutes les variables catégorielles en concaténant l'ensemble des vecteurs.

Comme on peut le constater sur la figure présentant l'architecture, la brique *softmax* qui était présente dans la prédiction à partir du texte (section précédente) n'y est plus car il n'y pas besoin de donner une probabilité à chacune des équipes pour la prédire. On regroupe le texte et les variables catégorielles en concaténant les vecteurs obtenus aux étapes précédentes (*BLSTM* et *One hot encoding*).

La dernière brique permettant la prédiction de l'équipe cite deux méthodes qui sont comparés : Le Support Vector Machine (SVM) et le Random Forest (RF). Ces méthodes sont des méthodes de classification classiques en Data Science. Elles sont présentées dans le rapport de recherche.

Redéfinition des tâches restantes

Avant de développer la dernière partie, je dois parler de la redéfinition des réalisations à faire avec la CNAF. En effet, après avoir développé une importante partie de la solution de prédiction d'équipe à partir du texte, plusieurs mois était écoulés et j'ai dû faire un choix. Après concertation avec les personnes de la CNAF, on a décidé de faire l'analyse de

prédiction d'équipe utilisant seulement le texte de façon complète et aboutie, d'où un rapport orienté recherche a été réalisé comparant différentes méthodes.

Je n'ai pas eu le temps non plus de développer jusqu'au bout la partie prédisant la similarité. Les parties de nettoyage de texte et la modification des données comme présentées précédemment afin de pouvoir appliquer le modèle ont été réalisées ainsi qu'un travail de réflexion et de recherche qui sera présenté dans la prochaine section. Le modèle présenté ensuite n'a donc pas été implémenté mais pourra faire l'objet d'une prochaine étude.

La redéfinition des tâches est importante car elle relève une partie difficile de mon alternance. En effet, étant en autonomie, il m'a fallu définir ce que j'allais faire étape après étape sans dispersion. De plus, il fallait estimer le temps nécessaire pour réaliser l'ensemble de mes projets. Ce fut une étape difficile qui s'est affinée au cours du temps car je me basais sur les travaux effectués. En effet, si je devais le refaire, j'essayerai d'identifier l'ensemble des étapes nécessaires à la réalisation d'une seule étape comme la prédiction d'équipe à partir du texte seul par exemple puis prévoir un temps d'imprévu que j'estimerai à 20% pour un travail comme celui-ci et dans le contexte d'ATOS. J'essayerai aussi d'estimer le temps nécessaire à lire des articles scientifiques pour définir la solution utilisée ainsi qu'un temps approximatif journalier (1h) pour continuer la recherche et affiner la solution. On peut résumer ceci par une amélioration de ma gestion du temps et de gestion du risque.

Il faut aussi noter l'utilisation de réseau de neurones avancés, qui ne sont pas des méthodes triviales. Bien que j'aie utilisé certaines méthodes précédemment, de nombreuses nouveautés et complexités se sont ajoutées. C'est un domaine qui me passionne mais qui nécessite parfois plusieurs heures de compréhension ce qui m'a parfois rendu la tâche difficile. Cependant, les résultats favorables de ces techniques justifient les heures passés et la motivation dont il a fallu faire preuve.

Prédiction de similarité

Utilisant le texte seulement

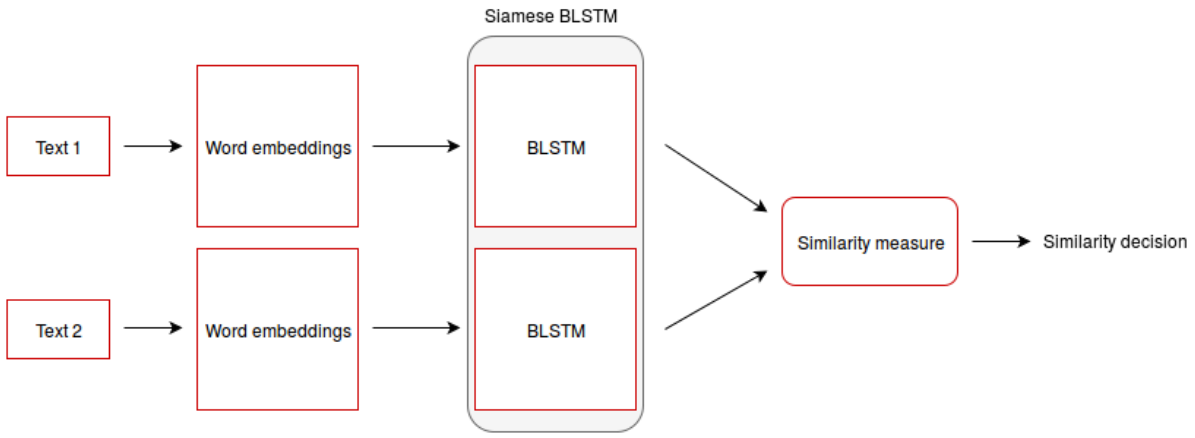

Figure 5 Modèle permettant la prédiction de similarité entre deux textes

Ce modèle ne présente que des solutions utilisées précédemment lors de la prédiction d'équipe. J'ai choisi ce modèle pour éviter un travail supplémentaire trop important car beaucoup de code pouvait être réutilisé pour cette solution.
On voit sur la figure ci-dessus, qu'il est écrit *Siamese BLSTM*[13] au-dessus des *BLSTM*. En effet, cette solution, provenant d'un papier de recherche[13], utilise deux *BLSTM* totalement similaires. Dans mon cas ils ont été pré-entrainés grâce à la solution prédisant l'équipe. Ils sont ensuite ré-entrainés pour être adapté à cette tâche de prédiction de similarité. J'ai eu l'idée du pré-entrainement étant donné que le travail précédent essaye lui aussi de retrouver la sémantique du texte pour prédire l'équipe. Une adaptation et un réentrainement succins devrait donc suffire. Plusieurs mesures de similarités peuvent être testées. La mesure choisie s'appelle *cosine similarity* car elle utilise le cosinus pour prédire la similarité entre deux vecteurs. La décision finale utilise un seuil, c'est-à-dire qu'au-dessus d'une certaine mesure de similarité, on décide que les textes sont similaires, et en-dessous de ce seuil on indique qu'ils ne le sont pas. Ce seuil reste à définir en fonction des résultats obtenus ainsi que les besoins du client.

Utilisant le texte et les variables catégorielles

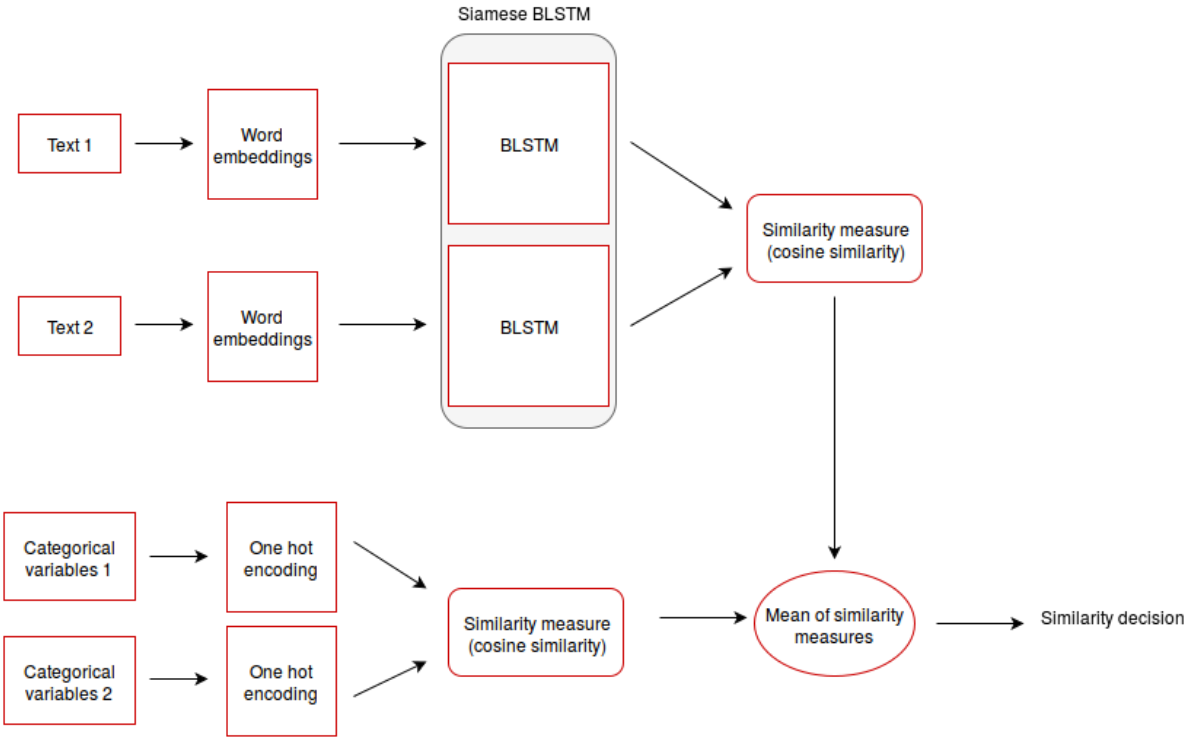

Figure 6 Architecture permettant de prédire la similarité entre deux tickets

Dans cette figure, on voit aussi qu'on ne réutilise que des briques connues sauf le « Mean of similarity measures » qui signifie qu'on prédit la similarité séparément grâce aux variables catégorielles et grâce au texte, puis on prend la moyenne des similarités prédites.

Travail de présentation, d'initiation à la data science et points de situations

Afin de présenter la data science et présenter mes travaux, j'ai réalisé plusieurs présentations. J'avais commencé celles-ci au premier semestre. On va tout d'abord rappeler les grands titres de présentations au premier semestre, et les présentations planifiés à la fin de mon alternance.

- **Présentation aux collaborateurs d'Atos travaillant dans l'équipe CNAF (Décembre) – 2h**

La première présentation se déroule après avoir accompli les travaux de visualisation, d'analyse tweets, et une brève approche pour la valorisation de l'Open Data (Décembre).
2h me sont attribuées pour présenter l'ensemble de mon travail.
Les collaborateurs présents lors de cette présentation sont:

- le directeur du marché CNAF à ATOS (Luc CHAMPALLE)
- le directeur technique du site, (Alain CIVEL)
- l'architecte logiciel du projet "Portail Partenaires" (Sébastien RENE)
- mes deux tuteurs d'alternance (Kévin Royère et Clémentine NEMO)

Je présente mon domaine, mes compétences ainsi que mes travaux réalisés.

- **Présentation à la Direction Technique des Opérations (DTO) (Janvier) – 15min**

Ma deuxième présentation s'est déroulée durant le comité opérationnel (ensemble des directeurs de projets du site Atos Sophia) de Janvier afin de leur présenter mon domaine d'activité et mes compétences dans le cas où ils auraient besoin de compétences telles que celle-ci au sein de leur projet.
Lors de cette réunion, une quinzaine de personnes étaient donc présentes. Cette présentation fut vraiment différente de la première, car j'avais seulement 15 minutes de parole qui m'ont permis de présenter mon domaine, mes compétences et d'énoncer une simple description de ce que j'avais fait à ce moment.
Les conclusions de cette réunion sont :

- Le domaine est intéressant, il faut poursuivre
- Mes travaux sur l'Open Data peuvent être intégrés dans le catalogue permettant de proposer de nouvelles idées aux clients
- Questionnement sur l'aspect commercial du travail sur l'Open Data

- **Présentation en clientèle à la CNAF (Janvier) – 1h**

Alain CIVEL ayant de nombreux contacts à la CNAF, il m'a permis de rencontrer et présenter mon travail à Laurence SALAGNAT (responsable fonctionnelle sur Sophia) et André DEWERDT qui était lui à distance (expert CNAF sur l'aspect données à Lyon).
Lors de cette réunion, j'ai eu 1h afin de présenter mon travail et les différentes propositions possibles.

- **Présentation en clientèle à la CNAF (Février) – 1h**

Lors de cette réunion, les mêmes personnes que précédemment étaient présentes ainsi que Christophe BACILIERE (expert en base de données CNAF et aide aux décisions à prendre concernant le logiciel de gestion de problèmes). J'ai récupéré l'extraction de données du logiciel de gestion de problèmes après celle-ci. Ils n'ont pas voulu me donner de détail sur les problèmes éventuels de l'outil de gestion des incidents afin que j'investigue les données seul puis que je leur présente mes solutions.

- **Présentation en clientèle à la CNAF (Mars) – 2h**

Avant cette réunion, plusieurs contacts avec Christophe BACILIERE ont eu lieu afin d'avoir des explications sur les données, ou des données supplémentaires. Il m'a aidé à comprendre le processus et les données de l'outil de gestion des incidents.
Lors de cette réunion, j'ai proposé l'idée de prédiction de similarité et d'équipe qui ont été retenues. Stéphane DONNE (responsable de la direction statistique de la CNAF) a assisté à cette réunion et a proposé de me mettre en contact avec Pierre COLLINET (statisticien à la CNAF).

- **Point de situation CNAF (Avril) – 2h**

Lors de cette réunion il y avait sept personnes, j'ai exposé les travaux qui étaient en cours. C'est suite à cette réunion que j'ai décidé de faire une analyse utilisant seulement le texte car de nombreuses applications de la CNAF pourraient en bénéficier. C'est aussi suite à cette réunion que j'ai décidé de comparer l'approche connue par les statisticiens de la CNAF et l'approche que je propose.
Lors de cette réunion, j'ai pu constater la difficulté de présenter mes travaux à sept personnes et de prendre les décisions qui y sont associés en tenant compte des remarques.

- **Echange technique avec la CNAF (Juin) – 2h**

Cet échange avec Pierre COLLINET fut nécessaire au bon déroulement de mon alternance. En effet, on a redéfini les priorités de mon travail. On a décidé de se focaliser sur la prédiction d'équipe à partir du texte et de faire un travail abouti plutôt que plusieurs parties non abouties. On a abordé l'ensemble des méthodes que j'utilisais, les rapports souhaités par les statisticiens et le partage du code.

- **Point de situation CNAF (Juin) - 1h**

Lors de cette réunion, j'ai exposé les décisions prises concernant le projet étant donné la contrainte de temps. J'ai donc expliqué que mon travail se focaliserait sur la prédiction d'équipe avec le texte.
Christophe BACILIERE et André DEWERDT sachant que j'avais de premiers résultats m'ont indiqués leur souhait de parler d'une future industrialisation.

- **Point de situation ATOS (Juillet) – 1h**

Suite à la demande d'industrialisation de la CNAF et du temps imparti restant, j'ai fait un point avec ATOS pour expliquer que je ne pourrai pas faire un travail d'industrialisation pour la CNAF mais qu'ils sont très intéressés par mon travail. L'industrialisation étant un travail complet pouvant être chiffré j'ai soumis l'idée de le transmettre à une personne d'ATOS. Cette idée n'a pas été retenue, car la direction CNAF ne pouvait pas investir pour un sujet qui loin du périmètre du contrat. Ce travail pourrait tout de même être repris par un étudiant pour continuer mes travaux.

- **Point de situation CNAF (Juillet) – 1h**

Dernier point de situation de mon projet avant le rendu fait à la CNAF. J'ai indiqué l'impossibilité de faire une industrialisation. La transmission de connaissances se fera donc aux statisticiens de la CNAF. De plus, une perspective de vulgarisation est proposée ainsi qu'un rapport technique et non technique.

- **Présentation de Deep learning et retour sur mes travaux à ATOS (Prévu en Septembre) – 2h**

Je vais présenter devant l'ensemble des collaborateurs intéressés de Sophia un retour sur la Deep Learning School qui a eu lieu dans l'enceinte de polytech au mois de juin (http://univ-cotedazur.fr/events/deep-learning-school) à laquelle j'ai participé dans le cadre de mes projets.
C'est aussi lors de cette réunion que je fais un retour sur mes travaux avec une partie vulgarisée et quelques détails techniques.

- **Retour sur mes travaux à la CNAF (Prévu en Septembre) – 2h**

Cette présentation en visio conférence présentera mes travaux à l'ensemble des intervenants de la CNAF qui m'ont aidé et de nombreux intéressés de la France entière. Cette présentation sera une présentation vulgarisée expliquant de façon simple les solutions proposées. Une introduction à la Data science et aux réseaux de neurones de 20 min sera donnée.

Ces deux dernières réunions sont l'occasion de présenter mes travaux devant des auditoires diverses assez nombreux et certaines personnes importantes de la CNAF et ATOS. Même si certaines personnes ont vu les possibilités de la *data science*, il m'est important de réussir ces présentations qui j'espère motiveront des collaborateurs à inclure la *data science* dans leurs projets.

Développement

Une partie importante de mon travail a nécessité de faire un code lisible, documenté et réutilisable. En effet, la CNAF m'ayant fait part de leur volonté de réutiliser le code et ATOS de possiblement prendre un nouvel étudiant sur la tâche, cela a nécessité du travail supplémentaire.

CNAF

Une problématique rencontrée est que les statisticiens de la CNAF travaillent avec un langage s'appelant R alors que mon travail s'effectue avec le langage python. Ils vont donc avoir un travail d'adaptation ou de conversion du code. De plus, ils n'ont jamais utilisés de réseaux de neurones.
Un travail de présentation et de transmission est donc nécessaire et sera effectué en septembre car ils pourront lire le code durant le mois d'août et me poser les questions à mon retour de vacances.

ATOS

La transmission avec un étudiant que je ne rencontrerai pas est difficile et demande des documentations, une simplicité dans le code et des annotations claires pour permettre à celui-ci de s'en imprégner.

Pour l'ensemble des parties, j'ai décidé de faire des *notebooks*, c'est-à-dire des codes dans des cellules que l'utilisateur peut lancer un à un accompagné de commentaires et de graphes. C'est un outil très utilisé pour présenter un travail car permet l'interactivité avec l'utilisateur qui peut modifier et tester chacun des points développé.

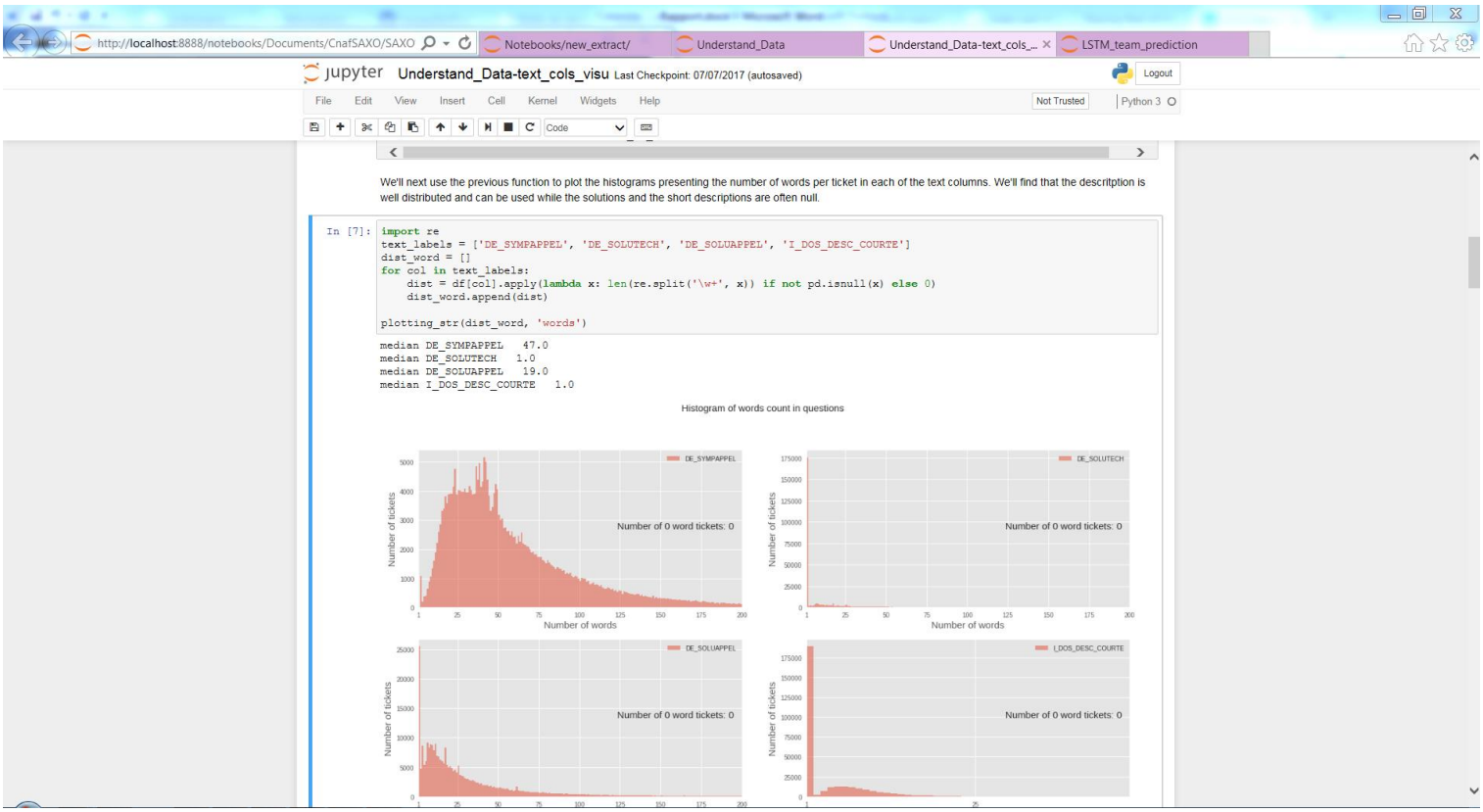

Figure 7 Exemple de Notebook utilisé pour expliquer mon travail

Partage

Pour faire un travail propre, même si j'étais seul sur ce travail, j'ai utilisé un logiciel de gestion de versions du code. L'ensemble de mon travail est donc répertorié.
De plus, j'ai demandé à la CNAF et à ATOS que mon travail soit accessible en open source sur mon compte github. Les demandes ont été acceptées tant que les données ne sont pas présentes car confidentielles. Cependant, l'ensemble du code ainsi que ma présentation de recherche et de vulgarisation seront présents.

Temps de calcul

Une des problématiques rencontrées lors de mon alternance est le temps nécessaire pour effectuer les calculs. En effet, les réseaux de neurones utilisés nécessitent des ressources appropriées pour s'en servir. Pour éviter ce problème j'ai effectué toute la première partie de mon travail que je testais régulièrement avec une toute petite partie des données. Cependant, lorsque mes algorithmes étaient prêts, il a fallu lancer l'expérience sur l'ensemble des données (Juin). Le temps de calcul sur mon ordinateur sans carte graphique

prenait plus de 12h et j'avais une nécessité de lancer environ 30 expériences pour mes comparaisons et trouver les bons paramètres.
Etant au fait de ce temps de calcul grâce à mon précédent stage, j'avais déjà demandé de la puissance de calcul précédemment.
ATOS m'a donc fourni un ordinateur avec une carte graphique qu'il m'a fallu configurer pour utiliser mes codes.
De plus, un travail de parallélisassions des calculs a aussi été nécessaire. Sans ces parallélisassions, le temps de calcul avec la carte graphique était de 5h. Avec la parallélisassions, le temps de calcul d'un modèle est devenu 2h.

Planning du semestre

Février
- Récupération des données
- Visualisation de données
- Lecture de recherche
- Recherche de problèmes
- Présentation

Mars
- Recherche des modèles pertinents
- Nettoyage simple des données
- Code de préparation des données
- Test de différentes librairies pour utiliser les réseaux de neurones
- création du projet sur github

Avril
- Implémenta-tion des tableaux de données utilisés par les modèles:
- données prêtes pour prédire la similraité
- données prêtes pour prédire l'équipe
- Essai des modèles à utiliser sur des bases simples
- Nettoyage des données texte
- Affinage des solutions proposées
- Travail de recherche continu pour les données texte

Mai
- Développement de la solution utilisant seulement le texte
- Nettoyage du code utilisé précédemment
- Réorganisation du code pour prévoir les différents modèles
- Présentation des différentes techniques à utiliser
- Vacances (1 semaine)

Juin
- Développement de la solution traditionnelle utilisant le texte seulement pour la comparaison
- Développement des diférents modèles à comparer utilisant le texte seulement
- Participation à la Deep learning school

Juillet
- Point de situation avec les différents intervenants
- Finalisation de la solution utilisant le texte seulement avec graphiques des résultats
- Ecriture du rapport de recherche

Août
- FInalisation du rapport de recherche
- Finalisation du code à partager sur github et transmission à la CNAF
- Vacances (3 semaines)

Septembre
- Présentations à ATOS et la CNAF
- Préparation d'un papier de 4-5 pages présentant ma solution de façon vulgarisée
- Nettoyage des codes et partage public sur github
- Si temps restant faire un poster pour présenter mon alternance de façon simple à ATOS.
- Fin d'alternance le 22.

Conclusion

Compétences acquises

Lors de cette alternance, il fut très intéressant d'être en autonomie pour me retrouver face à certaines difficultés pour lesquelles j'ai dû décider et trouver rapidement des solutions. Ceci m'a permis d'acquérir de nombreuses compétences.
Premièrement, les présentations demandent une très bonne préparation afin de tenir les délais et répondre aux questions souvent très pertinentes. De plus, animer une réunion n'est pas non plus évident, lors des premières réunions, l'auditoire pouvait souvent se disperser dans les explications ou commentaires que j'ai ensuite appris à recadrer tout au long de mon alternance afin d'aborder tous les points souhaités dans le temps imparti d'une réunion.
Un travail personnel de dépassement des doutes a aussi été nécessaire lors de ce travail. En effet, développer un algorithme depuis des approches de recherche récentes et utilisant des techniques complexes comme les réseaux de neurones est parfois difficile, il m'a donc fallu apprendre à outrepasser mes doutes, continuer mes recherches et parfois revoir mes objectifs afin d'obtenir des résultats. Ce fut un travail enrichissant techniquement bien qu'en autonomie ainsi qu'humainement. La vulgarisation demande une connaissance forte des techniques utilisées ainsi que l'explication des algorithmes utilisés à un auditoire de statisticiens qui demande d'apporter des justifications pour les approches proposés. Ceci m'a demandé un travail de recherche important.
Bien que de nombreuses parties soient nouvelles, j'ai tout de même pu me baser sur des librairies classiques et un langage connu car je les choisissais. J'ai aussi apprécié les mathématiques nécessaires à la compréhension des différentes méthodes utilisées qui m'a permis de mettre en avant les compétences acquises lors de mon cursus scolaire mathématiques.
Finalement, un travail sur mes compétences orales ont été nécessaires. En effet, la vulgarisation n'est pas évidente et nécessite de connaître son auditoire afin de trouver le niveau d'abstraction nécessaire. Il faut savoir être concis, claire et utiliser les bons mots pour satisfaire son auditoire. Suite à de nombreuses compétences acquises, je remercie une nouvelle fois l'ensemble des personnes qui ont su me faire confiance et m'ont aidés lors de cette alternance.

Projets

Le premier projet a été une bonne mise à l'épreuve pour présenter mes compétences et le champ des possibilités grâce à la *data science* même si les solutions proposées n'ont pas été retenues. Cela a été aussi pour moi l'occasion de mettre en œuvre certains cours que j'ai suivi lors de mes études à Polytech. De plus, la première partie de mon alternance m'a permis de montrer mes compétences et de trouver un sujet passionnant pour la deuxième partie d'alternance.
Le projet de la deuxième partie d'alternance a permis de montrer les possibilités de la *data science* ainsi que les résultats intéressant que l'on peut obtenir grâce à celui-ci. Ce fut aussi l'occasion de prouver mes capacités de recherche de problème, mise en place de sujet et résolution de celui-ci. Bien qu'avec le recul, de nombreux points seraient à améliorer pour avoir un travail plus abouti en cette fin d'alternance, la motivation montrée par les différents collaborateurs autour de la *data science* sont ma plus grande réussite.

Comme je l'ai présenté dans ce travail, il reste encore de nombreuses choses à faire dans ce second projet allant de la prédiction de similarité jusqu'à l'implémentation en contexte de production des solutions développées. Je souhaite à un futur stagiaire ou alternant de récupérer ce projet passionnant qui offre de nombreuses possibilités et un développement de soi important.

Abstract

My apprenticeship was about data science and semantic web. I was working for Atos and CNAF which is Atos' customer. I had to suggest new ideas to, at first, reinforce the partnership between CNAF and Atos, and then bring new skills into Atos Company. To do it, I have done a technical work that has been concluded by presentations, meetings and good results on text prediction.
Text prediction has been done using neural networks and this solution has been compared with traditional approach. Good results of my approach using neural networks satisfied CNAF which suggested me to develop the solution in a production context.
The project is still in progress and need a new student to work on it because multiple solutions have been suggested and validated but only one of them could be implemented due to the complexity of the task. Develop the algorithm in a production context is also a task which needs some time to implement.
This apprenticeship achieved goal of introducing data science at ATOS and even proposed new data science solutions unknown to CNAF statisticians.
Autonomy was a key point of this apprenticeship which came through doubt to lead to a running final solution which gives really good results.

Bibliographie

Articles lus (ces articles sont cités dans mon rapport de recherche que j'invite à lire):

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.
[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.
[3] R. JeffreyPennington and C. Manning, "Glove: Global vectors for word representation,"
[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, pp. 306–351. IEEE Press, 2001.
[5] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *CoRR*, vol. abs/1404.2188, 2014.
[6] A. Conneau, H. Schwenk, L. Barrault, and Y. LeCun, "Very deep convolutional networks for natural language processing," *CoRR*, vol. abs/1606.01781, 2016.
[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.
[8] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.
[9] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.
[10] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification.," 2016.
[11] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 1217–1229, Sept 2008.
[12] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196, 2014.
[13] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity.," 2016.
24
[14] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133–142, 2003.
[15] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," in *Text mining and its applications*, pp. 81–97, Springer, 2004.
[16] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification.," in *PACLIC*, 2015.
[17] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification.,"
[18] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification.," in *ACL (1)*, pp. 1555–1565, 2014.
[19] T. Shi and Z. Liu, "Linking glove with word2vec," *CoRR*, vol. abs/1411.5595, 2014.
[20] J. Suzuki and M. Nagata, "A unified learning framework of skip-grams and global vectors.," in *ACL (2)*, pp. 186–191, 2015.
[21] Y. Goldberg and O. Levy, "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.
[22] L. Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, 1991.
[23] G. Hinton, "Lecture 10 Recurrent Neural Networks," 2013.
[24] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
[25] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *CoRR*, vol. abs/1211.5063, 2012.
[26] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, 2016.
[27] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional {LSTM} and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602

– 610, 2005. {IJCNN} 2005.
[28] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
[29] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural Networks*, vol. 18, no. 5, pp. 602 – 610, 2005. IJCNN 2005.
[30] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[31] J. Weston and C. Watkins, “Multi-class support vector machines,” tech. rep., Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May, 1998.
25
[32] A. Statnikov, C. F. Aliferis, D. P. Hardin, and I. Guyon, *A Gentle Introduction To Support Vector Machines In Biomedicine: Volume 1: Theory and Methods*. World Scientific Publishing Co Inc, 2011.
[33] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT ’92, (New York, NY, USA), pp. 144–152, ACM, 1992.
[34] P. Neculoiu, M. Versteegh, and M. Rotaru, “Learning text similarity with siamese recurrent networks,” 2016.
[35] J. Mueller and A. Thyagarajan, “Siamese recurrent architectures for learning sentence similarity,” in *AAAI*, 2016.
[36] H. Schmid, “Probabilistic part-of-speech tagging using decision trees,” 1994.
[37] H. Schmid, “Improvements in part-of-speech tagging with an application to german,” in *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50, 1995.
[38] S. Ruder, “On word embeddings - Part 3: The secret ingredients of word2vec,” 2015.
[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013.
[40] J.-P. Fauconnier, “French word embeddings,” 2015.
[41] M. Sundermeyer, R. Schlüter, and H. Ney, “Lstm neural networks for language modeling.,” in *Interspeech*, pp. 194–197, 2012.

Principales newsletters (veille technologique) :
www.datascienceweekly.org
http://roundup.fishtownanalytics.com

Annexes

Présentation d'Atos

Atos est une entreprise internationale spécialisée dans les services informatiques et fait partie des 10 plus grandes SSII au niveau mondial. Elle est également le leader du paiement en ligne sécurisé pour les entreprises en France avec le système SIPS. L'entreprise a réalisé un chiffre d'affaire de 10 686 millions d'euros en 2015, et elle collabore avec environ 77 100 personnes dans 48 pays différents.

Figure 8 : Atos chiffres mondiaux

Atos fournit aux clients du monde entier des services transactionnels de haute technologie, des solutions de conseil et de services technologiques, d'intégration de systèmes et d'infogérance. Grâce à son niveau d'expertise technologique, Atos est présent dans différents secteurs d'activités : secteur public, secteur financier, industrie, énergie et télécom.

Complément d'information sur la CNAF

La sécurité sociale en France est divisée en quatre branches :

- Maladie
- Vieillesse
- Recouvrement
- Famille

La CNAF est la branche « Famille » de la sécurité sociale française et elle s'occupe de tout ce qui concerne les allocations familiales. Elle est présente sur tout le territoire au travers de 103 Caisses d'Allocations Familiales (CAF)

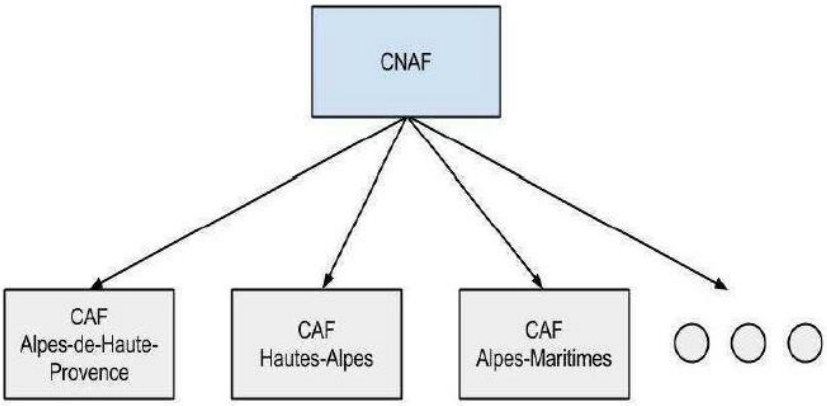

Figure 9 : Diagramme organisationnel CNAF et CAF

Chacune de ces CAF est un organisme de droit privé à compétence territoriale. Elles sont chargées de verser aux particuliers des aides financières à caractère familial ou social, dans des conditions déterminées par la loi, dites prestations légales. Chaque CAF est donc en partie indépendante et mène des actions sociale (compléments de revenus, de suivis, de conseils) qui peuvent différer des autres CAF.
Les particuliers percevant ces aides sont appelés « allocataires » et ils sont un peu plus de 11 millions.
Le site caf.fr a donc pour mission de permettre à la CNAF de mettre à disposition de ses millions d'allocataires des informations les concernant que ce soit au niveau national (CNAF) ou départemental (CAF).

Data Science

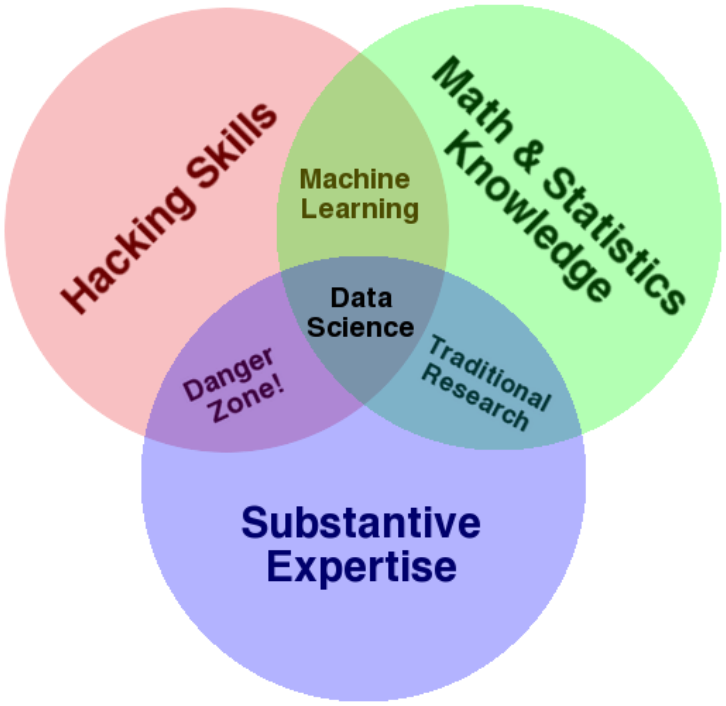

Figure 10 Venn Diagramm

Le « Venn Diagramm » représente les compétences nécessaires à l'utilisation de la « Data Science »

Apport technique

Voir le rapport de recherche écrit en anglais qui peut être fourni sur demande. Demande a effectuée à l'adresse mail suivante : turpaultn@gmail.com