

Exploration interne de la structure conceptuelle des données

Une image contenant texte

Description générée automatiquementHAUREL Maxime | L3 MIASHS parcours Sciences Cognitives | Mai - Juillet 2022

Sous la direction de : Mathieu D’Aquin

# Table des matières

[Table des matières 1](#_Toc108454620)

[Introduction 2](#_Toc108454621)

[Présentation de l’entreprise 3](#_Toc108454622)

[Présentation du projet global 4](#_Toc108454623)

[Présentation du travail réalisé 5](#_Toc108454624)

[Science des données 5](#_Toc108454625)

[Récolte des données 5](#_Toc108454626)

[Traitement des données 6](#_Toc108454627)

[Exploration des activations du modele 7](#_Toc108454628)

[logiciel de visualisation 7](#_Toc108454629)

[Conclusion 7](#_Toc108454630)

[Glossaire 8](#_Toc108454631)

[Annexes 9](#_Toc108454632)

[Remerciements 10](#_Toc108454633)

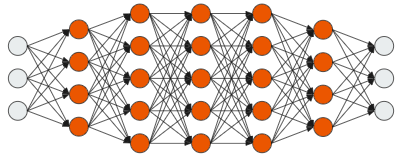
[Quatrième de couverture (changer titre) 11](#_Toc108454634)

# Introduction

Le domaine de l’intelligence artificielle (IA) est en plein essor et n’a presque jamais connu d’hiver. Le nombre de recherches en IA n’a jamais été aussi conséquent et les applications de l’IA touchent tous les domaines, de la santé à l’agriculture en passant par l’éducation.

Alors que les systèmes experts utilisant des règles bien définies étaient les systèmes rois il y a 20 ans, aujourd’hui c’est les modèles statistiques qui sont les plus utilisés dans l’industrie et les plus connus sont les réseaux de neurones artificiels. Ces modèles fonctionnent très bien pour un grand nombre de tâches et c’est la principale tendance en IA connue sous le nom de Deep Learning.

Bien que ces réseaux de neurones soient largement utilisés de partout, ils ont aussi un ensemble de désavantages dont un est **l’inexplicabilité**. En effet, les réseaux de neurones permettent d’abstraire les entrées reçues dans les couches cachées qui composent le réseau.



Ainsi, comme dans la psychologie behavioriste, on ne s’intéresse qu’aux entrées et aux sorties et les couches cachées du réseau (en orange ci-dessus) sont la boîte noire dont, à l’heure d’aujourd’hui, nous ne sommes pas en mesure de comprendre les rouages.

Ces réseaux peuvent être utilisés en production pour prédire le comportement d’un utilisateur. Mais bien que ce cas soit relativement basique, pour une utilisation dans le domaine de la santé, on aimerait savoir en détail pourquoi et comment le réseau a qualifié une radiographie pulmonaire comme laissant présager un cancer proche par exemple.

C’est donc presque dans cet intérêt que le stage s’inscrit. La vision de mon maître de stage est portée sur le comparaison des activations d’une couche cachée d’un réseau de neurones à une connaissance bien définie.  
Dans cette optique, il est possible de penser plus loin en essayant par exemple de corriger le réseau sur des individus qu’il a mal classifié.

# Présentation de l’entreprise

L’entité dans laquelle j’ai fait mon stage est le Laboratoire Lorrain de Recherche en Informatique et ses Applications, abrégé en [LORIA](https://www.loria.fr/fr/) et qui a été créé en 1997.  
Le laboratoire se situe à Villers-lès-Nancy et représente un des plus grands laboratoires en recherche informatique sur le territoire français et un des plus grands laboratoires toutes thématiques confondues dans la région lorraine.

Au sein du laboratoire, on compte 400 personnes dont la plupart sont des ingénieurs et des chercheurs faisant partie de l’Université de Lorraine, de l’INRIA et du CNRS et les recherches sont divisées en 5 départements :

* Département 1 : Algorithmique, calcul, image et géométrie
* Département 2 : Méthodes formelles
* Département 3 : Réseaux, systèmes et services
* Département 4 : Traitement automatique des langues et des connaissances
* Département 5 : Systèmes complexes, intelligence artificielle et robotique

Quant à ma position dans le laboratoire, je me place en tant que stagiaire dans une des équipes les plus récentes, [l’équipe K](https://k.loria.fr/), dont le chef d’équipe est Mathieu d’Aquin, et qui est centrée sur la science des données, l’ingénierie des connaissances et le raisonnement.  
Créée très récemment l’équipe K n’est pas une équipe typique puisque deux des membres permanents sont médecins de formations et travaillent aussi pour le CHRU de Nancy.

# Présentation du projet global

L’intuition scientifique de M. Mathieu d’Aquin a conduit à la création de ce stage. Cette intuition de pouvoir révéler des concepts au sein des réseaux de neurones artificiels à partir des activations des couches cachées est très inspirante et pourrait faire avancer de nombreux domaines de recherche.

Dans un premier temps, mon travail était de réaliser le processus de récolte de données, de traitement des données, de construction du réseau de neurones et enfin de son entraînement.  
L’intuition étant d’aligner les activations à des connaissances existantes, le premier jeu de données exploité comprenait une liste de films avec, pour chacun des films dans le jeu de données, sa description texte, ses revenus ainsi que la catégorie scénaristique auquel il appartient (par exemple ‘*American\_films’*).  
A partir de ce jeu de données récupéré depuis la base de données DBpedia via une requête dans le langage SPARQL, l’idée était de faire prédire une classe de revenu au réseau de neurones en se basant seulement sur sa description textuelle.

Ensuite, j’ai commencé à visualiser les activations couche par couche des neurones afin de les comparer aux activations moyennes pour une catégorie.

# Présentation du travail réalisé

## environnement de travail

Pendant ce stage, j’ai eu accès au laboratoire dans lequel je n’avais pas de place attitrée mais où une salle stagiaire (bien que parfois pleine) était disponible pour travailler librement.

Tout le long de la période de stage j’ai utilisé mon ordinateur portable ainsi que mon ordinateur fixe par souci de performances dans les étapes d’entraînement des modèles.

Pour travailler avec les données j’ai utilisé la suite Anaconda afin de créer un environnement virtuel qui m’a fourni Jupyter Notebook pour pouvoir créer des notebooks permettant de travailler avec les données étape par étape, de faire des tests sur les méthodes à employer mais aussi de pouvoir communiquer les résultats avec mon maître de stage.  
Le traitement, la visualisation des données et la construction du modèle en langage Python a impliqué l’utilisation de certaines librairies telles que

* Pandas, NumPy
* Scikit-learn
* Matplotlib, Seaborn
* ...

Pour partager le travail avec mon maître de stage, nous avons utilisé GitHub, outil très pratique afin de gérer les versions de développement, de commenter une progression et de synchroniser les avancées.

Dans la partie du développement logiciel dont je parle plus en détail plus bas, j’ai utilisé l’outil PyCharm Professional qui offre un environnement intégré de développement afin de pouvoir développer une application.

J’ai eu un confort d’utilisation de tous ces outils/environnements puisque je connaissais déjà chacun d’entre eux et j’avais une expérience forte en langage Python donc je ne me suis pas senti dépaysé et cela m’a permit de réaliser les étapes plus facilement.

## Science des données

### Récolte des données

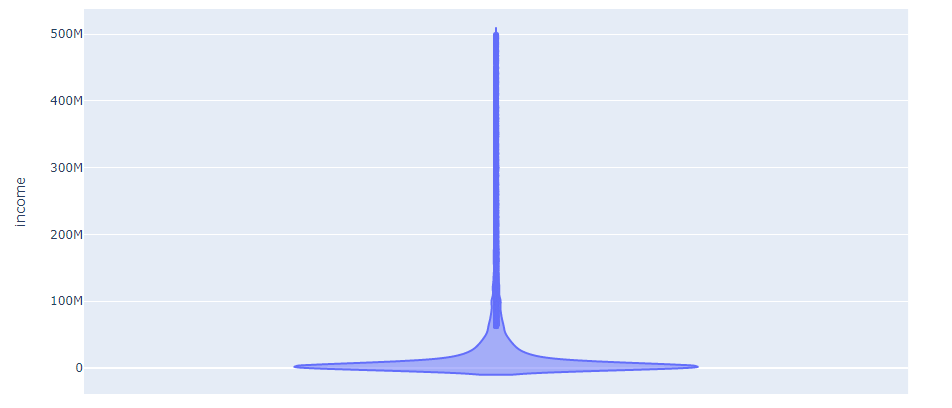
La première partie de mon stage consistait en la récolte de données. L’idée était de pouvoir récupérer des données publiques qui conviendrait au travail futur, c’est-à-dire un jeu de données sur lequel nous pourrions faire des prédictions à partir d’un attribut d’entrée ainsi qu’un autre attribut représentant des connaissances sur chaque individu.

Ainsi, le jeu de données que j’ai récolté comportait des films et pour chacun des films, sa description textuelle, ses revenus et la/les catégorie(s) à laquelle il appartient.  
Ce jeu de données nous permettait bien de pouvoir prédire à partir de la description d’un film, son revenu séparé en 3 classes :

* medium-low
* medium-high
* exceptional

Ce besoin de séparer les revenus en 3 classes distinctes relève du fait que nous ne pouvons pas prédire une valeur bien définie pour chaque film, c’est-à-dire faire une régression puisque les valeurs sont trop disparates.

[violin plot]



Ce jeu de données a été récolté par le biais de DBpedia qui est une base de données indexée sur Wikipédia donc avec beaucoup de données accessibles et ce, en faisant une requête dans le langage SPARQL.

select distinct ?film ?income ?cat ?desc where {  
?film a <http://dbpedia.org/ontology/Film> .  
?film <http://dbpedia.org/ontology/gross> ?income .  
?film <http://dbpedia.org/ontology/abstract> ?desc .  
 {  
 ?film <http://purl.org/dc/terms/subject> ?cat .   
 } UNION {  
 ?film <http://purl.org/dc/terms/subject> ?scat .   
 ?scat <http://www.w3.org/2004/02/skos/core#broader> ?cat.  
 }   
 filter (lang(?desc) = "en")  
 filter (lang(?film) = "en")  
} group by ?film ?cat ?desc LIMIT 3 OFFSET 100000

### Traitement des données

Les données étant récupérés sur DBpedia, elles ne sont pas prêtes à l’emploi. Elles ont besoin d’être traitées de manière réfléchie pour que le modèle les interprète correctement.

C’est notamment le cas des descriptions qui sont au format texte. On ne peut pas donner de texte à un réseau de neurones ou à tout autre modèle de Machine Learning puisque ces modèles sont avant tout des modèles statistiques donc qui fonctionnent avec des entrées numériques. Alors, afin de nettoyer le texte, j’ai appliqué des méthodes courantes de traitement automatique des langues comme :

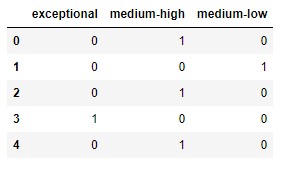
* Transformer chaque lettre en minuscule
* Enlever la ponctuation, les liens, les entités numériques, ainsi que les mots très fréquents (stopwords) du genre *the, a, been, ...*.

Précision : le jeu de données étant en anglais, les mots très fréquents sont évidemment les mots fréquents en langue anglaise.

Ensuite, j’ai dû remanier les revenus pour créer les classes de revenus que j’ai évoqué plus tôt. Puisque le modèle ne fonctionne qu’avec des valeurs numériques, les classes de revenus *(medium-low, medium-high et exceptional)* ont été encodé de la manière suivante :

* On crée une colonne dans le jeu de données pour chaque classe de revenus
* Pour chaque film et chaque colonne des classes de revenus on place un 0 si le film ne fait pas partie de cette classe, ou un 1 si justement le film fait partie de cette classe de revenus

On a alors un tableau présenté tel que :



[...]

## Exploration des activations du modele

La direction de la recherche de la présence de concepts au sein même du réseau de neurones a été orientée par l’intuition que les activations d’une couche cachée d’un réseau pouvaient être porteuses d’un ou de plusieurs concept(s).

Il n’existe pas de fonction préconstruite pour afficher les activations d’une couche d’un modèle au sein des librairies Tensorflow et Keras. Et pour cause, il suffit de reconstruire le réseau jusqu’à la couche que l’on veut explorer et on re-exécute des prédictions avec les entrées originales et on obtient les activations.

De cette manière il semble possible d’automatiser très facilement ce processus d’obtention des activations pour chaque couche et ce peu importe le modèle tant que l’on a accès aux données qui ont servies à son entraînement.

## logiciel de visualisation

# Conclusion

# Glossaire

# Annexes

# Remerciements

Je tiens à remercier chaleureusement M. Mathieu d’Aquin qui m’a permis de réaliser ma première expérience de recherche dans une entité que j’affectionne tout particulièrement.

# Quatrième de couverture (changer titre)

* Mathieu D’Aquin : [mathieu.daquin@loria.fr](mailto:mathieu.daquin@loria.fr)