كلية العلوم و التقنيات فاس
ⵜⴰⵖⵙⵓⵍⵜ ⵏ ⵜⵉⵡⵙⴷⵉⵏ ⴰ ⵜⵓⵙⵙⵔⵜⵙⵉ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵡⴰⵍⵜ ⵓⵙⴰⵙ ⵎⵓⵃⴰⵎⴷ ⵏ ⵄⴰⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

DÉPARTEMENT D'INFORMATIQUE

PROJET DE FIN D'ÉTUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

MISE EN PLACE D'UN FRAMEWORK D'ANALYSE DE TEXTE POUR L'EXTRACTION DES CONNAISSANCES À PARTIR DES DONNÉES NON-STRUCTURÉES

INDATACORE
Big Data Analytics & Data Science

LIEU DU STAGE : INDATACORE

Réalisé par :
- Abderrahman AL ACHARI

Encadré par :
- Pr. Azeddine ZAHI
- M. Soufiane EZGHARI

Soutenu le 12.06.2018 devant le jury composé de :

- Pr. Khalid ABBAD	Faculté des Sciences et Techniques de Fès	(Président)
- Pr. Ilham CHAKER	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. Arslane ZARGHILI	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. Azeddine ZAHI	Faculté des Sciences et Techniques de Fès	(Encadrant)

Année Universitaire 2017 – 2018

Dédicaces

Je dédie ce modeste travail à ma chère mère,

À mon père qui m'ont toujours soutenu

Et aidé à affronter les difficultés,

Mes Sœurs et Mes Frères

A tous les gens qui m'aiment

et qui ont contribué de près ou de loin à mon

Succès, je leur dédie ce travail

En leur souhaitant une vie pleine de

bonheur et de réussite.

Remerciements

Au terme de ce travail, j'exprime mes vifs remerciements à toute personne ayant contribué, de près ou de loin, à la réalisation de ce travail.

Je tiens à remercier tout d'abord les professeurs de la faculté des Sciences et Technique (FST) de Fès, notamment mes professeurs de l'équipe Master Système Intelligents et qui n'ont ménagé aucun effort afin de nous assurer ce parcours et nous transmettre leurs connaissances, leur expérience et leurs valeurs.

Ma profonde gratitude s'adresse à l'entreprise « INDATACORE » au nom de son directeur général Monsieur Amine FASSI FIHRI qui m'a accueilli et à Monsieur Hicham BOURAS le manager de l'équipe data science pour la confiance qu'ils m'ont 'accordé.

J'étais chanceux, pendant ma période de stage, de collaboré avec Monsieur Soufiane EZGHARI, mon tuteur de stage qui a dirigé mon projet de fin d'études, pour son engagement permanent, son soutien constant, son partage d'expérience, ses idées et sa disponibilité tout au long de cette période.

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance à mon encadrant pédagogique Monsieur Azzedine ZAHI pour ces conseils, ces orientations, et sa disponibilité tout au long de la période de stage.

Je tiens à remercier aussi les membres de l'équipe data science où j'ai effectué mon stage, d'avoir créé les conditions nécessaires à l'adaptation et à l'évolution dans le cadre de mon travail de réalisation de mon projet de fin d'études

Résumé

Les flux de données s'accroître de plus en plus et génère une quantité considérable des données non-structurées. Dans ce contexte l'analyse de textes est un outil qui permet d'analyser, d'indexer et de réaliser la fouille de textes des données non structurée. L'objectif de ce projet est de fournir un outil d'analyse de texte. Ce dernier doit contenir un système de flux de données qui permet d'extraire des textes de plusieurs sources. Ensuite, les textes récoltés doivent être indexées pour faciliter la recherche. Finalement, les données non-structurer seront utiliser pour élaborer un dictionnaire de données afin de réalisé des applications de la fouille de textes et d'analyse de sentiments et la Publicité contextuelle.

Mots clés : Données non structuré, Fouille de texte, Analyse de sentiment, Traitement du langage naturelle, Apprentissage automatique

Abstract

Data flows grow more and more and generates a considerable amount of unstructured data. In this context, text analysis is a tool for analyzing, indexing and performing text mining of unstructured data. The goal of this project is to provide a text analysis tool. This Last must contain a data flow system that makes it possible to extract texts from several sources. Then, the collected texts must be indexed to facilitate the search. Finally, the - unstructured data will be used to develop a data dictionary to carry out text mining, sentiment analysis and contextual advertising applications.

Keywords: unstructured data, Text mining, Sentiment, Analysis, Natural language processing, Machine learning

Table des Matières

Liste des Figures

Liste des Tableaux

Introduction générale

Actuellement, la communication sur le réseau internet prend une place de plus en plus importante dans notre vie personnelle, professionnelle et sociale. En effet, le réseau Internet offre, aux communautés ayant des centres d'intérêts communs, un espace pour se divertir communiquer et discuter sur divers sujets. Les réseaux sociaux constituent l'un des espaces les plus fréquentés par les internautes pour partager et donner leurs avis sur des sujets sociétales, politiques et économiques; les statistiques témoignent du grand nombre d'utilisation de ces réseaux. Dans ce contexte, beaucoup d'entreprises ont bien pris conscience de l'intérêt et l'importance des informations qui circulent sur les réseaux sociaux pour se développer et évoluer. En fait, les entreprises trouvent dans ces espaces de communication des données, souvent non-structurées telles que le texte, l'image, la vidéo, etc., qui représentent les perceptions qu'un client ou un prospect peut avoir à propos d'une entreprise, d'une marque, d'un produit ou d'un service. Ainsi, l'analyse de ces données se voit nécessaire pour les entreprises qui veulent se projeter dans un avenir de plus en plus compétitif et complexe.

Les données textuelles sont les données les plus pertinentes et aussi les plus faciles à exploiter pour caractériser et analyser la perception du client ou du prospect à propos d'une entreprise. Deux approches d'analyse peuvent être distinguées dans ce contexte: le traitement automatique du langage naturel (TALN) et la fouille de textes. Le TALN s'avère peu adapté au traitement de grandes quantités de textes ne respectant pas nécessairement les règles syntaxiques, mais potentiellement riches en informations utiles. Contrairement au TALN, la fouille de textes ne cherche pas à comprendre le sens profond des grandes quantités de textes mais à traiter efficacement certaines tâches précises et bien délimitées. Ces tâches servent à structurer le texte plus que les niveaux d'analyse linguistique. Plusieurs tâches sont considérées dans ce contexte, telles que *la recherche de l'information,* la *recommandation automatique* de documents, l'analyse des *sentiments*, la *classification* et *la catégorisation* de documents et l'analyse de la *personnalité;* Dans ce travail, nous nous intéressons aux trois dernières tâches. En plus, la fouille de textes s'inscrit dans les

problématiques de la *science* de *données* (data *science*). Cette dernière recouvre les problématiques et les techniques de la fouille de données considérées dans le contexte des données massives, non structurées et éventuellement générées en temps réel.

Ainsi, l'objectif de mon projet de stage, effectué à l'entreprise INDATACORE, consiste en la conception, la construction, la validation et la mise en place d'un Framework destiné à l'analyse de textes non structurées. Le future Framework doit offrir à ses utilisateurs les fonctionnalités suivantes :

- Collecte automatique des données à partir de plusieurs sources telles que les moteurs de recherche, les réseaux sociaux, etc.
- L'analyse de textes pour les tâches d'analyse des sentiments, la classification et la catégorisation de documents et l'analyse de la personnalité.
- La visualisation des résultats d'analyse.

Le présent rapport est organisé en quatre chapitres comme suit :

Le premier chapitre est consacré à la présentation du contexte général du projet. Il donne en premier lieu un aperçu sur l'organisme d'accueil. Ensuite, il décrit sommairement la science des données. Enfin il termine avec une description générale du projet. Le deuxième chapitre se focalise sur les éléments de base utilisés dans le développement du Framework. Nous commençons par une présentation du cycle de vie d'un projet science de données. Ensuite, nous décrivons les méthodes et outils de collecte des données textuelles sur le web. Enfin, nous présentons les différentes tâches de fouille de textes ainsi que les techniques d'analyse associées. Dans le troisième chapitre nous abordons la conception du Framework. Nous donnons en premier lieu la modélisation des besoins par un digramme de cas d'utilisation. Une description détaillée de l'architecture adoptée pour le Framework est ensuite présentée. Nous terminons avec une présentation des différents composants du Framework. Le quatrième chapitre fera Le sujet de la construction du Framework. Nous décrivons d'abord l'expérimentation conduite pour choisir le meilleur modèle à inclure dans le Framework. Ensuite nous présentons les résultats de l'application du Framework construit sur une étude de cas. Nous terminons ce rapport par une conclusion ainsi que les perspectives de ce travail.

Chapitre 1: Contexte général du projet

Introduction

Ce chapitre présente le contexte général qui permet de situer notre projet. Il présente dans un premier temps l'entreprise d'accueil dans laquelle ce projet a été réalisé ; l'accent est mis sur la structure de l'entreprise, ses domaines d'activités et ses réalisations. Ensuite, nous donnons un aperçu sur le domaine de la science de données (data science) suivi d'une présentation générale du projet en se focalisant sur la problématique et les objectifs. Enfin, nous présentons la démarche et le planning suivis pour la réalisation de ce projet.

1. Présentation de l'organisme d'accueil

Le projet présenté dans ce rapport a été réalisé dans le cadre d'un stage à la société INDATACORE qui siège à Casablanca. C'est une entreprise FINTECH spécialisée dans l'application des technologies autour des données pour repenser les services financiers et bancaires. En fait, le cœur du métier de INDATACORE s'articule autour des technologies des Big Data, la science de données et Business Intelligence. Grâce à ses équipes dédiées à la recherche et développement, INDATACORE adopte une approche scientifique pour proposer des modèles prédictive et descriptive pour leurs clients.

1.1. Structure de INDATACORE

La société INDATACORE se compose de plusieurs équipes :

- ***L'équipe Big data*** qui se compose de plusieurs membres pouvant jouer différents rôles : *Data Architect*, *Data Fonctionnel*, *Data Engineering* et *Data Scientist*. Cette équipe peut intervenir sur plusieurs champs tels que: la constitution du DATALAB et Gouvernance de données, définition de la ***stratégie*** technologique, architecture et Dimensionnement, installation et mise en place de l'écosystème Big Data, tests de performances de l'écosystème Big Data, Identification de différentes sources de données, Branchement de différentes sources de données avec l'écosystème Big Data, nettoyage et

traitement des données, préparation des données afin d'extraire de la valeur ajouté.

- ***L'équipe business*-intelligence :** les activités de cette équipe ont surtout destinées aux secteurs bancaire et télécom pour l'analyse client tels que le *Scoring*, la *connaissance*, la *rentabilité* et le *comportement*.
- ***L'équipe data science*** comporte à la fois des spécialistes en en analyse de données de grandes masses, des développeurs et des chercheurs. C'est une équipe nouvellement créée dans l'objectif de développer, pour ses clients des solutions, permettant de créer de la valeur ajoutée à partir des données. Elle requiert des compétences dans divers domaines scientifiques tels que *l'analyse de données*, *l'analyse et la fouille de Texte, le* traitement *d'image*, les *techniques d'apprentissage automatique* (Machine Learning) et *l'Intelligence Artificielle* (AI).

1.2. Réalisations de INDATACORE

INDATACORE a réalisé plusieurs solutions de type Big data, science de données science et Intelligence économique (Business Intelligence) pour les secteurs bancaire et télécom tels que :

- « Skybank » : Plateforme Omni-canal de banque en ligne complète
- « SkyCard » : Plateforme Omni-canal de cartes Prépayées
- « SkyAnalytics » : Une solution BI, Big Data et Data Science packagée « Clé en main » destinée au secteur Bancaire.

Actuellement, l'équipe data science, nouvellement créée au sein de INDATACORE, travaille sur le développement et la réalisation de prototypes de solutions data science; le Framework présenté dans ce rapport en fait partie. Ces prototypes sont destinés à la présentation aux clients afin de les convaincre de l'intérêt de ce genre de produits pour la création de la valeur et la prolifération de leurs entreprises.

2. Problématique

Les données circulant sur le web, et particulièrement sur les réseaux sociaux constituent aujourd'hui pour les entreprises une source d'information sur le degré de satisfaction de leurs clients. C'est toute la conversation qui se passe autour de l'entreprise, principalement

sur les médias sociaux, tweets et retweets, postes sur Facebook, Instagram, Pinterest, avis et évaluations, et blogs, etc. Ces canaux d'opinion, de priorités, de perspectives et d'auto-expression globale du client ont une forte influence sur le marché et par suite sur le succès et l'évolution de l'entreprise. En effet, un commentaire qu'il soit positif ou négatif peut rapidement se propager dans différentes communautés en ligne et peut influencer les décisions des autres clients ou prospects. Ainsi, une analyse profonde et pertinente de ces données peut permettre à l'entreprise de repérer les tendances, d'optimiser certaines opérations d'affaires telles que les stratégies marketing et enfin d'améliorer la prestation de services destinée aux clients. À un niveau plus personnalisé, l'analyse de données permettra aux spécialistes du marketing de se concentrer sur des conversations spécifiques et de proposer un contenu répondant directement aux désirs et aux besoins de leurs clients.

Les méthodes et le processus d'analyse de données classique semble présenter des limitations qu'on peut résumer sur les points suivants:

- Les données sont *non structurées* dans le sens où elles ne sont pas formellement collectées et générées par les méthodes traditionnelles telles que les enquêtes, les sondages et les systèmes d'information. En effet, les données cibles sont générées sous forme de commentaires, d'images et de vidéos sur les médias sociaux.
- La *collecte des données* d'analyse n'est pas contrôlée en raison de la quantité des données et aussi de la manière dont elles sont produites. En effet, les données sont générées en temps réel, en grandes quantités et sont disponibles sur les réseaux sociaux. Cela nécessite une collecte et un prétraitement automatique de ces données afin de pouvoir les exploiter par les méthodes d'analyse.
- La *non adéquation* des méthodes classiques d'analyse de données en raison de la nature des données, principalement la non structuration, et de la quantité des données qui sont générées en temps réel.

Dans ce projet, nous nous intéressons à l'analyse de données textuelles. Deux approches d'analyse peuvent être distinguées dans ce contexte: le traitement automatique du langage naturel (TALN) et la fouille de textes. Le TALN s'avère peu adapté au traitement de grandes quantités de textes ne respectant pas nécessairement les règles syntaxiques, mais potentiellement riches en informations utiles. Contrairement au TALN, la fouille de textes ne

cherche pas à comprendre le sens profond des grandes quantités de textes, mais à explorer de larges collections de ressources dans le but de générer de nouvelles informations utiles dans le traitement de certaines tâches précises et bien délimitées. L'exploration de texte, basée sur ces tâches, identifie les faits, les relations et les assertions qui, autrement, resteraient enfouis dans la masse des grandes données textuelles. Cela consiste à transformer le texte non structuré, plus que les niveaux d'analyse linguistique, en données structurées adaptées pour l'analyse, la visualisation (tableaux html, cartes mentales, graphiques, etc.), l'intégration avec des données structurées dans des bases de données ou des entrepôts, et le perfectionnement grâce aux systèmes d'apprentissage automatique. Plusieurs tâches sont considérées dans le contexte de la fouille de textes, telles que *la recherche de l'information,* la *recommandation automatique* de documents, l'analyse des *sentiments*, la *classification* et *la catégorisation* de documents et l'analyse de la *personnalité;* dans ce travail, nous nous intéressons aux trois dernières tâches.

L'entreprise INDATACORE et plus précisément l'équipe *Data Science*, donne une grande importance aux problématiques en relation avec l'analyse des données non structurées. Dans mon projet de stage, nous sommes intéressés à la fouille de textes dans le but de permettre aux entreprises clientes de INDATACORE d'avoir une idée claire et mesurable sur la satisfaction des leurs clients vis-à-vis de leurs produits. Dans ce contexte, l'objectif de mon stage à INDATACORE est de développer un Framework d'analyse de textes qui permettra aux analystes de mener une analyse de textes, principalement les tâches : analyse des sentiments, la classification et la catégorisation de documents et l'analyse de la personnalité. Le Framework doit répondre aux exigences fonctionnelles et techniques suivantes:

- Collecte automatique des données à partir de plusieurs sources telles que les moteurs de recherche, les réseaux sociaux, etc.
- Fouille de textes pour les tâches d'analyse des sentiments, la classification et la catégorisation de documents et l'analyse de la personnalité.
- Visualisation des résultats d'analyse.
- Fonctionnement sur plusieurs systèmes d'exploitation : Windows, Linux et MacOs.
- Le Framework doit être intégrable sur un écosystème Big Data.

- Traitement de plusieurs langages naturels, dans le sens où le Framework doit permettre de faire un apprentissage sur la langue souhaitée
- Visualisation et stockage des résultats de l'analyse des sentiments sous des formats graphiques appropriés à l'interprétation et la prise de décision.
- Le Framework doit être modulable dans le sens où il doit permettre d'intégrer d'autres outils et techniques de classification de documents, de prédiction de personnalité, etc.

3. Planification du projet

Dans l'objectif de bien mener le projet, j'ai commencé par établir le planning à suivre durant la période de stage. Pour ce faire, j'ai d'abords décomposé mon projet en phases, où chaque phase est définie par un certain nombre de tâches. Ensuite, j'ai élaboré une planification de ces phase sur la durée du projet, à l'aide d'un diagramme de Gantt.

3.1. Les phases du projet

Le projet se compose de quatre pahses :

- **Phase 1 : état de l'art sur l'analyse de texte.** cette pahse vise l'étude des differnts aspects liés au processus de l'analyse de textes. L'objectif est de réaliser un état de l'art sur la problématique de l'analyse de textes, les tâches cibles, ainsi que les méthodes d'anaylse. Un intérêt particulier doit être porté à l'analyse de sentiment, la classification et la catégorisation des documents, et la prédiction de la personnalité.
- **Phase 2 : familiarisation avec l'environnement de travail.** L'objectif de cette est de se familiariser avec les différents aspects liés à la gestion d'un projet mené dans le contexte de la science de données. Il s'agit du cycle de vie, des modèles de prédiction et des outils logiciels nécessaires pour le développement de notre projet. L'accent est mis sur les techniques et les outils liés à la probématique et les methodes évoqués dans la phase de l'état de l'art.
- **Phase 3 : exploration et préparation des données non-structurées de type texte. c**ette phase a pour objectif d'étudier les possibilités de collecter des données depuis le web à l'aide du web Scraping et des API des réseaux sociaux. En se

basant sur les conclusions des phases précédantes et des résultats de la phase en cours, un cahier de charge regroupant les besoins du Framework est fourni.

- **Phase 4 : développement et validation du framework.** Cette phase vise deux objectifs : le développement du Framework (conception et réalisation) et la validation sur une étude de cas. Le choix de certains composants de l'architecture proposée, particulièrement les modèles de prédiction, est effectué en conduisant une étude expérimentale selon le processus de la science de données.

3.2. Planning du projet

En analysant la description des phases élaborée précédemment, nous avons d'abords identifié les tâches de chaque phase, ensuite nous avons procédé à l'ordonnancement des ces tâches sur la durée du projet, enfin nous avons élaboré le planning à l'aide de l'outil Gantt Project. La figure 1 montre le digramme de Gatt du planning obtenu.

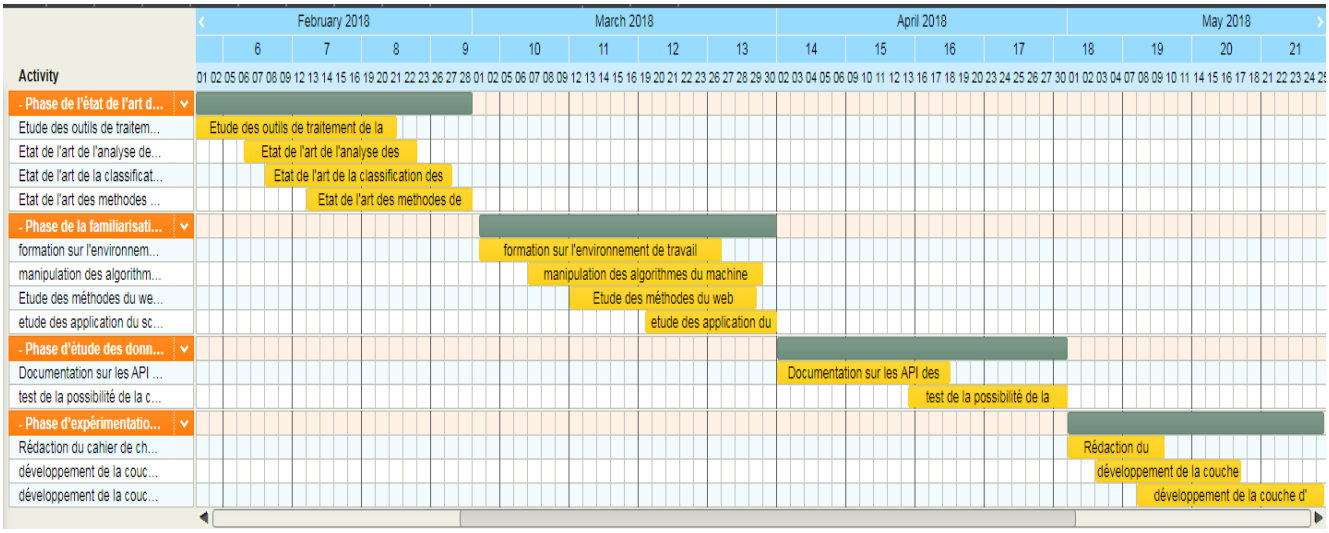

Figure 1 : Diagramme de Gant du projet

Conclusion

Dans ce chapitre nous avons mis en évidence le contexte général du projet. Nous avons présenté en premier lieu l'organisme d'accueil. Ensuite, nous avons donné une description de la problématique que notre projet veut résoudre. Enfin nous avons présenté la démarche et la planification à suivre durant la période de réalisation de notre projet. Dans le chapitre suivant nous allons présenter un aperçu de la science de données, l'accent est mis particulièrement sur le processus à suivre dans un projet science de données.

Chapitre 2: Vue générale sur la science de données

Introduction

Ces dernières années, la science de données a émergé comme étant un sujet récurrent et en plein essor destiné à l'analyse des données massives et variées. Elle est le résultat de fusion de la vison de l'ingénieur avec celle du scientifique dans un processus unique capable de produire des systèmes décisionnels fiables et efficaces. Ainsi, la science de données se veut un outil efficae qui peut être d'une grande utilité dans le choix des meilleurs techniques et modèles lors de la conception d'un support de décisions. Dans ce chapitre, nous allons présenter un aperçu sur la science de données, les caractériqtisues d'un projet de science de donnés principalement la nature des données considérées, ainsi que le cycle de vie d'un projet de science de données.

1. Qu'est ce que la science de données?

La science de données est le résultat d'une évolution technologique et du besoin acrrue de traiter et analyser de grandes quantités de données de differents types. Les motivations qui ont condu, les défis

1.1. Motivations et définition

Les développements technologiques réalisés dans le domaine de l'informatique et de la communication, ont largement contribué au changement de nos habitudes par rapport à la production et la consommation des données. Dans le modèle classique, la production des données est limitée à certaines organisation et la demande de consomation de ces données est élevée ; le producteur n'est pas forcement un consomateur. Aujourd'hui, nous sommes dans un modèle tout à fait différent où le consomateur des données est devenu lui aussi un producteur de son plein geré ; photos, vidéos, intéractions sur les réseaux sociaux, traces de d'opérations tels ques les achats, la navigation sur internet etc. Ces données sont générées à partir de différentes sources telles que la presse éléctronique, les supports multimédias, les capteurs et les instruments. Ceci a conduit à de nouveaux défits, à cause de l'augmentation considérable dans le volume de données produites, en termes de stockage, de traitement, et

d'anaylse et que les outils de l'intelligence économique(Bisiness Intelligence) sont incapables de résoudre. Ces défis sont associés à ce qu'on appelle aujourd'hui *Big data et science de données.*

La science de données a ainsi émergé comme un sujet récurrent destiné à l'analyse de données de grandes masses et de différents types (structurées ou non structurées). Elle recouvre des domaines à l'interface entre les *statistiques*, le *Machine Learning* ou

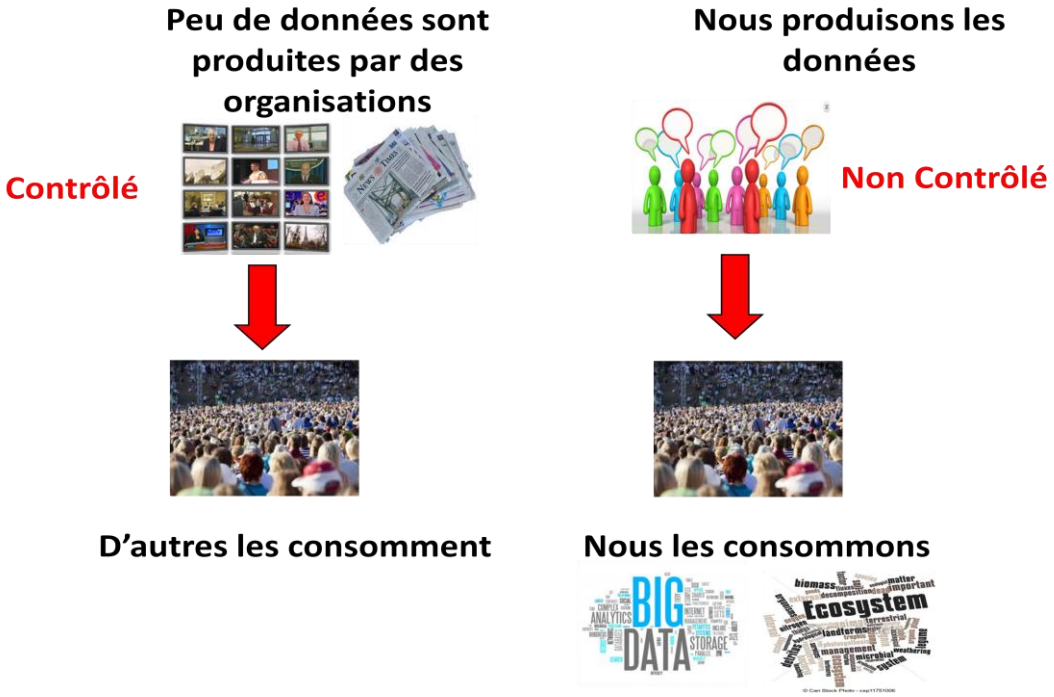

Figure 2: Modèles de production et de consommation des données

apprentissage automatique, la *fouille de données*, l'informatique et le *domaine métie***r**. L'objectif primaire de la science de données est de fournir des méthodes d'analyse influentes capables de créer la valeur pour certains problèmes tels que l'optimisation des opérations d'affaires, l'identification des risques et la prédiction de nouvelles opportunités. Les méthodes proposées sont construites en combinant les techniques avancés d'analyse telles que celles issues de fouille de données avec les technologies du Big data. Sur le plan méthodologique, la science de données intégre la vison de l'ingénieur et celle du scientifique dans un seul processus capable de conduire à des systèmes décisionnels fiables et efficaces. En effet, la science de données peut être d'une grande utilité dans la conception des supports de décisions ; les meilleurs techniques et modèles de décisions sont choisis à l'aide d'un processus expérimentale.

1.2. Intelligence économique vs science de données

L'intelligence économique et la science de données ont tous les deux pour objectif d'extraire de la valeur à partir de données pour des problèmes d'affaires. Cependant, ils diffèrent sur le type de problématique à résoudre et les techniques utilisés pour les résoudre, la nature et la tailles des données, et enfin les questions d'affaires [1].

La figure 3 illustre ces différences. L'intelligence économique se focalise sur l'utilisation d'un ensemble consistant de métriques pour mesurer les performances d'affaires afin d'expliquer le passé et d'identifier les éventuelles problèmes. Cela consiste à créer des indicateurs de performances qui informent les décideurs sur les bon déroulement des affaires. Ces mesures et indicateurs sont définies dans le schéma OLAP pour réaliser le reporting, les tableaux de bords et les alertes sur les mesures définies. L'intelligence économique utilise l'analyse exploratoire des données qui est basée sur les méthodes statistiques. Elle est destinée aux données structurées de petite à moyenne taille issues de sources typiques.

Au contraire, la science de données s'intéresse à l'analyse du passé et du présent afin de prédire le future. Elle consiste à explorer les données sous plusieurs angles afin d'en découvrir les patterns cachés et les intuitions inconnus auparavant sur le problème d'affaire considéré. La science de données combine les techniques d'analyse et d'apprentissage automatique pour la construction de modèles de prédictions, à partir des données, capables d'identifier l'occurrence d'un événement dans le futur. Ces méthodes consistent en des techniques de prédiction tels que les arbres de décision et la régression linéaire, des techniques d'optimisation et de simulation et des techniques de description tels que le clustering. La science de données est destinée à l'analyse de données structurées et non structurées en grandes quantités et issues de plusieurs sources.

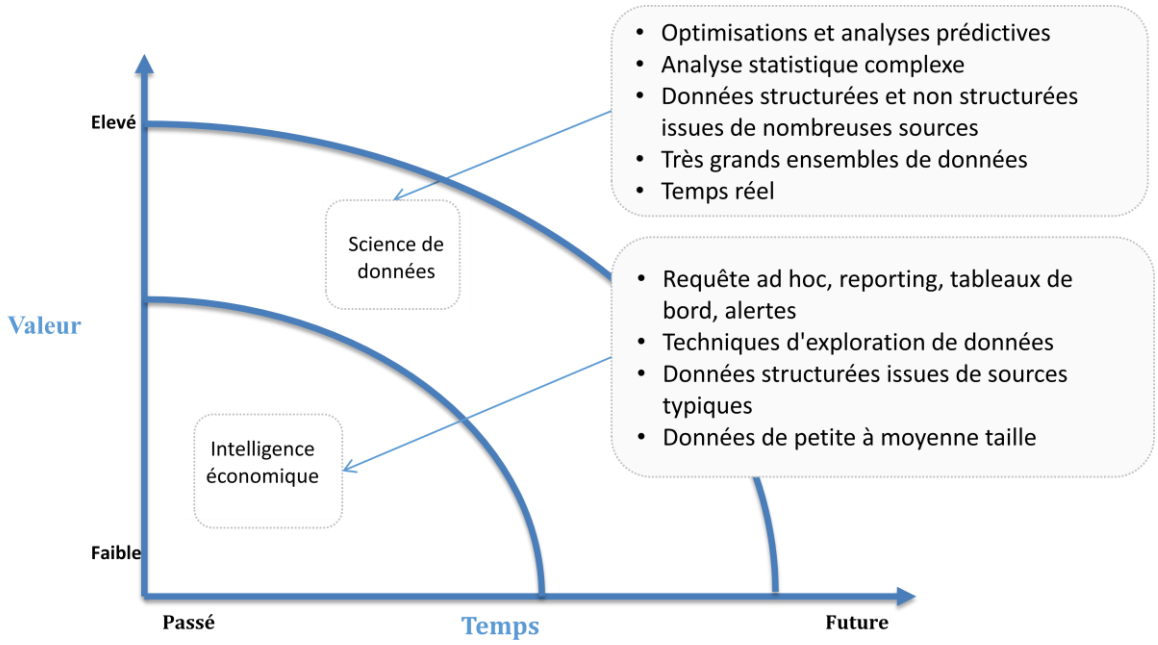

Figure 3 : Différence entre l'intelligence d'affaires et la science de données[1]

1.3. Domaines d'application

La science de données trouve applications dans plusieurs une large variété de domaines. La figure 4 montre les domaines ou la science de données peut intervenir.

- Que diriez-vous si vous pouviez comprendre les exigences précises de vos clients à partir des données existantes telles que l'historique de navigation, l'historique d'achat, l'âge et le revenu du client. Vous aviez sans doute toutes ces données plus tôt, mais maintenant, avec la quantité et la variété des données, vous pouvez former des modèles plus efficacement et recommander le produit à vos clients avec plus de précision. Ne serait-ce pas incroyable, car cela apportera plus de business à votre organisation ?
- Prenons un scénario différent pour comprendre le rôle de la science de données dans la prise de décision. Et si votre voiture avait l'intelligence de vous ramener à la maison ? Les voitures autonomes recueillent des données en direct à partir de capteurs, y compris des radars, des caméras et des lasers pour créer une carte de ses environs. Sur la base de ces données, il prend des décisions comme quand accélérer, quand dépasser, où prendre un virage - en utilisant des algorithmes avancés d'apprentissage automatique.

- Voyons comment Data Science peut être utilisé dans l'analyse prédictive. Prenons la prévision météorologique comme exemple. Les données des navires, des avions, des radars, des satellites peuvent être collectées et analysées pour construire des modèles. Ces modèles permettront non seulement de prévoir le temps, mais aussi de prévoir l'occurrence de calamités naturelles. Cela vous aidera à prendre des mesures appropriées à l'avance et à sauver de nombreuses vies précieuses.

Figure 4: Domaines ou la data science peut intervenir [1]

2. Caractéristiques d'un projet de science de données

Deux principales caractéristiques permettent de distinguer un projet d'analyse de données classique, qu'il soit dans le contexte de l'intelligence économique ou dans le contexte d'un autre métier, d'un projet de science de données: la nature des données et l'approche d'analyse. En effet, les projets de science de données opérent dans des enivrements Big data où les données sont générés en temps réel par des sources variées, en grandes quantités et dans formes variées telles que les enregistrements, le texte, l'image la vidéo, etc. Dans cette section, nous allons présenter les différents types de données ainsi qu'un apercu de l'approche d'analyse.

2.1. Nature des données

Les données d'un projet de science de données peuvent être divisées en quatre catégories : structurées, semi-structurées, quasi-structurées et non structurées. La figure 5 montre la hiérarchie des données selon le degré de structuration.

- ***Les données structurées*** font référence à toutes les données qui résident dans un champ fixe, un enregistrement ou un fichier. Cela inclut les données contenues dans les bases de données relationnelles, les feuilles de calcul ou des fichiers structurés tels que les fichiers csv, Excel, etc.
- ***Les données semi-structurées*** sont une forme de données structurées non conformes à la structure formelle des modèles de données associés aux bases de données relationnelles ou à d'autres formes de tableaux de données. Elles concernent les fichiers contenant des balises ou autres marqueurs pour séparer les éléments sémantiques. Par conséquent, il est également connu sous le nom de structure auto-descriptive. Dans les données semi-structurées, les entités appartenant à la même classe peuvent avoir des attributs différents même s'ils sont regroupés, et l'ordre des attributs n'est pas important.
- ***Les données quasi structurées*** sont les données textuelles avec des formats de données irréguliers pouvant être formatés avec effort, outils et temps, par exemple Les résultats de recherche Google couvrent tous les sites Web, mais sont difficiles à catégoriser sans accéder à la base de données Google elle-même.
- ***Les données non structurées*** (ou les informations non structurées) sont des informations qui n'ont pas de modèle de données prédéfini ou qui ne sont pas organisées de manière prédéfinie. Les informations non structurées sont généralement composées de texte, mais peuvent contenir des données telles que des dates, des chiffres et des faits. Il en résulte des irrégularités et des ambiguïtés qui rendent difficile la compréhension des programmes traditionnels par rapport aux données stockées sous forme de base de données ou annotées (sémantiquement étiquetées) dans les documents. Ce type de données concerne les documents textuels, les images et les vidéos.

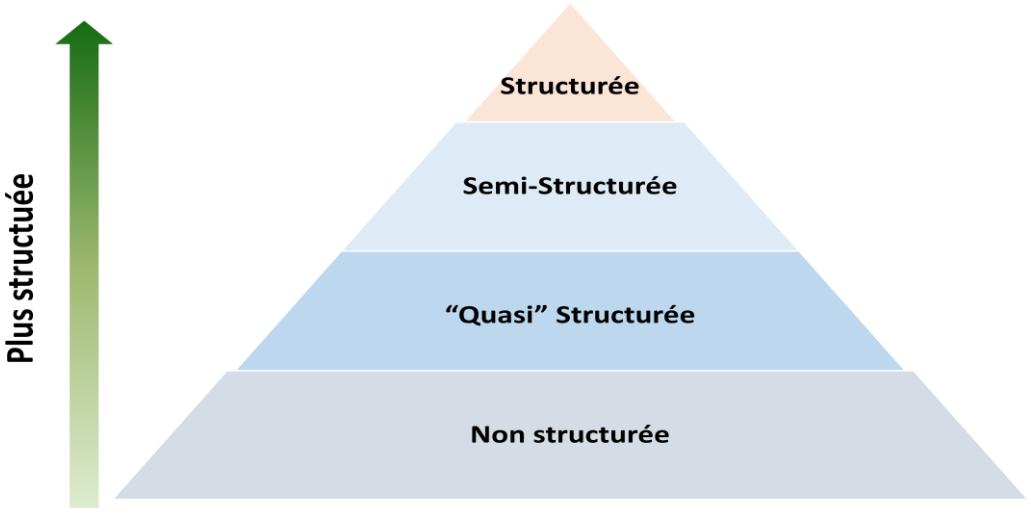

Figure 5 : nature des données dans un environnements Big data [2]

2.2. **Approche d’analyse**

Les projets de science de données diffèrent de la plupart des autres projets d’analyse dans la mesure où ils sont de nature exploratoire, analysent les données sans prétretemnt préalable, et analyse les données en mouvement. La figure 6 illustre ces propos.

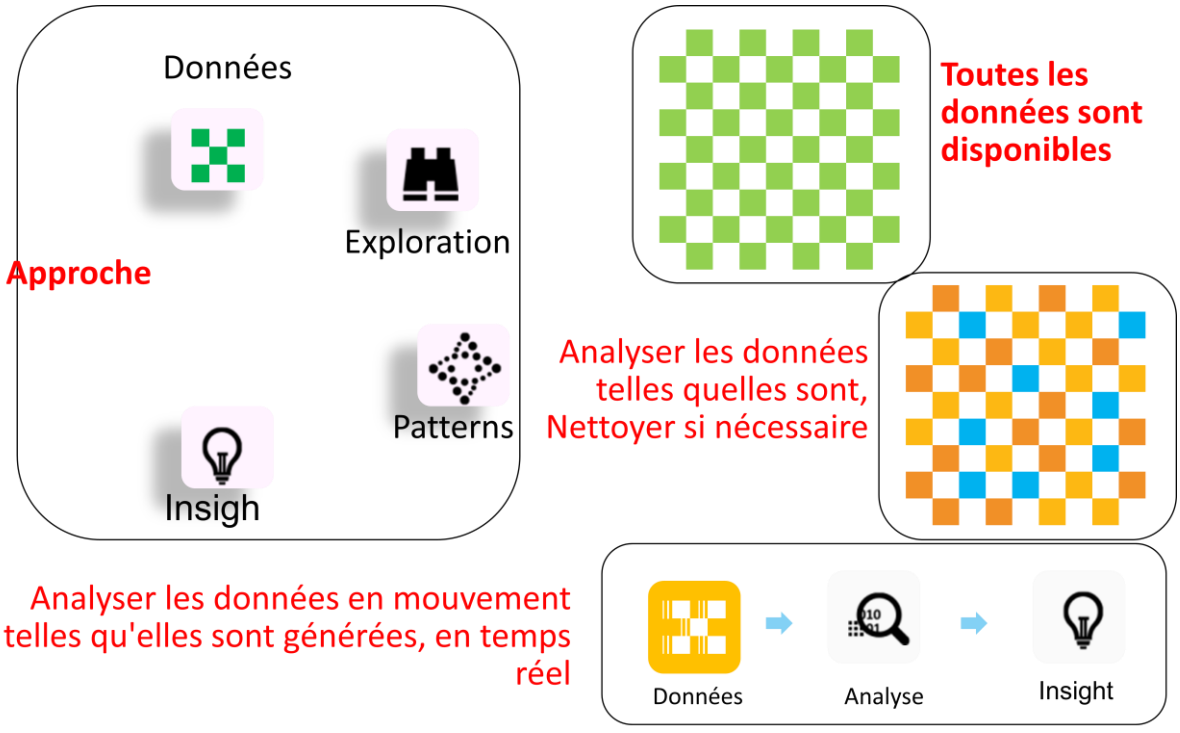

Figure 6:Approche d’analyse en science de données

3. Cycle de vie d'un projet en science de données

Dans section, nous allons présenter le cycle de vie proposé par DELL-EMC [1] pour les projets de science de données.

3.1. Contexte et aperçu du cycle de vie

Le cycle de vie de science de données définit les meilleures pratiques en matière de processus d'analyse, de la découverte à l'achèvement du projet. IL s'inspire des méthodes établies dans le domaine de l'analyse de données et de la science décisionnelle. Cette synthèse a été élaborée après le recueil des commentaires des scientifiques de données et consulter les approches établies, qui ont fourni des données sur des parties du processus.

La figure 7 présente une vue d'ensemble du cycle de vie qui comprend six phases. Pour la plupart des phases du cycle de vie, le mouvement peut être en avant ou en arrière. Cette représentation itérative du cycle de vie vise à mieux représenter un projet réel, dans lequel les aspects du projet avancent et peuvent revenir aux étapes précédentes à mesure que de nouvelles informations sont découvertes et que les membres de l'équipe apprennent davantage sur les différentes étapes du projet. Cela permet aux participants de se déplacer de manière itérative tout le long du processus et de conduire à l'opérationnalisation du travail du projet [1]. Sur la figure 7, les flèches circulaires représentent un mouvement itératif entre les phases et les légendes comprennent des exemples de questions à poser pour aider à déterminer si chacun des membres de l'équipe a suffisamment d'informations et a fait suffisamment de progrès pour passer à la phase suivante du processus. Notez que ces phases ne représentent pas des portes d'étapes formelles, ils servent plutôt de critères pour aider à déterminer s'il est logique de rester dans la phase actuelle ou de passer à la suivante.

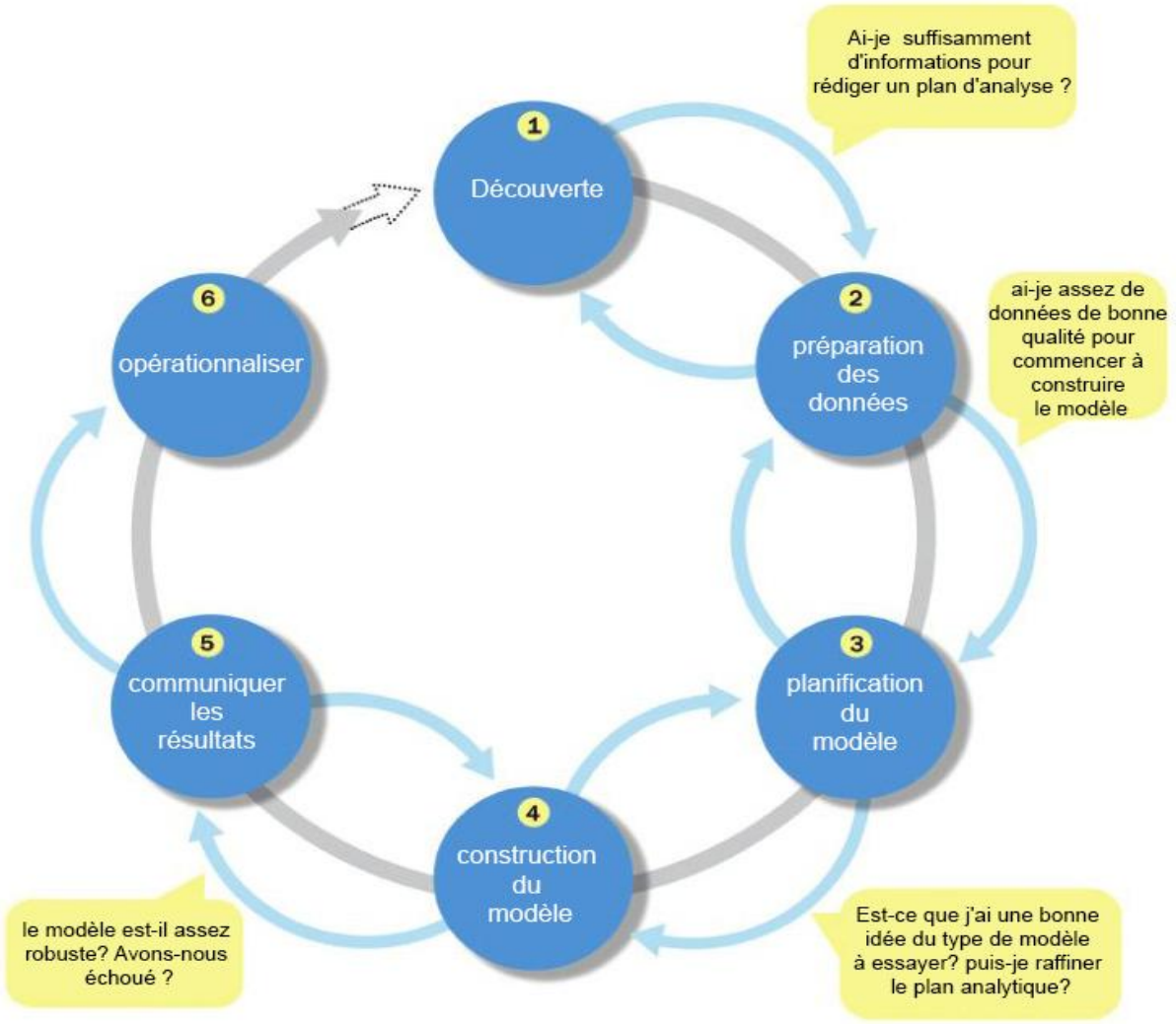

Figure 7 :: Cycle de vie d'un projet de science de données

3.2. **Phases du cycle de vie**

Le cycle de vie comporte six phases et le travail de projet peut se dérouler en plusieurs phases à la fois. Dans la suite, nous présentons les principales phases du cycle de vie :

- ***Phase de Découverte*** : Au cours de cette phase, l'équipe apprend le domaine métier, y compris l'historique pertinent, par exemple si l'organisation ou l'unité commerciale a tenté par le passé des projets similaires à partir desquels elle peut apprendre. L'équipe évalue les ressources disponibles pour soutenir le projet en termes de personnes, de technologie, de temps et de données. Les activités importantes dans cette phase incluent l'encadrement du problème commercial comme un défi d'analyse qui peut être traité dans les phases

suivantes et formuler des hypothèses initiales pour tester et commencer à apprendre les données.

- ***Phase de Préparation des données*** : cette phase nécessite la présence d'outils analytique, dans lequel l'équipe peut travailler avec des données et effectuer des analyses pendant la durée du projet. L'équipe doit exécuter, extraire, charger et transformer (ELT) ou extraire, transformer et charger (ETL) pour obtenir des données ; L'ELT et l'ETL sont parfois abrégés en ETLT. Les données doivent être transformées dans le processus ETLT afin que l'équipe puisse travailler avec et l'analyser. Dans cette phase, l'équipe doit également se familiariser avec les données et prendre des mesures pour conditionner les données
- ***Phase de Planification du modèle*** : cette phase concerne la planification du modèle, où l'équipe détermine les méthodes, les techniques et le flux de travail qu'elle entend suivre pour la phase suivante de construction du modèle. L'équipe explore les données pour en apprendre davantage sur les relations entre les variables et sélectionne ensuite les variables clés et les modèles les plus appropriés.
- ***Phase de Construction du modèle*** : Dans cette phase, l'équipe développe des ensembles de données d'entrainement et de test, De plus, dans cette phase, l'équipe construit et exécute des modèles basés sur le travail effectué dans la phase de planification du modèle. L'équipe considère également si ses outils existants seront suffisants pour exécuter les modèles, ou si elle aura besoin d'un environnement plus robuste pour exécuter des modèles et des workflows.(Par exemple, matériel rapide et traitement parallèle, sont appréciable).
- P***hase de Communication des résultats*** : Dans cette phase, l'équipe en collaboration avec les principales parties prenantes, détermine si les résultats du projet sont un succès ou un échec basé sur les critères développés dans la phase 1. L'équipe doit identifier les résultats clés, quantifier la valeur commerciale, et développer un récit pour résumer et transmettre les résultats aux parties prenantes.

- ***Phase d'Opérationnalisation*** : Dans cette phase, l'équipe fournit des rapports finaux, des briefings, du code et des documents techniques. En outre, l'équipe peut exécuter un projet pilote pour implémenter les modèles dans un environnement de production. Une fois que les membres de l'équipe ont exécuté des modèles et produit des résultats, il est essentiel d'encadrer ces résultats d'une manière adaptée au public qui a mobilisé l'équipe. De plus, il est essentiel de cadrer les résultats du travail d'une manière qui démontre une valeur évidente. Si l'équipe effectue une analyse techniquement exacte, mais ne parvient pas à traduire les résultats dans un langage qui résonne avec le public, les gens ne verront pas la valeur, et une grande partie du temps et des efforts sur le projet aura été gaspillée.

Conclusion

Dans ce chapitre, nous avons présenté un aperçu sur la science de données. Nous avons présenté en premier lieu les motivations qui ont permis l'émergence de cette approche d'analyse. Ensuite une description succinte des caractéristiques d'un projet science de données, à savoir les différents types de données et le caractère exploratoire de l'approche d'analyse. Enfin, une présentation du cycle de vie proposé par DELL EMC aisni que ses phases. Dans le chapitre suivant, nous allons présenter un état de l'art sur les tâches et les techniques de fouille de textes, ainsi que les approches de collecte de données.

Chapitre 3 : Fouille de texte

Introduction

Dans ce chapitre, nous allons présenter un état de l'art sur la fouille de textes. Nous commençons par le processus de la fouille de textes. Ensuite, nous abordons les méthodes existantes de la collecte les données textuelle depuis le web. Finalement nous donnons un aperçu sur l'analyse de texte en général et une description détaillée des tâches concernées par notre travail, à savoir l'analyse des sentiments, la classification et la catégorisation des documents et enfin la prédiction de la personnalité.

1. Processus d'analyse de textes non structurés

L'analyse de textes non structurés opère selon un processus en quatre étapes (Figure 8). La première étape est consacrée à la collecte des textes non structurés depuis le web ; cette étape peut être réalisée grâce a des robots de Scraping ou à l'aide des api dans le cas des données des réseaux sociaux. La deuxième étape s'occupe du stockage et de l'indexation qui transforme les textes collectés en données structurées facile à utiliser par les méthodes d'analyse. Dans la troisième étape, les méthodes d'analyse sont appliquées sur les données transformées. La dernière étape est dédiée à la visualiser sous forme de graphes et diagrammes pour faciliter la lecture et l'interprétation des résultats. Dans ce qui suit nous allons présenter en détails les différents composants qui entre jeux dans le processus de l'analyse de texte non structuré.

2. Collecte des données textuelles depuis le web

Les données non structurées croissent plus vite que les données structurées. Selon une étude IDC 2011[3], ces données représenteront 90% de toutes les données qui seront créées au cours de la prochaine décennie. Les données textuelles non structurées prennent une part importante. Elles se présentent sous formes de commentaires sur les réseaux sociaux ou documents. Ainsi, la collecte de ces données ne peut être qu'automatique. dans la suite, nous allons présenter les techniques les plus utilisées : le web scraping et les API.

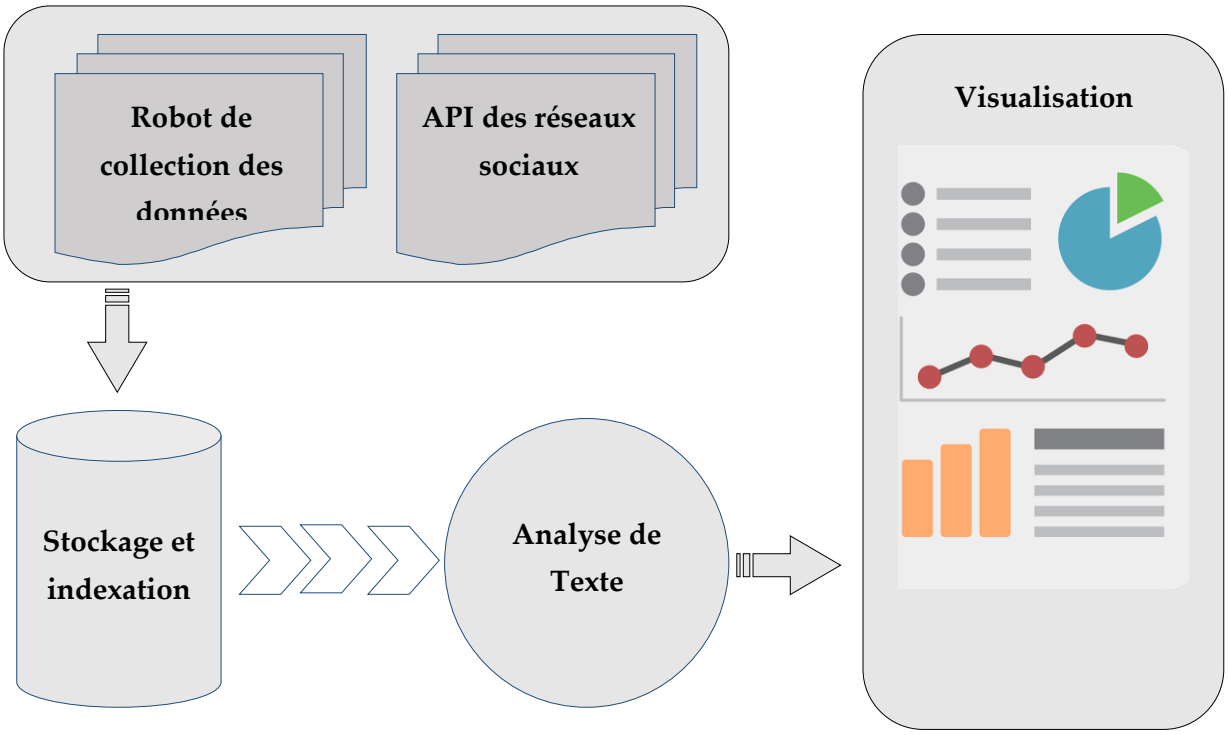

Figure 8 : processus d'analyse de texte non structuré

3.1. Le Web Scraping

Le Web Scraping ou l'extraction de données à partir de Web, est l'ensemble des opérations de récupération de données utilisées pour extraire des données de sites Web. Les logiciels du Web Scraping peuvent accéder directement au World Wide Web en utilisant le protocole de transfert hypertexte (HTTP) ou via un navigateur Web. Bien que le Web Scraping puisse être effectué manuellement par un utilisateur humain, le terme désigne généralement les processus automatisés mis en œuvre à l'aide d'un bot ou d'un robot d'indexation Web. C'est une forme de copie, dans laquelle des données spécifiques sont collectées et copiées à partir du Web, généralement dans une base de données locale centrale ou une feuille de calcul, pour une récupération ultérieure ou une analyse.

Le web Scraping permet à un individu ou une entreprise de promouvoir ses produits, de comprendre le marketing dynamique, de nouvelles promotions flottant sur Internet, etc. Il y a une tendance croissante des entreprises, organisations et particuliers à rassembler des informations à travers le web Scraping pour les utiliser dans leur meilleur intérêt. Les technologies de fouille de données passent par une énorme évolution et de nouvelles et meilleures techniques sont disponible tout le temps pour recueillir toutes les informations nécessaires. Les technologies de fouille des données Web ouvre des horizons non seulement

au niveau de la collecte de données, mais il soulève aussi beaucoup de préoccupations liées à la sécurité des données. Il y a une quantité d'informations personnelles disponibles sur Internet et l'exploration de données sur le Web a contribué à garder l'idée de la nécessité de sécuriser cette information au premier plan. Il existe de nombreux outils pour le web Scraping et l'extraction des données en ligne telles que :

- **FMiner** est un logiciel du Web Scraping, il prend en charge Windows et Mac OS X. Il intègre les meilleures fonctionnalités avec un outil de conception visuel intuitif. Pour rendre votre prochain projet de data mining un jeu d'enfant. Il suffit de choisir le format de sortie et d'enregistrer les étapes sur FMiner en fonction de l'objectif d'extraction d'un site web. Le puissant outil de conception visuelle de FMiner capture chaque étape et modélise une carte de processus qui interagit avec les pages du site cible pour capturer l'information[4].
- **Import.io** accorde aux utilisateurs une application desktop pour les aider à scraper toutes les données requises avec le nombre illimité de pages Web. Le service traite chaque page comme une source de données potentielle pour générer une API. Import.io vous guidera à travers le processus de création de la matrice de Scraping en construisant des connecteurs (pour la navigation) ou des extracteurs (pour extraire les données nécessaires). Le traitement d'extraction prend environ 24 heures pour obtenir les résultats. Les données de l'utilisateur sont privées et peuvent programmer des rafraîchissements automatiques à n'importe quelle période de temps choisie[5].
- **Easy Web extract** est un logicielle visuel pour extraire des données à des fins commerciales. Cet extracteur de texte extrait le contenu Web souhaité (texte, URL, image, HTML) depuis les Pages web avec un minimum d'effort. Easy Web Extract est parfait pour l'exportation et la génération des données textuelles dans les formats : Excel (CSV), texte, fichier XML, formats HTML, base de données MS Access, fichier script SQL, fichier script MySQL et formulaire de soumission HTTP et source de données ODBC. Un inconvénient de ce logicielle c'est qui' il prendre beaucoup de temps dans le Scraping[6].

3.2. Interface de programmation applicative (A.P.I) :

En programmation informatique, une (API) est un ensemble de définitions de sous-programmes, de protocoles et d'outils pour la construction de logiciels d'application. En terme générale, il s'agit d'un ensemble de méthodes de communication clairement définies entre différents composants logiciels. Une bonne API facilite le développement d'un programme informatique en fournissant tous les blocs de construction, qui sont ensuite assemblés par le programmeur. Une API peut être pour un système basé sur le Web, un système d'exploitation, un système de base de données, un matériel informatique. Une spécification d'API peut prendre de nombreuses formes, mais inclut souvent des spécifications pour des routines, des structures de données, des classes d'objets, des variables, sont des exemples de différentes formes d'API. La documentation de l'API est généralement fournie pour faciliter l'utilisation et la réimplémentation. Dans ce travail, nous nous intéressons au (API) qui nous permettront de collecter des informations depuis les réseaux sociaux. Dans la suite, nous parlerons des API des trois grandes plateformes en matière de réseaux sociaux, il s'agit de Facebook, Twitter et LinkedIn :

- **Facebook (API Graph)** : L'API Graph est le meilleur moyen d'insérer et de récupérer des données dans la plate-forme Facebook. Il s'agit d'une API HTTP de bas niveau qui permet aux applications d'avoir recours à la programmation pour interroger des données, publier de nouvelles actualités, gérer des publicités, importer des photos et réaliser un large éventail d'autres tâches. Le nom de l'API Graph s'inspire de l'idée d'un « graphe social » : une représentation des informations sur Facebook. Il est composé des éléments suivants :
    - Nœuds : représentent essentiellement des objets individuels, comme un utilisateur, une photo, une Page ou un commentaire ;
    - Arêtes : connexions entre une collection d'objets et un objet unique, comme des photos sur une Page ou des commentaires sur une photo ;
    - Champs : données concernant un objet, comme la date d'anniversaire d'un utilisateur ou le nom d'une Page.

Généralement, nous utilisons les nœuds pour obtenir des données sur un objet en particulier, les arêtes pour obtenir des collections d'objets sur un objet unique et les

champs pour obtenir des données sur un objet unique ou sur chaque objet d'une collection[7].

- **Twitter** : L'api Twitter permet plusieurs manipulations sur sa Platform[8], on peut citer :
    - La recherche des tweets par mot clé
    - L'envoi des messages
    - Réception des Tweets en temps réel
- **LinkedIn** : sur son site officielle[9], l api LinkedIn propose les services suivants :
    - Création de compte
    - Ajouter du contenu a votre profile
    - Partage du contenu sur votre mur
    - Gérer les pages d'entreprise

4. Extraction des caractéristiques du texte

A ce stade, on commence le traitement du texte recueilli. C'est pour cela qu'il est important d'aborder les techniques de numérisation des données textuelles. La différence entre les données numériques (vecteur 2D d'une image, vecteur 1D d'un signal vocal, …) et les données textuelles est que ces derniers ne peuvent pas être directement alimentés aux algorithmes d'apprentissage car la plupart d'entre eux attendent des vecteurs de caractéristiques numériques avec une taille fixe plutôt que les documents de texte bruts avec une longueur variable. Pour cela il est obligatoire de transformer les lignes de texte en format numériques. L'idée principale de la conversion est de donner un numéro unique (identifiant) à chaque mot. Les corpus d'entrainement et les données réel en général qu'une application de traitement d'image doit analyser sont en premier lieu des textes en paragraphe, alors il faut en premier lieu sera de d'extraire les mots des paragraphes, un traitement pareil on l'appelle segmentation du texte (Tokenization). Après cette étape on faire la vectorisation du texte en donnant un identifiant pour chaque « Token »

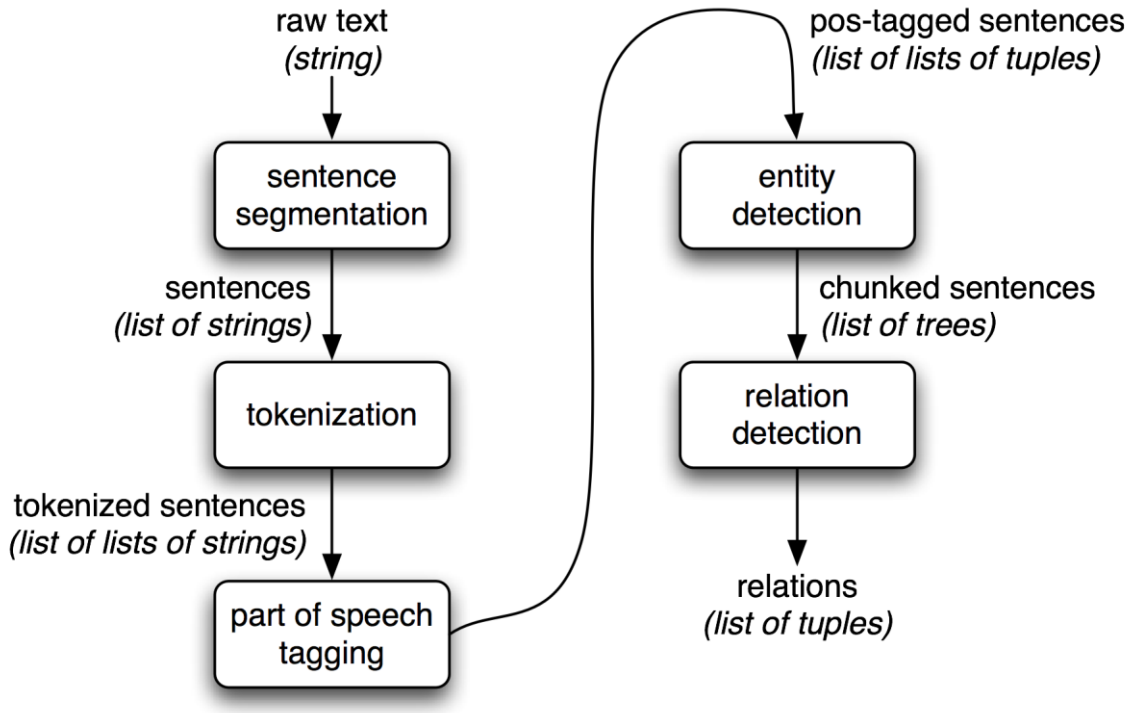

Figure 9: exemple de processus d'extraction des caractéristiques depuis le texte

On peut directement utiliser notre vecteur de Token comme caractéristique pour mener nos expérimentations mais il existe plusieurs vecteurs de caractéristique plus représentative que le vecteur des Token, on peut parler ici des deux vecteurs de caractéristiques les plus utilisé par la communauté scientifique:

- **TF-IDF (term frequency-inverse document frequency)** : Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur. La TF-IDF est le produit de deux statistiques, fréquence de terme (TF) et fréquence de document inverse (IDF) :
    - Fréquence de terme : La fréquence « brute » d'un terme est simplement le nombre d'occurrences de ce terme dans le document considéré (on parle de « fréquence » par abus de langage). On peut choisir cette fréquence brute pour exprimer la fréquence d'un terme.

- La fréquence inverse de document (inverse document frequency) est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma TF-IDF, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme (en base 10) de l'inverse de la proportion de documents du corpus qui contiennent le terme

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

. Où :

$|D|$: nombre total de documents dans le corpus

$|\{d_j : t_i \in d_j\}|$: nombre de documents où le terme t_i apparait

- N-gramme : est une sous-séquence de n éléments construite à partir d'une séquence donnée (. L'idée semble provenir des travaux de Claude Shannon en théorie de l'information [34]. Son idée était que, à partir d'une séquence de lettres donnée il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante. À partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour la prochaine lettre avec un historique de taille n. Nous nous intéressons à n = 3, le plus utilisé dans la littérature.

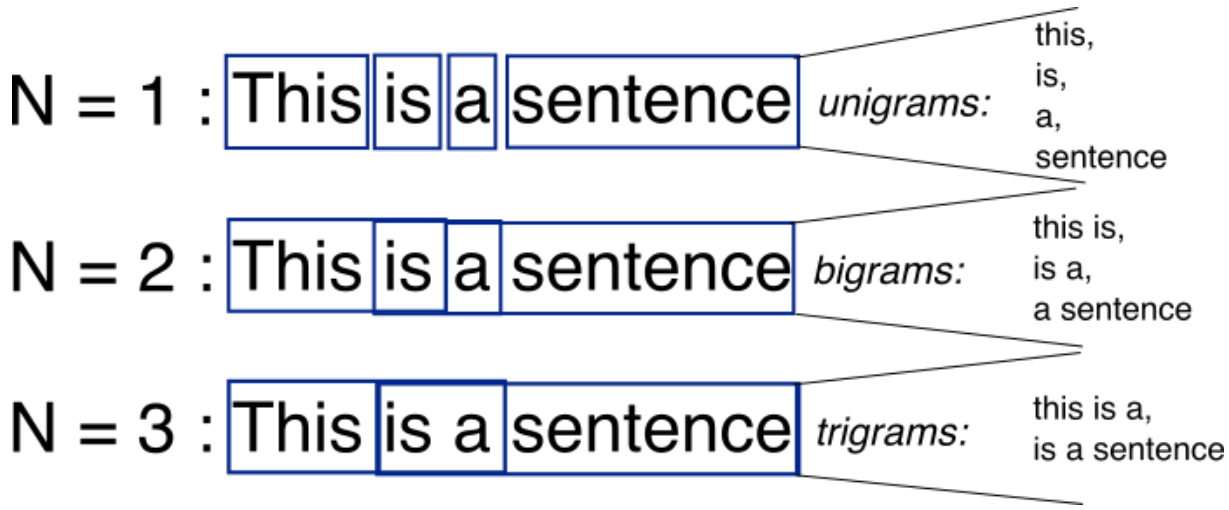

Figure 10 :: exemple de uni-gramme, bi-gramme et trigramme

5. Fouille de textes

L'analyse de texte, également appelée fouille de texte, consiste à examiner de large collections de ressources écrites afin de générer de nouvelles informations et à transformer le texte non structuré en données structurées pour les utiliser dans une analyse ultérieure. L'exploration de texte identifie les faits, les relations et les assertions qui, autrement, resteraient enfouis dans la masse des grandes données textuelles. Ces données sont extraites et transformées en données structurées, pour l'analyse, la visualisation (tableaux html, cartes mentales, graphiques, etc.), l'intégration avec des données structurées dans des bases de données ou des entrepôts, et le perfectionnement grâce aux systèmes d'apprentissage automatique. Les disciplines impliquées sont la linguistique calculatoire, l'ingénierie des langues, l'apprentissage artificiel, les statistiques et l'informatique.[10]. Les techniques du traitement du langage naturel ne seront pas utilisées pour générer du texte, mais pour déduire le sens présent dedans.

La fouille de textes ne cherche pas à comprendre le sens profond des grandes quantités de textes, mais les explorer dans le but de générer de nouvelles informations utiles dans le traitement de certaines tâches précises et bien délimitées. Plusieurs tâches sont considérées dans le contexte de la fouille de textes, telles que :

- ***Recherche de l'information*** c'est un sous-domaine de la fouille de texte; l'application la plus connue est les moteurs de recherche, qui passent également l'analyse des métadonnées et des liens entre les pages elles-mêmes.
- ***Reconnaissance d'entités nommées*** vise la détermination dans le texte, des noms propres, tels que des personnes ou des endroits, ainsi que les quantités, valeurs, ou dates.
- ***Classification et catégorisation des documents*** c'est une activité qui consiste à classer de façon automatique des ressources documentaires, généralement en provenance d'un corpus.
- ***Systèmes de tutorat intelligents*** **u**tilisés notamment pour l'enseignement des langues.

- ***Analyse des sentiments*** **v**ise à extraire le ressenti d'un texte, généralement positif ou négatif, en fonction des mots et du type de langage utilisé, d'indices typographiques ou de la personne qui l'a écrit.
- ***La recommandation automatique de documents*** consiste à extraire l'information importante d'une base de documents afin de les relier en « séries », afin de proposer ses éléments aux personnes intéressées par d'autres éléments de cette série.
- ***Prédiction de la personnalité*** c'est un nouveau champ dans la fouille du texte qui vise à prédire les traits de caractère d'une personne grâce à ces publications sur les réseaux sociaux.

Dans ce qui suit nous présentons les tâches de la fouille de texte qui seront intégrées dans notre Framework à savoir, la classification des documents par sujet, l'analyse des sentiments et la prédiction de la personnalité.

6. Analyse des sentiments

La fouille d'opinion (Opinion mining) (parfois appelée analyse de sentiment ou intelligence émotionnelle) fait référence à l'utilisation du traitement du langage naturel, de l'analyse de texte, de la linguistique computationnelle et de la biométrie pour identifier, extraire, quantifier et étudier les états affectifs et subjectifs. L'analyse des sentiments est largement appliquée à Voix du Client, comme les critiques et les réponses aux sondages, les réseaux sociaux, et tous les applications allant du marketing au service à la clientèle[11].

De manière générale, l'analyse des sentiments vise à déterminer l'attitude d'un locuteur, d'un écrivain ou d'un autre sujet par rapport à un sujet ou à la polarité contextuelle globale ou à une réaction émotionnelle à un document, une interaction ou un événement. L'attitude peut être un jugement ou une évaluation, un état affectif (c'est-à-dire l'état émotionnel de l'auteur ou du locuteur) ou la communication émotionnelle voulue (c'est-à-dire l'effet émotionnel voulu par l'auteur ou interlocuteur)[12].

5.1 Importance de l'analyse de sentiment

Il y a des millions d'utilisateurs en ligne, qui écrivent et lisent en ligne et utilisent l'Internet autour du monde. Les sentiments quotidiens en ligne deviennent un facteur

conducteur important dans la prise de décisions. Selon des études menée par Dimensional Research, le pourcentage de la confiance des avis des clients en ligne augmente de manière considérable chaque année [13]. Ci-dessous des phrases qui expriment les sentiments **:**

- Café ALFARAH est le meilleur café dans la ville de Fès.
- FST de Fès a des salles de TP bien équipé
- Je n'aime pas le café

Cependant il ya beaucoup de phrases **difficiles à interpréter, par exemple :**

- Je ne déteste pas les ornions (Action de négation)
- Le fait de ne pas aimer basketball n'est pas vraiment mon truc (Négation, ordre des mots inversé)
- Parfois je n'aime pas les vacances (adverbial, modifies the sentiment)
- J'adorerais vraiment sortir par ce temps ! (Peut-être sarcastique)
- Python est meilleurs que java (Deux attitudes, deux noms de marque).
- Python est meilleurs et java, mais la communauté de java est plus grande (Deux noms de marque, identifier la cible de l'attitude est difficile).
- Le film est surprenant avec plein de scènes troublantes (Terme négatif utilisé dans un sens positif dans certains domaines).
- J'aime mon ordinateur mais je ne vous le recommande pas (Sentiment positif qualifié, difficile à catégoriser)

5.2 Les défis de l'analyse des sentiments

En plus du problème illustré par les exemples de la section précédente, on trouve d'autres problèmes tels que [14] :

- ***La subjectivité*** **:** La recherche dans le domaine de l'analyse des sentiments a commencé avec l'étude du problème de la classification de la subjectivité et de la classification des sentiments, la classification de la subjectivité est le champ qui identifie si un document texte donné contient Des informations factuelles (des faits) ou informations opiniâtres
- ***Analyse des sentiments basés sur les caractéristiques*** **:** Ce modèle découvre d'abord les cibles sur lesquelles les opinions sont exprimées dans une phrase, puis

détermine si les opinions sont positives, négatives ou neutre. Les cibles sont des objets, ainsi que leurs composants, attributs et fonctionnalités. Un objet peut être un produit, service, individu, organisation, événement, sujet, etc. Par exemple, dans une phrase de révision de produit, il identifie les fonctionnalités du produit qui ont été commentées par le réviseur et détermine si elles commentaires sont positifs ou négatifs. Par exemple, dans la phrase, "La durée de vie de la batterie de cet appareil photo est trop petit, "le commentaire est sur" la vie de la batterie "de l'objet caméra et l'avis est négatif. Beaucoup d'applications de la vie exigent ce niveau d'analyse détaillée, car afin de faire des améliorations du produit il faut savoir quels composants et / ou caractéristiques du produit sont aimés et détestés par consommateurs (Figure 11). Une telle information n'est pas découverte par la classification du sentiment et de la subjectivité.

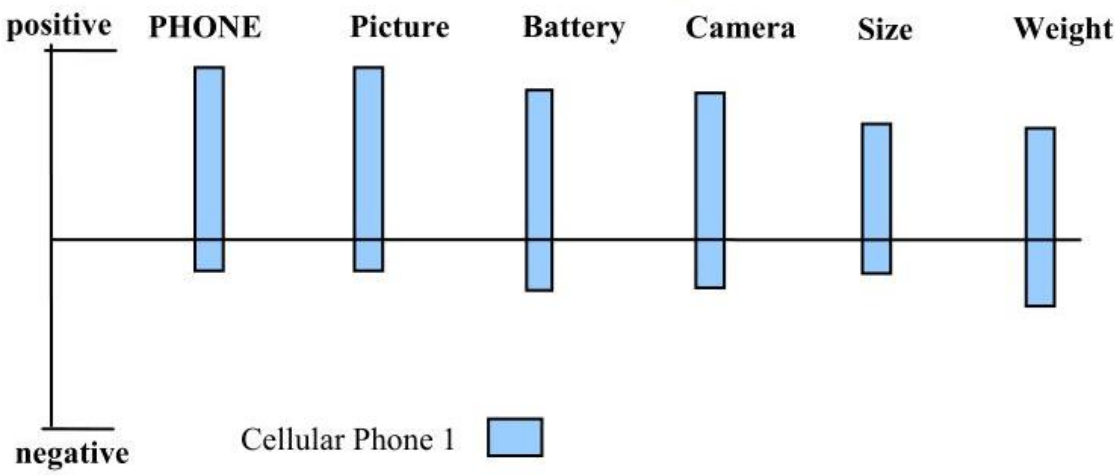

Figure 11: Visualisation des opinions basées sur les caractéristiques d'un téléphone cellulaire

- **Analyse de sentiment de phrases comparatives** : L'évaluation d'un objet peut être faite de deux manières principales, évaluation directe et comparaison. L'évaluation directe, appelée opinion directe, donne des résultats positifs ou négatifs opinion sur l'objet sans mentionner d'autres objets similaires. Comparaison signifie comparer l'objet avec d'autres objets similaires (par exemple, des produits concurrents (Figure 12)). Par exemple, " la qualité de L'image de cet appareil photo est médiocre "exprime une opinion directe, alors que" La qualité d'image de cet appareil est mieux que celle de Caméra-X.

"exprime une comparaison. De toute évidence, il est utile d'identifier de telles phrases, extraire les opinions comparatives exprimées en eux et déterminer quels objets sont préférés par les auteurs de phrase (dans l'exemple ci-dessus, Camera-x est préféré par rapport à la qualité d'image).

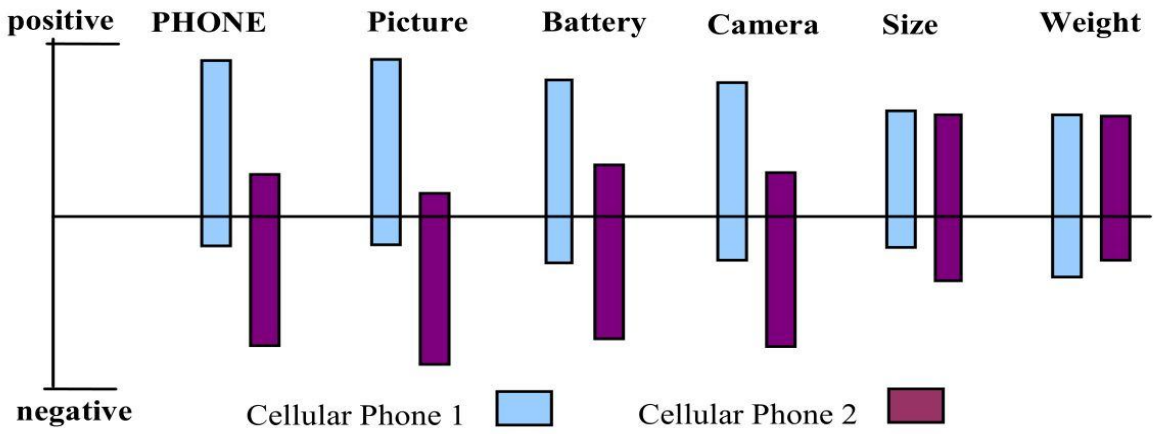

Figure 12 : Comparaison d'opinion de deux téléphones cellulaires

- **Recherche d'opinion :** Depuis la recherche générale sur le Web a été un tel succès dans de nombreux aspects, il n'est pas difficile d'imaginer que la recherche d'opinion sera également très utile. Par exemple, donné un mot-clé interroger "télétravail ", on veut trouver des avis positifs et négatifs sur la question à partir d'un avis moteur de recherche. Pour une telle requête, deux tâches doivent être effectuées :

  (1) récupérer des documents ou des phrases

  (2) identifier et classer des documents ou des phrases ceux-ci récupérés.

  La recherche d'opinion est donc une combinaison de recherche d'information et d'analyse de sentiment.
- **Opinion spam et utilité des opinions :** Comme les opinions sur le Web sont importantes pour de nombreuses applications, Il n'est pas surprenant que les gens ont commencé à jouer le système. Le spam d'opinion fait référence au faux opinions qui tentent délibérément d'induire en erreur les lecteurs ou les systèmes automatisés en donnant des résultats positifs non méritants des opinions sur certains objets cibles afin de promouvoir les objets et / ou en donnant des

opinions à d'autres objets afin d'endommager leur réputation. La détection de tels spam sont très importants pour les applications.

L'utilité des opinions renvoie à l'utilité ou à la qualité des opinions. L'attribution automatique de valeurs d'utilité aux opinions est utile car les opinions peuvent ensuite être classées en fonction de leurs valeurs d'utilité. Avec le classement, le lecteur peut se concentrer sur ces opinions de qualité.

5.3 Les approches de l'analyse de sentiment

Dans la littérature on trouve plusieurs approches pour aborder l'analyse des sentiments dans le cas de données textuelle : on trouve une approche qui se base la construction d'un dictionnaire avec des informations sur les mots et les phrases qui sont positifs et négatifs. Par Exemple, SentiWordNet[15] est une ressource lexicale ouvertement disponible dans laquelle chaque mot est attribué trois scores numériques décrivant comment objectif, positif, et négatif ,et les classificateurs sont ensuite entrainé pour classer un nouveau lot de mots ou de phrases.

Il y a d'autres approches pour analyser les sentiments se concentrent sur l'extraction de phrases ou de documents entiers, plutôt que de dépendre de la parité de mots. Cette approche fonctionne généralement avec des corpus de documents texte. L'essentiel problème avec la classification des documents (classification de polarité) qui est qu'il doit déterminer les caractéristiques générales de sentiment d'un document entier, tandis que sentiment exprimé peut être inclus dans une seule phrase ou un mot, on parle souvent de classification du sentiment niveau de mot, de niveau de phrase et de niveau de document.

D'autre part il y'a une autre approche dans l'exploitation du sentiment est sur le web. L'extraction d'opinion sur Internet vise à extraire un résumé et à suivre divers aspects des informations subjectives sur le Web [16]. Cela peut s'avérer utile pour la publicité des entreprises ou observateurs de tendances. Par un résumé de la défection de l'analyse des sentiments (aussi appelé extraction d'opinion) qui fait référence à l'utilisation du traitement du langage naturel, l'analyse de texte, et la linguistique computationnelle (CL) pour identifier et extraire les informations subjectives dans les sources. L'analyse de sentiment est largement utilisée pour critiques en ligne et les médias sociaux pour une variété d'applications, allant du marketing au service client.

3. Classification et catégorisation de documents

La classification et catégorisation de documents est l'un des champs du Traitement automatique des langues naturelles qui consiste à classer de façon automatique des ressources documentaires, généralement en provenance d'un corpus. Cette classification peut prendre une infinité de formes[17]. On citera ainsi la classification par genre, par thème, ou encore par opinion. La tâche de classification est réalisée avec des algorithmes spécifiques, mis en œuvre par des systèmes de traitement de l'information. C'est une tâche d'automatisation d'un processus de classement, qui fait le plus souvent appel à des méthodes numériques (c'est-à-dire des algorithmes de recherche d'information ou de classification de type mathématique).

L'activité de classification de documents est essentielle dans de nombreux domaines économiques : elle permet d'organiser des corpus documentaires, de les trier, et d'aider à les exploiter dans des secteurs tels que l'administration, l'aéronautique, la recherche sur internet, les sciences. Le déploiement d'un système de classification repose sur plusieurs étapes. On peut les schématiser ainsi[17] :

- Définition des classes (exemple : catégories "Sport", "Politique", "Diplomatie", ou encore Opinion "bonne/mauvaise")
- Apprentissage des classes avec un système de classification en utilisant un corpus d'apprentissage
- Évaluation des performances du système avec un corpus de test

Dans la littérature de classification des documents [18], plusieurs algorithmes ont donné de bonnes résultat, on peut citer par exemples : les réseaux bayésien naïve, l'arbre de décision, K-plus proche voisin, Machine a vecteur de support et les réseaux de neurones.

4. Prédiction de la personnalité

La personnalité est une façon dont la personne réagit à une situation particulière. C'est Une combinaison de caractéristiques qui rendent un individu unique. L'évaluation de la personnalité au cours des deux dernières décennies dans diverses recherches ont révélé que la personnalité peut être définie par cinq dimensions connues comme Big Five traits de personnalité. En général, l'étude de la personnalité considérée comme une recherche

psychologique basée sur les études ou le questionnaire. Mais cela limite les données de recherche à moins de personnes. D'où il y a un besoin de quelque chose à travers lequel nous pouvons augmenter le nombre de personnes impliquées dans les études et pour automatiser le processus.

4.1. La théorie du Big five

En psychologie, les Big Five sont cinq traits centraux de la personnalité empiriquement proposé par Goldberg [19],puis développé par Costa et McCrae dans les années 1987-1992[20]. Ils constituent non une théorie mais un repère pour la description et l'étude théorique de la personnalité. Il est parfois question du « modèle OCEAN » suivant les différentes dimensions du modèle :

- (O) Ouverture : appréciation de l'art, de l'émotion, de l'aventure, des idées peu communes, curiosité et imagination ;
- (C) Conscienciosité (conscience) : autodiscipline, respect des obligations, organisation plutôt que spontanéité ; orienté vers des buts ;
- (E) Extraversion : énergie, émotions positives, tendance à chercher la stimulation et la compagnie des autres, fonceur ;
- (A) Agréabilité (amabilité) : une tendance à être compatissant et coopératif plutôt que soupçonneux et antagonique envers les autres ;
- (N) Neuroticisme ou névrosisme : contraire de stabilité émotionnelle : tendance à éprouver facilement des émotions désagréables comme la colère, l'inquiétude ou la dépression, vulnérabilité.

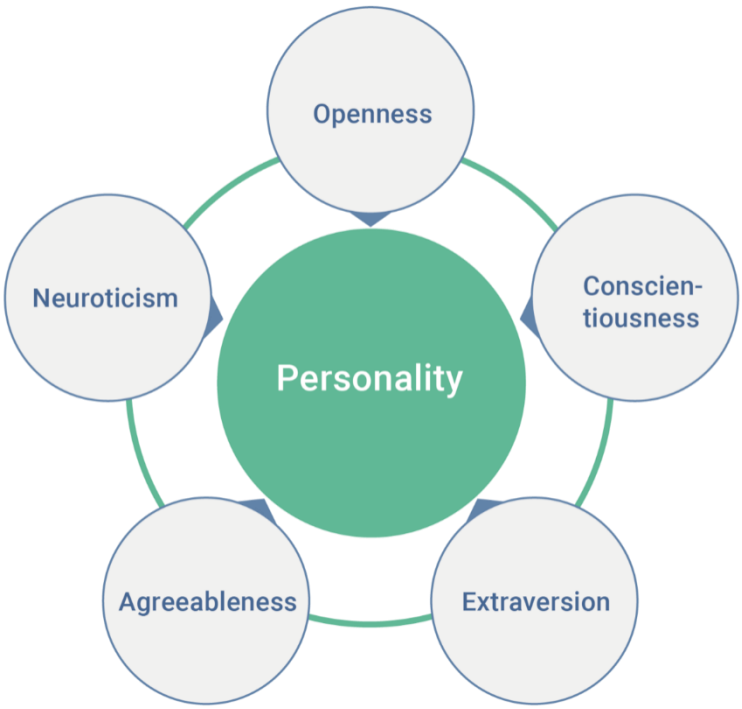

Figure 13 : Les cinq grands traits de personnalité

4.2. Approche de l'évaluation automatique de la personnalité

Dans la littérature de l'évaluation automatique de la personnalité il existe plusieurs approches[21] :

- Questionnaires de personnalité.
- Approche linguistique
- Approche Vocabulaire Ouvert.
- Approche Basé sur l'utilisation d'Internet et la tendance à utiliser les réseaux sociaux
- Lien et analyse basée sur le contenu des données de médias sociaux.
- Structure / analyse de contenu de données de médias sociaux.
- Comportement / analyse basée sur l'activité des données de médias sociaux.

Questionnaires de personnalité

La plupart des évaluations de personnalité se basait sur des questionnaires avec des échelles conçues pour une application pratique spécifique ou pour mesurer les constructions dérivées de la théorie de la personnalité. Plusieurs chercheurs ont utilisé des questionnaires

de 28 question [22], d'autre on utilisé 124 question y compris l'usage des réseaux sociaux [23]. Cette approche est la plus simple et la plus utilisé par la communauté scientifique, mais il présente plusieurs limites :

- Les gens peuvent falsifier le test.
- Beaucoup de temps (évaluation personne par personne).
- Cher (embauche d'un professionnel) que les professionnels facturent par les candidats.
- Exactitude car ils sont jugés par la personne, alors parfois la qualification et le talent de la personne comptent ici. De plus, les humains ont une tendance naturelle à préjuger et sont donc exposé à des erreurs humaines.

Approche linguistique

Les auteurs de [24] et [25] ont montré que les caractéristiques linguistiques peuvent être utilisées pour prédire les traits de personnalité. Ces méthodes peuvent aider à prédire la personnalité en fonction du texte de l'utilisateur sur les sites de médias sociaux. C'est l'une des approches d'analyse les plus simples pour l'évaluation de la personnalité. L'outil d'enquête linguistique et de comptage de mots (LIWC) [26] est un outil d'analyse psycholinguistique qui traite un document texte et produit le pourcentage de mots correspondant à des catégories prédéfinies. Il produit des statistiques sur 81 caractéristiques différentes du texte dans cinq catégories, y compris les comptes standards, les processus psychologiques, la relativité, les préoccupations personnelles et d'autres dimensions. Il compte également les mots en fonction de la partie du discours et de la longueur moyenne des mots. Le texte de l'utilisateur dans les réseaux sociaux comme la mise à jour du statut de Facebook peut être analysé avec l'outil LIWC pour analyser les traits de personnalité. Les chercheurs ont utilisé LIWC et fait des conclusions sur les traits de personnalité.

Approche Vocabulaire Ouvert

Ceci est une extension de l'approche linguistique. Dans l'approche précédente, les publications décrites utilisent la technique du vocabulaire fermé. Dans la technique du vocabulaire fermé, les psychologues choisissent une liste de mots fermée pour déterminer les traits de personnalité. Dans [27] la technique du vocabulaire ouvert; la collecte de mots, de phrases et de sujets basée sur les données est extraite. Le lexique dépend des mots du texte

analysé. Dans cet article, l'auteur a utilisé 700 millions de mots, de phrases et de sujets recueillis auprès de 75000 utilisateurs de Facebook et les a corrélés avec le sexe, l'âge et la personnalité. L'auteur a montré que cette approche fournit des informations supplémentaires par rapport aux techniques de vocabulaire fermées classiques.

Approche Basé sur l'utilisation d'Internet et la tendance à utiliser les réseaux sociaux

Beaucoup de travaux ont été élaboré en utilisant cette approche, surtout sur la Platform Facebook, on peut citer le papier qui a montré que certains des cinq grands traits de personnalité sont associés à l'utilisation totale d'Internet et à la propension des utilisateurs à utiliser les médias sociaux et le site de réseautage social [28]. Le document qui a fait des hypothèses basées sur l'utilisation de Facebook[23]. Dans cette recherche, il a prouvé que l'individu avec un score plus élevé sur l'extraversion et le narcissisme est plus susceptible d'être un utilisateur de Facebook ; alors que les gens très consciencieux sont des non-utilisateurs de Facebook. Les auteurs de [29] qui ont mené des recherches sur 132 étudiants et conclu que ceux qui sont émotionnellement moins stables ont tendance à passer plus de temps sur Facebook. Le travail présenté dans [30] qui a conclu que l'extraversion et l'ouverture à l'expérience sont positivement liée à l'utilisation des applications sociales sur internet, la stabilité émotionnelle était négativement associée. Il a prouvé que l'âge et le genre jouent également un rôle dans ces dynamiques. Les hommes et les femmes extravertis sont susceptibles d'être plus fréquent les utilisateurs des médias sociaux, alors que seuls les hommes ayant une faible stabilité émotionnelle étaient des utilisateurs réguliers. Les résultats de l'extraversion étaient particulièrement valables chez les jeunes adultes et l'ouverture est apparue comme un indicateur important de l'utilisation des médias sociaux pour le segment mature. Enfin le papier [31] qui a montré que l'ouverture est positivement corrélée à la volonté d'utiliser Facebook comme outil de communication.

Les limitations de cette approche résident dans la concentration sur la quantité de temps passé au lieu de la façon dont les individus utilisent ces derniers.

Lien et analyse basée sur le contenu des données de médias sociaux.

Contrairement aux autres approches, les données sur les médias sociaux comme la mise à jour du statut de Facebook permettent aux chercheurs d'observer le comportement des gens

lorsqu'ils s'expriment librement dans leurs propres mots. En raison de la grande popularité des médias sociaux, il fournit une quantité sans précédent de données aux chercheurs en termes de contenu (texte, image audio et vidéo), de profil, d'informations sur la structure et les liens, etc. De nombreux chercheurs ont utilisé ces données pour prédire les traits de personnalité. La plupart des recherches utilisent des approches multiples telles que la combinaison de caractéristiques linguistiques[32], de données d'utilisation d'Internet, de questionnaires et de données de couplage pour prédire les traits de personnalité[33]. Il y a deux approches pour analyser le contenu de Facebook.

- Structure / analyse de contenu de données de médias sociaux.
- Comportement / analyse basée sur l'activité des données de médias sociaux.

Conclusion

Dans ce chapitre nous avons fait le tour de l'état de l'art sur les composantes de la fouille de texte. L'accent a été mis sur les outils de collecte de données textuelles depuis le web ; les tâches de la fouille de textes et les méthodes de résolution. Dans le chapitre suivant, nous allons aborder la conception et la validation de notre Framework.

Chapitre 4 : Framework d'analyse de textes

Introduction

Dans ce chapitre, nous allons présenter l'analyse, la conception et la réalisation de notre Framework. Nous présentons en premier lieu le cahier des charges, ensuite une description détaillée de l'architecture proposée et enfin les composants du Framework dans un digramme de packages.

1. Besoins fonctionnelles et techniques du Framework

L'objectif principal de ce projet est le développement d'un Framework d'analyse de textes issus des réseaux sociaux. Après une analyse approfondie de l'état de l'art, nous avons pu élaborer un cahier charge qui regroupe toutes les fonctionnalités du Framework résumés comme suit :

- Appelle des API des réseaux sociaux.
- Recherche et collecte des données textuelles concernant un sujet précis sur les réseaux sociaux ;
- Stockage de données collectées ;
- Classification de données collectées ;
- Indexation de données collectées ;
- Lacement d'une analyse des sentiments ;
- Recherche et collecte des informations textuelles sur une personne ;
- Lacement d'une analyse de personnalité ;
- Classification des documents par sujet ;
- Alimentation du Framework par des corpus sur des sujets précis ;
- Transformation des résultats de l'analyse en représentation des graphiques ;
- Visualisation des résultats finaux sur des Dashboard de graphes.

Ces besoins ont été ensuite modélisés par un diagramme de cas d'utilisation présenté par la figure 14

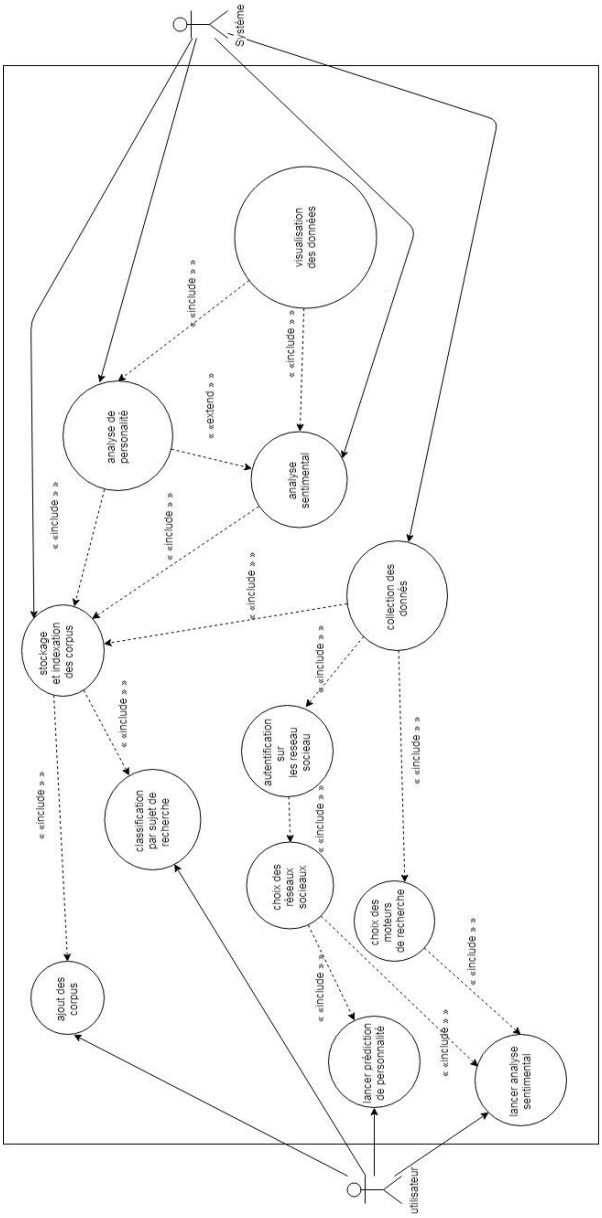

Figure 14 : Diagramme de cas d'utilisation du Framework proposé

En plus, le Framework doit satisfaire un certain nombre d'exigences techniques résumées comme suit :

- Fonctionnement sur plusieurs systèmes d'exploitation : Windows, Linux et MacOs.
- Le Framework doit être intégrable sur un écosystème Big Data.

- Traitement de plusieurs langages naturels, dans le sens où le Framework doit permettre de faire un apprentissage sur la langue souhaitée
- Visualisation et stockage des résultats de l'analyse des sentiments sous des formats graphiques appropriés à l'interprétation et la prise de décision.
- Le Framework doit être modulable dans le sens où il doit permettre d'intégrer d'autres outils et techniques de classification de documents, de prédiction de personnalité, etc.

2. Architecture du Framework

Afin de répondre aux besoins techniques présentés dans le cahier des charges, nous proposé une architecture en couhes montré dans la Figure15. Nous avons proposé quatre couches : la couche collection de données textuelles depuis le web ; la couche stockage et indexation ; la couche d'analyse de texte et la couche outils contenant le corpus, le lexique et les modèles de prédiction. Dans la suite nous présentons en détails les différentes couches.

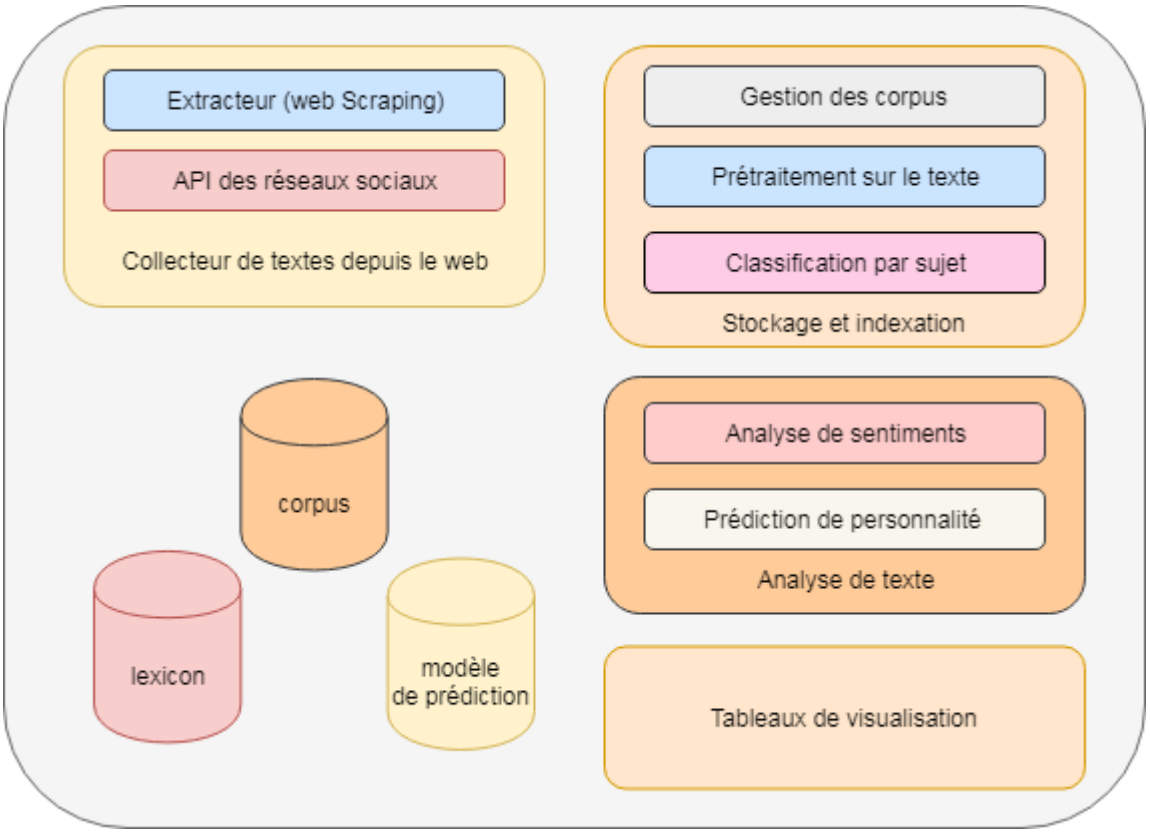

Figure 15 : l'architecture du Framework proposé

2.1. Couche de collection de données textuelles

Ce composant implique le processus d'exploration et de collecte des résultats des moteurs de recherche (blog, forums, …) à partir de différents sites Web. Le processus comprend aussi l'appel des API des réseaux sociaux (Facebook, tweeter). Le composant inclut aussi le nettoyage des données (tag html, images, vidéos, …).

Dans cette couche nous avons fait appel, en premier lieu, aux APIs de Facebook et Twitter. Malheureusement, l'api de Facebook présente des limitations à extraire des informations comme les commentaires d'une publication. Pour résoudre cette limitation, nous avons fait appel à l'outil Selenium.

Selenium est un Framework de tests développé en Java pour l'automation des tests fonctionnelle d'une application web, il supporte plusieurs langage de programmation notamment java et python. Son principe de fonctionnement est très simple, il lance un navigateur en arrière-plan qui exécute un script qui peut contenir par exemple, des tests sur le bon fonctionnent des champs de saisi d'une page de login. Cet avantage que Selenium offre en terme d'interaction avec les applications web est utilisé dans le web Scraping, puisque plusieurs site web cachent les données pour inciter les utilisateurs à interagir avec l'interface graphique. Dans ce contexte, nous l'avons utilisé pour extraire les informations textuelles depuis Facebook ainsi que le moteur de recherche de Google. L'avantage avec Selenium c'est qu'il permet de d'extraire les données cachées d'une page web (les cas où l'utilisateur doit interagir avec l'interface graphique pour afficher des résultats) par exemple les données où un utilisateur doit cliquer sur les objets visuels. Dans le cas de Facebook quand une page de publication est chargée les commentaires par rapport à cette publication ne sont pas tous affichés, par défaut Facebook affiche N commentaire (N entre 10 et 20) il faut donc cliquer sur un bouton « afficher plus » (voir Figure16) pour faire glisser les autres commentaires. Un seul clique sur ce bouton ne suffit pas pour afficher tous les commentaires puisqu'il charge seulement 10 commentaires, donc il faut cliquer jusqu' à ce que tous les commentaires soient chargés.

Figure 16 : le bouton afficher plus sur Facebook

2.2. Couche de stockage et indexation

La tâche principale cette couche est d'extraire des termes ou des phrases clés d'un texte donné. Les termes sont des expressions qui sont statistiquement significatifs dans le corpus Une autre tâche consiste à filtrer et classer un les textes collectés par sujet. Lors de la collecte de textes depuis les différentes sources du web, il est très commun de constater que de nombreux textes collectés ne sont pas pertinents pour les marques ou les produits surveillés. Par conséquent, un modèle de classification pourrait être développé pour filtrer les textes pertinents du corpus. Après avoir obtenu les textes pertinents, un autre modèle de classification pourrait être formé pour classer chaque texte dans un ensemble prédéfini de sujets. Par exemple, dans le domaine des services mobiles, les sujets peuvent inclure la qualité de signal, les promotions et le service client. Dans cette étape on prépare les données pour la couche suivante en nettoyant des données bruitées par les tag html et XML pour la donnée collectée à l'aide du web Scraping d'un coté, et la réorganisation des données collecté à l'aide des APIdes réseaux sociaux.

Pour cette étape nous avons utilisé la bibliothèque Scikit-learn et Tensorflow d'une part pour construire les modèles de classification des documents, et les bibliothèques NLTK pour préparer les textes et supprimer les informations non pertinent : les espaces, les ponctuations ; les mots vides : de, la, le …. La bibliothèque Scikit-learn permet aussi de transformer le texte en format numérique et l'extraction de ses caractéristiques. Nous avons aussi utilisé dans cette partie la bibliothèque Beautifulsoup pour extraire les informations

pertinentes et le nettoyage des fichiers web (html, XML, …) et aussi la bibliothèque Pandas pour sauvegarder les documents de façon efficace.

2.3. Couche d'analyse de texte

C'est la principale composante de notre Framework, il est constitué essentiellement de deux modules d'analyse : l'analyse des sentiments, et la prédiction de personnalité, l'appel des deux modules peut de faire dans le même processus ou bien de façon séparé. (Le cas séparé peut paraitre évidant lorsque on n'a pas d'information sur propriétaire d'une publication, c.-à-d. : on peut pas lancer une analyse de personnalité) cette couche exploite les données collectées par la couche de collection de données. Dans cette couche, nous avons utilisés la bibliothèque Scikit-learn et Tensorflow d'une part pour construire les modèles d'analyse des sentiments et de prédiction de personnalité. Scikit-learn est utilisé aussi pour la numérisation et l'extraction des caractéristiques. Pandas est exploité dans cette couche pour la sauvegarde des résultats de l'analyse de texte afin de les réutiliser à n'importe quel moment.

2.4. Couche de visualisation

Cette couche est le résultat final de notre Framework ; ce composant présente les résultats de l'analyse de texte sous formes de graphes adéquats. Ces graphes permettent aux utilisateurs du Framework d'avoir une idée générale sur leurs produits ou leurs marques, afin qu'ils puissent prendre des discisions par la suite. Pour la mise en œuvre de cette couche nous avons fait appel à Matplotlib : une bibliothèque de visualisation open source écrite en langage python.

3. Développement du Framework

Le Framework implémentant l'architecture ci-dessus, a étét développé en deux étapes :

- ***Conception des composants*** : cette étape consiste à choisir les éléments adéquats pour notre architecture. Dans notre cas, il s'agit de choisir les meilleurs modèles de prédiction, en termes de performances, et la meilleure implémentation de la phase d'extraction des caractéristiques. Pour ce faire, nous avons utilisé le processus de la science de données qui nous a fournit un cadre méthodologique bien maitrisé pour atteintre ce but. Ainsi, nous avons implémenté notre

architecture avec différents modèles et conduit une étude expérimentale afin de choisir les meilleurs modèles à inclure dans la version finale. L'étude expérimentale est présentée dans le chapitre suivant.

- ***Réalisation du Framework*** **:** consiste à implémenter la version finale du Framework avec les choix conceptuelles validés lors de l'experimentation.

3.1. Environnement Matériel et logiciel

Le Framework a été développé dans les environnements suivants :

Environement matériel

- Lenovo Thinkpad T410 :
  - RAM : 8 GB
  - Processeur : Intel(R) Core(TM) i5-CPU M520 @ 2,40GHZ 2,40GHZ
  - GPU : Intel(R) HD Graphics Family
  - Disque dur : 500 GB

- HP Z620 Desktop Workstation
  - RAM : 64 GB
  - Processeur :: Intel(R) Xeon(R) CPU E5-2640 @ 2.50GHz * 24
  - GPU :GetForce GTX 1070/PCIe/SSE2

Environement logiciel

- Système d'exploitions linux (Ubuntu 16.04 LTS).
- Langage de programmation : Python 3.6.
- Jupyter notebook outil d'expérimentions des modèles.
- Spyder IDE de développement.

3.2. Organisation du Framework

Le diagramme de packages de la Figure 17 présente l'organisation physique de notre Framework rganisation du Framework, il montre les liens de généralisation et de dépendance entre les packages . Les modules du Framework contiennent des modules, où chaque module qui contient des scripts codant les fonctionnalités par rapport un traitement particulier.

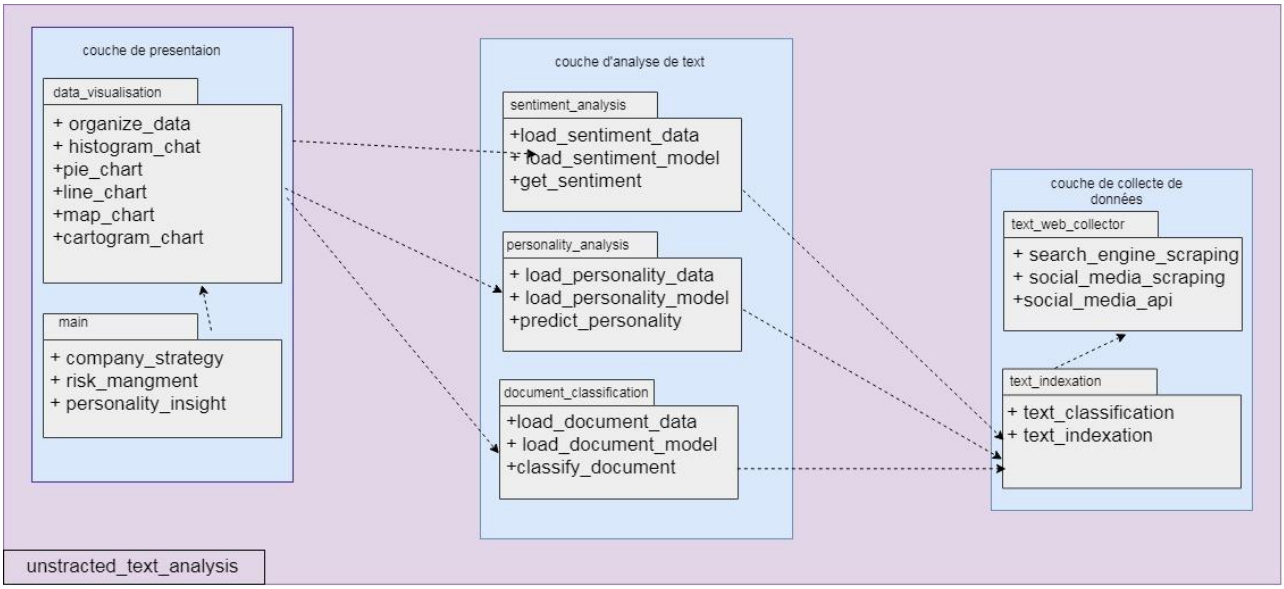

Figure 17 : Diagramme de package de notre Framework

Conclusion

Dans ce chapitre, nous avons présenté les étapes de développement de notre Framework. Nous avons présentés le cahier des charges l'architecture proposée et l'implémentation du Framework. Nous avons aussi expliqué que le choix des éléments de chaque composant est effectué à l'aide du processus science de données. Ainsi, dans le chapitre suivant, nous allons présenter les résultats des expérimentations conduites pour cette fin.

Chapitre 5 : Etude expérimentale

Introduction

Ce chapitre présent la partie développement de notre projet dans un cadre projet data science. Le chapitre inclut une partie d'expérimentation où on va choisir les meilleurs algorithmes d'apprentissage que on va intégrer dans notre Framework. A la fin de ce chapitre une section d'étude de cas ou on mettra des captures d'écran qui montrerons le bon fonctionnement du Framework.

1. Etude Expérimental

A ce stade nous avons presque tous les éléments pour mener nous expérimentations avant d'aborder la partie du développement.

1.1. Source de données

Dans le deuxième chapitre on a parlé de plusieurs outils de traitement de langage naturelle, mais on s'focalisé dans notre recherche sur le sentiment analyse, la classification des documents, et la prédiction de la personnalité, donc en premier temps on a cherché des base d'apprentissage pour ces trois champs d'analyse de textes, dans le tableau ci-dessous une description des base d'apprentissage que on a utilisé :

Classification des document par sujet	
Nom de la base	20 Newsgroups
Langue	Anglais
Description	Cette base de données comprend 20000 messages/articles provenant de 20 groupes de discussion. 20 sujets : religions, politiques,Matériel informatique, pistolets …
Sources / propriétaire	Tom Mitchell Universitéde Carnegie Mellon tom.mitchell@cmu.edu

Format de la base	Chaque sujet est stocké dans un sous-répertoire, chaque article étant stocké dans un fichier distinct.
Analyse de sentiment	
Nom de la base	Large Movie Review Dataset
Langue	Anglais
Description	Il s'agit d'un ensemble de données pour la classification des sentiments binaires contenant 25 000 critiques de films très polaires pour le training, et 25 000 pour les tests.
Sources / propriétaire	Maas Andrew et al. Association for Computational Linguistics
Format de la base	Deux dossier (TRAIN, TEST) qui contient deux sous dossier (POS, NEG) qui contient chacun 12500 fichier (critiques positive ou négative)
Prédiction de personnalité	
Nom de la base	English stream-of-consciousness texts by students
Langue	Anglais
Description	Il se compose d'un total de 2468 essais ou des soumissions d'écriture quotidienne de 34 étudiants en psychologie. Il y a un total de 29 femmes et 5 hommes âgés de 18 à 67 ans
Sources / propriétaire	James Pennebaker et Laura King (1999)[25]
Format de la base	Fichier CSV contient 2468 lignes, chaque line est composé d'une colonne qui contient le texte du volontaire et 5 colonne des valeur des Big 5 (YES ou NO).

Tableau 1 : Les Bases de données utilisé pour l'apprentissage automatiques

1.2. Mesure d'évaluation des modèles

Pour tester la qualité de tout système de classification, il est nécessaire d'effectuer certaines mesures d'évaluation. La procédure typique pour l'apprentissage automatique comprend les phases suivantes :

- Choisir un algorithme approprié et définir les options initiales.
- Former le modèle sur des données compatibles.
- Création de prédictions à l'aide de nouvelles données, en fonction du modèle.
- Évaluer le modèle pour déterminer l'exactitude des prédictions,le nombre d'erreurs, et s'il y a un sur-apprentissage.

Pour valider les modèles que on va utiliser dans notre Framework, on choisit plusieurs métriques, ci-dessous une description des métriques utilisées :

- Accuracy (exactitude):Dans la classification multi label, cette fonction calcule la précision des sous-ensembles: l'ensemble des étiquettes prédites pour un échantillon doit correspondre exactement à l'ensemble des étiquettes correspondant.
- Confusion matrix (matrice de confusion) : Une matrice de confusion est une table qui est souvent utilisée pour décrire la performance d'un modèle de classification (ou « classificateur ») sur un ensemble de données de test pour lesquelles les vraies valeurs sont connues. La matrice de confusion elle-même est relativement simple à comprendre, mais la terminologie associée peut être source confusion [35]. Commençons par un exemple de matrice de confusion pour un classificateur binaire (bien qu'il puisse facilement être étendu au cas de plus de deux classes)

	valeur prédit = 0	Valeur prédit = 1
Label = 0	50	10
Label = 1	5	100

Tableau 2 : exemple de teste de classificateur binaire

  - Il y a deux classes prédites possibles : "1" et "0". Si nous prédisions la présence d'une maladie, par exemple, «1 » signifierait qu'ils ont la maladie et «0 » signifierait qu'ils n'ont pas la maladie.
  - Le classificateur a fait un total de 165 prédictions (par exemple, 165 patients étaient testés pour la présence de cette maladie).

- Sur ces 165 cas, le classificateur prédit «1 » 110 fois, et «0 » 55 fois.
- En réalité, 105 patients dans l'échantillon ont la maladie, et 60 patients ne le font pas.

Définissons maintenant les éléments de la matrice de confusion, qui sont des nombres entiers (pas des taux) :

- Vrais positifs (VP): Ce sont des cas dans lesquels nous avons prédit oui (ils ont la maladie), et ils ont la maladie.
- Vrais négatifs (VN): Nous avons prédit non, et ils n'ont pas la maladie.
- Faux positifs (FP): Nous avons prédit oui, mais ils n'ont pas réellement la maladie.
- Faux négatifs (FN): Nous avons prédit non, mais ils ont effectivement la maladie.

- **Precision and recall** (précision et rappel) : ces deux mesures sont toujours confondues l'une à l'autre, pour mieux comprendre la différence entre ces deux mesures, considérant l'exemple suivant [36]: Disons que vous avez cherché sur Google "qu'est-ce que la précision et le rappel ?" Et en moins d'une minute, vous avez environ 15 600 000 résultats. Disons que sur ces 15,6 millions de résultats, les liens pertinents à votre question étaient d'environ 2 millions. En supposant qu'il y avait aussi environ 6 millions de résultats supplémentaires qui étaient pertinents mais qui n'ont pas été retournés par Google, pour un tel système nous dirions qu'il a une précision de 2M / 15.6M et un rappel de 2M / 8M. Cela implique que la probabilité que l'algorithme de Google récupère tous les liens pertinents était de 0,25 (rappel) et la probabilité que tous les liens récupérés soient pertinents est de 0,13 (précision).
- **ROC AUC** : pour comprendre cette mesure il faute d'abord comprendre la courbe du ROC[37].
  - **Courbe ROC** : Une courbe ROC (receiver operating characteristic) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs :
    - Taux de vrais positifs

- Taux de faux positifs

Le taux de vrais positifs (TVP) est l'équivalent du rappel. Il est donc défini comme suit :

$$TVP = \frac{VP}{VP + FN}$$

Le taux de faux positifs (TFP) est défini comme suit :

$$TFP = \frac{VP}{FP + VN}$$

Une courbe ROC trace les valeurs TVP et TFP pour différents seuils de classification. Diminuer la valeur du seuil de classification permet de classer plus d'éléments comme positifs, ce qui augmente le nombre de faux positifs et de vrais positifs. La figure ci-dessous représente une courbe ROC classique.

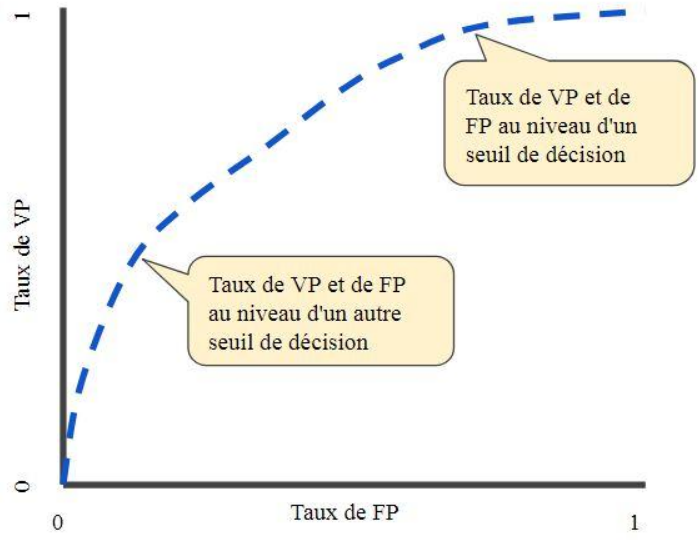

Figure 18 : Taux de VP et de FP pour différents seuils de classification

- **AUC** : **AUC** signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire à deux dimensions située sous l'ensemble de la courbe ROC (par calculs d'intégrales) de (0,0) à (1,1)

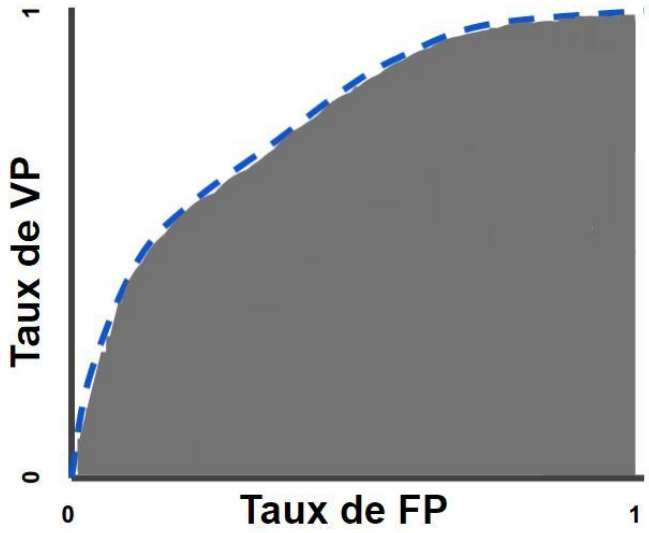

Figure 19: AUC (aire sous la courbe ROC)

L'AUC fournit une mesure agrégée des performances pour tous les seuils de classification possibles. On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire.

- **F1-score** : Dans l'analyse statistique de la classification binaire, le score F1 (également F-score ou F-mesure) est une mesure de la précision d'un test. Il considère à la fois la précision p et le rappel r du test pour calculer le score. Le score F1 est la moyenne harmonique de la précision et du rappel, où un score F1 atteint sa meilleure valeur à 1 (précision parfaite et rappel) et le pire à 0.

Remarque : pour notre cas on va s'intéresser seulement à l'exactitude (Accuracy) puisque on se n'intéresse pas à un label tout particulier (0 ou 1), par exemple dans une situation où on veut protéger un système d'identification par visage on s'assure que le système éjectera les imposteurs on veut donc minimiser (Faux positifs), dans ce que on peut par exemple d'intéresser à la précision ou à la matrice de confusion tout court

1.3. Choix de la base entrainement / test

Le bon choix des échantillons de la base d'entrainement et ceux de la base de test est très important pour la construction du model. En effet le choix de mauvais échantillon peut

influencer sur l'efficacité du model, et peut mener à un sur-apprentissage. Comme il n'y a jamais assez de données pour construireles modèles,et supprimer une partie pour la validation pose un problème de sous apprentissage En réduisant les données d'apprentissage, nous risquons de perdre d'important pattern dans les de données, ce qui augmente à son tour l'erreur et diminue l'exactitude. Donc, ce dont nous avons besoin, c'est d'une méthode qui fournit suffisamment de données pour la formation du modèle et qui laisse aussi beaucoup de données pour la validation. K Fold validation croisée fait exactement cela.

Dans la validation croisée K Fold, les données sont divisées en k sous-ensembles. Maintenant, la méthode de division est répétée k fois, de sorte qu'à chaque fois, l'un des k sous-ensembles est utilisé comme ensemble de test / validation et les autres sous-ensembles k-1 sont assemblés pour former un ensemble d'apprentissage. L'estimation de l'erreur est moyennée sur tous les k essais pour obtenir l'efficacité totale de notre modèle.

Pour s'assurer que le changement des positionnent du coupage influence sur l'efficacité du model on calcule le meilleur score obtenu, le minimum, la moyenne, et l'écart type, ce dernier permet de déterminer si les score pour chaque itération est proches des autres.

1.4. Modèles utilisés

Comme résultat de la phase de l'état de l'art nous avons pu choisir les algorithmes d'apprentissages les plus utilisé par la communauté scientifique. Ci-dessous une description non détaillée des algorithmes utilisé dans la phase d'expérimentation

- **Régression logistique** : Ne soyez pas confus par son nom ! C'est une classification et non un algorithme de régression. Il est utilisé pour estimer des valeurs discrètes (valeurs binaires telles que 0/1, oui / non, vrai / faux) sur la base d'un ensemble donné de variable (s) indépendante (s). En termes simples, il prédit la probabilité d'occurrence d'un événement en ajustant les données à une fonction LOGIT. Par conséquent, il est également connu sous le nom de régression LOGIT. Depuis, il prédit la probabilité, ses valeurs de sortie se situent entre 0 et 1 (comme prévu).

- **Arbre de décision** : C'est un type d'algorithme d'apprentissage supervisé qui est principalement utilisé pour les problèmes de classification. Étonnamment, cela fonctionne à la fois pour les variables dépendantes catégorielles et continues. Dans cet algorithme, nous divisons la population en deux ou plusieurs ensembles homogènes. Ceci est fait sur la base des attributs les plus significatifs / variables indépendantes à faire en tant que groupes distincts que possible.

- **Machine a vecteur de support** (SVM) : C'est une méthode de classification. Dans cet algorithme, nous traçons chaque élément de données comme un point dans un espace à n dimensions (où n est le nombre d'entités que vous avez) avec la valeur de chaque entité étant la valeur d'une coordonnée particulière.
  Par exemple, si nous n'avions que deux caractéristiques comme la hauteur et la longueur des cheveux d'un individu, nous commencerions par tracer ces deux variables dans un espace bidimensionnel où chaque point a deux coordonnées (ces coordonnées sont connues sous le nom de vecteurs de support)

- **Forêt d'arbres décisionnels** : est un terme utilisé pour désigner un ensemble d'arbres de décision. Dansla Forêt d'arbres décisionnels, nous avons une collection d'arbres de décision (connus sous le nom de "Foret"). Pour classer un nouvel objet basé sur des attributs, chaque arbre donne une classification et nous disons l'arbre "votes" pour cette classe. La forêt choisit la classification ayant le plus de votes (sur tous les arbres de la forêt).

- **Gradient boosting** : algorithme de boost utilisé lorsque nous traitons avec beaucoup de données pour faire une prédiction avec un pouvoir de prédiction élevé. Boosting est en fait un ensemble d'algorithmes d'apprentissage qui combine la prédiction de plusieurs estimateurs de base afin d'améliorer la robustesse par rapport à un seul estimateur. Il combine plusieurs prédicteurs faibles ou moyens à un prédicteur fort de construction.

- **Bagging** : est un algorithmequi diminue la variance de la prédiction en générant des données supplémentaires pour la formation à partir d'un ensemble de données original

en utilisant des combinaisons avec des répétitions pour produire des multi ensemble de la même cardinalité / taille que les données originales. En augmentant la taille de l'ensemble d'entraînement, vous ne pouvez pas améliorer la force prédictive du modèle, mais simplement diminuer la variance, en ajustant étroitement la prédiction aux résultats attendus.

- **Les réseaux neuronaux artificiels (RNA)** : sont des modèles statistiques directement inspirés et partiellement modélisés sur des réseaux neuronaux biologiques. Ils sont capables de modéliser et de traiter les relations non linéaires entre les entrées et les sorties en parallèle. Les algorithmes associés font partie du domaine plus large de l'apprentissage automatique, et peuvent être utilisés dans de nombreuses applications

  Les réseaux neuronaux artificiels sont caractérisés en ce qu'ils contiennent des poids adaptatifs le long des chemins entre les neurones qui peuvent être ajustés par un algorithme d'apprentissage qui apprend à partir des données observées afin d'améliorer le modèle. En plus de l'algorithme d'apprentissage lui-même, il faut choisir une fonction de coût appropriée.

  La fonction de coût est ce qui est utilisé pour apprendre la solution optimale au problème à résoudre. Cela implique de déterminer les meilleures valeurs pour tous les paramètres du modèle accordable, les poids adaptatifs du trajet neuronal étant la cible principale, ainsi que les paramètres d'ajustement de l'algorithme tels que le taux d'apprentissage. Cela se fait généralement par des techniques d'optimisation telles que la descente en gradient ou la descente en gradient stochastique.

Le tableau présente la configuration utilisé dans nos expérimentations

Algorithmes	Configuration utilisé
Régressionlogistique	solver : newton-cg'
Machine a vecteur de support	Kernel : linear
Arbre de discision	Minimum sample split : 20
Forêt d'arbres décisionnels	Estimator number : 100
Gradient boosting	Estimator number: 500

	Minimum sample split : 30
Bagging	Base estimator: logistic regression Estimator number : 50
Réseaux de neurones	Hidden layer sizes: 20 Optimizer: Adam Activation : sigmoid

Tableau 3 : configuration des algorithmes utilisé

1.5. Expérimentation

La procédure de l'expérimentation se fait comme suit :

1. Pour chaque base de donnée on effectue plusieurs prétraitement (extraction de caractéristiques)
2. On divise la base de données en deux bases : entrainement et test à l'aide du validation croisé K-Fold
3. Pour chaque prétraitement on lance plusieurs Algorithmes d'apprentissage,
4. Le meilleur algorithme est celui qui a les meilleurs résultats dans tout métrique de Scoring

Le tableau suivant contient les résultats des expérimentations :

Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Classification des documents par sujet		20 Newsgroups		TF-IDF
Ouverture				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.856764	0.859403	0.862041	0.002639
Forêt d'arbres décisionnels	0.783554	0.790235	0.796917	0.006681
Bagging	0.837666	0.845468	0.853270	0.007802
Gradient boosting	0.417331	0.427498	0.437666	0.010167
Arbre de décision	0.554907	0.561611	0.568315	0.006704

Machine a vecteur de support	0.862599	0.865776	0.868953	0.003177
Régression logistique	0.838727	0.845201	0.851675	0.006474
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		TF-IDF
Ouverture				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.580802	0.589784	0.600243	0.008005
Forêt d'arbres décisionnels	0.516403	0.535076	0.551766	0.014505
Bagging	0.619684	0.621811	0.626066	0.003008
Gradient boosting	0.507898	0.551302	0.582217	0.031600
Arbre de décision	0.516403	0.535076	0.551766	0.014505
Machine a vecteur de support	0.623329	0.625458	0.627284	0.001628
Régression logistique	0.617254	0.621814	0.629720	0.005612
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		3-gram
Ouverture				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.600487	0.610045	0.623329	0.009690
Forêt d'arbres décisionnels	0.585871	0.598692	0.606318	0.009120
Bagging	0.596833	0.612067	0.626974	0.012307
Gradient boosting	0.556638	0.582063	0.609964	0.021840
Arbre de décision	0.504263	0.528963	0.545565	0.017807
Machine a vecteur de support	0.573691	0.586532	0.594168	0.009134

Régression logistique	0.587089	0.597478	0.603888	0.007413
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		TF-IDF
Conscience				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.515188	0.529806	0.545676	0.012478
Forêt d'arbres décisionnels	0.507898	0.531833	0.548112	0.017287
Bagging	0.540705	0.567113	0.600487	0.024898
Gradient boosting	0.507917	0.530991	0.555286	0.019357
Arbre de décision	0.484812	0.508339	0.544458	0.025927
Machine a vecteur de support	0.540705	0.561440	0.598051	0.025964
Régression logistique	0.546780	0.567925	0.602923	0.024926
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		3-gram
Conscience				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.543135	0.548040	0.555420	0.005312
Forêt d'arbres décisionnels	0.534714	0.543973	0.550425	0.006714
Bagging	0.539490	0.551281	0.557716	0.008349
Gradient boosting	0.538368	0.539926	0.540705	0.001102
Arbre de décision	0.493317	0.508734	0.532278	0.016912
Machine a vecteur de support	0.537150	0.540735	0.543135	0.002583
Régression logistique	0.537060	0.541956	0.544458	0.003462

Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		TF-IDF
Extraversion				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.543135	0.565879	0.586375	0.017724
Forêt d'arbres décisionnels	0.548662	0.552089	0.558394	0.004464
Bagging	0.565693	0.575191	0.580292	0.006722
Gradient boosting	0.529769	0.550880	0.574209	0.018210
Arbre de décision	0.504866	0.524518	0.539490	0.014517
Machine a vecteur de support	0.572993	0.579245	0.583942	0.004603
Régression logistique	0.570560	0.578840	0.585158	0.006119
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		3-gram
Extraversion				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.534629	0.542768	0.549878	0.006268
Forêt d'arbres décisionnels	0.540146	0.546818	0.552311	0.005036
Bagging	0.551095	0.554115	0.559611	0.003892
Gradient boosting	0.532847	0.544380	0.557716	0.010232
Arbre de décision	0.513382	0.536276	0.551095	0.016421
Machine a vecteur de support	0.539490	0.543171	0.549878	0.004750
Régression logistique	0.547445	0.548440	0.549878	0.001042
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**

Prédiction de personnalité Agréabilité		English stream-of-consciousness texts by students		TF-IDF
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.535844	0.539927	0.544350	0.003481
Forêt d'arbres décisionnels	0.520049	0.534656	0.551640	0.013006
Bagging	0.552855	0.558582	0.570037	0.008099
Gradient boosting	0.487242	0.520466	0.558931	0.029501
Arbre de décision	0.492102	0.499400	0.509135	0.007164
Machine a vecteur de support	0.546894	0.549653	0.551640	0.002013
Régression logistique	0.546780	0.552905	0.561510	0.006264
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité Agréabilité		English stream-of-consciousness texts by students		3-gram
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.544350	0.555335	0.561510	0.007788
Forêt d'arbres décisionnels	0.538275	0.548440	0.557716	0.007962
Bagging	0.534629	0.552103	0.571255	0.014999
Gradient boosting	0.511543	0.531034	0.561510	0.021828
Arbre de décision	0.505481	0.517218	0.529769	0.009932
Machine a vecteur de support	0.527339	0.541568	0.565164	0.016803
Régression logistique	0.528554	0.547652	0.576127	0.020522
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-		TF-IDF

Agréabilité		consciousness texts by students		
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.535844	0.539927	0.544350	0.003481
Forêt d'arbres décisionnels	0.520049	0.534656	0.551640	0.013006
Bagging	0.552855	0.558582	0.570037	0.008099
Gradient boosting	0.487242	0.520466	0.558931	0.029501
Arbre de décision	0.492102	0.499400	0.509135	0.007164
Machine a vecteur de support	0.546894	0.549653	0.551640	0.002013
Régression logistique	0.546780	0.552905	0.561510	0.006264
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		TF-IDF
Neuroticisme				
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.531630	0.556949	0.576642	0.018802
Forêt d'arbres décisionnels	0.512758	0.547643	0.570560	0.025069
Bagging	0.558394	0.569112	0.579075	0.008460
Gradient boosting	0.513382	0.554514	0.581509	0.029555
Arbre de décision	0.498783	0.504657	0.516403	0.008306
Machine a vecteur de support	0.562044	0.573164	0.580292	0.007967
Régression logistique	0.557178	0.569515	0.575942	0.008726
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Prédiction de personnalité		English stream-of-consciousness texts by students		3-gram
Neuroticisme				

Algorithme d'apprentissage	Minimum accuracy	Maximum accuracy	Mean accuracy	Standard deviation
Réseaux de neurones	0.557178	0.563033	0.571776	0.006300
Forêt d'arbres décisionnels	0.553528	0.558167	0.561361	0.003357
Bagging	0.557178	0.571134	0.583232	0.010718
Gradient boosting	0.544350	0.549252	0.557178	0.005656
Arbre de décision	0.481752	0.507900	0.523114	0.018572
Machine a vecteur de support	0.546780	0.555334	0.569343	0.009986
Régression logistique	0.546229	0.557358	0.570560	0.010041
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Classification des documents par sujet		20 Newsgroups		TF-IDF
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.856764	0.859403	0.862041	0.002639
Forêt d'arbres décisionnels	0.783554	0.790235	0.796917	0.006681
Bagging	0.837666	0.845468	0.853270	0.007802
Gradient boosting	0.417331	0.427498	0.437666	0.010167
Arbre de décision	0.554907	0.561611	0.568315	0.006704
Machine a vecteur de support	0.862599	0.865776	0.868953	0.003177
Régression logistique	0.838727	0.845201	0.851675	0.006474
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Classification des documents par sujet		20 Newsgroups		3-gram
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.816663	0.829443	0.824556	0.001032
Forêt d'arbres décisionnels	0.785612	0.791209	0.791243	0.005692

Bagging	0.826346	0.841234	0.852260	0.003601
Gradient boosting	0.474531	0.485497	0.483456	0.021227
Arbre de décision	0.553908	0.561416	0.565845	0.005743
Machine a vecteur de support	0.861234	0.861297	0.861298	0.002137
Régression logistique	0.858827	0.859201	0.851445	0.003434
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Analyse de sentiment		Large Movie Review Dataset		TF-IDF
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.897763	0.899453	0.894543	0.000135
Forêt d'arbres décisionnels	0.8228	0.83204	0.8362	0.004812
Bagging	0.8828	0.88464	0.8888	0.002129
Gradient boosting	0.8488	0.86048	0.8750	0.010840
Arbre de décision	0.6882	0.69824	0.7056	0.006309
Machine a vecteur de support	0.8898	0.89276	0.8990	0.003208
Régression logistique	0.8854	0.88720	0.8906	0.001931
Outil d'analyse de texte		**Base de données**		**Vecteur de caractéristiques**
Analyse de sentiment		Large Movie Review Dataset		3-gram
Algorithme d'apprentissage	**Minimum accuracy**	**Maximum accuracy**	**Mean accuracy**	**Standard deviation**
Réseaux de neurones	0.87793	0.87955	0.87224	0.000136
Forêt d'arbres décisionnels	0.8243	0.83123	0.8362	0.004812
Bagging	0.8536	0.86664	0.8888	0.002129
Gradient boosting	0.8345	0.84038	0.8750	0.012690
Arbre de décision	0.7022	0.71334	0.7056	0.001209

Machine a vecteur de support	0.8619	0.87116	0.8990	0.003242
Régression logistique	0.8634	0.87660	0.8796	0.001431

Tableau 4 : Résultat des expérimentions

1.6. Discussion des résultats

- **Prédiction de personnalité :**

Le tableau 2, montre que l'ouverture à l'expérience est le trait le plus facile identifier quelles que soient les algorithmes d'apprentissage utilisé et les caractéristiques choisi les traits restants, classés du plus facile au plus difficile à identifier, sont : laConscienciosité, le Neuroticisme, l'Extraversion et l'Agréabilité

Y'a pas une grande différence entre les scores des modèles l'exactitude est entre 0.6% et 0.5%

L'écart type montre qu'il n'y a pas de grandes différence dans l'exactitude des model dans chaque K-fold.

- **Analyse de sentiment :**

On remarque que les réseaux de neurones sont les performant et plus précisément dans la caractéristique de la TF-IDF

Pour chaque model la partition des bases de test et d'apprentissage n'as pas trop influencé sur la différence entre les scores

- **Classification des doucement par sujet :**

Le SVM est l'algorithme qu'a donné les meilleurs de résultat en terme de 'exactitude, et comme les modèles précédant l'écart type n'apporter aucune nouvelle information

2. Etude de Cas sur la marque Adidas

D'après les expérimentations que on a faites on a pu conclure les modèles fiables que on va utiliser dans notre Framework, malheureusement le model de prédiction de personnalité a un taux d'erreur qui monte jusqu'à 0 .45. Ainsi ce model ne sera pas présent

dans cette première version du Framework. Puisque le Framework n'est pas intégrer pour le moment dans une interface home machine, on va lui faire appelle dans l'enivrement Jupyter, qui fournit des outils de visualisation performantes.

Pour tester le bon fonctionnement de notre Framework on va faire une analyse sentimentale de La marque Adidas. On va chercher les publications sur Adidas entre janvier 2017 et janvier 2018. D'après le résultat de la collecte de donnée et l'analyse de sentiment, on peut conclure plusieurs informations qu'on représentera dans les graphes suivant :

- **Pourcentage des publications positives et négatives :**

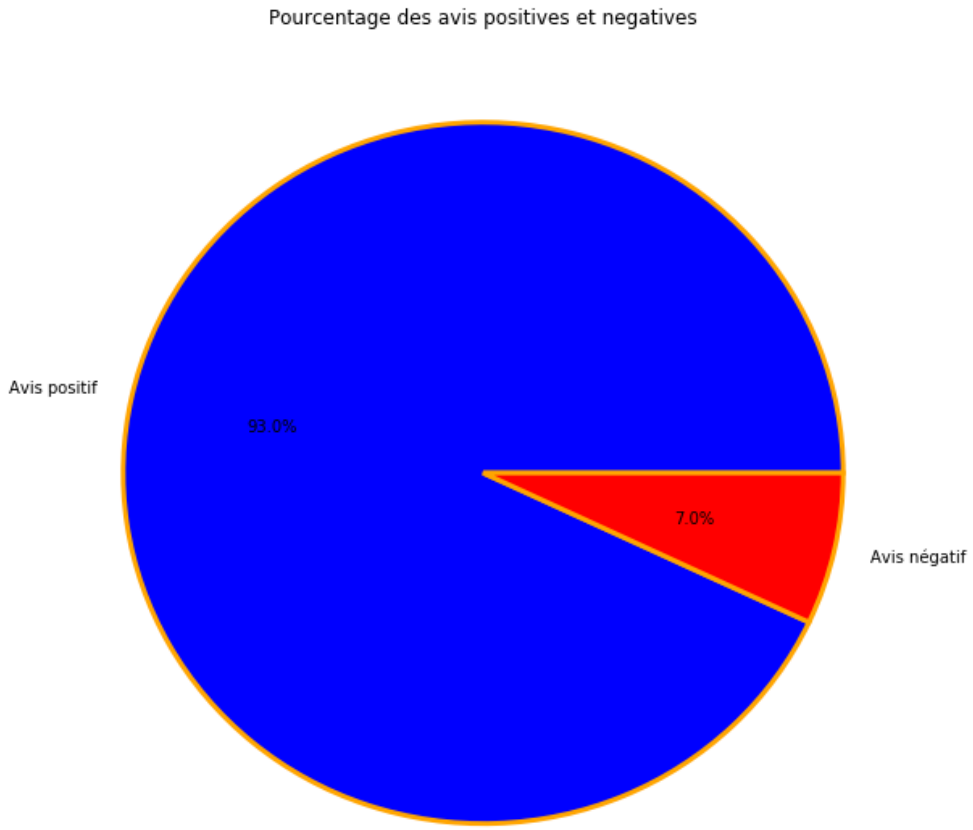

Figure 20 : pourcentages des avis positives et négative de la marque Adidas entre janvier 2017 et janvier 2018

- **Nombre des publication positives et négatives par rapport au temps**

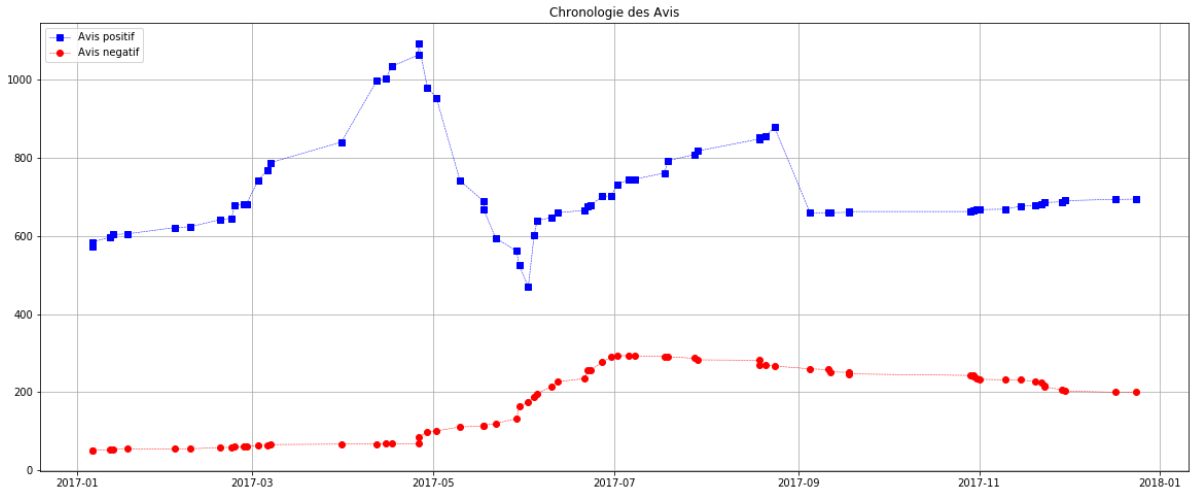

Figure 21 : Nombre des publication positives et négatives par ordre chronologiques

On remarque dans la Figure 21 que les publications positives sont plus nombreuses par rapport à celle qui sont négative, même si au mois de juin 2017 les publication positives on vécut une grosse chute mais sa rester toujours en dessuedes publications négatives

- **Sources de données :**

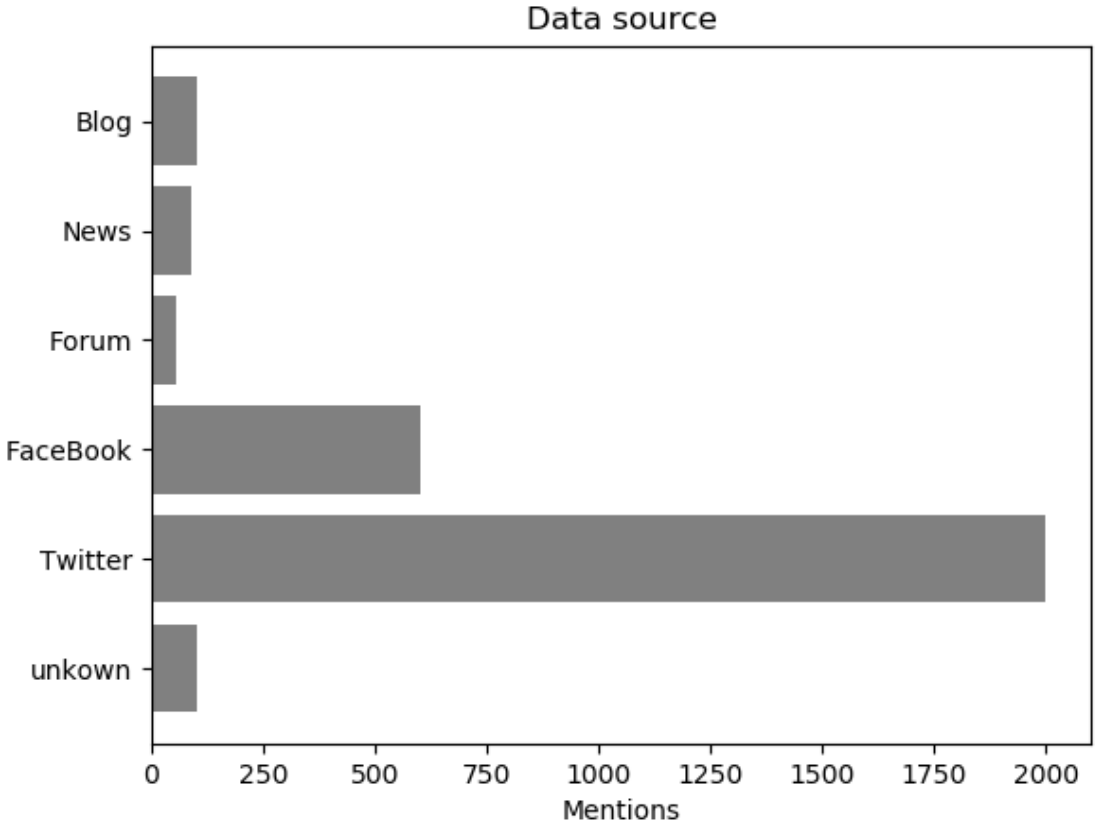

Figure 22: sources de données collecter pour la marque Adidas

- Géo distribution des publications :

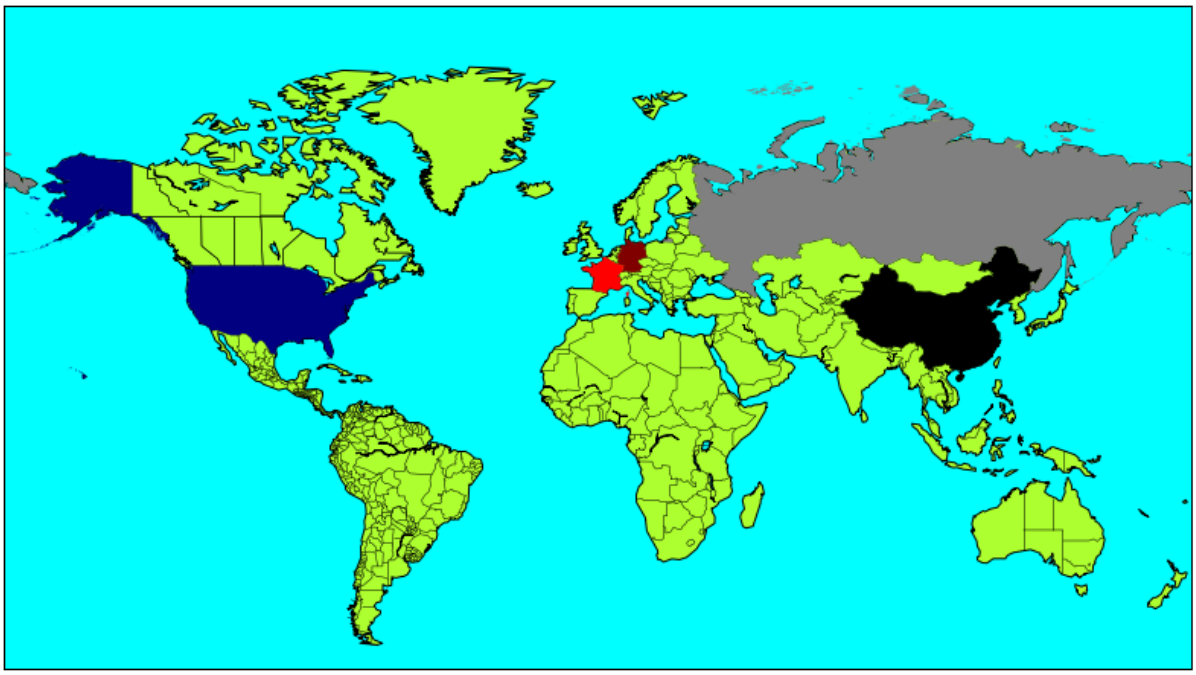

Figure 23 : Géo distribution des publications

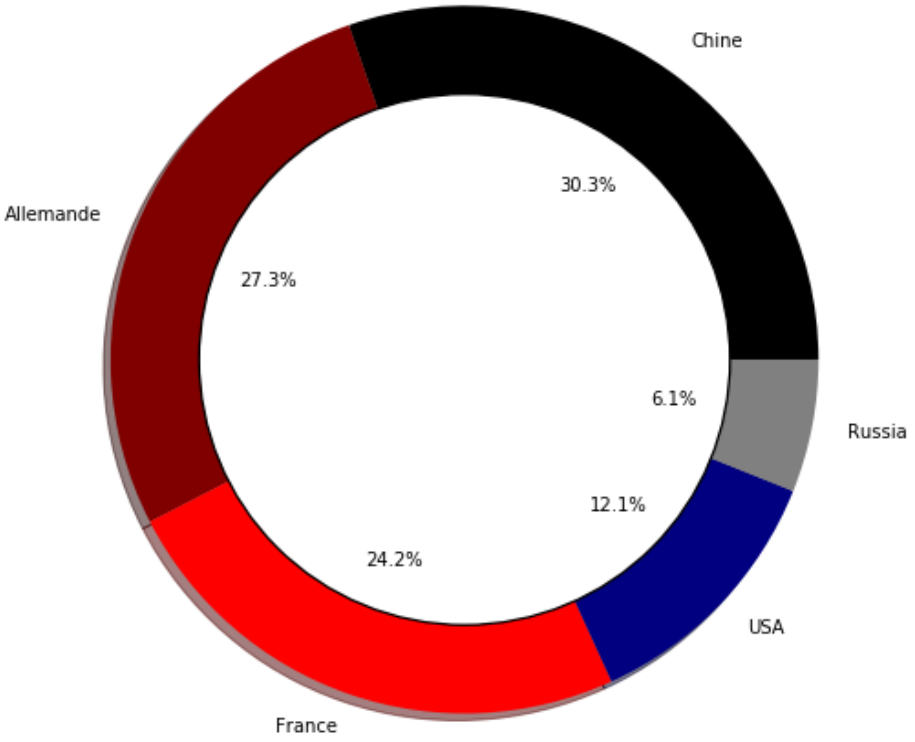

Figure 24 : pourcentage des publications par paye

Conclusion :

Dans ce chapitre, nous avons présenté en détails les expérimentations que nous avons mené,on a fait en sort de respecté le cycle de vie que on a présenté dans le premier chapitre. Nous avons aussi testé notre Framework sur un cas d'utilisation, ou on a représenté le résultat final de notre projet.

Conclusion Générale

Dans le cadre de mon stage au sein de l'entreprise INDATACORE, nous étions amenés à développer un Framework pour l'analyse de texte à partir des données non structurées, dans le but de réaliser des applications de fouille de texte et d'analyse de sentiments.

Dans un premier temps nous avons fait une étude sur les méthodes de collecte des données non structuré depuis le web et un état de l'art sur les méthodes d'analyse de texte plus précisément, la classification des document, l'analyse des sentiments et la prédiction de la personnalité. Nous avons relevé les difficultés par rapport à ces trois tâches de la fouille de texte ainsi que les algorithmes d'apprentissage les plus utilisé pour les mettre en œuvre. Cette étape nous a permis de définir les fonctionnalités essentielles de notre Framework, ce qui nous a aidé par la suite dans la phase de conception. Finalement à l'aide des expérimentations que on a mené, on a pu choisir les meilleurs algorithmes à inclure dans la version finale.

Comme perspectives de notre travail, nous proposons utiliser des données de l'analyse des sentiments, classification de document et de prédiction de personnalité dans différentes langues notamment dans l'arabe, le français et pourquoi pas la DARIJA, afin que on puisse intégrer le Framework dans des applications réelles.

Références

[1] « What Is Data Science? A Beginner's Guide To Data Science | Edureka », *Edureka Blog*, 05-janv-2017. .

[2] « Data Science & Big Data Analytics | Wiley Online Books ». [En ligne]. Disponible sur: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119183686. [Consulté le: 23-mai-2018].

[3] J. Gantz et D. Reinsel, « Extracting value from chaos », *IDC iview*, vol. 1142, n° 2011, p. 1–12, 2011.

[4] « Documentation ». [En ligne]. Disponible sur: http://www.fminer.com/documentation/. [Consulté le: 03-juin-2018].

[5] « Import.io Extract - Import.io - Web Scraping, Web Scraper, Data Extraction, Web Extraction, Web Data, Web Harvesting », *Import.io*. .

[6] « Web Scraper, Web Extractor, Screen Scraper, Web Ripper ». [En ligne]. Disponible sur: http://webextract.net/. [Consulté le: 03-juin-2018].

[7] « API Graph - Documentation », *Facebook for Developers*. [En ligne]. Disponible sur: https://developers.facebook.com/docs/graph-api. [Consulté le: 23-mai-2018].

[8] « Docs — Twitter Developers ». [En ligne]. Disponible sur: https://developer.twitter.com/en/docs. [Consulté le: 23-mai-2018].

[9] « Home | LinkedIn Developer Network ». [En ligne]. Disponible sur: https://developer.linkedin.com/. [Consulté le: 23-mai-2018].

[10] « Fouille de textes », *Wikipédia*. 04-janv-2018.

[11] B. Pang et L. Lee, « Opinion mining and sentiment analysis », *Foundations and Trends® in Information Retrieval*, vol. 2, n° 1–2, p. 1–135, 2008.

[12] B. Liu et L. Zhang, « A survey of opinion mining and sentiment analysis », in *Mining text data*, Springer, 2012, p. 415–463.

[13] « 88% Of Consumers Trust Online Reviews As Much As Personal Recommendations », *Search Engine Land*, 07-juill-2014. [En ligne]. Disponible sur: https://searchengineland.com/88-consumers-trust-online-reviews-much-personal-recommendations-195803. [Consulté le: 23-mai-2018].

[14] N. Indurkhya et F. J. Damerau, *Handbook of natural language processing*, vol. 2. CRC Press, 2010.

[15] A. Esuli et F. Sebastiani, « SentiWordNet: a high-coverage lexical resource for opinion mining », *Evaluation*, p. 1–26, 2007.

[16] B. Pang et L. Lee, « A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts », in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271.

[17] M. W. Berry et M. Castellanos, « Survey of text mining », *Computing Reviews*, vol. 45, nº 9, p. 548, 2004.

[18] V. Korde et C. N. Mahender, « Text classification and classifiers: A survey », *International Journal of Artificial Intelligence & Applications*, vol. 3, nº 2, p. 85, 2012.

[19] L. R. Goldberg, « An alternative" description of personality": the big-five factor structure. », *Journal of personality and social psychology*, vol. 59, nº 6, p. 1216, 1990.

[20] P. T. Costa Jr et R. R. McCrae, « Four ways five factors are basic », *Personality and individual differences*, vol. 13, nº 6, p. 653–665, 1992.

[21] P. K. Atrey et A. K. Tripathi, « Personality Prediction with Social Behavior by Analyzing Social Media Data-A Survey », 2010.

[22] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, et R. R. Orr, « Personality and motivations associated with Facebook use », *Computers in Human Behavior*, vol. 25, nº 2, p. 578-586, mars 2009.

[23] T. Ryan et S. Xenos, « Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage », *Computers in human behavior*, vol. 27, nº 5, p. 1658–1664, 2011.

[24] F. Mairesse, M. A. Walker, M. R. Mehl, et R. K. Moore, « Using linguistic cues for the automatic recognition of personality in conversation and text », *Journal of artificial intelligence research*, vol. 30, p. 457–500, 2007.

[25] J. W. Pennebaker et L. A. King, « Linguistic styles: Language use as an individual difference. », *Journal of personality and social psychology*, vol. 77, nº 6, p. 1296, 1999.

[26] J. W. Pennebaker, M. E. Francis, et R. J. Booth, « Linguistic inquiry and word count: LIWC 2001 », *Mahway: Lawrence Erlbaum Associates*, vol. 71, nº 2001, p. 2001, 2001.

[27] H. A. Schwartz *et al.*, « Personality, gender, and age in the language of social media: The open-vocabulary approach », *PloS one*, vol. 8, nº 9, p. e73791, 2013.

[28] T. Correa, A. W. Hinsley, et H. G. De Zuniga, « Who interacts on the Web?: The intersection of users' personality and social media use », *Computers in Human Behavior*, vol. 26, nº 2, p. 247–253, 2010.

[29] C. Wang et G. Ching, « A study on the relationship of Facebook and EFL learners' personality », *International Journal of research studies in educational technology*, vol. 2, nº 2, 2013.

[30] B. Zhong, M. Hardin, et T. Sun, « Less effortful thinking leads to more social networking? The associations between the use of social network sites and personality traits », *Computers in Human Behavior*, vol. 27, nº 3, p. 1265–1271, 2011.

[31] Y. Amichai-Hamburger et G. Vinitzky, « Social network use and personality », *Computers in human behavior*, vol. 26, nº 6, p. 1289–1295, 2010.

[32] G. Farnadi, S. Zoghbi, M.-F. Moens, et M. De Cock, « Recognising personality traits using Facebook status updates », in *Proceedings of the workshop on computational personality*

recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13), 2013.

[33] M. Kosinski, D. Stillwell, et T. Graepel, « Private traits and attributes are predictable from digital records of human behavior », *Proceedings of the National Academy of Sciences*, vol. 110, n° 15, p. 5802–5805, 2013.

[34] C. E. Shannon, « A mathematical theory of communications », *Bell Systems Technical Journal*, vol. 27, p. 379–423, 1948.

[35] « Simple guide to confusion matrix terminology », *Data School*, 26-mars-2014. [En ligne]. Disponible sur: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Consulté le: 02-juin-2018].

[36] T. Afonja, « Model Evaluation I: Precision And Recall », *Towards Data Science*, 10-janv-2017. [En ligne]. Disponible sur: https://towardsdatascience.com/model-evaluation-i-precision-and-recall-166ddb257c7b. [Consulté le: 02-juin-2018].

[37] « Classification : ROC et AUC | Cours d'initiation au machine learning », *Google Developers*. [En ligne]. Disponible sur: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=fr. [Consulté le: 02-juin-2018].

Mise en place d'une Framework d'analyse de textes pour l'extraction des connaissances à partir des données non-structurées

Résumé

Les flux de données s'accroître de plus en plus et génère une quantité considérable des données non-structurées. Dans ce contexte l'analyse de textes est un outil qui permet d'analyser, d'indexer et de réaliser la fouille de textes des données non structurée. L'objectif de ce projet est de fournir un outil d'analyse de texte. Ce dernier doit contenir un système de flux de données qui permet d'extraire des textes de plusieurs sources. Ensuite, les textes récoltés doivent être indexées pour faciliter la recherche. Finalement, les données non-structurer seront utiliser pour élaborer un dictionnaire de données afin de réalisé des applications de la fouille de textes et d'analyse de sentiments et la Publicité contextuelle.

Mots clés : Données non structuré, Fouille de texte, Analyse de sentiment, Traitement du langage naturelle, Apprentissage automatique

Setting up text analysis framework for the extraction of knowledge from unstructured data

Abstract

Data flows grow more and more and generates a considerable amount of unstructured data. In this context, text analysis is a tool for analyzing, indexing and performing text mining of unstructured data. The goal of this project is to provide a text analysis tool. This Last must contain a data flow system that makes it possible to extract texts from several sources. Then, the collected texts must be indexed to facilitate the search. Finally, the - unstructured data will be used to develop a data dictionary to carry out text mining, sentiment analysis and contextual advertising applications.

Keywords: unstructured data, Text mining, Sentiment, Analysis, Natural language processing, Machine learning

Master Systèmes Intelligents & Réseaux

Département d'Informatique

Faculté des Sciences et Techniques de Fès

A.U. 2017 - 2018