



**Baptiste  
O'Jeanson**

**Rapport de stage de  
spécialité**

**Étudiant:** Baptiste O'JEANSON

**Maître de stage:** Christian  
FRISCH

**Entreprise:** Data Publica

**Année:** 2014-2015

**Dates:** du 01/06/2015 au  
28/08/2015

**Lieu:** Paris, France

**Classification de signaux  
entreprises avec une  
approche machine learning**

INFORMATION RETRIEVAL, TEXT MINING AND NATURAL  
LANGUAGE PROCESSING





# Remerciements

Premièrement, j'aimerais remercier François Bancilhon<sup>1</sup>, directeur général de Data Publica, et Christian Frisch<sup>2</sup>, directeur technique de Data Publica, de m'avoir donné l'opportunité de découvrir le monde du travail en start-up. Je voudrais aussi les remercier de m'avoir fait confiance et de m'avoir confié une mission très intéressante.

Ensuite, je voudrais remercier Samuel Charron, Clément Chastagnol et Guillaume Lebourgeois qui m'ont encadré, suivi et conseillé durant mon stage, et qui ont enrichi mes connaissances du monde professionnel ainsi que mes connaissances en informatique.

J'aimerais aussi remercier tout particulièrement Thomas Dudouet, qui a fait suite à ma candidature spontanée, et qui m'a donc offert cette chance de pouvoir m'entretenir avec Clément Chastagnol puis Christian Frisch.

Je souhaiterais également remercier l'ensemble de l'équipe de développeurs et l'ensemble de l'équipe de marketing de m'avoir accueilli si chaleureusement.

Enfin, merci à mon tuteur de stage, Nicolas Malandain, pour sa disponibilité.

---

1. François Bancilhon, présenté ici A.2

2. Christian Frisch, présenté ici A.1

# Table des matières

<b>Remerciements</b>	<b>3</b>
<b>1. Présentation de l'entreprise</b>	<b>6</b>
1.1. L'entreprise . . . . .	6
1.1.1. L'activité de Data Publica . . . . .	6
1.1.2. L'équipe de Data Publica . . . . .	7
1.2. C-Radar . . . . .	7
1.2.1. Présentation commerciale de C-Radar . . . . .	7
1.2.2. Présentation technique de C-Radar . . . . .	7
<b>2. Présentation du sujet</b>	<b>11</b>
2.1. La fonctionnalité de C-Radar . . . . .	11
2.2. Ma mission chez Data Publica . . . . .	12
2.2.1. Inscription dans le domaine du <i>Big Data</i> . . . . .	12
2.2.2. Synthèse du travail à réaliser . . . . .	12
2.3. Présentation des signaux . . . . .	13
<b>3. Travail effectué</b>	<b>15</b>
3.1. Démarche de travail . . . . .	15
3.1.1. Mes acquis à l'INSA . . . . .	15
3.1.2. Déroulement du stage . . . . .	16
3.2. Travaux réalisés en Java, le <i>Text Mining</i> et le <i>Natural Language Processing</i> avec la bibliothèque de Stanford . . . . .	16
3.2.1. Présentation de mon environnement de travail . . . . .	16
3.2.2. La bibliothèque : Stanford Natural Language Processing . . . . .	17
3.2.3. Première mise en pratique de la bibliothèque de Stanford . . . . .	20
3.2.4. Amélioration de ma première application et prétraitement . . . . .	23
3.3. Travaux réalisés en Python . . . . .	28
3.3.1. Modules implémentés . . . . .	29
3.3.2. Plugin de classification . . . . .	31
3.4. Bilan sur mes travaux . . . . .	31
<b>4. Conclusion</b>	<b>33</b>
<b>A. Annexes</b>	<b>35</b>
A.1. L'équipe de Data Publica . . . . .	35
A.1.1. L'équipe technique . . . . .	35
A.1.2. L'équipe commerciale . . . . .	36
A.2. Dataviz . . . . .	36
A.3. La validation croisée et les hyper-paramètres . . . . .	36
A.3.1. Les hyper-paramètres . . . . .	36
A.3.2. La validation croisée . . . . .	36
A.3.3. La validation croisée stratifiée . . . . .	38

A.4. Les métriques de mesure de la qualité d'une classification . . . . .	38
A.4.1. Terminologies . . . . .	38
A.4.2. La précision . . . . .	38
A.4.3. Le rappel . . . . .	38
A.5. Les stratégies de modèles de classification en contexte multi-classe . . . . .	38
A.5.1. One versus All . . . . .	38
A.5.2. One versus One . . . . .	38
A.6. Le QA ou Quality Assessment . . . . .	38
A.7. Modèle de classifieur . . . . .	39
A.7.1. Le modèle génératif . . . . .	39
A.7.2. Le modèle discriminant . . . . .	39
A.8. Les transducteurs . . . . .	39

# 1. Présentation de l'entreprise

L'évolution des technologies et leurs usages ont fait exploser la quantité de données générées. Selon IBM, 2.5 milliard de gigabytes (GB) de données a été générée tout les jours de l'année 2012. De plus, cette quantité de données, double tous les deux ans. Cependant, seules 0,05% de ces données sont analysées.

L'exploration ou la fouille de données (« data mining ») consiste à en extraire des informations utiles, et ceci peut s'avérer très fructueux. La question principale qui se pose est de savoir comment utiliser intelligemment cette immense masse de données pour en tirer une plus-value ? C'est le rôle des entreprises spécialisées dans l'exploitation de ces données.

## 1.1. L'entreprise

La société Data Publica a été fondée en juillet 2011 par François Bancilhon et Christian Frisch, respectivement l'actuel directeur général et l'actuel directeur technique.

### 1.1.1. L'activité de Data Publica

Data Publica est un des précurseurs de l'open data en France. Cette société , qui a bénéficié d'investissements technologiques faits en 2010 dans le cadre d'un projet de R&D, a été financée initialement par un groupe de business angels et le fonds d'amorçage **IT Translation**. Data Publica est une start-up spécialisée dans les données entreprises, l'open data, le big data et la dataviz. C'est une société relativement jeune, axée R&D. Son leitmotiv, alimenté par une équipe très dynamique et compétente, est la recherche constante du dépassement technique.

Historiquement, Data Publica ne faisait que de l'open-data. C'est-à-dire que la société se servait de données accessibles à tous (provenant d'institutions gouvernementales notamment) pour créer des jeux de données sur mesure pour des entreprises. Ainsi, la société s'est spécialisé dans l'identification des sources de données, leur extraction et leur transformation en données structurées.

Depuis quelques années, Data Publica se spécialise dans les données sur les entreprises françaises en dépit de son activité open-data qu'elle a progressivement mis de côté. Les services qu'elle propose ne sont plus tout-à-fait les mêmes. En effet, Data Publica réutilise les données open-data concernant les entreprises française dans son produit phare. Ce produit est lui même conçu pour les entreprises du B2B. Le produit est décrit en partie 1.2.

Data Publica participe également à de nombreux projets de recherche français et européens tels que XDATA, Diachron ou Poqemon, en partenariat avec l'INRIA.

### 1.1.2. L'équipe de Data Publica

Data Publica emploie quatorze personnes réparties en deux équipes : une équipe commerciale (quatre personnes) et une équipe technique (dix développeurs). Les deux équipes travaillent chacune dans son open-space. Pendant mon stage, j'ai été immergé au sein de l'équipe technique.

**L'équipe technique :** Elle est composée de dix développeurs (ordonnés par ancienneté visible en annexe A.1.1) :

- Christian Frisch, directeur de l'équipe
- Thomas Dudouet, développeur Back-end Java
- Guillaume Lebourgeois, chef de produit C-Radar
- Samuel Charron, Data scientist Python et mon maître de stage
- Loïc Petit, développeur Back-end Java (JBM)
- Clément Chastagnol, Data scientist Python et mon maître de stage
- Clément Déon, développeur Front-end
- Fabien Bréant, développeur Back-end
- Jacques Belissent, développeur Back-end
- Vincent Ysmal, développeur Back-end Java

**L'équipe commerciale :** Elle est composée de cinq commerciaux (ordonnés par ancienneté visible en annexe A.1.2) :

- François Bancelhon, directeur général
- Benjamin Gans, Responsable Communication et Marketing
- Emmanuel Jouanne, Business Development Manager
- Philippe Spenato, Ingénieur d'affaire
- Justine Pourrat, Responsable Communication et marketing

## 1.2. C-Radar

### 1.2.1. Présentation commerciale de C-Radar

Son produit est un moteur de recherche B2B (Business to Business). Celui-ci a pour objectif de permettre aux services ventes et marketing des entreprises B2B de vendre plus et mieux. Ce moteur de recherche, appelé C-Radar, est un produit de vente prédictive construit sur une base de référence des entreprises françaises. Il regroupe beaucoup d'informations sur les entreprises françaises, dont notamment leurs informations administratives, financières et toutes celles qui découlent de leur communication sur les réseaux sociaux et le web.

C-Radar est un concentré de technologies du big data. En effet, il utilise diverses technologies comme le crawling, le scraping ou encore le machine learning. Ceci afin d'offrir à l'utilisateur diverses fonctionnalités : moteur de recherche d'entreprises, fiche d'activité d'entreprises avec contacts commerciaux, détection de nouveaux prospects, scoring de prospects existants, segmentation automatique d'entreprises, identification de marché, etc.

### 1.2.2. Présentation technique de C-Radar

#### Les technologies utilisées par C-Radar

Pour répondre aux problématiques auxquelles Data Publica se confronte, la société a acquis 4 expertises majeures :

- Le web crawling / web scraping : la récupération des données ;
- Le data mining / text mining : l'analyse, l'extraction et l'enrichissement des données ;
- Le machine learning : l'apprentissage automatique à partir de données structurées ;
- La dataviz : la visualisation des données.

**Le crawling :** Le crawling est l'action réalisée par un programme informatique, appelé le web crawler, qui va de site en site afin d'en extraire automatiquement toute l'information qui est présente sur les différentes pages. Cette technique est utilisée notamment pour l'extraction de données non structurées : la structure du site n'est pas connue à l'avance, l'extraction des données se fait directement sur le contenu (c'est à dire le code HTML) de la page crawlée. Ce processus est « brutal ».

**Le scraping :** Le scraping est l'action réalisée par un programme informatique pour extraire des unités d'information structurées d'un site web. Contrairement au crawling, il est question d'extraire des données précises, et pas la totalité des données disponibles sur le site. Le site « scrappé » et sa structure doivent donc être connus et analysés à l'avance afin d'adapter le scraper au site. Ce processus est « intelligent ».

**Le data mining / text mining :** Une fois des sites web crawlés et scrapés, ou que des flux (RSS ou réseaux sociaux) ont été captés, on peut commencer à analyser le contenu récupéré à la recherche d'informations sous forme de motifs particuliers. On analyse afin de normaliser les encodages, et on extrait des dates, des numéros de téléphone, etc. Ces normalisations et extractions sont effectuées sous forme de liste.

Principalement, cette phase consiste à regarder les données « dans le fond des yeux »<sup>1</sup> afin de voir leur fond mais aussi leur forme.

**Le machine learning :** L'idée du machine learning est de pouvoir extraire automatiquement des informations d'une nouvelle donnée et ainsi prédire une classe de donnée.

Quand les données sont correctement formatées et normalisées, on peut construire des applications capables d'apprendre automatiquement de ces données, de les classer. Ainsi quand on aura une nouvelle donnée l'application sera capable de prédire sa classe d'après ses caractéristiques.

Exemple (tiré de Wikipédia<sup>2</sup>) : La reconnaissance de caractères manuscrits est une tâche complexe, car deux caractères similaires ne sont jamais exactement égaux. On peut concevoir un système d'apprentissage automatique qui apprend à reconnaître des caractères en observant des « exemples », c'est-à-dire des caractères connus.

**La dataviz :** C'est la dernière étape. Elle présente les données de manière visuelle et interprétable. Ainsi, on peut comprendre plus facilement et rapidement les informations extraites des données. Un exemple de dataviz est présenté en annexe A.2.

## Construction des fiches entreprises dans C-Radar

Pour construire les fiches entreprises, on part d'un fichier de SIREN<sup>3</sup> contenant les données administratives de l'entreprise (nom, SIREN, adresse, ...). Ensuite, on va essayer d'associer au

1. Expression empruntée à Stéphane Canu.

2. Wikipedia. Apprentissage automatique. url : [https://fr.wikipedia.org/wiki/Apprentissage\\_automatique](https://fr.wikipedia.org/wiki/Apprentissage_automatique).

3. Système informatique du répertoire des entreprises. C'est un code Insee unique servant à identifier une entreprise française.



numéro SIREN de l'entreprise un site web. Une fois le site web détecté, on va le crawler et en extraire de l'information structurée. Enfin, on enrichie encore les données de l'entreprise, en cherchant de l'information sur des sites web externes (tels que [linkedin](#) et [viadeo](#)). Il ne reste plus qu'à agréger toutes ces données pour former une fiche d'entreprise complète.

## L'architecture technique de C-Radar

L'architecture de C-Radar peut être divisée en plusieurs parties (visible en figure 1.1) :

- Différentes bases de données pour différents stockages (une base de données Mongo, une base de données Cassandra et une base de données PostgreSQL) ;
- Un moteur de recherche sémantique (Elastic Search) ;
- Un gestionnaire de file (RabbitMQ) ;
- Différents plugins Python s'interfaçant avec le gestionnaire de file RabbitMQ ;
- Un gestionnaire de flux entre les différents composants précédents, le Workflow ;
- Une application Java s'interfaçant avec Elastic Search et les bases de données Mongo et PostgreSQL.

**Le JBM ou Java Base Manager :** Le JBM est le projet Java qui rassemble le Workflow et l'application [app.c-radar.com](#). Ce projet a été conçu et construit au dessus de Spring. Le framework Spring est une plate-forme Java qui fournit une architecture complète permettant de développer des applications Java. Spring gère l'architecture de manière à ce que le développeur n'ait à s'occuper que de l'application. Il prend en charge énormément de tâches dont notamment le modèle MVC, la sécurité de l'application, l'interface avec les gestionnaire de file, etc. (Pour plus d'information, voir <http://spring.io/>).

**Le Workflow :** Le Workflow est le gestionnaire de flux permettant de lancer les différents processus de récupération et d'analyses des données. Il gère tout ce qui est « computing ». C'est lui qui lance des tâches en les écrivant dans RabbitMQ. Ensuite les plugins Python écoutent RabbitMQ pour obtenir de nouvelles tâches à effectuer. Une fois les sites webs crawlés, le plugin Python va les stocker dans Cassandra sans passer par le Workflow mais en lui indiquant que le crawl est terminé (c'est la seule fois où un plugin Python écrira directement dans une base). Il gère tout ce qui est exécution des plugins Python (crawling et scraping du web, capture et catégorisation des signaux, etc), stockage des données produites à l'issue de ces exécutions et indexation dans Elastic Search.

Pour résumer, il s'occupe de l'orchestration de tous les processus permettant de passer d'un SIREN à une fiche d'entreprise complète, présentable à l'utilisateur via l'application.

**L'application [app.c-radar.com](#) :** L'application, « présente » les données aux utilisateurs. Elle fait le reporting des données produites par le Workflow. Elle permet de rechercher des entreprises, de voir leur répartitions géographiques, de créer des listes d'entreprises, etc. C'est l'application visible et utilisée par l'utilisateur.

**Les bases de données :** Les bases de données stockent différents types de données : La base MongoDB a pour avantage de pouvoir agréger des documents de sources hétérogènes avec beaucoup de souplesse (base de données NoSQL orientée documents). Elle stocke les données administratives des entreprises, les données issues des sites web, les signaux, ... RabbitMQ stocke les tâches du Workflow et Cassandra stocke des données de crawl.

**Les plugins Python :** Enfin, les plugins Python sont des applications Python isolées, qui exécutent des tâches sur des données. Ils sont isolés, le plus possible, pour éviter les effets de bords. Cela permet de paralléliser au maximum les processus, et, en cas d'erreur dans le traitement, d'être capable de répéter l'opération sans effet sur le reste du système. On utilise le modèle acteur<sup>4</sup>. Ces données sont reçues depuis d'autres composants et le plugin leur retourne les résultats de sa tâche.

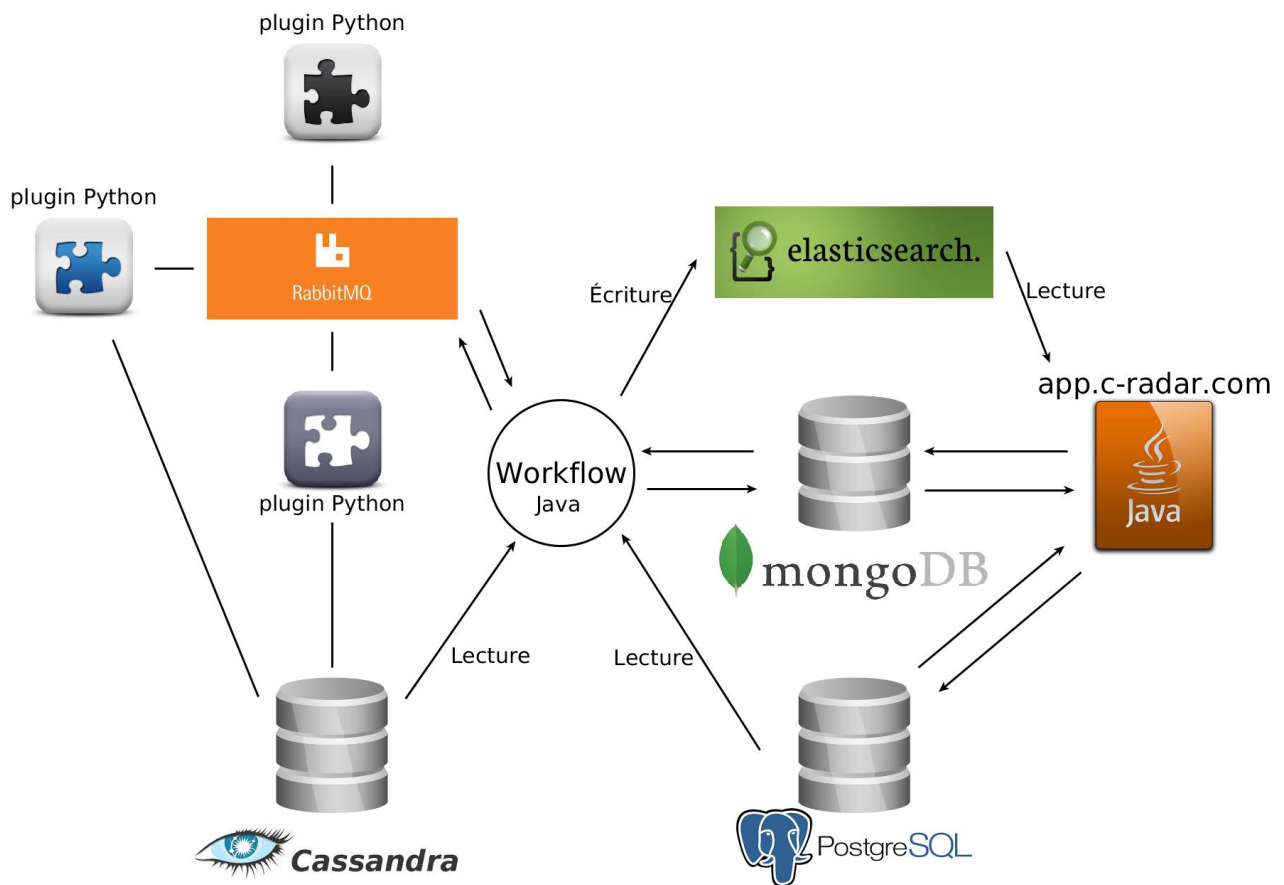


Figure 1.1. – L'architecture générale de C-Radar

4. Wikipedia. Actor model. url : [https://en.wikipedia.org/wiki/Actor\\_model](https://en.wikipedia.org/wiki/Actor_model).

## 2. Présentation du sujet

À la suite d'une candidature spontanée et de premiers contacts avec Thomas Dudouet et Clément Chastagnol, je me suis entretenu avec Christian Frisch afin de savoir quelle pourrait être ma contribution chez Data Publica. Ayant suivi une formation à l'INSA plutôt orienté big data / data mining, Christian Frisch m'a alors proposé de travailler sur l'une des fonctionnalités du produit C-Radar.

### 2.1. La fonctionnalité de C-Radar

L'objectif de mon stage serait le suivant : améliorer la fonctionnalité de C-Radar permettant d'être au courant de l'intégralité de la communication des entreprises françaises et belges simplement en s'abonnant à une newsletter.

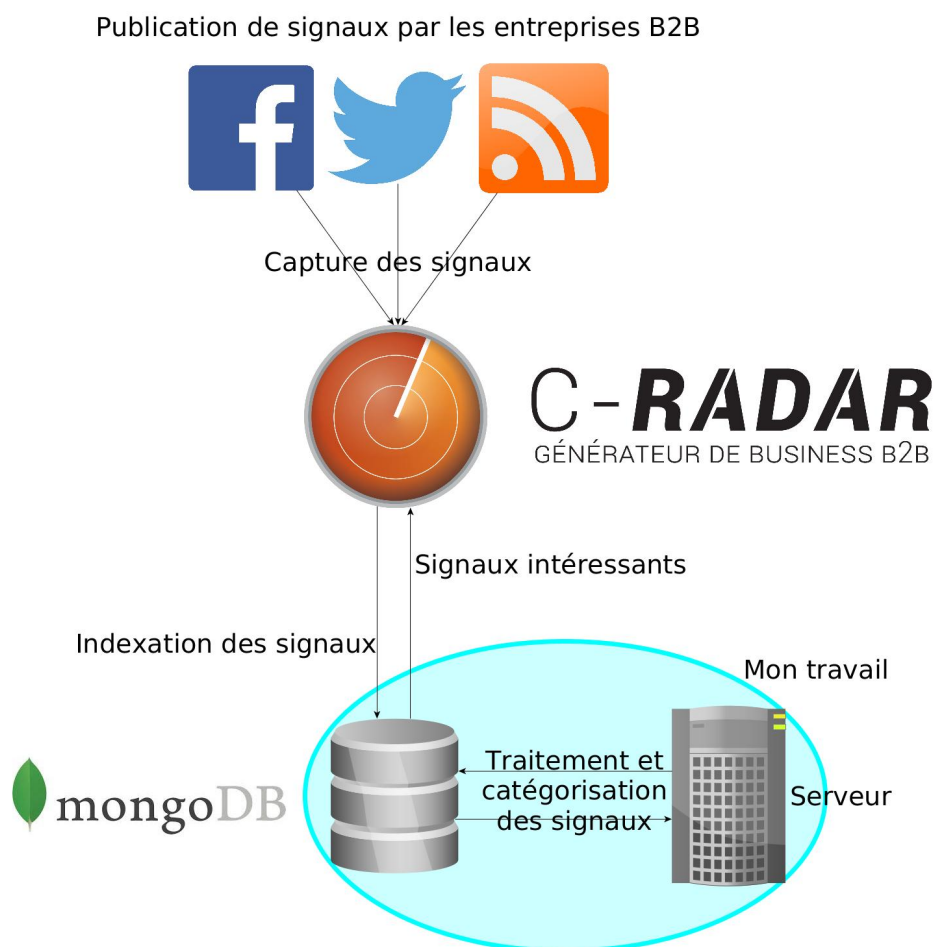


Figure 2.1. – Le fonctionnement général de cette fonctionnalité.

De l'offre d'emploi, à la participation à des salons, en passant par les nominations de personnel ainsi que les présentations des derniers produits ou encore des éventuelles levées de fonds

ou investissements, la communication des entreprises n'est plus équivoque mais bel et bien ordonnée grâce à cette fonctionnalité. On peut désormais savoir quels sont les derniers postes à pourvoir chez Orange, par exemple, ou encore les dernières nominations qui ont eu lieu à la Banque Postale.

Le fonctionnement général de cette fonctionnalité est le suivant (visible en figure 2.1) :

1. C-Radar capte tous les signaux émis par les entreprises sur les réseaux sociaux (Facebook et Twitter), dans les médias ou via des flux RSS. Ces signaux sont ensuite stockés dans une base de données Mongo.
2. Une fois ces signaux capturés et stockés, il faut les traiter afin d'identifier l'intérêt potentiel de leur contenu.

C'est sur ce second point que j'interviens.

## 2.2. Ma mission chez Data Publica

Ma mission est de construire une chaîne de traitement automatique, un plugin Python, récupérant la liste des signaux émis par les entreprises en entrée depuis un point d'API, leur appliquer les traitements nécessaires afin de fournir, en sortie, la liste des signaux intéressants ainsi que leur catégorie respective.

Il s'agit donc de construire une application, un plugin Python, capable de classer un signal dans une catégorie par la seule connaissance de son contenu (éventuellement un titre). L'application traitera donc des documents textuels.

Le travail se divise en deux étapes :

1. Appliquer une série de prétraitements sur le contenu des signaux afin de sélectionner les features jugés porteurs d'information. En effet, comme les données manipulées sont textuelles, il faut filtrer certaines chaînes de caractères considérées comme du bruit.
2. Construire un classifieur à partir des features sélectionnés parmi les données que l'on juge porteuses d'information.

### 2.2.1. Inscription dans le domaine du *Big Data*

Les tâches qui découlent de ma mission, et notamment tous les prétraitements, sont directement liées à la discipline de la fouille de textes (*Text Mining*) (et de l'*Information Retrieval*) pour trouver des informations dans le contenu des signaux. Elles sont également liées à la discipline du traitement automatique du langage naturel (*Natural Language Processing*). Enfin, la construction du classifieur automatique implique des compétences en *Machine Learning*.

**Remarque :** Un classifieur est un outils d'apprentissage automatique qui prend des données et les classe dans k classes.

### 2.2.2. Synthèse du travail à réaliser

Si l'on devait résumer le travail à effectuer (visible en figure 2.1), on pourrait le synthétiser comme ceci : Construire une application (plugin Python), qui prend en entrée un signal émis par une entreprise et répond aux questions suivantes :

1. Ce signal a-t-il de l'intérêt ?

2. Si oui, quel est le sujet du signal ? Parle -t-il d'une offre d'emploi, d'une nomination, d'une levée de fond, etc ?

La figure 2.2 montre ce que l'application (le plugin Python) doit être capable de faire.

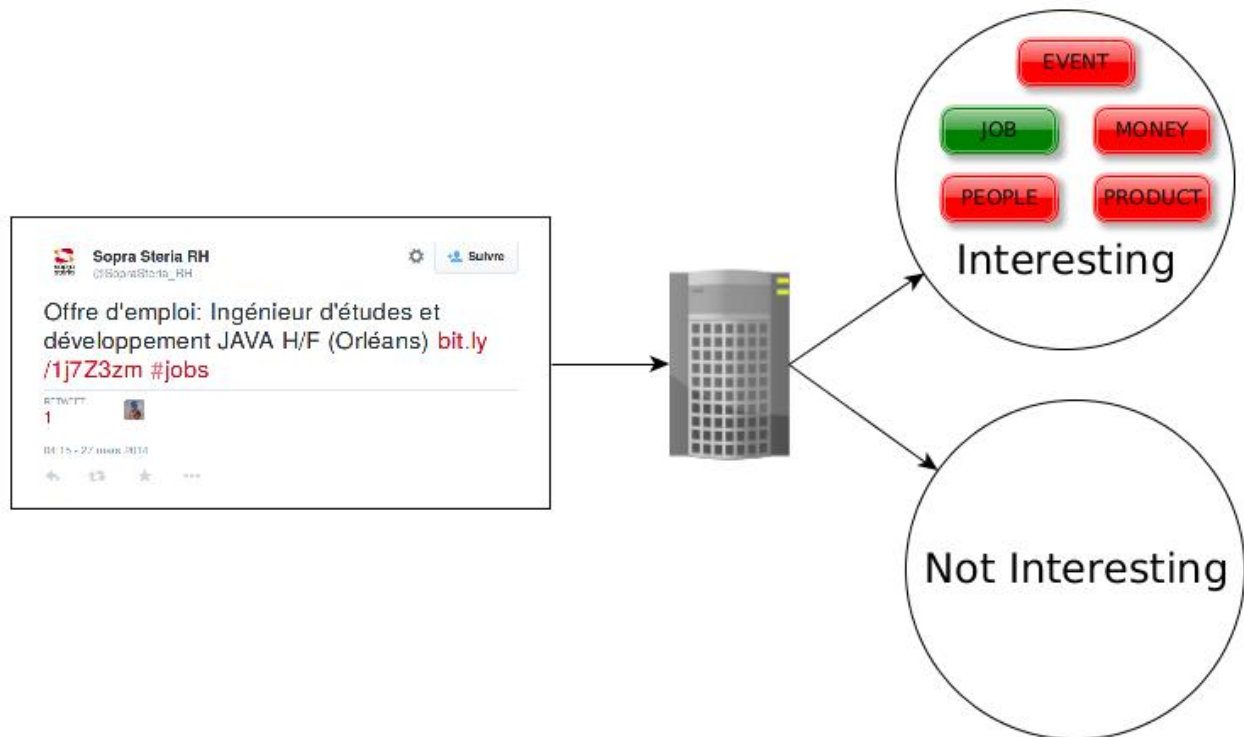


Figure 2.2. – La tâche du plugin Python.

## 2.3. Présentation des signaux

Les signaux sont des posts Facebook, des tweets ou bien des flux RSS publiés par des entreprises.

**Hypothèses de départ :** On considérera qu'un signal est intéressant si son contenu a pour sujet :

- Une offre d'emploi (ou un stage) à pourvoir au sein de l'entreprise qui l'a postée. Par la suite, on associera le tag **JOB** à cette catégorie.
- Un évènement auquel l'entreprise participe (un salon par exemple). Par la suite, on associera le tag **EVENT** à cette catégorie.
- Un produit que l'entreprise vient de présenter. Par la suite, on associera le tag **PRODUCT** à cette catégorie.
- Une nomination d'un employé dans l'entreprise ou d'une entreprise vers une autre entreprise. Par la suite, on associera le tag **PEOPLE** à cette catégorie.
- Une levée de fonds, un investissement, ou encore une déclaration de résultats ou de chiffre d'affaire.. Par la suite, on associera le tag **MONEY** à cette catégorie.

**Exemple de signal type pour chaque catégorie :**

- **JOB** : « Offre d'emploi : Ingénieur d'études et développement JAVA H/F (Orléans) <http://t.co/LOvN8rLH1e> #jobs »

- **EVENT** : « Fraispertuis sera présent dès demain jusqu'à dimanche inclus au salon « Tour-rissimo » de Strasbourg Parc des Expositions, Hall 21 Stand B40, venez rendre visite au capt'ain Fraisp ! »
- **PRODUCT** : « Découvrez quelques unes de nos réalisations de Pergola Biotempérée pour notre clientèle de Toulouse et sa région. Plus d'informations sur notre site [www.pergola-biotemperee.com](http://www.pergola-biotemperee.com) »
- **PEOPLE** : « Matthieu Frairot a été nommé Directeur Associé au sein de l'agence FullSIX France, l'agence marketin <http://t.co/Z2ENeeWZmF> »
- **MONEY** : « Keyrus : Publication des résultats annuels 2013. <http://t.co/k4TPJ11fW4> »

**État de l'ensemble des signaux :** Les travaux a réalisé implique de l'apprentissage supervisé. Ainsi, une base de signaux a été validée manuellement. Au 8 juin 2015, seul 1426 signaux étaient validés manuellement avec potentiellement un tag (JOB, EVENT, PRODUCT, MONEY ou PEOPLE), dans le cas où le signal est intéressant.

Le nombre de signaux validés par classes était le suivant :

- la classe **JOB** : 123 ;
- la classe **EVENT** : 100 ;
- la classe **MONEY** : 14 ;
- la classe **PRODUCT** : 6 ;
- la classe **PEOPLE** : 0 ;
- le reste (soit 1183) est dit « inintéressant ».

## 3. Travail effectué

Des travaux initiaux avaient été réalisés en Python par Samuel Charron. Il avait réalisé un plugin récupérant les signaux, construisant un classifieur naïf bayésien multinomial avec et permettant de classifier de nouveaux signaux. cependant les performances n'étaient pas suffisantes. N'étant pas formé au Python, j'ai préféré commencer mes travaux en utilisant le Java avec l'accord de Samuel. Je savais, en m'orientant vers le Java, qu'une fois que l'application obtiendrait de bonnes performances, j'aurais à implémenter son fonctionnement général en Python sous forme de plugin pour pouvoir l'intégrer à l'architecture de C-Radar.

### 3.1. Démarche de travail

Ce projet s'inscrit parfaitement dans le type de projet R&D. De ce fait, l'avancement est très difficile à planifier dans le temps. Surtout lorsque l'on ne connaît pas les différentes notions sous-jacentes au projet et qu'il y a une bonne part d'auto-formation avant de pouvoir développer une application.

#### 3.1.1. Mes acquis à l'INSA

Les connaissances générales que j'avais en *Data Science*, avant le début du stage, concernaient le *Data Mining* en contexte **numérique** et étaient les suivantes :

- Concepts en analyse et normalisation de données : Analyse en Composantes Principales (ACP), centrage et réduction de données numériques ;
- Concepts d'apprentissage non-supervisé : méthodes de regroupement des données (Clustering : Classification Hiérarchique Ascendante, algorithmes des K-Means et EM) ;
- Base de l'optimisation : méthodes du gradient et de Newton, introduction à l'optimisation sous contraintes convexe ;
- Concepts d'apprentissage supervisé : méthodes pour la discrimination de données (Décision Bayésienne, Régression logistique, SVM linéaire) et notions de validation croisée.

Ce projet ne permet pas de mettre mes connaissances en apprentissage non-supervisé en avant. Cependant, mes notions d'apprentissage supervisé telles que : la démarche à suivre pour construire un classifieur, les concepts liés à la validation des performances (validation croisée) et la notion de sur-apprentissage ; ont été fort utiles.

Mes connaissances en *Text Mining* n'étaient pas suffisamment étoffées pour pouvoir choisir le classifieur le plus adapté à un contexte donné (binaire ou multi-classe). En effet, ma formation (à l'INSA) est axée manipulation et traitement de données **numériques**. De ce fait, une formation en *Text Mining* m'était donc indispensable avant de pouvoir commencer le développement d'une application.

### 3.1.2. Déroulement du stage

Ainsi, durant les deux premiers mois, j'ai exploré le domaine du *Text Mining* et du *Natural Language Processing* au travers de la bibliothèque de Stanford implémentée en Java (*Stanford Natural Language Processing*). Conjointement, j'ai étudié les cours associés, et construit une première application Spring répondant aux contraintes évoquées en partie 2.2 (sauf le critère du langage). Le travail en ressortant est décrit en partie 3.2.

Ensuite, lors du dernier mois, j'ai implémenté le comportement général de cette application sous la forme d'un plugin Python (partie 3.3). Certains composants n'existaient pas en Python, je les ai donc ré-implémentés.

## 3.2. Travaux réalisés en Java, le *Text Mining* et le *Natural Language Processing* avec la bibliothèque de Stanford

### 3.2.1. Présentation de mon environnement de travail

Dans C-Radar tous les traitements de type « computing » (calcul) sont réalisés en Python (cf partie 1.2.2). De ce fait, créer une application permettant de classifier les signaux, devrait être fait en Python sous la forme d'un plugin. Ayant fait le choix de commencer mes recherches en Java, il a fallu m'initialiser un environnement de travail un peu différent de l'environnement de travail Python.

**Spring et MongoDB :** Loïc Petit m'a créé une application Spring de base permettant de me connecter en local à une base de données Mongo. Ce projet Spring est géré avec Maven. L'application me fournit un environnement de travail dans lequel construire le classifieur à partir des signaux validés récupérés depuis une base de données Mongo. Cette base de données contient mon ensemble de signaux permettant de construire et tester mon classifieur. Les 1426 signaux validés y sont stockés ainsi que 350.000 autres signaux non validés (voir le dernier paragraphe de la partie 2.3).

Voici comment les signaux sont stockés dans Mongo :

```
{
  "id" : "TWITTER:agencenetdesign:329129810423083009",
  "content" : "Salon eCom Genève : l'équipe ND est en place au stand E1 \ :) #ecomSITB",
  "publicationDate" : ISODate(2013-04-30T07:07:16Z),
  "sourceId" : "TWITTER:agencenetdesign",
  "source" : {
    "type" : "TWITTER",
    "resourceId" : "agencenetdesign"
  },
  "externalSignalId" : 329129810423083009,
  "validated" : false,
  "validatedTags" : [ ],
  "tags" : [
    "EVENT"
  ],
  "url" : "http://twitter.com/agencenetdesign/status/329129810423083009"
}
```



Le signal comporte :

- un identifieur unique *id* ;
- un contenu *content* ;
- une date de publication *publicationDate* ;
- l'identifieur de la source *sourceId* ;
- la source *source* composée :
  - du type de réseau dont provient le signal *type* ;
  - de l'identifieur du publieur dans ce réseau *resourceId* ;
- un identifieur externe *externalSignalId* ;
- un booléen spécifiant si le signal a été manuellement validé ou non *validated* ;
- la liste des tags s'il a été validé *validatedTags*, la liste des tags potentiels (trouvés par le classifieur) *tags* ;
- l'url là où a été publié le signal *url*.

**Formation Spring et Mongo :** Je me suis rapidement formé à Spring et Mongo pour pouvoir interfacier ces deux composants ensemble. Pour cela, j'ai suivi les tutoriels de Spring disponibles sur <https://spring.io/guides/gs/accessing-data-mongodb/>.

Grâce aux tutoriels, j'ai appris à créer mes premiers points d'API permettant de faire des requêtes dans Mongo. Ces requêtes sont relativement basiques : récupérer tous les signaux sous forme de liste, récupérer uniquement les signaux validés (également sous forme de liste), savoir combien de signaux sont stockés dans ma base, etc. La documentation de Mongo disponible sur <https://docs.mongodb.org/manual/> m'a également bien aidé. Les concepts de base de Mongo ne sont pas très compliqués à comprendre quand on a des notions de base de données.

Dès que j'ai été capable de faire ce type de requête auprès de ma base de données, il fallait me concentrer sur le traitement des signaux, et donc construire un classifieur. C'est à ce moment que je me suis intéressé à la bibliothèque de Stanford.

### 3.2.2. La bibliothèque : Stanford Natural Language Processing

Samuel Charron m'a conseillé de me former en *Text Mining* et en *Natural Language Processing* au travers de cette bibliothèque implémentée en Java. Celle-ci propose un ensemble d'outils pour le traitement automatique de la langue anglaise, chinoise et espagnole.

#### Les cours de l'université de Stanford sur le *Natural Language Processing*

Ces cours accessibles librement sur Internet, m'ont permis de découvrir de nombreux concepts (expliqués ci-après et en partie 3.2.3) :

- Le modèle du *bag of words* ;
- Le modèle *maximum entropy* ;
- Les prétraitements comme la lemmatisation.

Ces cours proviennent du livre *Introduction to Information Retrieval*.<sup>1</sup>

En parallèle, j'ai visionné sur Coursera les vidéos que cette même université avait diffusé suite à un MOOC sur le *Natural Language Processing*. Grâce à ces cours, j'ai pris conscience de l'importance de bien choisir son modèle de classifieur. De plus, j'ai également compris l'intérêt

---

1. Hinrich Schütze Christopher D. Manning Prabhakar Raghavan. Information Retrieval. url : <http://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html>.

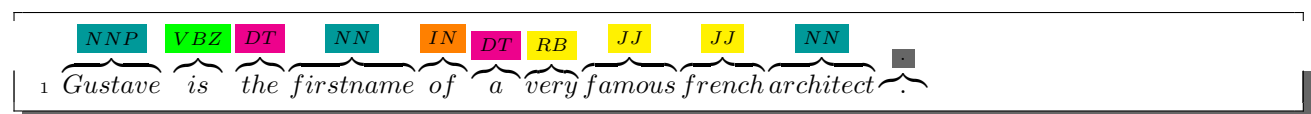
des prétraitements, permettant de normaliser et formater les données textuelles afin d'augmenter les performances et la capacité de généralisation du classifieur.

Durant ma formation au *Natural Language Processing*, j'ai également lu les livres *Natural Language Processing with Python*<sup>2</sup> et *Python 3 Text Processing with NLTK 3 Cookbook*,<sup>3</sup> ainsi que la version web du livre *Introduction to Information Retrieval*.<sup>4</sup>

### Stanford POS Tagger (Part-Of-Speech) :

Le *POS Tagger* permet d'indiquer la fonction grammaticale de chaque mot d'une phrase. Celui-ci fonctionne aussi pour le français.

**Exemple :** Entrée « Gustave is the firstname of a very famous french architect. »  
Sortie :

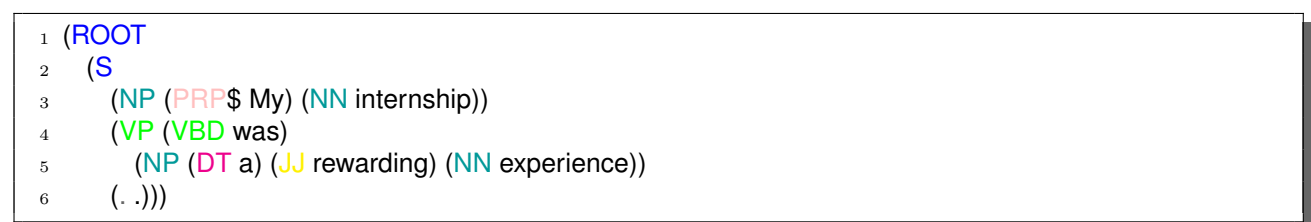


**NNP** signifie qu'il s'agit d'un nom propre singulier, **VBZ** signifie qu'il s'agit d'un verbe à la 3ème personne du singulier au présent, **DT** signifie qu'il s'agit d'un déterminant, **NN** signifie qu'il s'agit d'un nom commun singulier, **IN** signifie qu'il s'agit d'une préposition ou d'une conjonction de subordination, **RB** signifie qu'il s'agit d'un adverbe et **JJ** signifie qu'il s'agit d'un adjectif.

### Stanford Parser :

Le *Parser* permet de connaître la structure grammaticale d'une phrase, à savoir quel(s) groupe(s) de mots forme(nt) le sujet, le verbe et le complément. Cet outil est une sur-couche du *POS Tagger* puisqu'il réutilise son résultat.

**Exemple :** Entrée « My internship was a rewarding experience. »  
Sortie :



### Stanford Named Entity Recognizer :

Le *Named Entity Recognizer* permet d'identifier les groupes de mots qui sont des noms de personnes, d'entreprises, de gènes, etc.

2. Edward Loper Steven Bird Ewan Klein. *Natural Language Processing with Python*. 2009. isbn : 978-0-596-51649-9.

3. Jacob Perkins. *Python 3 Text Processing with NLTK 3 Cookbook*. 2010. isbn : 978-1-78216-785-3.

4. Christopher D. Manning, cf. note 1.

**Exemple :** Entrée « François Bancilhon is the CEO of Data Publica, a Startup located in Paris. »

Sortie :

1 François Bancilhon is the CEO of Data Publica, a Startup located in Paris.

Les mots surlignés en magenta comme **François** **Bancilhon** sont potentiellement des noms de personnes (**PERSON**). Ceux surlignés en orange comme **Data** **Publica** sont potentiellement des noms d'organisations (**ORGANIZATION**). Enfin, les mots surlignés en violet comme **Paris** sont potentiellement des noms de lieux (**LOCATION**).

## Stanford Classifier :

Le **Classifier** permet de construire un classifieur automatique pour la catégorisation de texte. Un classifieur **maximum entropy** et un naïf bayésien sont disponibles. Cet outils propose une classe (**ColumnDataClassifier**) qui permet de construire n'importe quel type de classifieur, simplement en lui fournissant un fichier de configurations, et des données d'apprentissage et de test. Le soucis de cette classe est que les données doivent se conformer à un format bien précis.

## Stanford CoreNLP :

Le **CoreNLP** permet de construire une suite de traitements automatiques sur de l'anglais, du chinois, de l'espagnol. Un exemple de chaîne de traitement est visible en figure 3.1. La particularité de cet outil est qu'il permet de réutiliser les traitements précédents les uns à la suite des autres (sauf le **Stanford Classifier**). De plus, il permet aussi d'appliquer des traitements supplémentaires à ceux énoncés jusque là. En effet, celui-ci permet de faire de la recherche morphologique telle que lemmatisation ou stemming. Le **POS Tagger** et le **Tokenizer** sont réutilisés pour ces traitements.

## Lemmatisation et stemming :

La lemmatisation consiste à trouver les lemmes de chaque mot d'une phrase. Ce traitement est détaillé dans le paragraphe **La lemmatisation** en partie 3.2.4. Ce traitement nécessite que la phrase soit préalablement découpée en tokens (les mots de la phrase) par un **Tokenizer** et que chacun d'entre eux soit tagué par le **POS Tagger**. Ainsi, en couplant la fonction grammaticale d'un mot avec un dictionnaire, il est possible de retrouver le lemme du mot.

**Exemple :** Entrée « No better example than these could have been found. »

Sortie :

1 no good example than this can have be find  
No better example than these could have been found.

**Le stemming :** Le stemming est un traitement assez similaire à la lemmatisation. Celui-ci consiste à trouver les stems et non les lemmes. Ce traitement est détaillé dans le paragraphe **Le stemming** en partie 3.2.4.

**Exemple :** Entrée « Les chevaliers chevauchent leur chevaux, le cavalier son cheval. »

Sortie :

1 <sup>le</sup> Les <sup>cheva</sup> chevaliers <sup>cheva</sup> chevauchent <sup>leur</sup> leur <sup>cheva</sup> chevaux ; <sup>le</sup> le <sup>caval</sup> cavalier <sup>son</sup> son <sup>cheva</sup> cheval ;

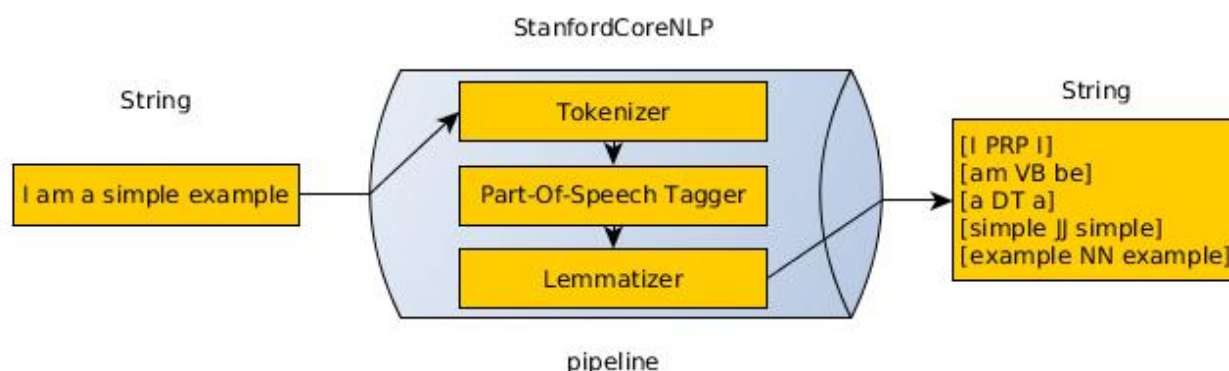


Figure 3.1. – Trois traitements sont assignés au pipeline (entité de la classe *StanfordCoreNLP*) : Tokenization, POS-Tagging et Lemmatization. En sortie, une chaîne de caractère présente le résultat de chaque traitement en colonne.

## Premier bilan sur la bibliothèque

Dans un premier temps, mon objectif était seulement de réutiliser la partie *Stanford Classifier* de la bibliothèque pour créer un classifieur binaire. Celui-ci devait être capable de classifier les signaux relatifs aux offres d'emploi et de stage (soit le tag *JOB*). La classe *ColumnDataClassifier* m'a été d'une grande aide, même si elle exige des données dans un format particulier et qu'elle ne gère pas certains prétraitements.

Des traitements comme la lemmatisation ou le stemming ont été utiles, lors de la normalisation des données. Pouvoir utiliser le *StanfordCoreNLP* pour normaliser et sélectionner les features utiles aurait été l'idéal, avant de les fournir au classifieur pour construire un modèle.

## 3.2.3. Première mise en pratique de la bibliothèque de Stanford

### Première application Spring

J'ai construit une application réalisant les actions suivantes (visible en figure 3.2 et expliquée ci-après) :

- Récupérer les signaux stockés dans Mongo sous forme de liste ;
- Créer un ensemble de données à partir des signaux validés manuellement ;
- Diviser aléatoirement cet ensemble de données en deux ensembles (un pour entraîner le classifieur et un pour le tester) tout en gardant la proportion de chaque classe dans les deux ensembles ;
- Entraîner un classifieur binaire naïf bayésien ;
- Fixer les hyper-paramètres du classifieur par validation croisée pendant la phase d'apprentissage ;
- Évaluer la qualité du classifieur construit (l'erreur de généralisation) en calculant sa précision et son rappel sur un ensemble de données de test (qui n'ont pas « été vues » jusqu'ici par le classifieur).

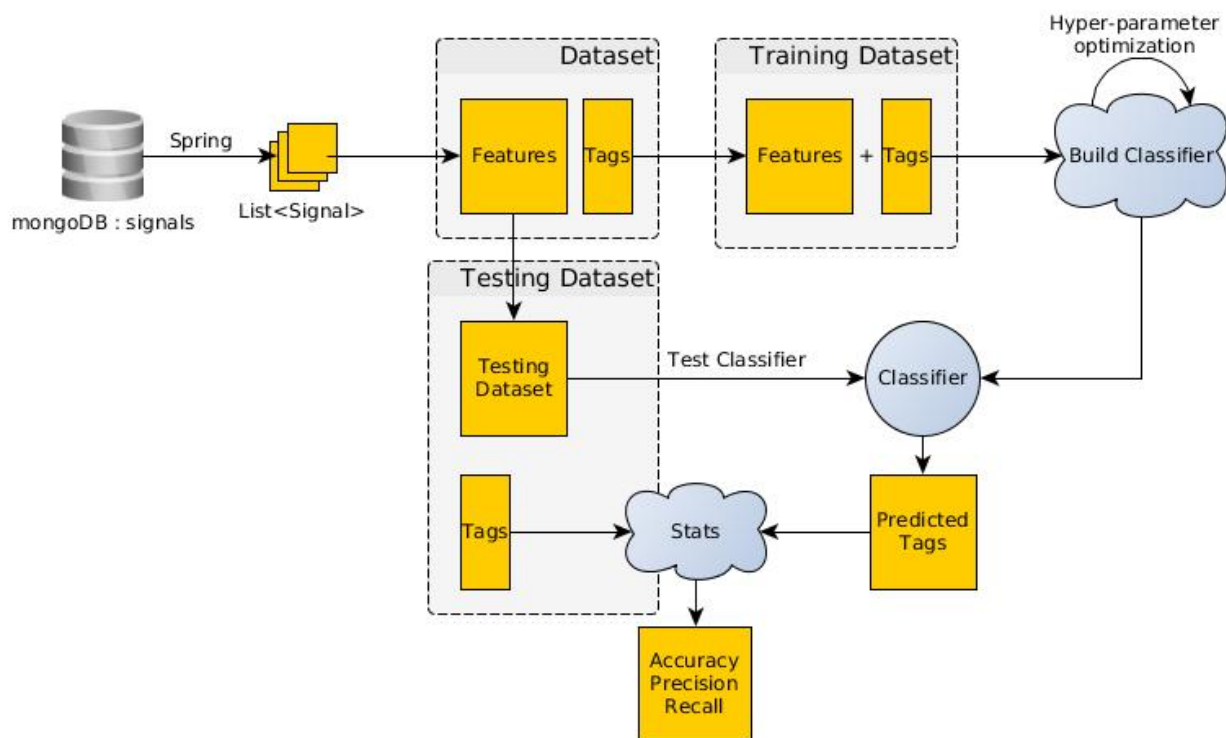


Figure 3.2. – La construction du classifieur.

## Le modèle de classifieur

Quel type de classifieur construire ? Un SVM ? Une régression logistique ? Un modèle naïf bayésien ?

Dans la littérature de la classification de texte, comme la détection de spam dans les emails ou l'analyse des sentiments (savoir si un texte est critique ou élogieux), il est plutôt commun de construire des classifieurs naïfs bayésiens avec comme caractéristique la fréquence des mots. Ainsi, j'ai choisi de construire un tel classifieur pour catégoriser mes signaux. (C'est également ce qu'avait fait Samuel Charron en Python.)

## Construction de l'ensemble de données

Ensuite, s'est posée la question de comment utiliser les données labellisées pour construire un ensemble de données pour l'apprentissage et la validation. Sur ce point, j'ai choisi de construire mon ensemble de données selon le modèle du *bag of words*.

Dans ce modèle, un texte (une phrase ou un document) est représenté comme un sac (*bag*), un ensemble de ses mots (au sens mathématique), sans se préoccuper de la grammaire ou de l'ordre des mots, mais en gardant la multiplicité. Ce modèle est communément utilisé en classification de document, quand la fréquence, l'occurrence des mots est utilisée comme caractéristique. C'est le cas du modèle naïf bayésien.

Ainsi, un signal est caractérisé par la liste des occurrences des mots de son titre et son contenu.

**Exemple :** Modélisation de deux signaux à l'aide du *bag of words* :

1 Offre d'emploi : Ingénieur d'études et développement Java.

À partir de ces deux signaux, une liste de mot est construite à l'aide d'un tokenizer :

```
[ "Offre", "d'", "emploi", ":", "Ingénieur", "études", "et", "développement",
  "Java", ".", "de", "stage", "Classification", "signaux", "entreprises",
  "ou", "Python" ]
```

Celle-ci contient 17 mots distincts. En utilisant son index, on peut représenter chaque signal comme un vecteur de taille 17 :

```
1 [ 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0 ]
2 [ 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 2, 1, 1, 1, 1, 1, 1 ]
```

Chaque ième composante du vecteur représente le nombre de fois que le ième mot de la liste est présent dans le signal. Par exemple, dans le premier vecteur (qui représente le contenu du premier signal), les deux premières composantes sont « 1, 2 ». La première composante correspond au mot « Offre » qui est le premier mot de la liste, et sa valeur est « 1 » car « Offre » est présent une fois dans le premier signal. De la même façon, la deuxième composante correspond au mot « d' », le deuxième mot de la liste et sa valeur est « 2 » car il est présent deux fois dans le signal.

La méthode de construction de l'ensemble de données d'apprentissage est restée la même pour la suite des expérimentations. Pour mon premier classifieur binaire, les 123 signaux validés *JOB* de l'ensemble des 1426 signaux validés (présentés dans le dernier paragraphe de la partie 2.3), forment la classe *JOB* contre le reste. J'ai utilisé le *PTB Tokenizer* de *Stanford Classifier* (détails au paragraphe **La segmentation ou tokenisation** de la partie 3.2.4) pour découper les phrases en mots.

Enfin, le seul pré-traitement des données mis en place est le fait de garder les éléments du *bag of words* apparaissant au moins cinq fois. Ceci afin de supprimer les mots rares qui n'apportent pas d'informations, quand on fait de la classification textuelle (les mots mal orthographiés, les mots très spécifiques, les noms propres, etc).

## Construction du classifieur

Pour construire mon classifieur binaire naïf bayésien, j'ai divisé mon ensemble de données en deux ensembles de données dans les proportions suivantes :

- 75% de l'ensemble de données pour l'ensemble d'apprentissage ;
- 25% de l'ensemble de données pour l'ensemble de test.

À noter que la méthode de la bibliothèque, permettant de diviser l'ensemble de données en deux ensembles, ne garantit pas que la proportion des classes soit conservée dans les deux ensembles résultant. De plus, l'ensemble n'est pas mélangé à chaque nouvelle construction du classifieur, il y donc un risque de sur-apprentissage. Christian Frisch me l'a fait remarqué, lors d'une réunion de suivi. J'ai donc repris le code source de la méthode pour ré-implémenter le bon comportement.

## Optimisation des hyper-paramètres par validation croisée

Une méthode de la bibliothèque se charge d'optimiser l'hyper-paramètre<sup>5</sup> du classifieur par validation croisée<sup>6</sup>. Dans un premier temps, j'ai simplement réutilisé cette méthode, car les

5. Définition en annexe A.3.1.

6. Définition en annexe A.3.2.

performances en étaient bonifiées.

## Évaluation de la qualité du classifieur

Enfin pour évaluer la qualité de mon classifieur, je l'ai testé sur un ensemble de test, non utilisé jusqu'ici (les 25% de l'ensemble initial). J'ai calculé la précision et le rappel obtenus par la classe **JOB** et par la classe **USELESS** représentant le reste (pas **JOB**). Ceux-ci sont visibles en table 3.1. Pour rappel, ce classifieur a été construit avec comme exemples pour la classe **JOB**

	JOB	USELESS
Précision	0,84	0,99
Rappel	0,91	0,98

Table 3.1. – Performances du classifieurs.

les 123 signaux validés **JOB** de l'ensemble des 1426 signaux validés (présentés dans le dernier paragraphe de la partie 2.3), contre le reste des signaux.

**Remarques :** Ici, on met l'accent sur le fait d'avoir une précision<sup>7</sup> très élevée quitte à avoir un rappel<sup>8</sup> un peu diminué (car quand l'un augmente l'autre diminue et vice versa). On préfère se tromper très rarement dans notre classification (précision élevée) quitte à rater des signaux (rappel moyen). Autrement dit, on préfère qu'un utilisateur voit peu de signaux intéressants plutôt qu'il voit beaucoup de signaux dont l'intérêt reste à démontrer.

## Premier bilan

Le classifieur construit est prometteur. Cependant, la généralisation à plusieurs classes, va amener à faire de nouveaux choix :

- Faut-il adopter une stratégie de type *One versus All*<sup>9</sup> ou *One versus One*<sup>10</sup> ?
- Est-ce que le modèle naïf bayésien va bien se généraliser ?
- Faut-il un prétraitement global et/ou spécifique à chaque classe ?

## 3.2.4. Amélioration de ma première application et prétraitement

La démarche de construction du classifieur présentée dans la partie précédente (3.2.3) reste la même. Cependant, j'ai repris la classe principale **ColumnDataClassifier** de la bibliothèque **Stanford Classifier**, car elle intègre, notamment, un vrai processus de validation croisée en k plis. Je l'ai donc refactorisée à plusieurs fins :

- Pour qu'elle récupère mes données directement depuis Mongo ;
- Pour qu'elle leur applique des prétraitements avant de construire l'ensemble de données ;
- Pour qu'elle réalise de la validation croisée stratifiée<sup>11</sup> sur dix plis lors de l'apprentissage du classifieur.

Ce dernier point, me permet d'avoir une idée de la capacité de généralisation de mon classifieur.

---

7. Définition en annexe A.4.2.

8. Définition en annexe A.4.3.

9. Définition en annexe A.5.1.

10. Définition en annexe A.5.2.

11. Définition en A.3.3.



## Agrandissement de mon ensemble de signaux validés

Avant de pouvoir passer à un classifieur multi-classe, il fallait que mon ensemble de signaux validés grandisse. En effet, celui-ci souffre d'un fort déséquilibre entre les classes : il y a beaucoup plus de signaux inintéressants qu'intéressants. Ainsi, considérer les prédictions d'un classifieur construit avec ces données comme correctes était impossible. Il a donc été nécessaire d'en valider d'autres manuellement.

Durant mon stage, j'ai donc organisé plusieurs QAs<sup>12</sup> (*Quality Assessment*) pour approfondir l'ensemble des signaux d'apprentissage et de test.

**Proportions des classes dans les signaux validés :** La proportion des signaux sans intérêt est énorme (plus de 80% sur environ 300.000).

Pour les QAs, j'ai donc fait une pré-sélection des signaux à valider. J'ai réutilisé le premier classifieur implémenté en Python de Samuel Charron, construit avec les 1426 signaux. Celui-ci a classifié des signaux non validés. Ce sont les signaux appartenant potentiellement aux catégories **EVENT**, **JOB** et **PRODUCT** (« selon le classifieur ») qui ont été sélectionnés pour être validés manuellement. Pour les classes **MONEY** et **PEOPLE**, ils ont été sélectionnés dans Mongo à l'aide d'expressions rationnelles sur des termes comme « levée de fonds », « chiffre d'affaire », « nommer », « nomination », etc.

De cette manière, 2574 nouveaux signaux ont été validés. Ce type de méthode biaise les proportions des classes de signaux intéressantes par rapport à celle sans intérêt, mais autrement il mettait impossible d'obtenir plus d'exemple.

Au 27.07.2015, il y avait donc 4000 signaux validés manuellement :

- 488 catégorisés **EVENT** soit 12,2%
- 258 catégorisés **JOB** soit 6,4%
- 118 catégorisés **MONEY** soit 3%
- 83 catégorisés **PRODUCT** soit 2,1%
- 49 catégorisés **PEOPLE** soit 1,3%
- 3004 validés mais considérés comme inintéressant soit 75%

## Le passage au multi-classe

Pour le passage au multi-classe, la bibliothèque de Stanford impose une stratégie *One versus All*. Concernant le modèle de classifieur, deux options s'offraient à moi :

- Un modèle génératif<sup>13</sup> et notamment un classifieur naïf bayésien ;
- Un modèle discriminant<sup>14</sup> et notamment un classifieur qui maximise l'entropie.

Jusque là, j'avais opté pour le modèle bayésien car c'est ce qui ressortait le plus souvent dans la littérature. J'ai donc repris mon application précédente et j'ai construit un classifieur naïf bayésien multinomial<sup>15</sup> avec les 4000 signaux tagués de la même manière que précédemment. Les performances obtenues (visibles en table 3.2) avaient grandement baissées (la classe **PRODUCT** n'y figure pas car elle n'était pas suffisamment représentée).

---

12. Définition en annexe A.6.

13. Définition en annexe A.7.1.

14. Définition en annexe A.7.2.

15. qui suit plusieurs lois binomiales.



	JOB	EVENT	PEOPLE	MONEY
Précision	0,63	0,51	0,46	0,44
Rappel	0,91	0,88	0,63	0,87

Table 3.2. – Performances du classifieur naïf bayésien multinomial.

## Le modèle discriminant maximisant l'entropie

Par curiosité, j'ai changé le modèle pour le modèle maximisant l'entropie. À ma grande surprise, ce modèle obtint de bien meilleures performances que le modèle bayésien dans les mêmes conditions (construction de l'ensemble de données identique, prétraitements identiques, etc). Les performances sont visibles en table 3.3.

	JOB	EVENT	PEOPLE	MONEY
Précision	0,96	0,82	0,73	0,81
Rappel	0,73	0,60	0,50	0,49

Table 3.3. – Performances du classifieur maximisant l'entropie.

**Le *MaxEnt* ou *Maximum Entropy* :** Les modèles *maximum entropy* sont aussi connus sous le nom de *softmax classifiers* et sont équivalents aux modèles de régression logistique multi-classe (avec des paramètres différents). Il s'agit de la généralisation de la régression logistique au cas multinomial. Dans ce cas, maximiser la vraisemblance revient à maximiser l'entropie. Il s'agit ni plus ni moins que du passage du cas binaire au cas N classes. Ce type de modèle (*MaxEnt*) est avantageux dans le cas où les données sont *sparse*<sup>16</sup>. Ce qui est le cas, d'où des résultats meilleurs qu'avec le modèle bayésien.

## Les travaux de prétraitement global

Les traitements, expliqués ci-après, sont réalisés dans la construction de l'application, en amont de la construction de l'ensemble de données selon le modèle du *bag of words*.

**La segmentation ou tokenisation :** La tokenisation consiste à découper une phrase en tokens, dans l'idéal représentant des mots. Cette tâche est très spécifique à chaque langue. Les difficultés principales sont surtout liées aux contractions. Il est nécessaire de s'attarder sur cette phase de la normalisation pour minimiser la perte sémantique, car beaucoup de traitements s'appuient sur la tokenisation. Dans mon application, j'ai utilisé le *PTB Tokenizer* disponible dans *Stanford Classifier*.

**Exemple :** Entrée : « Je n'ai pas d'argent. En as-tu ? »

Sorties possibles :

"Je", "n", "ai", "pas", "d", "argent", "En", "as", "tu"

"Je", "n", "'", "ai", "pas", "d", "'", "argent", ".", "En", "as", "-", "tu", "?"

"Je", "n'", "ai", "pas", "d'", "argent", ".", "En", "as", "-", "tu", "?"

Le résultat du *PTB Tokenizer* dans l'exemple précédant est le premier.

16. beaucoup de dimensions pour décrire les données. Pour un exemple, la plupart des coordonnées sont nulles (ce qui est le cas pour les données textuelles avec un modèle de type bag of words parce que un texte utilise une faible proportion du vocabulaire existant).

**La casse et la ponctuation :** Pour normaliser les mots, une manière simple consiste à réduire la casse de tous les mots. Ainsi, on réduit notre ensemble de mot. Cela est très facile à effectuer mais cela a quand même quelques défauts :

- Les noms propres perdent leur différences par rapport aux noms communs ;
- Les noms d'organisation (comme l'OTAN) perdent leur sens en minuscule.

De plus, supprimer la ponctuation permet aussi d'épurer les données car celle-ci n'apporte pas d'information. Encore une fois, c'est simple à réaliser mais certains mots perdent leur sens.

- Les mots composés comportant des tirets perdent leur sens (exemple : « après-midi ») ;
- Les mots composés comportant des apostrophes perdent leur sens (exemple : « aujourd'hui ») ;
- Les acronymes comportant des points de séparation perdent leur sens (exemple : « U.S.A. »)

**Les stopwords :** Les stopwords sont les mots d'une phrase inutiles à la compréhension de celle-ci. Ils ne portent pas d'information et sont présents dans tous les documents textuels. Il est avantageux de les supprimer pour réduire le bruit et la taille du vocabulaire. Une liste de stopwords contient majoritairement des pronoms, des prépositions et des déterminants comme : « a », « au », « ce », « de », « le », « mon », etc.

**La recherche morphologique :** Enfin, les deux derniers moyens permettant de normaliser du texte sont la lemmatisation et le stemming. Ces deux traitements ont pour objectif de faire baisser le nombre de forme infléchi. Une forme infléchi est appelé un lexème en morphologie. Il faut savoir qu'un lexème est composé de différentes types de morphèmes : les stems et les affixes. Voici leur définition à l'aide d'un exemple :

Le lexème « chanteurs » est composé de trois morphèmes : « chant », « eur » et « s ». Parmi ces trois morphèmes, l'un est un stem « chant » et les deux autres des affixes « eur » et « s ».

**Le stemming :** Le stemming réduit les lexèmes en leur stems : « chanteurs », « chanteuse » transformés en « chant ».

Un des désavantage du stemming est que les stems ne sont pas toujours des lemmes, c'est à dire la forme canonique d'un lexème (son entrée dans le dictionnaire), et donc pas un mot.

Exemple : le stem de « chercheur » est « cherch » (n'est pas un mot).

**La lemmatisation :** La lemmatisation réduit les lexèmes en lemmes, la forme canonique d'un lexème. Le lemme d'un verbe correspond au verbe à l'infinitif, le lemme d'un nom commun est ce nom commun au masculin singulier, etc. Les lemmes correspondent aux entrées dans le dictionnaire de tous les lexèmes.

Exemple : le lemme de « préférées » est « préférer ».

**Remarques :** Le stemming fait perdre plus d'information à nos features que la lemmatisation, j'ai donc choisi de garder les lemmes des mots.

La bibliothèque de Stanford ne propose pas de Lemmatizer pour la langue française. J'ai utilisé une [bibliothèque externe](#) de Ahmet Aker, que j'ai rajoutée dans le projet Spring grâce à Maven en ajoutant une dépendance.

## Bilan sur les prétraitements globaux

Dans l'idéal, j'aurais aimé pouvoir réaliser les prétraitements avec le *Stanford CoreNLP* mais il ne gère pas bien le français et il est très gourmand en mémoire.

En outre, les prétraitements permettent de normaliser le texte et de réduire la taille du vocabulaire, ce qui facilite la tâche du classifieur. Il est important de réduire la disparité des features pour pouvoir calculer leur fréquence par la suite.

Un parallèle pourrait être fait entre la lemmatisation et la compression de données numériques par ACP (Analyse en Composantes Principales).

## Feature engineering (ou les travaux de prétraitement spécifiques)

Il est possible et important de traiter nos données plus spécifiquement, afin d'augmenter les performances du classifieur par la suite. Pour cette tâche, il faut d'avantage s'attarder sur le fond des données de chaque classe, et se demander si des informations (tel que des motifs particuliers) ne seraient pas être détruites par les traitements précédents. Tous les traitements qui suivent sont réalisés en amont de la tokenisation.

**Les signaux issus de Twitter :** Ceux-ci comportent souvent des références « @pseudo » et des mentions comme « #job ». Les références n'apportent pas d'information dans la majorité des cas. Les supprimer permettrait d'éliminer du bruit. Quant aux mentions comme « #job », celles-ci portent de l'information et dans un processus de normalisation, il serait bien de seulement garder le mot suivant le dièse (#).

**Les emails et les URLs** À l'image des pseudonymes Twitter, les emails et les URLs (présents dans certains signaux) ne portent aucune information et sont mal tokenisés du fait de leur formes. Ainsi, il serait bien de les supprimer.

**Les signaux de la classe *JOB* :** Souvent dans les offres d'emploi ou de stage, la mention « H/F » signifiant « homme ou femme » est présente. Cependant, lors de la tokenisation, ce type de motif est détruit à cause du caractère de ponctuation slash (/). Ainsi, il serait intéressant de pouvoir les détecter et de les remplacer par une chaîne de caractère sans ponctuation comme le terme « hommeoufemme ».

**Les signaux de la classe *MONEY*** Une caractéristique des signaux de cette classe est que, souvent lorsque ces signaux parlent d'une levée de fonds, une somme est annoncée avec une devise. Exemple : « L'INSA a levée 5 k€ pour construire un amphithéâtre ».

Il serait donc intéressant de conserver l'unité et le symbole de la devise suivant le montant qui est caractéristique de ce type de signal (k€, m€, etc).

**Mise en place de ces traitements et bilan :** J'ai implémenté ces traitements assez facilement à l'aide d'expressions rationnelles. Ce travail d'inspection du contenu des signaux est important car c'est là que l'on détecte des informations caractéristiques.

## Performances obtenues par mon application avec les prétraitements

Les performances, présentées en table 3.5, ont été obtenues en ajoutant tous les prétraitements présentés dans cette partie au processus de construction du classifieur.

**Le problème des FP (faux positifs) :** Les signaux qui peuvent avoir potentiellement plusieurs labels engendrent une quantité non négligeable de faux positifs, comme par exemple :

- Signal : "Venez découvrir nos nouveautés produits du 20 au 22 Mai 2014 au salon SEPÉM, Hall 4 - Stand F13 <http://t.co/vS7gT2x4yp> #marquagepermanent" Label : *PRODUCT* ou *EVENT*

	JOB	EVENT	PEOPLE	MONEY
Précision	0,96	0,82	0,94	0,81
Rappel	0,81	0,63	0,60	0,60

Table 3.4. – Performances du classifieur maximisant l'entropie.

	JOB	EVENT	PEOPLE	PRODUCT	MONEY
Précision	0,98	0,70	0,83	0,56	0,51
Rappel	0,92	0,80	0,97	0,75	0,95

Table 3.5. – Performances du classifieur sur des données non validées.

- Signal : "Nous embauchons ! Étudiants de HEC Paris, nous sommes aujourd'hui au #CarrefoursHEC. Venez découvrir nos offres d'emploi dans les domaines du #digital #data #marketing" Label : **JOB** ou **EVENT**

Une solution potentielle serait d'autoriser le classifieur à attribuer plusieurs label. C'est ce qu'avait fait Samuel Charron en Python mais les résultats n'étaient pas bons.

**Le problème des FN (faux négatifs) :** Les signaux ne sont pas toujours très directs et synthétiques. Parfois en plus d'un contenu intéressant le signal peut contenir du bruit, ce qui engendre une quantité non négligeable de faux négatifs. Ce qui baisse les rappels. Exemple :

- Signal : "Dans le cadre de son développement, EXCELIUM a fait l'acquisition du fonds de commerce de la société lyonnaise SES Vidéo, spécialisée dans l'installation de systèmes vidéo et d'alarme." Label : Aucun alors que c'est **MONEY**
- Signal : "A l'occasion de la fête de la gastronomie, nous vous invitons à visiter notre nouveau laboratoire à Tétéghem samedi 27 septembre de 10 h à 18h. Visite, dégustation et chèques cadeaux offerts pour toute commande passée sur le site cette semaine !" Label : aucun alors que c'est **EVENT**

### Performances du classifieur sur un ensemble de données non vus (obtenus grâce à un QA) :

341 718 signaux ont été tagués automatiquement par le classifieur entraîné sur 4000 signaux validés. (Certaines sources de signaux sont blacklistées et leurs signaux émis ne sont donc pas tagués par le classifieur : Les entreprises d'interim, les chasseurs de tête, les e-commerçants, etc). Parmi tous ces signaux, 388 ont été validés manuellement lors d'un QA. Ces signaux validés ont été sélectionnés aléatoirement parmi les 341 718 signaux tagués. Les performances du classifieur sur ces signaux sont visibles en 3.5.

Malgré tout, grâce à tout ces prétraitements, mon classifieur a obtenu de meilleures performances, jugées suffisantes par mon maître de stage Samuel Charron pour que je puisse passer à l'implémentation en Python. Les classes **PRODUCT** et **MONEY** manquent d'exemples représentatifs, d'où leurs mauvais résultats. Un QA supplémentaire sur ces classes leur permettrait de se bonifier.

## 3.3. Travaux réalisés en Python

À la suite de ces deux mois de formations au *Text Mining*, au *Naturel Language Processing* et à l'*Information Retrieval*, j'étais à présent capable d'identifier les tâches à réaliser en Python :

- Construire un module Python permettant de réaliser tous les prétraitements expliqués précédemment ;

- Reprendre le plugin Python de Samuel Charron afin d'appliquer ces prétraitement en amont de la construction de l'ensemble de données ;
- Changer de modèle de classifieur.

Les difficultés majeures rencontrées lors de ce dernier mois de stage, étaient le langage Python, lui-même, que je connaissais peu et le besoin de pallier l'inexistence de lemmatiser pour le français en Python.

Concernant le langage Python, j'ai assez facilement surmonté cette difficulté grâce à mes connaissances en Java, en algorithmique ainsi que mes bases en Scala.

Le problème du lemmatiser a été plus difficile mais très intéressant et stimulant.

Les livres *Natural Language Processing with Python*<sup>17</sup> et *Python 3 Text Processing with NLTK 3 Cookbook*<sup>18</sup> m'ont énormément aidé.

### 3.3.1. Modules implémentés

#### Module de prétraitement spécifique aux signaux

Ce module de prétraitement (basé sur des expressions rationnelles) prend en charge la détection des motifs précédents à savoir : les emails, les URLs, les pseudonymes et les références Twitter, les mentions « H/F », les devises, la ponctuation et les stopwords ; et les traite spécifiquement (voir le paragraphe **Feature engineering** partie 3.2.4).

#### Module de tokenisation

Pour ce module, j'ai réutilisé le module [NLTK](#) (*Natural Language Toolkit*) en Python qui propose un tokenizer de texte en phrase pour le français. Pour la tokenisation des phrases en mots, il n'y a pas de tokenizer spécifique pour le français. J'ai donc repris un tokenizer *TreebankWordTokenizer* auquel j'ai apporté quelques règles pour gérer les apostrophes et les tirets.

#### Module de lemmatisation

En Python 3, il n'existe pas de module permettant de faire de la lemmatisation de la langue française. J'ai réutilisé le *SnowballStemmer* disponible dans la bibliothèque NLTK. Voyant que les résultats étaient décevants, je me suis alors demandé comment le lemmatiser utilisé en Java fonctionnait. Celui-ci couplait le *POS Tagging* des mots à un autre outils, le HFST (*Helsinki Finite State Transducer*), permettant de trouver les différents lemmes possibles d'un mot.

**Exemple d'utilisation du HFST :** Entrée : « suis »

Sortie :

- 1 être + verb + singular + indicative + present + firstPerson
- 2 suivre + verb + singular + imperative + present + secondPerson
- 3 suivre + verb + singular + indicative + present + firstPerson
- 4 suivre + verb + singular + indicative + present + secondPerson

17. Steven Bird, cf. note 2.

18. Perkins, cf. note 3.

**Idee du lemmatizer Java :** J'ai donc reproduit l'idée du lemmatizer Java pour l'implémenter en Python :

1. Pour chaque phrase, calculer le *POS Tagging* de chaque mot via un POS Tagger ;
2. Pour chaque mot de chaque phrase, trouver tous ses potentiels lemmes via le HFST ;
3. Choisir le bon lemme grâce à l'information apporter par son *POS Tagging*.

## Helsinki Finite-State Transducer Technology (HFST)

Le *Helsinki Finite-State Transducer software* est une implémentation de plusieurs outils d'analyse morphologique et d'autres outils basés sur les transducteurs (éventuellement pondérés) à états finis. Il s'agit d'un [grand projet](#) portant sur les transducteurs littéralement en anglais « transform reducer ».

Le HFST propose un module Python permettant de trouver tous les lemmes d'un mot grâce notamment à un modèle de transducteur du français.

## Stanford POS Tagger

Pour le POS Tagging, il existe un wrapper Python du *Stanford POS Tagger*. Cet outil fournit également un modèle français pour taguer des phrases en français.

**Qualité du POS Tagger :** Pour vérifier la qualité de celui-ci, j'ai utilisé des corpus de texte français. Ceux-ci présentent un texte dans un format XML, sous forme de liste de paires mot/-fonction grammaticale. J'ai dû transformer le corpus en un texte composé de phrases afin de pouvoir le passer à mon *POS Tagger* Python. J'ai également dû transformer le corpus car les séparateurs utilisés créaient un décalage entre la sortie de mon POS Tagging et cette liste de paires.

Une fois les problèmes de formatage résolus pour un texte du corpus, j'ai testé le *POS Tagger* dessus et celui-ci a obtenu un taux d'erreur jugé suffisant (inférieur à 10%).

## Lemmatiser Python :

J'ai donc combiné ces deux outils pour mettre au point un lemmatiser en Python 3. Avec Clément Chastagnol, on a pu le déployé, mais un problème inattendu est survenu. En effet, il arrivait que le Tagger (jar appelé par NLTK) freeze. Dans ce cas, le programme s'arrêtait, car Python ne reprenait pas la main (100% du CPU sur le processus Java lancé). Le problème a été résolu temporairement en modifiant directement NLTK dans l'environnement virtuel Python *virtualenv* en intégration. On a rajouté un timeout sur le subprocess lançant la commande Java.

En plus de ce problème, on sait aperçu que la lemmatisation était extrêmement lente (facteur 100 sur la durée de traitement d'un signal) malgré une parallélisation des tâches (multithread), car à chaque appelle de la méthode de POS Tagging, le processus lancé par le wrapper Python relançait une nouvelle JVM.

La lemmatisation permettait d'améliorer les performances du classifieur de 2% environ. Comme ce progrès était infime, on a retiré la lemmatisation pour la remplacer par le stemming.

### 3.3.2. Plugin de classification

Le plugin de classification permet de récupérer les 4000 signaux depuis un point d'API, de les normaliser et de construire un ensemble de données avec, selon le modèle du *bag of words*, et enfin de créer un classifieur *maximum entropy* (une régression logistique multinomiale) *One vs All* à partir de l'ensemble de données construit. Ce dernier est très efficace de par ses performances en classification, mais aussi de part la rapidité de construction de son modèle (30 secondes environs) et de son déploiement.

Sa qualité est évalué par une validation croisée sur dix plis. Il se révèle aussi performant que le modèle créé en Java à peu de choses près (2% d'écart selon les classes maximum).

Une fois le classifieur créé, ce plugin est déployable et permet de taguer de nouveaux signaux.

Ce plugin utilise les modules Python 3 [scikit-learn](#) et [numpy](#). Numpy est une bibliothèque d'algèbre vectorielle. Scikit-learn est un module de machine learning construit au dessus de numpy.

L'utilisation de scikit-learn est relativement facile car la documentation est très bien faite mêlant explications aux exemples d'utilisation.

## 3.4. Bilan sur mes travaux

J'ai réussi à mettre au point un plugin Python fonctionnel, capable de classifier les signaux captés par C-Radar. Celui-ci obtient de meilleures performances que le précédent, ce qui était l'objectif, et celles-ci sont satisfaisantes. De plus, pour souligner l'aboutissement de ce stage, ce plugin devrait être déployé en production dans quelques semaines.

Les principales difficultés rencontrées en première partie de stage sont liées à mon manque de connaissances en *Text Mining* et *Natural Language Processing*. J'ai cependant pu surmonter ces difficultés grâce à un bon accompagnement de la part de mes tuteurs et grâce à de l'expérience acquise, notamment, lors du projet de type R&D pour Orange Vallée au semestre 8.

J'ai rencontré d'autres difficultés. Celles-ci liées au le manque d'outils de *Naturel Language Processing* disponible pour la langue française. En effet, pour l'anglais les outils sont très bien aboutis et nombreux, mais pour le français, il est difficile d'en trouver. Je pense, notamment, au Lemmatiser qui n'existe pas pour le français en Python 3. D'autres outils existent mais en Python 2. (De grandes différences existent entre Python 2 et Python 3. Parfois, des wrappers permettant de faire le pont entre les deux, mais leur utilisation ne sont pas de bonnes pratiques.)

Les pistes d'amélioration possibles sont les suivantes :

- Introduire un outil permettant de détecter la langue du signal (comme [Tika](#)) afin d'adapter les traitements en aval ;
- Utiliser un détecteur d'entités nommées *Named Entity Recognizer* afin de gérer autrement les noms d'entreprises, de personnes qui sont généralement caractéristiques des signaux *MONEY* (exemple : « Data Publica a levé 10 m€ en janvier auprès du fond d'investissement français constitué du groupe Lagardère... »). Pour le moment l'information apportée par ces noms propres n'est pas utilisée puisqu'ils ne sont pas retenus dans le vocabulaire du fait de leur rareté.



- Éventuellement, redéfinir la classe **MONEY** en deux classes, comme suit, pour voir si le classifieur s'améliore :
  - **MONEY** : Déclaration de résultats financiers, de CA (chiffre d'affaire), de levée de fonds ;
  - **BUSINESS** : Déclaration de partenariat, de rachat ou d'acquisition d'entreprises, de « contrat gagné », etc.
- Se tourner vers des outils statistiques comme le traducteur de Google pour exécuter la lemmatisation.



## 4. Conclusion

Ce stage a été une expérience très enrichissante tant sur le plan humain et relationnel, que sur le plan technique. Je souhaitais absolument découvrir l'univers de la start-up et c'est chose faite grâce à Data Publica et ses équipes.

Sur le plan humain, j'ai eu la chance de rencontrer des personnes extrêmement sympathiques et compétentes, mais aussi ouvertes à la discussion et toujours disponibles. L'équipe de Data Publica m'aura toujours bien encadré. Que ce soit Samuel Charron (et Christian Frisch) pendant la première partie stage ou Clément Chastagnol (et Guillaume Lebourgeois) en deuxième partie, j'aurais constamment eu quelqu'un à mon écoute. En effet, lorsque je rencontrais une difficulté, j'avais, à chaque moment, quelqu'un de disponible à « déranger » pour mettre à l'épreuve son expérience ou ses connaissances.

Un excellent stage sur le plan relationnel aussi. En effet, tout le monde m'a accordé du temps à mon arrivée pour se présenter, et me dire leurs tâches et responsabilités chez Data Publica. De plus, partager des déjeuners et des pots conviviaux a permis de créer de premiers liens avec les collègues puis de les renforcer.

J'ai également beaucoup aimé cet esprit de partage et ces discussions ouvertes, lors des réunions des équipes pour différents points.

Sur le plan technique, j'ai énormément progressé dans le domaine du *machine learning* et énormément appris dans le domaine du *text mining* que je ne connaissais pas. Tous ces concepts de normalisation de textes, le langage Python et ses modules disponibles (scikit-learn, numpy, nltk, etc). J'ai découvert une facette de la *data science* que j'ignorais et qui m'a beaucoup plu.

Ce stage aura donc été très riche, techniquement parlant, par la découverte de domaines inconnus et passionnants, mais aussi humainement parlant, par la rencontre de personnes formidables et poussant constamment à la réussite et au dépassement de soi.

# Bibliographie

Christopher D. Manning Prabhakar Raghavan, Hinrich Schütze. Information Retrieval. url : <http://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html>.

Perkins, Jacob. Python 3 Text Processing with NLTK 3 Cookbook. 2010. isbn : 978-1-78216-785-3.

Steven Bird Ewan Klein, Edward Loper. Natural Language Processing with Python. 2009. isbn : 978-0-596-51649-9.

Wikipedia. Actor model. url : [https://en.wikipedia.org/wiki/Actor\\_model](https://en.wikipedia.org/wiki/Actor_model).

- Apprentissage automatique. url : [https://fr.wikipedia.org/wiki/Apprentissage\\_automatique](https://fr.wikipedia.org/wiki/Apprentissage_automatique).
- Modèle génératif vs. modèle discriminant. url : [https://fr.wikipedia.org/wiki/Classifieur\\_lin\%C3%A9aire\#Mod.C3.A8le\\_g.C3.A9n.C3.A9ratif\\_vs.\\_mod.C3.A8le\\_discriminant](https://fr.wikipedia.org/wiki/Classifieur_lin\%C3%A9aire\#Mod.C3.A8le_g.C3.A9n.C3.A9ratif_vs._mod.C3.A8le_discriminant).
- Transducteur fini. url : [https://fr.wikipedia.org/wiki/Transducteur\\_fini](https://fr.wikipedia.org/wiki/Transducteur_fini).

# A. Annexes

## A.1. L'équipe de Data Publica

### A.1.1. L'équipe technique

L'équipe technique de Data Publica est visible en figure A.1.



Figure A.1. – L'équipe technique de Data Publica

### A.1.2. L'équipe commerciale

L'équipe commerciale de Data Publica est visible en figure A.2.

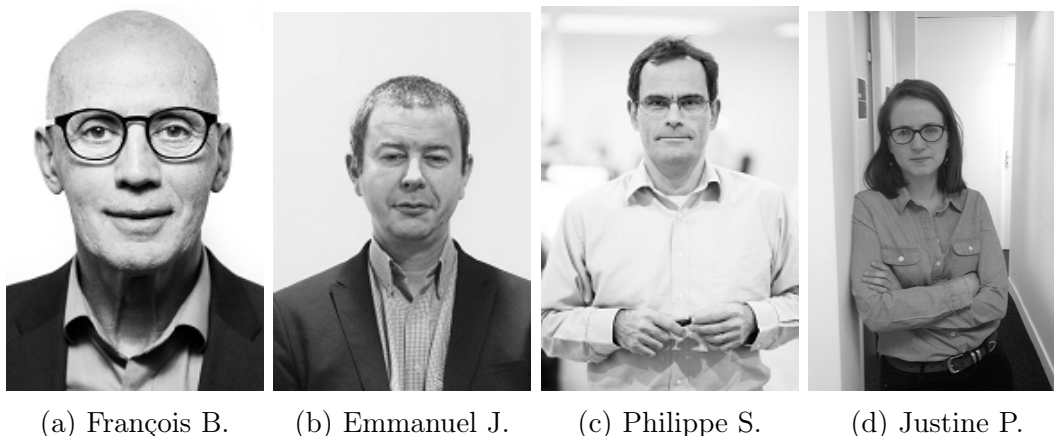


Figure A.2. – L'équipe commerciale de Data Publica

## A.2. Dataviz

Un exemple de dataviz est présenté en figure A.3.

## A.3. La validation croisée et les hyper-paramètres

### A.3.1. Les hyper-paramètres

### A.3.2. La validation croisée

La validation croisée est une technique de validation indiquant comment les résultats d'une analyse statistique se généralisent à un ensemble de données indépendant. Cette dernière est particulièrement utilisée lorsque l'on souhaite estimer la précision d'un modèle prédictif en pratique. En générale, dans les problèmes de prédiction, un modèle est construit avec un ensemble de données d'apprentissage, et ensuite un autre ensemble de données (jamais vu du modèle) sont utilisées pour tester le modèle. Le but de la validation croisée est de définir un ensemble de données utilisé pour tester le modèle durant sa phase d'entraînement (de construction). L'ensemble de données de validation est utilisé pour éviter les problèmes de sur-apprentissage et pour donner un aperçu de comment le modèle va se généraliser à un ensemble de données indépendant (un ensemble de données inconnu).

Un plis de validation croisée consiste à partitionner un échantillon des données en deux sous-ensembles. Le premier permet d'effectuer l'analyse sur celui-ci (appelé l'ensemble d'apprentissage), et le second de valider l'analyse sur celui-là (appelé l'ensemble de validation ou de test). Pour réduire l'incertitude, de multiples plis de validation croisée sont effectués en partitionnant l'ensemble de données en plusieurs partitions utilisées à chaque plis. Les résultats de la validation sont moyennés sur les différents plis.

La validation croisée est importante pour évaluer la qualité et la capacité de généralisation d'un modèle surtout lorsque les données sont rares ou difficiles à collecter.

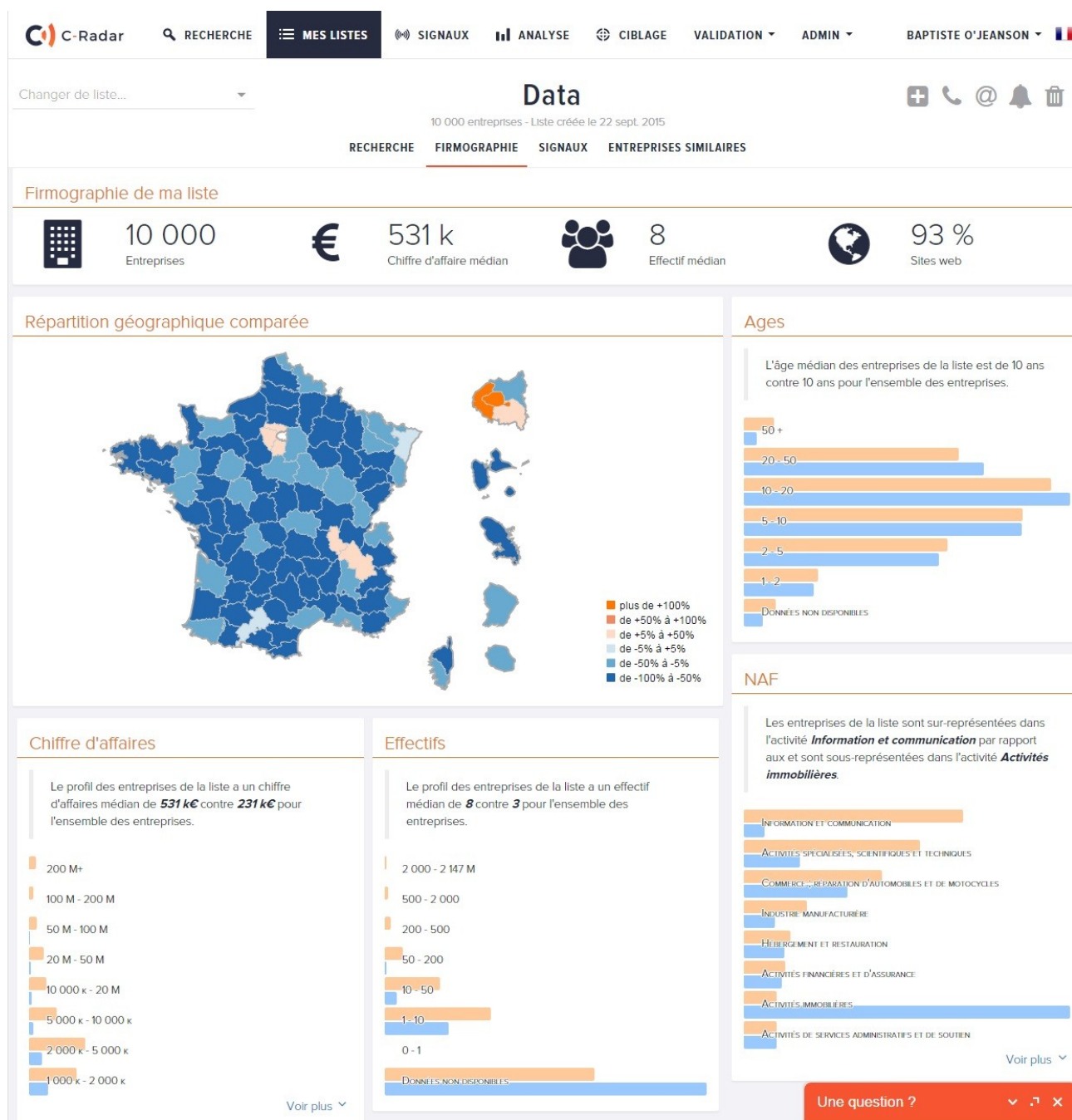


Figure A.3. – Exemple de Dataviz : Firmographie dans C-Radar

L'une des raisons principales pour laquelle on l'utilise, plutôt qu'une méthode plus traditionnelle (consistant à partitionner l'ensemble de données en deux ensembles, 70% de l'ensemble pour la phase d'apprentissage et 30% pour la phase de test), est la suivante : Dans une validation traditionnelle, l'erreur sur l'ensemble d'apprentissage n'est généralement pas un bon estimateur de la performance du modèle, de même que l'erreur sur l'ensemble de test. Celles-ci ne décrivent pas totalement la performance du modèle. C'est le cas lorsqu'il n'y a pas suffisamment de données disponibles ou lorsque la distribution et la disparité des données à partitionner ne sont pas bonnes en phase d'apprentissage et de test.

Les deux techniques de validation croisée s'appellent :

- **Leave-p-out cross-validation** : méthode la plus traditionnelle ;
- **k-fold cross-validation** : autre méthode décrite ci-dessus fonctionnant par plis.

### A.3.3. La validation croisée stratifiée

Lors du partitionnement de l'ensemble de données en  $k$  sous-ensembles, la validation croisée stratifiée garantit que la proportion des classes soit conservée dans les  $k$  sous-ensembles résultant. De plus, l'ensemble est mélangé à chaque nouvelle construction du modèle avant d'être divisé en  $k$  sous-ensembles.

## A.4. Les métriques de mesure de la qualité d'une classification

### A.4.1. Terminologies

Pour une classe donnée, voici la signification de ces terminologies :

- **TP** : True positive, bonne assignation à la classe ;
- **TN** : True negative, bon rejet de la classe ;
- **FP** : False positive, mauvaise assignation à la classe ;
- **FN** : False negative, mauvais rejet de la classe.

### A.4.2. La précision

Pour une classe donnée, la précision est le nombre de documents correctement assignés à cette classe rapporté au nombre de documents total assignés à cette classe par le classifieur.

$$\text{Précision} = \frac{TP}{TP+FP}$$

### A.4.3. Le rappel

Pour une classe donnée, le rappel est défini par le nombre de documents correctement assignés à cette classe au regard du nombre total de documents appartenant réellement à cette classe. Le rappel mesure le fait que le classifieur ait trouvé tous les documents d'une classe.

$$\text{Rappel} = \frac{TP}{TP+FN}$$

## A.5. Les stratégies de modèles de classification en contexte multi-classe

### A.5.1. One versus All

### A.5.2. One versus One

## A.6. Le QA ou Quality Assessment

L'objectif du QA est de demander la contribution d'un maximum de personnes sur une tâche de validation manuelle pénible.

## A.7. Modèle de classifieur

Les points d'explication qui suivent, sont en partie tirés de Wikipédia.<sup>1</sup>

### A.7.1. Le modèle génératif

Le modèle génératif consiste à modéliser les probabilités conditionnelles soit  $P(\text{donnée}|\text{classe})$ . Voici quelques exemples d'algorithmes de ce type :

- Classifieur naïf bayésien implique une distribution de probabilité conditionnelle de type binomiale (ou multinomiale) ;
- Analyse discriminante linéaire (LDA). Elle implique l'existence d'un modèle discriminant basé sur une distribution de probabilité de type Gaussienne.

Le modèle génératif maximise la vraisemblance de la probabilité jointe  $P(\text{classe}, \text{donnée})$ .

### A.7.2. Le modèle discriminant

Le modèle discriminant cherche à maximiser la qualité de la classification. Une fonction de coût va réaliser l'adaptation du modèle de classification (en minimisant les erreurs). Voici quelques exemples d'algorithmes de ce type :

- Régression logistique, maximisation de la vraisemblance en considérant que les données suivent un modèle binomial ;
- Support Vectort Machine (SVM), maximisation de la marge entre l'hyperplan séparant les données.

Le modèle discriminant maximise la vraisemblance de la probabilité conditionnelle  $P(\text{classe}|\text{donnée})$ .

## A.8. Les transducteurs

Définition tirée de Wikipédia.<sup>2</sup>

En informatique théorique, en linguistique, et en particulier en théorie des automates, un transducteur fini (appelé aussi transducteur à états finis par une traduction maladroite de l'anglais *finite state transducer*) est un automate fini avec sorties. C'est une extension des automates finis. Ils opèrent en effet sur les mots sur un alphabet d'entrée et, au lieu de simplement accepter ou refuser le mot, ils le transforment, de manière parfois non déterministe, en un ou plusieurs mots sur un alphabet de sortie. Ceci permet des transformations de langages, et aussi des utilisations variées telles que notamment l'analyse syntaxique des langages de programmation, et l'analyse morphologique ou l'analyse phonologique en linguistique.

D'autres explications détaillées sont disponibles [ici](#).

---

1. Wikipedia. Modèle génératif vs. modèle discriminant. url : [https://fr.wikipedia.org/wiki/Classifieur\\_lin%C3%A9aire#Mod.C3.A8le\\_g.C3.A9n.C3.A9ratif\\_vs.\\_mod.C3.A8le\\_discriminant](https://fr.wikipedia.org/wiki/Classifieur_lin%C3%A9aire#Mod.C3.A8le_g.C3.A9n.C3.A9ratif_vs._mod.C3.A8le_discriminant).

2. Wikipedia. Transducteur fini. url : [https://fr.wikipedia.org/wiki/Transducteur\\_fini](https://fr.wikipedia.org/wiki/Transducteur_fini).



Ce stage obligatoire s'effectue en fin de quatrième année. Au cours de ce stage l'étudiant devra mettre en pratique les connaissances acquises au cours de sa formation et devra approfondir son savoir-faire au sein de l'entreprise. Il faudra qu'à la fin de son stage l'étudiant réalise un rapport écrit. La validation du stage dépend de la qualité du travail réalisé, du rapport et de la fiche d'évaluation du tuteur industriel.

Ce stage a eu lieu chez Data Publica, un des précurseurs de l'open data en France. Data Publica est une jeune start-up (fondée en juillet 2011) spécialisée dans les données entreprises, l'open data, le big data et la dataviz. Data Publica emploie quatorze personnes très dynamiques et compétentes. Data Publica développe C-Radar, un produit de vente prédictive construit sur une base de référence des entreprises françaises regroupant informations administratives, financières, web, réseaux sociaux et media. C-Radar est un concentré de technologies du big data (crawling, scraping ou encore machine learning).

Ma mission, chez Data Publica, est de construire une chaîne de traitement automatique, un plugin Python, récupérant une liste de documents textuels en entrée et fournissant, en sortie, une liste de ces documents labellisés selon leur catégorie d'intérêt (offre d'emploi, participation à des événements, nomination de personnel, levée de fond, etc).

Les tâches qui en découlent sont directement liées au *Text Mining* et au *Natural Language Processing* (aussi à l'*Information Retrieval*) dans la découverte d'informations dans le contenu des documents. Des compétences en *Machine Learning* sont également requises.

Ce stage s'est conclue de belle manière, puisque le plugin Python est fonctionnel et capable de classifier de tels documents.

This internship is obligatory and carried out at the end of the fourth year. During this internship, the student shall put into practice the knowledge acquired during his training and shall get better skills within the company. At the end of the intership, the student must write a report. The internship validation is depending on the quality of the work, the report and the evaluation sheet of the industrial tutor.

This internship took place at Data Publica, an open data pioneer in France. Data Publica is a young start-up (founded in July 2011) specialized in business data, open data, big data and DataViz. Data Publica employs fourteen dynamic and competent people. Data Publica is developing C-Radar, a predictive selling product built on French enterprises informations, like administrative, financial, web, social networks and media. C-Radar is made of big data technologies (crawling, scraping or machine learning).

My mission at Data Publica, is to build an automatic processing chain, a Python plugin, which giving a list of textual records as input, get back a list of tagged documents according to their interest subject (job offer, event participation, staff nomination, fundraising, etc).

The directed tasks linked to the detection of information in the records are Text Mining and Natural Language Processing (also Information Retrieval). Skills in Machine Learning are also required.

This internship completed well, because the Python plugin is running and able to classify records.

**INSA Rouen**  
Campus du Madrillet  
685 avenue de l'Université – BP 08  
76801 SAINT-ÉTIENNE-DU-ROUVRAY cedex  
[www.insa-rouen.fr](http://www.insa-rouen.fr)



## RENSEIGNEMENTS

Département ASI  
02 32 95 97 79  
[asi@insa-rouen.fr](mailto:asi@insa-rouen.fr)

Membre de



Normandie Université

Financeurs institutionnels



MINISTÈRE  
DE L'ÉDUCATION NATIONALE,  
DE L'ENSEIGNEMENT SUPÉRIEUR  
ET DE LA RECHERCHE

