RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ECHAHID HAMMA LAKHDAR - EL OUED

FACULTÉ DES SCIENCES EXACTES

Département D'Informatique

Mémoire de Fin D'étude

Présenté pour l'obtention du Diplôme de

MASTER ACADEMIQUE

Domaine : **Mathématique et Informatique**
Filière : **Informatique**
Spécialité : **Systèmes Distribués et Intelligence Artificielle**

Thème :

Machine Learning pour la détection des niveaux d'éducation sur les réseaux sociaux.

Présenté par :

GOURZI Sabrina

Soutenue lexx-xx- 2020 Devant le jury:

M.	……	MCA	Président
M.	……	MAA	Rapporteur
M.	Med charaf eddine meftah	MAA	Encadreur

Année universitaire : 2019/2020

Remerciements

On remercie Dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de, Mr. Med charaf eddine meftah, on le remercie pour la qualité de son encadrement exceptionnel, pour son patience, son rigueur et son disponibilité durant notre préparation de ce mémoire

Nos profonds remerciements vont également à tous les professeurs et chaque personne

qui nous a aidés dans nos recherches.

Dédicace

Louange à Dieu, Seigneur des mondes, et les prières et la paix soient sur le plus honorable des prophètes et messagers, notre noble prophète Muhammad.

Je dédie ce travail au soleil qui a illuminé le ciel de mon âme, la prunelle de mes yeux, et la lumière de ma vie est la chérie de mon cœur, celle qui m'a élevé et éclairé mon chemin et m'a aidé avec des prières et des invitations, à la personne la plus précieuse de cette existence, ma mère bien-aimée.

À l'âme de mon grand-père et de ma grand-mère Taher et Djannete

À l'esprit de mon père qui m'accompagne toujours

À ma petite mère, Ma compagne, et mon âme sœur, ma sœur, Fouzia

À mes filles, la vue de mes yeux, MaySs et DJannEt

Pour ma chère amie Saida

À mon professeur superviseur, je suis reconnaissant pour tous ses efforts, Professeur Muhammad Sharaf El Din Meftah

Sabrina

Résumé

Les réseaux sociaux sont un moyen important de communication sociale, et le data mining, en particulier le text mining, est très important pour extraire des informations sur les utilisateurs de réseaux sociaux. Le but de cette mémoire est de connaître le niveau d'éducation des utilisateurs des réseaux sociaux en appliquant des algorithmes d'apprentissage automatique supervisé, en utilisant le langage Python

Mots clés : réseaux sociaux, data mining, text mining, apprentissage automatique, apprentissage automatique supervisé

ABSTRACT

Social networks are an important means of social communication, and data mining, in particular text mining, is very important in extracting information on users of social networks. The goal of this note is to know the educational level of users of social networks by applying machine learning algorithms under supervision, using the Python language.

Keywords: social networks, data mining, text mining, machine learning, supervised machine learning.

ملخص

الشبكات الاجتماعية وسيلة مهمة للتواصل الاجتماعي كما ان التنقيب عن البيانات وبشكل خاص التنقيب عن النصوص يعد جد مهما في استخراج المعلومات عن مستخدمي الشبكات الاجتماعية والهدف من هذه المذكرة معرفة المستوى التعليمي لمستخدمي الشبكات الاجتماعية بتطبيق خوارزميات التعليم الآلي تحت اشراف وذلك باستخدام لغة البايثون.

الكلمات المفتاحية: الشبكات الاجتماعية، التنقيب عن البيانات، التنقيب عن النصوص، التعلم الآلي، التعلم الآلي تحت اشراف.

Liste Des Figures

Figure .01: apprentissage supervisé 13
Figure .02: methods k plus proches voisins 14
Figure .03: Illustration de la différence entre régression linéaire et classification linéaire 15
Figure .04: réseaux de neuronnes 16
Figure .05: classification binaire 17
figure .06: L'apprentissage non supervisé 21
Figure .7: La représentation noeuds-liens 32
Figure .08: La représentation matricielle 33
Figure .09: Text Mining 35
Figure .10: Le processus d'exploration de texte 36
Figure .11: La déclaration et l'importation des bibliothèques nécessaires 69
Figure .12: Importation des données 70
Figure .13: compte les mots 70
Figure .14: Word count list 70
Figure .15: Aprés l'ajout d'une colonne COUNT WORD. 71
Figure .16: tableau des vecteurs 71
Figure .17: Le résultat 71
Figure .18: table de caractéristiques 72
Figure .19: verification de correspendance 72
Figure .20: resultat de verification 73
Figure .21: resultat 73
Figure .22: insertion des resultat dans la table de dataset 74
Figure .23: resultat 75
Figure .24: entrainnement du programme 75
Figure .25: les correspondances avec des caractéristiques 76
Figure .26: resultat 76
Figure .27: partie d'application des algorithmes 77
Figure .28: resultat 77
Figure .29: SVM meilleur resultat 78
Figure .30 : Resultat 78
Figure .31 : Resultat 78

Liste des tableaux

Tableau 1 : Nombre de commentaires. 47
Tableau 2 : Nombre de commentaires. 47
Tableau 3 : Statistiques de notre dictionnaire. 48
Tableau 4 : dictionnaire des lycéens. 51
Tableau 5 : dictionnaire des universitaires 59
Tableau 6 : Analyse des resultats. 60

Sommaire

Remerciements ..... **i**

Dédicace ..... **ii**

Résumé ..... **iii**

ABSTRACT ..... **iv**

ملخص ..... **v**

Liste Des Figures vi

Liste des tableaux vii

Sommaire ..... **viii**

Introduction générale **1**

Chapitre I : Généralité sur l'apprentissage automatique

Introduction *5*

I.L' Intelligence artificiel *6*

1. Définition : 6

2. Domaines de l'intelligence artificielle : 6

2.1. Systèmes experts : 6

2.2. Calcul formel (opposé au calcul numérique) : 6

2.3. Représentation des connaissances : 7

2.4. Simulation du raisonnement humain : 7

2.5. Traitement du langage naturel : 7

2.6. Résolution de problèmes : 7

2.7. Reconnaissance de la parole : 7

2.8. Reconnaissance de l'écriture : 7

2.10 Reconnaissance des visages : 7

2.11 Robotique : 7

2.12. L'apprentissage : 8

2.13. Réseaux neuronaux : 8

II. L'apprentissage automatique et la classification *9*

1. L'apprentissage automatique 9

1.1. Définition 1 9

1.2. Définition 2 9

2. Origines de l'Apprentissage Automatique 9

3. Domaines de l'Apprentissage Automatique 9

4. Définitions et Particularités 10

4.1. Apprentissage 10
4.2. Adaptation 10
5. Dilemme de l'Apprentissage : 11
5.1. Précision *Vs* Généralisation 11
6. Intelligibilité 11
7. Définition de la classification : 11
8. Domaines d'application de la classification 12
9. Les différentes méthodes de l`apprentissage : 12
9.1. L`apprentissage supervisée 12
9.1.1. Définition 12
9.1.2. Les methods de classification supervisee 13
g)- Machines à Vecteurs de Support – Séparateurs à Vaste Marge SVM : 19
10. L'apprentissage non supervisé " le clustering" 21
a)-Définition 21
b)-Les approches différentes du clustering 21
b).1- Le clustering partionnel : 21
b).2- Le clustering hiérarchique 22
c)-Les algorithmes d'apprentissage non supervisé 22
11. L'apprentissage semi supervisé : 23
11.1. L'apprentissage semi-supervisé classique : 23
11.2. L'apprentissage semi-supervisé par Regroupement : 24
11.3. Les algorithmes d'apprentissage semi supervisé : 24
12. L'apprentissage par renforcement 24
a) -Définition : 24
b) -Les algorithmes d'apprentissage par renforcement : 24

Chapitre II : Analyse des reseaux sociaux numeriques et text mining (la fouille de textes)

Introduction 27
1. Analyse des réseaux sociaux 28
1.1. Les réseaux sociaux : 28
1.1.1– Définition des réseaux sociaux 28
1.1.2-La différenciation du réseau social et du média social 28
1.2-Les principales catégories de réseaux sociaux et leurs usages 29
A-Les médias sociaux à usage professionnel : 29

B-Les réseaux sociaux de partage de contenus : 29
C-les médias sociaux de « loisirs » : 29
D-les médias sociaux destinés au partage d'expression : 29
1.3. Les principaux réseaux sociaux plus utilisés : 29
1.4 La notion de « communauté virtuelle » 31
A-Définition : 31
B- Caractéristiques des communautés virtuelles : 31
1.5 Analyse des réseaux sociaux : 31
a)- Définition 31
b)- Les concepts d'analyse des réseaux sociaux 32
c)- Les représentations des réseaux sociaux 32
1.6 Fondements théoriques de l'analyse des réseaux sociaux 33
1.6.1 Méthodes d'analyse des réseaux sociaux 33
1.6.2. Méthodes traditionnelles 33
Objectifs et applications 36
Bag of words 37

Conclusion 44

Chapitre III : L'approche proposée : DataSet, Features, Apprentissage, Résultats et Analyses

1. Présentation 46
2. Contributions 46
2.1 DataSet : 46
A- Définition : 46
B- Source des commentaires 46
2.2 Les Caractériqtiques proposés : 47
A-Les profils des lycéens : 47
B-Les profils des universitaires : 48
2.3 Création des dictionnaires : 48
A-Dictionnaire des lycéens : 48
B-Dictionnaire des universitaires : 51
3. Expériences et résultats : 59
3.1 Analyse des résultats: 59
3.2 Résultats (analyse et discussion) 60
3.2.1 Analyse des résultats 60

3.2.2 Discussion sur les résultats: 62
3.3 Discussion sur les résultats : 63
Conclusion *64*

Chapitre IV : L'implémentation

Introduction *66*
1. Python *67*
2. A quoi sert Python ? *67*
2.1 Développement web 67
2.2 Scripting 67
2.3 Science des données (apprentissage automatique, analyse et présentation du Big Data) 67
3. Notebook Jupyter *67*
4. Editeur Spyder *67*
5. Les bibliothèques de Python : *68*
5.1. Bibliothèque Requests : 68
5.2. Bibliothèque Pluggy : 68
5.3. Bibliothèque SQLAlchemy : 68
5.4. Bibliothèque Numpy : 68
5.5. Bibliothèque SciPy : 68
5.6. Bibliothèque Skit-learn 68
5.7. Bibliothèque Pandas : 69
5.8. Bibliothèque CSV : 69
6. Matériel de configuration : *69*
7. Les étapes d'exécution : *69*
7.1 La déclaration et l'importation des bibliothèques nécessaires : 69
7.2 Importation des données (les commentaires et les dictionnaires) sont lus et répertoriés dans des tableaux pour être traités ultérieurement 70
7.1.3 On compte les mots dans chaque commentaire du Dataset .la séparation entre le mot et le mot est l'espace et la fonction Split () qui effectuent cette opération 70
7.1.4 Deuxièmement: 71
Conclusion générale **80**
Références bibliographiques **82**

Introduction générale

Un groupe de réseaux sociaux a commencé à apparaître à la fin des années 1990 pour se connecter avec des collègues c'est classmate.com, et sur ce site sont apparus des profils d'utilisateurs et un service de messagerie privé pour un groupe d'amis.

Depuis lors, les réseaux sociaux ont permis aux utilisateurs d'échanger des données grâce auxquelles ils peuvent obtenir des informations privées sur les utilisateurs concernant leurs sentiments et leurs opinions sur des sujets économiques, politiques et autres.

Mais les méthodes d'obtention et d'analyse de ces données peuvent poser un grand défi, car les données sont considérées comme l'une des ressources les plus importantes des médias sociaux, et la politique de chaque site diffère de l'autre en termes de mise à disposition des données pour les chercheurs.

L'exploration de données (data mining) est l'une des techniques disponibles pour analyser et structurer des informations à grande échelle au sein des réseaux sociaux

L'exploration de données est utilisée pour identifier les désirs des consommateurs et leur évaluation des produits, et pour identifier les positions politiques des citoyens et ainsi que l'exploration de données est utilisée à des fins de sécurité pour protéger la société contre les crimes, la violence et les maux sociaux. Le niveau d'éducation de groupe cible de citoyens et des gens joue un rôle fondamental dans tous les éléments mentionnés précédemment. Par conséquent, à travers ce travail, nous visons à utiliser des techniques d'apprentissage automatique pour classer les textes et commentaires utilisés dans les réseaux sociaux en fonction du niveau d'éducation du rédacteur du texte ou du commentateur.

Alors peut-on appliquer des algorithmes d'apprentissage automatique supervisé dans l'exploration de données pour déterminer le niveau d'éducation des utilisateurs dans les réseaux sociaux?

Pour cette raison, dans le premier chapitre, nous allons parlé de l'intelligence artificielle et de ses domaines, y compris l'apprentissage automatique, puis des types d'apprentissage automatique et de ses importants algorithmes et leurs domaines d'application

Dans le deuxième chapitre, nous allons parlé des réseaux sociaux, de leur définition des méthodes d'analyse, et du data mining, en particulier du text mining.

Dans le troisième chapitre, nous allons parlé de l'approche intelligente que nous appliquerons en collectant des données (dataset) et en créant des dictionnaires et nous avons appliqué quatre algorithmes à quatre classes de données équilibrées et non équilibrées et nous analysons les résultats. .

Enfin, en utilisant Python, nous allons appliqué 4 algorithmes d'apprentissage automatique supervisé pour obtenir le meilleur modèle pour déterminer le niveau d'éducation des utilisateurs dans les réseaux sociaux.

Chapitre I : Généralité sur l'apprentissage automatique

Introduction

L'intelligence artificielle (IA) est le domaine public qui couvre tout ce qui concerne l'acquisition de machines en tant qu '«intelligence», dans le but de mimer les capacités de pensée logique uniques de l'homme. L'apprentissage automatique représente une catégorie dans le domaine plus large de l'intelligence artificielle, et il est spécialisé dans la capacité des machines à "apprendre". Ceci est réalisé en utilisant des algorithmes qui peuvent détecter des modèles et générer des idées à partir des données qui leur sont présentées, pour les appliquer aux processus de prise de décision et aux prévisions futures, un processus qui évite d'avoir à programmer les étapes d'une manière spécifique à chaque action possible seule.

I.L' Intelligence artificiel

1. Définition :

L'I.A. est une science qui date d'une trentaine d'années. Son objet est de reconstituer à l'aide de moyens artificiels – presque toujours des ordinateurs – des raisonnements et des actions intelligentes. Les difficultés sont a priori de deux types :

1. Pour la plupart de nos activités nous ne savons pas nous-mêmes comment nous nous y prenons. Nous ne connaissons pas de m´méthode précise – pas d'algorithme disent aujourd'hui les informaticiens – pour comprendre un texte écrit ou reconnaitre un visage, démontrer un théorème, établir un plan d'action, résoudre un problème, apprendre...
2. Les ordinateurs sont a priori très loin d'un tel niveau de compétence. Il faut les programmer depuis le tout début. Les langages de programmation ne permettent, en effet, que d'exprimer des notions très élémentaires.

L'I.A. est, de ce double point de vue, une science expérimentale : expériences sur ordinateurs qui permettent de tester et d'affiner les modèles exprimés dans les programmes sur de nombreux exemples ; observations sur l'homme (en général le chercheur lui-même) pour découvrir ces modèles et mieux comprendre le fonctionnement de l'intelligence humaine.

2. Domaines de l'intelligence artificielle :

2.1. Systèmes experts :

Soit un logiciel capable de simuler le comportement d'un humain effectuant une tâche très précise. C'est un domaine où l'intelligence artificielle est incontestablement un succès, dû au caractère très précis de l'activité demandée à simuler.

2.2. Calcul formel (opposé au calcul numérique) :

Traiter les expressions symboliques. Des logiciels sur le marché, comme Mathematica, Maple, etc., effectuent tous des calculs formels.

2.3. Représentation des connaissances :

On entend la représentation symbolique de la connaissance pour que le logiciel soit capable de la manipuler. Un des secteurs de recherche en intelligence artificielle qui est le plus important.

2.4. Simulation du raisonnement humain :

Tenter de mettre au point des logiques qui formalisent le mode de raisonnement (logiques modales, floues, temporelles, etc.).

2.5. Traitement du langage naturel :

C'est la compréhension qui reste le problème majeur à la traduction ou au résumé d'un texte dans une autre langue. De grands progrès ont été faits pour obtenir une représentation sous une forme indépendante de la langue dans laquelle l'original est écrit. Certains traducteurs orientés dans des domaines spécialisés donnent de meilleurs résultats grâce à leurs spécificités.

2.6. Résolution de problèmes :

Représentation, analyse et résolution de problèmes concrets. C'est le cas des jeux de réflexion tels que les échecs, le backgammon ou les dames. Dans le cas du backgammon le champion du monde est un programme. Ils restent quelques jeux qui résistent aux efforts des programmeurs.

2.7. Reconnaissance de la parole :

Beaucoup de progrès ont été effectués. Un logiciel comme Naturaly Speaking permet la dictée. Cependant, la compréhension d'un mot ou d'une phrase requiert une grande quantité d'informations extralangagières (contexte, connaissance du sujet, etc.).

2.8. Reconnaissance de l'écriture :

La reconnaissance de l'écriture dactylographiée n'est pas facile (malgré des logiciels assez performants sur le marché), mais l'écriture manuscrite pose des problèmes beaucoup plus complexes. Cela n'est pas étonnant dans la mesure où cette tâche peut nous poser à nous aussi des problèmes insolubles. Certains chercheurs essaient de reconstituer à partir du texte le mouvement de la main afin de comprendre ce qui est écrit.

2.10 Reconnaissance des visages :

Considérée de longue date comme un des problèmes de l'intelligence artificielle le plus difficile, les résultats récents deviennent intéressants avec les réseaux neuronaux.

2.11 Robotique :

La robotique dans les usines est déjà fortement répendue. La première génération est capable d'exécuter une série de mouvements préenregistrés, la deuxième génération est dotée

de capteurs de perception permettant de prendre certaines décisions et la troisième génération possède une plus grande autonomie, elle peut se déplacer dans un environnement.

2.12. L'apprentissage :

Un logiciel devrait avoir des capacités d'apprentissage autonome pour pouvoir être véritablement qualifié d'intelligent. Douglas Lenat travaille actuellement à la constitution d'une gigantesque base de données censée contenir toutes les connaissances partagées par les humains d'un même groupe pour leur communication. Il souhaite adjoindre un module d'apprentissage à sa base de données lui permettant de travailler seule pour collecter des informations nouvelles et réorganiser l'architecture de ses connaissances.

2.13. Réseaux neuronaux :

Un réseau de neurones formels est un modèle rudimentaire du cerveau humain. Une cellule neuronale possède une sortie et des entrées reliées à d'autres neurones. Ces réseaux partagent des propriétés importantes avec le cerveau humain. Cela requiert une programmation non explicite et la répartition de l'information sur l'ensemble du réseau.[9]

II. L'apprentissage automatique et la classification

1. L'apprentissage automatique

1.1. Définition 1

L'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques [30].

1.2. Définition 2

L‟apprentissage, ou Learning en anglais, c‟est le processus de construire un modèle général à partir de données (observations) particulières du monde réel. Ainsi, le but est double :

- Prédire un comportement face à une nouvelle donnée.

- Approximer une fonction ou une densité de probabilité. Deux (02) branches d‟apprentissage existent :

- Apprentissage symbolique, issue de l‟IA.

- Apprentissage numérique, issue des statistiques.

Dans la pratique, le mot entraînement (Training, en anglais) est souvent synonyme de : apprentissage.

Ainsi, en science cognitive, l‟apprentissage est définit comme étant la capacité à améliorer les performances au fur et à mesure de l‟exercice d‟une activité [30].

2. Origines de l'Apprentissage Automatique

La discipline de l'apprentissage automatique (AA) possède de riches fondements théoriques. il est aisé de deviner que les fondements de cette discipline, il proviennent de diverses sciences :

- Des mathématiques pour l'informatique : algèbre linéaire, la probabilité, la logique, l'analyse élémentaire, …
- La théorie statistique de l'estimation,
- L'apprentissage Bayésien,
- L'inférence grammaticale ou l'apprentissage par renforcement, et tant d'autres.

3. Domaines de l'Apprentissage Automatique

Les principaux domaines d'applications de l'apprentissage automatique (AA) sont les fouilles de données et l'intelligence artificielle.

- La fouille de données (Data Mining, en anglais) est le processus d'extraction de la connaissance : il consiste à sélectionner les données à étudier à partir de bases de données (BDs) (hétérogènes ou homogènes), à épurer ces données et enfin à les utiliser en apprentissage pour construire un modèle.
- L'intelligence artificielle, la vision par ordinateur, la robotique, l'analyse et la compréhension des images, la reconnaissance de formes, reconnaître des objets dans les vidéo et extraire des contenus sémantiques des images sont autant d'applications qui requièrent la construction de modèles par apprentissage automatique.

4. Définitions et Particularités

Parmi les principales caractéristiques et facultés adoptées par les modèles d'apprentissage, nous citons : l'entrainement, la généralisation, l'adaptation, l'amélioration, l'intelligibilité et la prédiction.

4.1. Apprentissage

L'apprentissage, ou Learning en anglais, c'est le processus de construire un modèle général à partir de données (observations) particulières du monde réel. Ainsi, le but est double:

- Prédire un comportement face à une nouvelle donnée.
- Approximer une fonction ou une densité de probabilité.

Deux (02) branches d'apprentissage existent :

- Apprentissage symbolique, issue de l'IA.
- Apprentissage numérique, issue des statistiques. Dans la pratique, le mot entraînement (Training, en anglais) est souvent synonyme de : apprentissage. Ainsi, en science cognitive, l'apprentissage est définit comme étant la capacité à améliorer les performances au fur et à mesure de l'exercice d'une activité.

Exemple, un joueur de jeu d'échec : assimile (par expérience, s'entraîne) et raisonne (ceci lui procure une certaine intelligence ou puissance de raisonnement pour qu'il puisse progresser) c'est le cas pour un algorithme intelligent.

4.2. Adaptation

Elle peut être vue comme étant la disposition du modèle (algorithme ou système) à corriger son comportement ou à remanier sa réponse (ex., prédiction) par rapport à de nouvelles situations. Pour les tâches de perception, en vision artificielle, on accumule les bonnes et mauvaises expériences, et à partir d'elles, on peut faire évoluer les règles pour mieux effectuer la tâche, c'est le phénomène d'adaptation ou d'amélioration.

5. Dilemme de l'Apprentissage :

5.1. Précision *Vs* Généralisation

- Précision : c‟est l'écart entre une valeur mesurée ou prédite par le modèle et une valeur réelle.
- Généralisation : c‟est la capacité de reconnaître de nouveaux exemples jamais vus auparavant.

NB : Souvent, un seuil de généralisation est utilisé et est propre à chaque modèle pendant

l‟apprentissage.

6. Intelligibilité

C‟est amélioré la **compréhension des résultats d'apprentissage**, afin que le modèle puisse fournir une **connaissance claire** et **compréhensible**, au sens **interprétable** (en anglais, on parle de comprehensibility ou understandability) [30].

7. Définition de la classification :

La classification est la tâche consistant à attribuer une classe à une donnée qu'on veut classer, ou autrement dit, à assigner une donnée à une classe de données.

Plus formellement, soit $X \subseteq R$ d un ensemble représentant un espace à d dimensions, appelé l'espace des instances.

La donnée $x \in X$ est appelée une instance et représente un point dans l'espace X.

L'instance x est présentée sous forme d'un vecteur de taille d, x = (x (1), ..., x(d)), où chaque composante $x (i) \in R$ est une valeur discrète ou continue.

Soit Y un ensemble fini de classes où chaque classe $y \in Y$ est présentée sous forme d'une valeur discrète appelée étiquette de classe (un nom ou identifiant unique pour la classe).

Le classifieur se présente alors sous forme d'une fonction de classification h (appelée aussi modèle de classification) permettant d'associer une donnée $x \in X$ à une étiquette de classe $y \in Y$ (équation 2.1).

$$h: X \longrightarrow Y \quad x \longmapsto y = h(x) \tag{2.1}$$

Notez que pour le problème de classification, l'espace des réponses $Y \subset N$ est discret et fini, vu que chaque $y \in Y$ représente une classe.

Lorsque Y est continu (c.à.d. $Y \subset R$), on parle alors du problème de régression qui sert à estimer la relation entre une ou plusieurs variables $x (i) \in R$ et une autre variable $y \in R$. [26]

8. Domaines d'application de la classification

Les techniques de la machine learning sont utilisées dans de très nombreux domaines. A titre d'exemple, il peut s'agir de l'industrie des transports pour le développement d'un système de navigation sans conducteur, du secteur médical où les machines aident à diagnostiquer le cancer, ou bancaire où l'on cherche à estimer la capacité d'une personne à rembourser un prêt. L'utilisation de l'apprentissage automatique se répand rapidement dans le marketing digital. Les réseaux de neurones sont appliqués notamment pour la reconnaissance d'images, et le système immunitaire artificiel est largement utilisé dans le domaine de la détection de la fraude.[1]

9. Les différentes méthodes de l`apprentissage :

9.1. L`apprentissage supervisée

9.1.1. Définition

L'apprentissage supervisé (supervised learning en anglais) est une tâche d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples annotés, au contraire de l'apprentissage non supervisé. On distingue les problèmes de régression des problèmes de classement1. Ainsi, on considère que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification.

Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien généraliser, c'est-à-dire d'apprendre une fonction qui fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage. [3]

[1] "https://www.wizaly.fr/blog/post/machine-learning-quelles-applications"

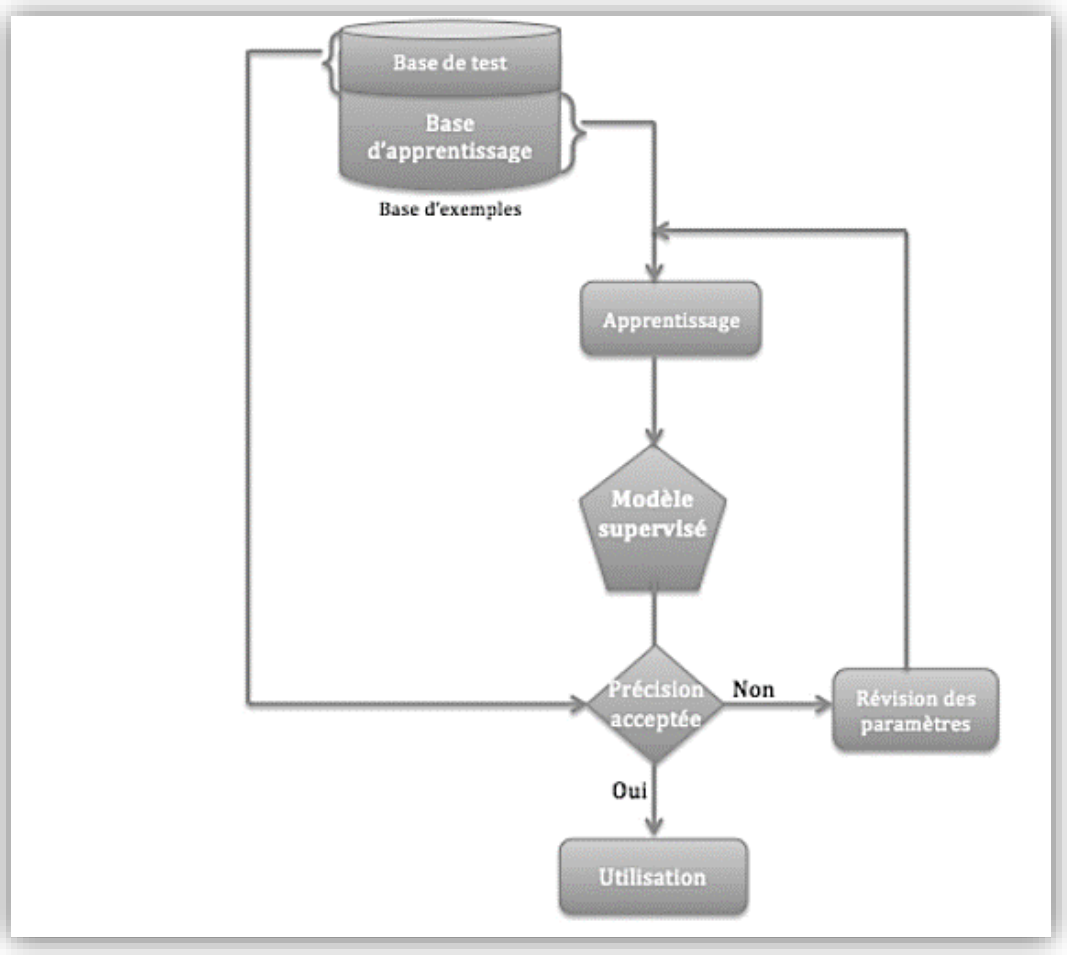


Figure .01: apprentissage supervisé

9.1.2. Les methods de classification supervisee

Il existe de nombreuses méthodes de classification supervisée :

a) k plus proches voisins
b) analyse (factorielle) discriminante.
c) régression logistique.
d) réseaux de neuronnes.
e) arbres de décision
f) forêts aléatoires (random forest)
g) SVM
h) NB : Naïve Bayes

a)- k plus proches voisins

La méthode des k plus proches voisins (kNN pour k nearest neighbour) est très simple :

I pour un nouvel x*, il suffit de rechercher les k voisins les plus proches, au sens d'une certaine distance (euclidienne par exemple)

I on compte le nombre nc de ces voisins appartenant à la classe c

I on estime alors la probabilité que y = c par :

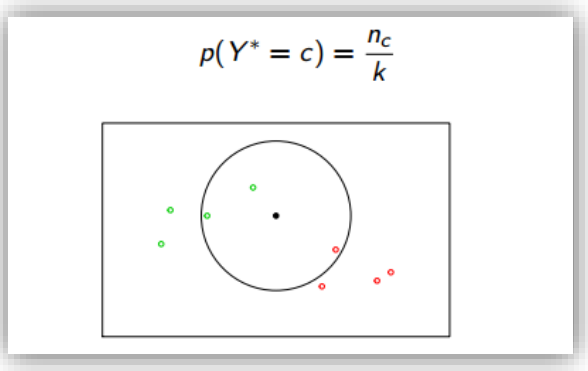


Figure .02: methods k plus proches voisins

Remarque :

- le choix de k se fera par validation croisée
- plus n est grand, plus on peut se permettre de prendre un k grand
- cette méthode est très simple à mettre en oeuvre mais pas toujours des plus efficaces. En particulier en grande dimension, où l'espace est généralement vide (les voisins sont tous très éloignés), cette méthode n'est que peu efficace.[1]

b)-analyse (factorielle) discriminante

L'analyse factorielle discriminante (AFD) ou simplement analyse discriminante est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire…) d'un ensemble d'observations (individus, exemples…) à partir d'une série de variables prédictives (descripteurs, variables exogènes…).

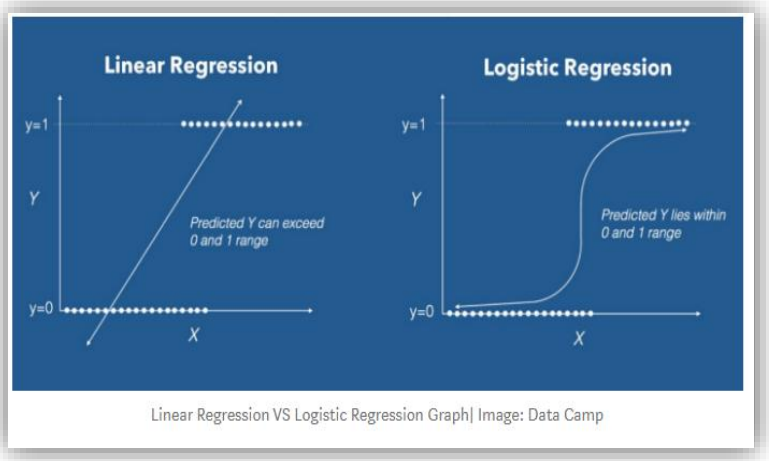
On distingue généralement deux grandes approches :

- L'analyse discriminante peut être une technique descriptive. On parle dans ce cas d'analyse factorielle discriminante (ou analyse discriminante descriptive
- L'analyse discriminante peut être prédictive. Il s'agit dans ce cas de construire une fonction de classement (règle d'affectation, …) qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives [7]

[1] I "PLS-DA Classification Supervisée Julien JACQUES 26/09/2018"

c)- régression logistique

La régression logistique est un algorithme d'apprentissage automatique qui est utilisé pour les problèmes de classification, c'est un algorithme d'analyse prédictive et basé sur le concept de probabilité.



Linear Regression VS Logistic Regression Graph| Image: Data Camp

Figure .03: Illustration de la différence entre régression linéaire et classification linéaire

L'hypothèse de régression logistique tend à limiter la fonction de coût entre 0 et 1. Par conséquent, les fonctions linéaires ne la représentent pas car elles peuvent avoir une valeur supérieure à 1 ou inférieure à 0 ce qui n'est pas possible selon l'hypothèse de régression logistique.[1]

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

[1] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

d)- réseaux de neuronnes

Cette méthode repose sur la notion de neurone formel. Un neurone formel est un modèle caractérisé par des signaux d'entrée (les variables explicatives par exemple) , une fonction d'activation f ,

$f(\alpha 0+\sum i\ \alpha i \times xi)$

f peut être linéaire, à seuil, stochastique et le plus souvent sigmoïde. Le calcul des paramètres se fait par apprentissage.

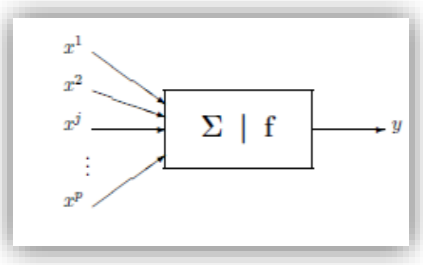


Figure .04: réseaux de neuronnes

Les neurones sont ensuite associés en couche. Une couche d'entrée lit les signaux entrant, un neurone par entrée xj , une couche en sortie fournit la réponse du système. Une ou plusieurs

couches cachées participent au transfert. Un neurone d'une couche cachée est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche suivante.

De façon usuelle et en régression (Y quantitative), la dernière couche est constituée d'un seul neurone muni de la fonction d'activation identité tandis que les autres neurones (couche cachée) sont munis de la fonction sigmoïde.

En classification binaire, le neurone de sortie est muni également de la fonction sigmoïde tandis que dans le cas d'une discrimination à m classes (Y qualitative), ce sont m neurones avec fonction sigmoïde, un par classe, qui sont considérés en sortie.

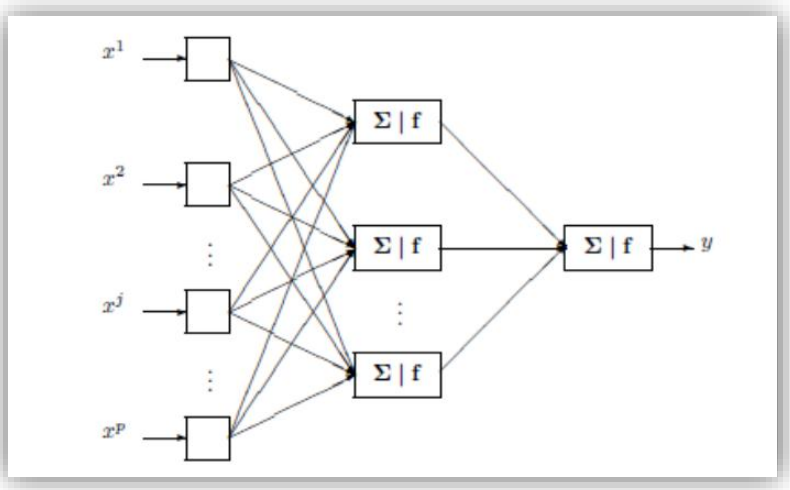


Figure .05: classification binaire

E)- arbres de décision

Algorithme récursif de découpage de l'espace de représentation. Découpage forcément parallèle aux axes. Modèle linéaire par morceaux, non-linéaire au final.

Un des enjeux de la construction des arbres est de détecter la profondeur « optimale » correspondant au meilleur arbitrage entre biais et variance.[1]

Avantage :

- connaissances « intelligibles » -- validation d'expert (si arbre pas trop grand)
- traduction directe de l'arbre vers une base de règles
- sélection automatique des variables pertinentes, robuste face aux variables redondantes
- non paramétrique
- traitement indifférencié des différents types de variables prédictives
- robuste face aux données aberrantes, solutions pour les données manquantes
- rapidité et capacité à traiter des très grandes bases
- enrichir l'interprétation des règles à l'aide des variables non sélectionnées
- possibilité pour le praticien d'intervenir dans la construction de l'arbre

[1]http://www.unitheque.com/Livre/springer/Regression_avec_R-37175.html
http://cedric.cnam.fr/~saporta/multicolinearite.pdf
http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/tp_reg-penal_ukcomp.pdf
http://apiacoa.org/publications/teaching/data-mining/m2p6/supervised-slides.pdf

Inconvénients :

- problème de stabilité sur les petites bases de données (feuilles à très petits effectifs)
- recherche « pas-à-pas » : difficulté à trouver certaines interactions (ex. xor)
- peu adapté au « scoring »
- performances moins bonnes en général par rapport aux autres méthodes (en réalité, performances fortement dépendantes de la taille de la base d'apprentissage)

f)- forêts aléatoires (random forest)

Les forêts d'arbres décisionnels (ou forêts aléatoires de l'anglais random forest classifier) ont été premièrement proposées par Ho en 19952 et ont été formellement proposées en 2001 par Leo Breiman3 et Adèle Cutler4. Elles font partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de bagging. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.

Algorithme :

La base du calcul repose sur l'apprentissage par arbre de décision. La proposition de Breiman3 vise à corriger plusieurs inconvénients connus de la méthode initiale, comme la sensibilité des arbres uniques à l'ordre des prédicteurs, en calculant un ensemble de {\displaystyle {B}} arbres partiellement indépendants.

Une présentation rapide de la proposition peut s'exprimer comme suit :

Créer {\displaystyle {B}} nouveaux ensembles d'apprentissage par un double processus d'échantillonnage :

1. Sur les observations, en utilisant un tirage avec remise d'un nombre {\displaystyle {N}} d'observations identique à celui des données d'origine (technique connue sous le nom de bootstrap), et sur le {\displaystyle {p}} prédicteurs, en n'en retenant qu'un échantillon de cardinal {\displaystyle m< {\sqrt {p}}} (la limite n'est qu'indicative).
2. Sur chaque échantillon, on entraîne un arbre de décision selon une des techniques connues, en limitant sa croissance par validation croisée.
3. On stocke le {\displaystyle {B}} prédictions de la variable d'intérêt pour chaque observation d'origine.
4. La prédiction de la forêt aléatoire
5. est alors un simple vote majoritaire (Ensemble learning).

Le principal revers de cette méthode est que l'on perd l'aspect visuel des arbres de décision uniques.

g)- Machines à Vecteurs de Support – Séparateurs à Vaste Marge SVM :

Est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs dans l'ensemble des exemples. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Cela garantit une généralisation du principe car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être situés d'un coté ou l'autre de la frontière. L'intérêt de cette méthode est la sélection de vecteurs supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode.

Classifieur linéaire :

Un classifieur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire en. On peut exprimer une telle fonction par :

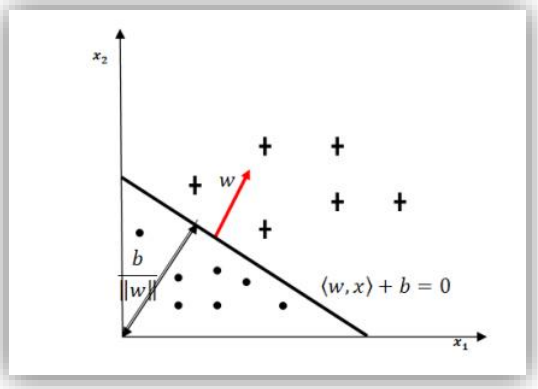
$$h(x) = \langle w, x \rangle + b = \sum_{i=1}^{n} w_i x_i + b$$

Où w ($\in R^n$)est le vecteur de poids et b($\in R^0$) le biais, alors que x est la variable du problème. X est l'espace d'entrée et qui correspond à R^n , où n est le nombre de composantes des vecteurs contenant les données. Notons que l'opérateur < >désigne le produit scalaire usuel dans R^n .w et b sont les paramètres à estimer de la fonction de décision h(x) .

Pour décider à quelle catégorie un exemple estimé x1 appartient, il suffit de prendre le signe de la fonction de décision :

Y = sign(h(x1))

La fonction sign () est appelée classifieur.

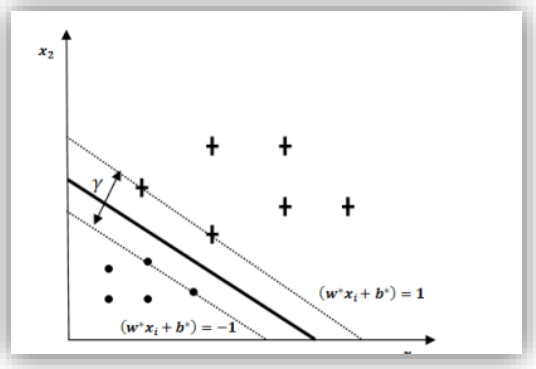


Hyperplan séparateur <w,x> +b= 0

Marge maximale de l'hyperplan :

La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge.

La marge géométrique représente la distance euclidienne prise perpendiculairement entre l'hyperplan et l'exemplex_i . En prenant un point quelconque x_p se trouvant sur l'hyperplan, la marge géométrique peut s'exprimer par :



$$\frac{w}{\|w\|} \cdot (x_i - x_p)$$

Hyperplans canoniques et marge maximale

10. L'apprentissage non supervisé " le clustering"

a)-Définition

Pour ce type d'apprentissage la base de données d'apprentissage ne contient pas de variable cible (comme on l'a vu en apprentissage supervisé). Il y a seulement un ensemble de données collectées en entrée. L'algorithme doit découvrir par lui-même la structure en fonction des données. On utilise cette technique pour partitionner les données en groupes d'éléments homogènes. La distance est souvent la plus utilisée comme mesure de similarité entre les groupes [4].

Former des groupes homogènes é l'intérieur d'une population

Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux, trouver des groupes (classes, segments, clusters) tels que :

Les points ´l'intérieur d'un même groupe sont très similaires entre eux.

Les points appartenant é des groupes différents sont très dissimilaires.

Le choix de la mesure de similarité est important. [6]

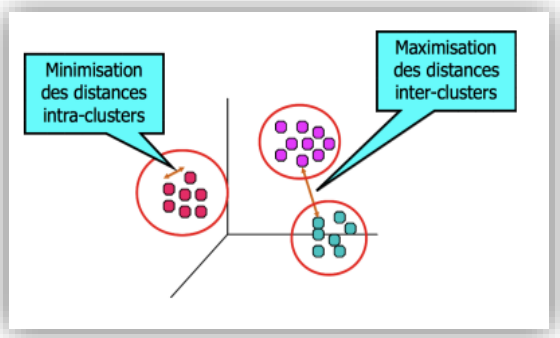


figure .06: L'apprentissage non supervisé

b)-Les approches différentes du clustering

Il existe deux approches différentes dans le clustering

- Le clustering partionnel (non-hiérarchique)
- Le clustering hiérarchique.

b).1- Le clustering partionnel :

vise à obtenir directement une seule partition de la collection d'éléments en clusters (la division d'un ensemble de données en K groupes) jusqu'à obtenir une similarité satisfaisante, qui possèdent les propriétés suivantes : – Homogénéité dans les groupes (intra-classe) : les données appartenant à un même cluster doivent être les plus similaires possibles.

Hétérogénéité entre groupe (inter-classe) : les données appartenant à différents clusters doivent être les plus dissemblables possibles.

L'inertie inter-cluster mesure "l'éloignement" des centres des clusters entre eux. Plus cette inertie est grande, plus les clusters sont bien séparés. Il faut minimiser l'inertie intra-cluster et maximiser l'inertie inter-cluster. Bien qu'elles permettent de classifier un ensemble volumineux de données, mais il reste qu'il faut fixer au départ le nombre de classes : choix du nombre de k clusters.

b).2- Le clustering hiérarchique

Vise à obtenir une hiérarchie de clusters, appelée dendrogramme, qui montre comment les clusters sont liées les uns aux autres. Ces méthodes procèdent soit par :

– Les méthodes ascendantes : (Agglomérations successives) Un algorithme de clustering hiérarchique est fondé sur l'union entre les deux plus proches clusters c'est-à-dire : consiste à trouver des clusters successifs utilisant des clusters précédemment établis.

La première condition est de mettre, au début, chaque objet dans un cluster distinct et les fusionner en clusters successivement plus grand.

– Les méthodes descendantes : Divisions successives en deux groupes différents.

Au départ tous les individus sont dans le même groupe.

A chaque étape, un groupe est séparé en deux.

Il faut un critère de séparation. Après quelques itérations on atteint le Cluster final voulu qui regroupe tous les sous-clusters (sous-partitions)[1].

c)-Les algorithmes d'apprentissage non supervisé

- K-means clustering (K-moyenne)
- Dimensionality Reduction (Réduction de la dimensionnalité)
- Neural networks (Réseaux de neurones) / Deep Learning
- Principal Component Analysis (Analyse des composants principaux)
- Singular Value Decomposition (Décomposition en valeur singulière)
- Independent Component Analysis (Analyse en composantes indépendantes)
- Distribution models (Modèles de distribution)
- Hierarchical clustering (Classification hiérarchique) [2]

[1] - MÉMOIRE DE FIN D'ÉTUDES pour obtenir le grade de MASTER EN GÉNIE BIOMÉDICAL Spécialité : Informatique Biomédicale présenté et soutenu publiquement par Lakhdari Salsabil & Saidi Amari

[2] -https://fr.wikipedia.org/wiki/Apprentissage_non_supervis%C3%A9

11. L'apprentissage semi supervisé :

L'apprentissage semi-supervisé est un bon compromis entre les deux types d'apprentissage « supervisé » et « non-supervisé », car il permet de traiter un grand nombre de données sans avoir besoin de toutes les étiqueter, et il profite des avantages des deux types mentionnés.

D'autre part la labellisation a priori de toutes les données nécessite l'intervention d'un expert humain.

C'est une opération difficile voire fastidieuse lorsque le nombre de données est important. Dans des applications concrètes, il est souvent impossible que l'expert puisse assigner toutes les données d'apprentissage aux classes en présence.

Le contexte semi-supervisé qui se situe à l'intersection entre le contexte supervisé et le contexte non supervisé, est alors une solution alternative [2].

Il se caractérise par la présence de quelques informations disponibles sur l'ensemble des données. Ces informations sont représentées soit sous la forme de quelques données labellisées, soit sous la forme de ressemblance ou dissemblance au sein de couples de données. Le contexte semi-supervisé utilise des connaissances partielles qui sont soit incomplètes (par exemple, le cas où des relations entre certains individus sont connues) ou tout simplement les exemples étiquetés ne sont pas en quantité suffisante pour que l'on puisse appliquer des algorithmes supervisés. Parmi les nombreuses méthodes d'apprentissage récemment apparues, de nouvelles méthodes de classification connaissent des succès importants sont les méthodes semi-supervisées, où l'on dispose à la fois d'un (petit) ensemble de données étiquetées, et d'un (grand) ensemble de données non-étiquetées [27].

Les approches d'apprentissage semi-supervisé peuvent être globalement divisés en deux catégories :

- L'apprentissage semi- supervisé classique
- L'apprentissage par Regroupement.

11.1. L'apprentissage semi-supervisé classique :

L'ensemble d'apprentissage est constitué de données libellées et non libellées mais la tâche d'apprentissage qui est essentiellement supervisée. La classification est effectuée premièrement par un apprentissage avec des données libellées, en second lieu, un renforcement de l'apprentissage est réalisé avec des données nonlibellées. L'objectif est le même que pour l'apprentissage supervisé mais en tirant profit des observations non labellisée.

11.2. L'apprentissage semi-supervisé par Regroupement :

L'apprentissage semi- supervisée par regroupement ou Semi-Supervised Clustering (SSC) en anglais, est une extension de la classification non-supervisée. Le Clustering semi-supervisé introduire des connaissances expertes partielles dans un algorithme du clustering relève du domaine du clustering semi-supervisé. [24]

11.3. Les algorithmes d'apprentissage semi supervisé :

- Self-Training
- Generative methods, mixture models
- Graph-based methods
- Co-Training
- Semi-supervised SVM

12. L'apprentissage par renforcement

a) -Définition :

En intelligence artificielle, plus précisément en apprentissage automatique, l'apprentissage par renforcement consiste, pour un agent autonome (robot, etc.), à apprendre les actions à prendre, à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps. L'agent est plongé au sein d'un environnement, et prend ses décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers d'expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal, en ce sens qu'il maximise la somme des récompenses au cours du temps.

b) -Les algorithmes d'apprentissage par renforcement :

Apprentissage par Renforcement (Reinforcement learning -RL-) est un domaine de l'apprentissage machine qui s'intéresse à la façon dont les agents logiciels devraient prendre des mesures dans un environnement de manière à maximiser une certaine notion de récompense cumulative.

L'apprentissage par renforcement est l'un des trois paradigmes fondamentaux de l'apprentissage automatique, aux côtés de l'apprentissage supervisé et de l'apprentissage non supervisé.

Il diffère de l'apprentissage supervisé en ce qu'il n'est pas nécessaire de présenter les paires étiquetées entrée / sortie et que les actions sous-optimales ne doivent pas être explicitement corrigées. Au lieu de cela, l'objectif est de trouver un équilibre entre l'exploration (d'un territoire inexploré) et l'exploitation (du savoir actuel).

L'environnement est généralement formulé comme un processus de décision de Markov (MDP), car de nombreux algorithmes d'apprentissage par renforcement pour ce contexte utilisent des techniques de programmation dynamiques. La principale différence entre les méthodes classiques de programmation dynamique et les algorithmes d'apprentissage par renforcement réside dans le fait que ces derniers ne supposent pas la connaissance d'un modèle mathématique exact du PDM et s'adressent à de grands PDM où les méthodes exactes deviennent infaisables.[14]

Chapitre II : Analyse des reseaux sociaux numeriques et text mining (la fouille de textes)

Introduction

Ce chapitre analyse la technologie des réseaux sociaux numériques. Il présente leurs spécificités dans le cadre de la pratique curation de contenu. Il propose un état de l'art des théories à disposition pour évaluer l'utilité des fonctionnalités des réseaux sociaux dans le cadre d'un usage de curation de contenu.

Et nous allons tenter d'appliquer les outils linguistiques à mon disposition à la "fouille de textes". L'enjeu de ce domaine est la manipulation rapide et efficace de grandes quantités de textes, pour lesquels il n'est pas envisageable de procéder à une analyse syntaxico-sémantique complète.

.

1. Analyse des réseaux sociaux

1.1. Les réseaux sociaux :

1.1.1– Définition des réseaux sociaux

Le concept de « social » a été inventé en 1954 par un anthropologue du nom de. John A. Barnes Le principe de réseau se définit par deux éléments : les contacts et les liaisons entre les contacts. Plus nous avons de contacts plus notre réseau est important et donc plus nous sommes utiles (la notion d'utilité ici se résume à la capacité à transmettre des informations). [29]

D'autre part, les réseaux sociaux sont considérés comme des services web qui permettent aux individus de construire un profil public ou semi-public liée avec une liste d'autres profils qui nécessite une confirmation bidirectionnelle pour l'amitié, mais parfois unidirectionnels comme fans ou abonnés. [5]

1.1.2-La différenciation du réseau social et du média social

A.Le média peut être défini comme « un procédé permettant la distribution, la diffusion ou la communication d'œuvres, de documents, ou de messages sonores ou audiovisuels (presse, cinéma, affiche, radiodiffusion, télédiffusion, vidéographie, télédistribution, télématique, télécommunication) »

B.Les réseaux sociaux sont des sites web ouverts à n'importe qui par le biais d'une inscription gratuite. Ils mettent à la disposition de leurs membres des fonctionnalités de partage de contenu, leur permettant de tisser des liens . Au début du 21e siècle, les réseaux Friendster, MySpace, et Friends Reunited sont ainsi mis en place.

C.Les termes de média social et de réseau social se différencient en le sens que 'réseau social' peut se définir comme n'importe quelle plateforme permettant la mise en place d'un réseau ; lorsque la désignation de 'média social' est beaucoup plus large et s'étend à l'ensemble des plateformes auxquelles l'internaute peut participer en créant du contenu. La force de la communauté n'est pas forcément au centre, en termes de business model, les réseaux sociaux s'identifient comme tels lorsqu'il constitue le modèle de base du business, permettant à la plateforme de survivre. Nous pouvons ainsi nous poser la question : sans les membres du réseau de chaque participant, la plateforme existerait-elle ? Dans le cas négatif, il s'agit simplement la plateforme de média social, il faut considérer le média social comme un outil au centre des usages quotidiens régissant le fonctionnement de notre société actuelle tout en participant à l'amélioration, ou du moins à l'évolution, des liens sociaux. Mais le terme de réseau social reste trop large. Car plusieurs types de réseaux sociaux sont disponibles, et chacun d'eux utilisés pour faire de la curation génère des résultats très différents.

1.2-Les principales catégories de réseaux sociaux et leurs usages

Ils sont devenus incontournables : ce sont les réseaux sociaux :

A-Les médias sociaux à usage professionnel :

Viadéo, Linkedin, Xing… déstinés à la mise en relation professionnelle mais également Digg, Reddit… qui permettent le partage de liens. Sans oublier bien sûr Slideshare ou Scribd pour le partage de documents

B-Les réseaux sociaux de partage de contenus :

Ils intègrent les sites de partage de liens et de documents cités ci-dessus mais également les sites de partage de vidéos comme Youtube et Dailymotion ainsi que ceux de partage de photographies comme Picasa et Flickr. Il est a noter que si les sites de partage de vidéos et de photographies sont largement utilisés par les entreprises, par les professionnels, ils ont également conquis la sphère des loisirs et touchent par conséquent le grand public

C-les médias sociaux de « loisirs » :

Ce sont les plate-formes sociales comme l'incontournable facebook, mais également Copains d'avant. L'usage concerne principalement « l'individu » mais également le « professionnel » qui peut se cacher derrière l'individu.

Ainsi, Facebook propose la création de pages destinées aux entreprises, organisations, associations… Dans cette catégorie s'intégrent également les réseaux sociaux de joueurs comme Habbo ou Zynga utilisés par des profils personnels

D-les médias sociaux destinés au partage d'expression :

Ils intègrent les plate-formes sociales évoquées dans le paragraphe précédent mais également les forums, les blogs, les microblogs comme twitter, les wikis…

1.3. Les principaux réseaux sociaux plus utilisés :

Facebook : Considéré comme synonyme de «réseaux sociaux» par certains, Facebook est le seul site où vous êtes susceptible de trouver des amis, des collègues et des parents autour de vous tellement la plateforme est devenu populaire. Bien que Facebook soit principalement axé sur le partage de photos, de liens et de la vie quotidienne, les utilisateurs peuvent également montrer leur soutien aux marques ou aux organisations en devenant des fans. [15]

Twitter : Peut-être la plus simple de toutes les plateformes de médias sociaux, Twitter se juste être l'un des plus fun et intéressants à utiliser. Les messages sont limités à 140 caractères ou moins, mais c'est plus que suffisant pour afficher un lien, partager une image ou même échanger des idées avec votre célébrité préférée ou un influenceur. La création d'un nouveau profil ne prend que quelques minutes et l'interface de Twitter est facile à prendre en main. [11]

Google+ : En combinant le meilleur de Facebook et Twitter dans une seule plateforme, Google a donné aux utilisateurs un réseau social qui a un petit quelque chose de tout le monde. Vous pouvez ajouter du nouveau contenu, mettre en surbrillance des sujets avec des hashtags et même séparer des contacts en cercles. Même si un profil G + ne prend que quelques minutes pour être configuré, le succès de Google + n'est surement pas ce qui était espéré par la firme de MountainView. [11]

Snapchat : Il s'agit d'une plate-forme sociale de de partage photos et de vidéos disponible sur smartphone et qui permet de discuter avec des amis en utilisant des images. La particularité de ce réseau social est que chaque image ou vidéo envoyée ne peut être visible que durant une période de temps par son destinataire. Il vous permet d'explorer les nouvelles et même de consulter des histoires en direct qui se passent dans le monde.

YouTube : En tant que service de partage de vidéos, YouTube est devenu si populaire que son catalogue de milliards et de milliards de vidéos est devenu le «deuxième moteur de recherche mondial». Les utilisateurs peuvent partager, évaluer et commenter toutes les vidéos qu'ils visionnent La plateforme qui a été rachetée par Google et faisant parti à part entière de son écosystème est le leader incontestable des réseaux sociaux de vidéo.

Instagram : Instagram permet de partager ses photo graphies ou vidéos prises avec son smartphone avec ses amis sur la plateforme ou via les autres réseaux sociaux notamment Facebook qui l'a racheté. Vous pouvez aussi choisir parmi une variété de filtres photo et inviter des amis à les commenter.

1.4 La notion de « communauté virtuelle »

A-Définition :

Le terme de « communauté virtuelle » est apparu pour la première fois dans l'ouvrage de Howard Rheingold en 1995.

Les communautés virtuelles sont des regroupements socioculturels qui émergent du réseau lorsqu'un nombre suffisant d'individus participent à ces discussions publiques pendant assez de temps en y mettant suffisamment de cœur pour que des réseaux de relations humaines se tissent au sein du cyberespace.

B- Caractéristiques des communautés virtuelles :

Preece et Maloney-Krishmar (2003) proposent cinq caractéristiques inspirées de Whittaker, Isaacs et O'Day (1997) :

1. Les membres ont un but, un intérêt, une activité ou un besoin commun qui constitue la raison principale de leur appartenance à la communauté ;
2. Les membres s'engagent dans une participation répétée, active et il y a souvent des interactions intenses, des liens émotionnels forts et des activités communes entre les participants ;
3. Les membres ont accès à des ressources partagées et des règles déterminent les modalités d'accès à ces ressources ;
4. La réciprocité des informations, des soutiens et des services entre les participants est importante ;
5. Il y a un contexte de conventions sociales, d'utilisation du langage et de façons de se comporter. [8]

1.5 Analyse des réseaux sociaux :

a)- Définition

Analyse des réseaux sociaux (ARS) est la cartographie et la mesure des relations et des flux entre personnes, groupes, organismes, ordinateurs, sites web ou toute autre entité de traitement d'information ou de connaissances. Les nœuds dans le réseau sont les personnes ou les groupes tandis que les liens montrent les relations ou les flux entre les nœuds. L'analyse des réseaux sociaux permet une analyse visuelle et mathématique des relations humaines.

b)- Les concepts d'analyse des réseaux sociaux

Selon Wassermann et Faust, trois concepts sont également retenus dans cette analyse des réseaux sociaux :

1. Les acteurs et leurs actions sont considérés comme des entités indépendantes.
2. L'environnement des acteurs procure des opportunités et exerce des contraintes sur leurs actions individuelles.
3. Les structures sociales, politiques, économiques, …etc. ont une influence sur les formes de relations entre les acteurs. [28]

c)- Les représentations des réseaux sociaux

❖ **La représentation noeuds-liens**

Au milieu du vingtième siècle, Cartwright et Harary sont les premiers à avoir appliqué la théorie des graphes à l'analyse des réseaux sociaux. Le graphe est devenu par la suite la représentation adoptée par toutes les sciences manipulant l'analyse des réseaux sociaux, dont la sociologie, les mathématiques et l'informatique.

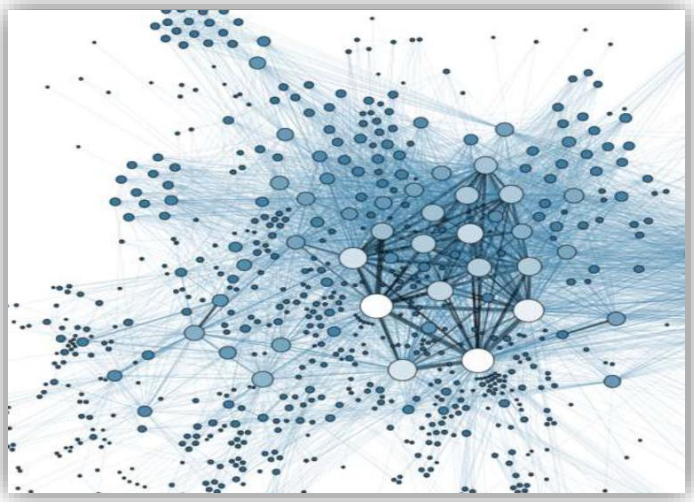


Figure .7: La représentation noeuds-liens

❖ **La représentation matricielle**

La représentation matricielle des graphes est une représentation beaucoup moins utilisée que les diagrammes noeuds-liens. Un graphe peut être représenté sous la forme d'une matrice A à n lignes et n colonnes représentant un tableau. Chaque case de ce tableau est notée aij avec i et j les numéros respectifs de ligne et de colonne de la case.

S	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	0	1	1	1	0	0	0	0	0
v2	1	0	1	0	0	0	0	0	0
v3	1	1	0	1	0	0	0	0	0
v4	1	0	1	0	1	1	0	0	0
v5	0	0	0	1	0	1	1	1	0
v6	0	0	0	1	1	0	1	1	0
v7	0	0	0	0	1	1	0	1	1
v8	0	0	0	0	1	1	1	0	0
v9	0	0	0	0	0	0	1	0	0

Figure .08: La représentation matricielle

1.6 Fondements théoriques de l'analyse des réseaux sociaux

La paternité de l'analyse des réseaux sociaux en tant que théorie à part entière est attribuée à Simmel (1917) qui la définit comme le fondement même de la sociologie, science des structures des relations sociales (Forsé, 2002). Vandenberghe (2001) complète cette analyse en parlant de sociologie formaliste (axée sur la forme des relations et non sur leur contenu) et dualiste (appréhendant les phénomènes sociaux comme issus des interactions individuelles mais également comme influençant et conditionnant la nature de ces interactions). Enfin, l'analyse des réseaux sociaux a puisé dans les théories mathématiques que sont celles des graphes et des matrices le pouvoir calculatoire et de représentation qui lui faisait défaut jusque là. [1]

1.6.1 Méthodes d'analyse des réseaux sociaux

On distingue alors deux grandes familles : traditionnelle et fouille de données

1.6.2. Méthodes traditionnelles

La centralité : Les acteurs importants sont ceux qui sont liés et impliqués avec les autres acteurs d'une façon extensive. Après avoir modélisé les contacts par des liens dans un réseau, nous pouvons définir un acteur central comme étant celui qui est impliqué dans plusieurs liens**.**

Il existe différents types de liens entre les noeuds (acteur), ce qui a permet de générer plusieurs types de centralités dont :

Degré de Centralité : Soit n le nombre total de noeud (acteurs) dans un réseau (graphe). Le degré de centralité d'un acteur i noté CD(i) est le degré du noeud acteur (le nombre d'arêtes) noté d(i) normalisé par le degré maximal n-1

$$CD(i)=\frac{d(i)}{(n-2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots 2.3$$

La centralité de proximité(closenesscentrality) indique si le sommet est situé à proximité de l'ensemble des sommets du graphe et s'il peut rapidement interagir avec

$$C_c(v) = \frac{1}{\sum_{u \in V \setminus \{v\}} d_G(u,v)} \ldots\ldots\ldots\ldots 2.4$$

Ces sommets. Il s'écrit formellement :

avec dG (u,v) la distance entre les sommets u et v.[15]

La centralité d'intermédiarité (betweenesscentrality) : est un des concepts les plus importants. Il mesure l'utilité du sommet dans la transmission de l'information au sein duréseau. Le sommet joue un rôle central si beaucoup de plus courts chemins entre deux sommets doivent emprunter ce sommet. Elle s'écrit :

$$C_B(v) = \sum_{\substack{i,j \\ i \neq j \neq v}} \frac{\sigma_{ij(v)}}{\sigma_{ij}} \ldots\ldots\ldots 2.5$$

Avec σij(v) le nombre de chemins entre i et j qui passent par v.

Le prestige: La notion de prestige est une autre manière de mesurer l'importance d'un noeud.

Un noeud prestigieux est un noeud à lequel un grand nombre d'autres noeud se lient- il reçoit un grand nombre de liens entrants. Nous distinguons les mesures suivantes :

Le degré de prestige : avec dI(i) le degré entrant du noeud i, n le nombre noeuds dans le réseau, le degré de prestige est donné par la relation :

$$\text{PD(i)} = \frac{dI(i)}{(n-1)} \ldots\ldots\ldots\ldots 2.6$$

Le prestige de tri : Les mesures proposées jusqu'au là sont fondées sur les liens entrants et sortants d'un acteur donné. La mesure du prestige du tri considère la réputation et I importance des acteurs choisissant l'acteur i. Ce type d'algorithme est utilisé pour trier les résultats de recherche comme Google (algorithme Rank Page). [23]

Fouille de données dans les réseaux sociaux: Les réseaux sociaux amassent de très nombreuses données : les amis, les messages, les images, la fréquence d'utilisation...etc. Dès lors se pose le problème de l'exploitation de cette masse d'informations. La fouille de données s'avère être un outil incroyablement riche et puissant lorsqu'il est appliqué aux réseaux

sociaux, puisque cette méthode nous permet de modéliser le réseau sous forme mathématique, l'analyser pour tirer le maximum d'information mais aussi prédire en partie l'évolution future du réseau.

1. Définition Data Mining

Le *Data Mining* est en fait un terme générique englobant toute une famille d'outils facilitant l'exploration et l'analyse des données contenues au sein d'une base décisionnelle de type Data Warehouse ou DataMart. Les techniques mises en action lors de l'utilisation de cet instrument d'analyse et de prospection sont particulièrement efficaces pour extraire des informations significatives depuis de grandes quantités de données. [17]

2. Text Mining

Est l'exploration de texte et l'extraction de l'implicite, jusqu'alors inconnue et des informations potentiellement utiles provenant de quantités de ressources textuelles.

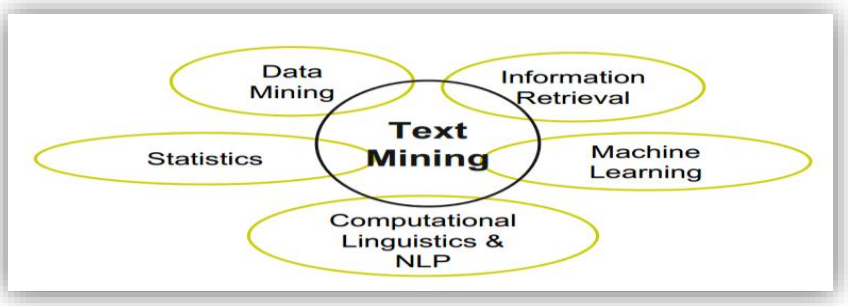


Figure .09: Text Mining

2.1. Les applications d'exploration de texte :

1. Classification des actualités ou des pages Web
2 .Filtrage des e-mails et des actualités / détection du SPAM
3 .Analyse des sentiments
4 .Regroupement de documents ou de pages Web
5 .Saisie automatique des termes de recherche
6. Extraction d'information

2.2. Le processus d'exploration de texte

1 .Prétraitement du texte Syntaxique et / ou analyse sémantique
2 .Génération de fonctionnalités
   - Sac de mots, intégration de mots
3 .Sélection des fonctionnalités
   - Réduisez un grand nombre

4 .Exploration de données

- Clustering
- Classification
- Association une analyse

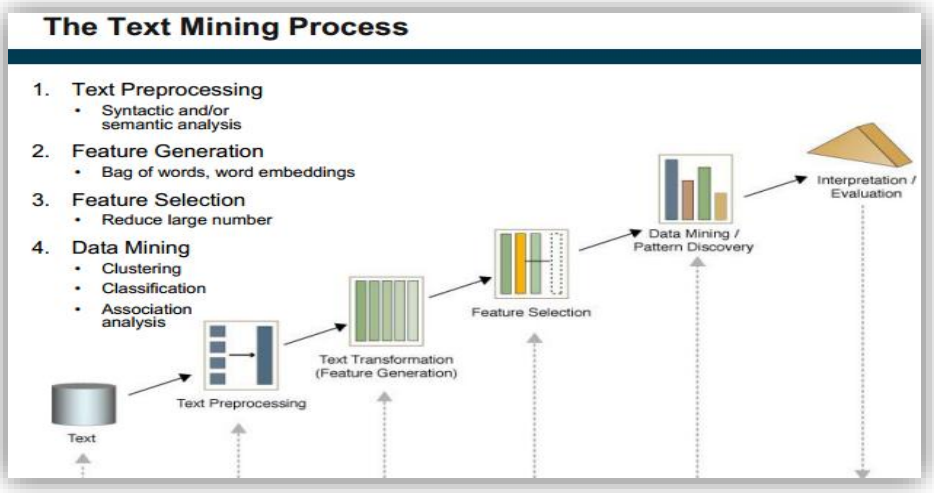


Figure .10: Le processus d'exploration de texte

Objectifs et applications

Parmi les principaux objectifs de la fouille de données textuelles, les suivants :

1. L'identification de thèmes, qui vise à regrouper des textes (ou parties de textes, ou ensembles de *tags*, etc.) en thèmes inconnus *a priori*. Cela implique en général l'utilisation de méthodes de classification automatique sur des données textuelles dont la représentation a été adaptée, par exemple par le passage à des représentations vectorielles.
2. Le classement de textes, qui vise à affecter des textes (ou parties de textes) à des catégories (classes) prédéfinies. Cela passe en général par l'application de modèles décisionnels, obtenus par apprentissage supervisé, à des représentations vectorielles de textes . La détection d'entités nommées, à représenter séparément, peut être utile.
3. L'extraction d'informations, qui vise à mettre en correspondance des textes avec des « schémas » d'interprétation prédéfinis. Un schéma d'interprétation regroupe plusieurs variables qui reçoivent des valeurs à partir du texte traité. Ces variables sont ensuite utilisées pour la fouille, conjointement avec d'autres variables (quantitatives, nominales) décrivant la même population.

Bag of words

Définition

Bag of words (Un sac de mots) est une technique de traitement du langage naturel de la modélisation de texte. En termes techniques, on peut dire qu'il s'agit d'une méthode d'extraction de caractéristiques avec des données textuelles. Cette approche est un moyen simple et flexible d'extraire des fonctionnalités à partir de documents.

Pourquoi sac de mots !

L'un des plus gros problèmes avec le texte est qu'il est désordonné et non structuré, et les algorithmes d'apprentissage automatique préfèrent des entrées de longueur fixe structurées et bien définies.En utilisant la technique du sac de most, nous pouvons convertir des textes de longueur variable en une longueur fixe de plus, à un niveau beaucoup plus granulaire, les modèles d'apprentissage automatique fonctionnent avec des données numériques plutôt qu'avec des données textuelles. Donc, pour être plus précis, en utilisant la technique du sac de mots (BoW), nous convertissons un texte en son vecteur équivalent de nombres.

Comprendre le sac de mots avec un exemple :

Voyons un exemple de la façon dont la technique du sac de mots convertit le texte en vecteurs

Exemple (1) sans prétraitement :

Sentence 1: "Welcome to Great Learning, Now start learning"

Sentence 2: "Learning is a good practice"

Sentence 1	Sentence 2
Welcome	Learning
to	is
Great	a
Learning	good
,	practice
Now	
start	
learning	

Étape 1: Parcourez tous les mots du texte ci-dessus et faites une liste de tous les mots de notre vocabulaire modèle.

- Welcome
- To
- Great
- Learning
- ,
- Now
- start
- learning
- is
- a
- good
- practice

Notez que les mots ‘Learning’ et‘ learning’ ne sont pas les mêmes ici en raison de la différence entre leurs cas et sont donc répétés. Notez également qu'une virgule «,» est également prise dans la liste.

Parce que nous savons que le vocabulaire a 12 mots, nous pouvons utiliser une représentation de document de longueur fixe de 12, avec une position dans le vecteur pour noter chaque mot.

La méthode de notation que nous utilisons ici est de compter la présence de chaque mot et de marquer 0 pour l'absence. Cette méthode de notation est utilisée plus généralement.

La notation de la phrase 1 ressemblerait à ceci:

Word	Frequency
Welcome	1
to	1
Great	1
Learning	1
,	1
Now	1
start	1
learning	1
is	0
a	0
good	0
practice	0

Ecrire les fréquences ci-dessus dans le vecteur Phrase 1→ [1,1,1,1,1,1,1,1,0,0,0]

Maintenant pour la phrase 2, le score voudrait

Word	Frequency
Welcome	0
to	0
Great	0
Learning	1
,	0
Now	0
start	0
learning	0
is	1
a	1
good	1
practice	1

De même, écrire les fréquences ci-dessus sous forme vectorielle

Phrase 2→ [0,0,0,0,0,0,0,1,1,1,1,1]

Sentence	Welcome	to	Great	Learning	,	Now	start	learning	is	a	goo
Sentence1	1	1	1	1	1	1	1	1	0	0	0
Sentence2	0	0	0	0	0	0	0	1	1	1	1

Mais est-ce la meilleure façon d'exécuter un sac de mots. L'exemple ci-dessus n'était pas le meilleur exemple de la façon d'utiliser un sac de mots. Les mots Learning et learning, bien qu'ayant le même sens, sont pris deux fois. De plus, une virgule "," qui ne transmet aucune information est également incluse dans le vocabulaire.

Apportons quelques changements et voyons comment nous pouvons utiliser «sac de mots de manière plus efficace

Exemple (2) avec prétraitement :

Phrase 1 : "”Welcome to Great Learning, Now start learning"

Phrase 2 : «Learning is a good practice»

Étape 1 : Convertissez les phrases ci-dessus en minuscules car la casse du mot ne contient aucune information.

Étape 2 : supprimez les caractères spéciaux et les mots vides du texte. Les mots vides sont les mots qui ne contiennent pas beaucoup d’informations sur le texte comme ‘is’, ‘a’, et bien d’autres.

Après avoir appliqué les étapes ci-dessus, les phrases sont remplacées par

Phrase 1: "welcome great learning now start learning"

Phrase 2: «learning good practice»

Bien que les phrases ci-dessus n'aient pas beaucoup de sens, le maximum d'informations est contenu dans ces mots uniquement.

Étape 3: Parcourez tous les mots du texte ci-dessus et faites une liste de tous les mots de notre vocabulaire modèle.

- welcome
- great
- learning
- now
- start
- good

- practice

Maintenant que le vocabulaire ne comporte que 7 mots, nous pouvons utiliser une représentation de document de longueur fixe de 7, avec une position dans le vecteur pour noter chaque mot.

La méthode de notation que nous utilisons ici est la même que celle utilisée dans l'exemple précédent. Pour la phrase 1, le nombre de mots est le suivant :

Word	Frequency
welcome	1
great	1
learning	2
now	1
start	1
good	0
practice	0

Ecrire les fréquences ci-dessus dans le vecteur

Phrase 1→ [1,1,2,1,1,0,0]

Maintenant, pour la phrase 2, le score serait comme

Word	Frequency
welcome	0
great	0
learning	1
now	0
start	0
good	1
practice	1

De même, écrire les fréquences ci-dessus sous forme vectorielle

Phrase 2 → [0,0,1,0,0,1,1]

Sentence	welcome	great	learning	now	start	good	practice
Sentence1	1	1	2	1	1	0	0
Sentence2	0	0	1	0	0	1	1

L'approche utilisée dans l'exemple deux est celle qui est généralement utilisée dans la technique du sac de mots, la raison étant que les ensembles de données utilisés dans l'apprentissage automatique sont extrêmement volumineux et peuvent contenir un vocabulaire de quelques milliers, voire des millions de mots. Par conséquent, le prétraitement du texte avant d'utiliser un sac de mots est une meilleure façon de procéder.

Il existe différentes étapes de prétraitement qui peuvent augmenter les performances de Bag-of-Words.

Création d'un modèle de sac de mots avec Sklearn (Python) :

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
sentence_1="This is a good job.I will not miss it for anything"
sentence_2="This is not good at all"
CountVec = CountVectorizer(ngram_range=(1,1), # to use bigrams ngram_range=(2,2)
stop_words='english')
#transform
Count_data = CountVec.fit_transform([sentence_1,sentence_2])
#create dataframe
cv_dataframe=pd.DataFrame(Count_data.toarray(),columns=CountVec.get_feature_na
mes())
print(cv_dataframe)
```

Output:

```
   good  job  miss
0     1    1     1
1     1    0     0
```

Limitations du sac de mots

Bien que Bag-of-Words soit assez efficace et facile à mettre en œuvre, cette technique présente néanmoins quelques inconvénients qui sont indiqués ci-dessous :

Le modèle ignore les informations de localisation du mot. Les informations de localisation sont une information très importante dans le texte. Par exemple, «aujourd'hui est éteint» et «est aujourd'hui éteint», ont exactement la même représentation vectorielle dans le modèle BoW.

Sac de modèles de mots ne respecte pas la sémantique du mot. Par exemple, les mots «football» et «football» sont souvent utilisés dans le même contexte. Cependant, les vecteurs correspondant à ces mots sont assez différents dans le modèle du sac de mots. Le problème devient plus sérieux lors de la modélisation des phrases. Ex: «Acheter des voitures d'occasion» et «Acheter des voitures anciennes» sont représentés par des vecteurs totalement différents dans le modèle du sac de mots. [16]

Conclusion

Dans ce chapitre, nous avons parlé des réseaux sociaux et des réseaux les plus importants utilisés par les internautes, Facebook, Twitter et autres.

Nous avons également parlé des communautés virtuelles, de l'analyse des réseaux sociaux et des théories d'analyse de réseau les plus importantes

Nous avons conclu le chapitre avec un exemple d'application en Python.

Chapitre III : L'approche proposée : DataSet, Features, Apprentissage, Résultats et Analyses

1. Présentation

Dans ce chapitre, nous allons recueilli les commentaires des étudiants universitaires et des lycéens et compilé un ensemble de caractéristiques qui sont représentées dans un groupe de mots et d'expressions les plus fréquemment utilisés dans les comm[entaires de chaque classe du secondaire et universitaire.

Nous voulions ajouter à la fois le nombre d'icônes et le niveau linguistique en tant que fonctionnalités, mais nous n'avions pas assez de temps pour les programmer en tant que fonctionnalités et obtenir leurs résultats et les analyser.

2. Contributions

Mes contributions sont les suivantes:

- La Collection des données "data set" qui sont des commentaires des étudiants universitaires et des élèves du secondaire, et leur nombre atteint 779 commentaires.
- Profilage et création de dictionnaires.
- L'application des algorithmes d'apprentissage automatique supervisé "SVM , NB, RF,DT".
- Création des 4 modèles pour la classification des commentaires.

2.1 DataSet :

A- Définition :

Est un ensemble de valeurs (ou données) où chaque valeur est associée à une variable (ou attribut) et à une observation. Une variable décrit l'ensemble des valeurs décrivant le même attribut et une observation contient l'ensemble des valeurs décrivant les attributs d'une unité (ou individu statistique)1 et c'est la première étape d'un algorithme de Supervised Learning consiste à importer un Dataset qui contient les exemples que la machine doit étudier.

Ce Dataset inclut 2 types de variables :

- Une variable objectif (***target***)
- Une ou plusieurs variables caractéristiques (***features)***

B- Source des commentaires

Des commentaires ont été recueillis à partir de pages Facebook et de groupes d'étude pour les éléves du lycée et des étudiants d'université .

Nous avons également recueilli des commentaires pour les niveaux primaire et moyenne, mais nous avons trouvé une grande difficulté de définir les caractéristiques de ces

deux niveaux, ce qui signifie que la plupart des commentaires sont neutres et n'ont pas leurs propres caractéristiques.

La collecte des commentaires a pris beaucoup de temps pour les lycéens et les universitaires, cela est dû à la diversité des commentaires dans les groupes sur Facebook et à la difficulté de déterminer le véritable niveau d'éducation du commentateur.

Nous avons collecter deux types de dataset qui sont équilibrées et déséquilibrées en termes de nombre de commentaires.

- **Dataset équilibrée :**

Niveau éducatif	**Lycéens**	**Universitaires**	**Total**
Nombre de commentaires	388	391	779

Tableau 1 : Nombre de commentaires.

- Dataset non équilibrée :

Niveau éducatif	**Lycéens**	**Universitaires**	**Total**
Nombre de commentaires	388	513	946

Tableau 2 : Nombre de commentaires.

2.2 Les Caractériqtiques proposés :

La principale chose qui distingue les niveaux educatifs est l'utilisation de certains mots et expressions ... Etc.

Nous avons identifié les caractéristiques de chaque niveau que nous expliquerons ci-dessous.

A-Les profils des lycéens :

Étant donné que la majorité des commentaires ont été recueillis auprès de groupes d'étude sur Facebook, les fonctionnalités de niveau secondaire ont été caractérisées par les caractéristiques suivantes :

- Utilisation de la langue vernaculaire.
- Parler de matières étudiées dans le programme.
- Grande utilisation des autocollants.
- Parler de personnalités célèbres dans le domaine de l'éducation et du sport.

B-Les profils des universitaires :

Idem pour les étudiants universitaires, Les commentaires sont recueillis auprès des groupes scolaires

Les caractéristiques des universitaires sont représentées dans:

- L'utilisation du langage formel
- Parler de questions politiques
- Parler du matériel et des programmes d'enseignement supérieur

2.3 Création des dictionnaires :

Il s'agit d'un ensemble de mots et de phrases qu'un classificateur peut utiliser pour évaluer la polarité d'un texte afin d'obtenir une classification plus précise et de réduire la marge d'erreur.

Nous construisons notre propre dictionnaire basé sur ce que nous avons trouvé dans les commentaires et sur les fonctionnalités. Il a fallu près de cinq semaines de travail pour capturer le large éventail de mots et d'expressions présentés.

Niveau educatif	**Lycéens**	**Universitaires**
Nombre de mots et de phrases	**206**	**513**

Tableau 3 : Statistiques de notre dictionnaire.

A-Dictionnaire des lycéens :

Les profils des lycéens	**Les profils des lycéens**
افليكسولي	الأشتقاقية
احتجاج	شتقاقية
الجدول	تمارين
الأدبية	فرض
العلمية	السؤال
الادبية	سؤال
الباك	شجرة الاحتمالات
الكورونا	الاحتمالات
الدراسة	بلييز
مرسي	الوظيفة
البياني	معدل

البيلتا	تحليل
الادب	نص
الاستقصاء بالرفع	مقالة
لامتحانات	عولمة
الامتحانات	المقاومة
فصل	رياضي
تحليل نص	فهموني
المعدل	زوايا
مراجعة	نهايات
الأكسدة والأرجاع	أستاذ
الأكسدة	يااستاذ
الأرجاع	الفيزياء
تمارينات	الأدب
الزوايا الموجهة	الادب
تمارين	المات
جداء سلمي	والمات
نشاط	البرنامج
الكمياء العضوية	التأجيل
الكمياء	التاجيل
تراجعوا	شعبة
يقراو	ميكانيك
استمارة التسجيل	طرائق
المعدل السنوي	كهرباء
نقطة	دروس
ماسنجر	مادة
الانجليزيه	هندسة طرائق
القراية	تمرين
سبور	تقني رياضي
قراية	كونكسيو
مكعبات	راهم يقراو
متتاليات	باطل
تحاكي	استاذة
هندسة في الفضاء	استاذ

هندسة	الشيخ
السوجي	نكونكتي
عدد زوجي	فيسبوك
القواعد الازوتية	الفلسفة
النكليوتيدات	شعبة علوم
البيام	علوم
عام ابيض	فلسفة
الاستوائي	تاريخ
فيديو	جغرافيا
الفصل الثالث	تسيير
الخلايا	البكالوريا
روتانا	باكالوريا
تلفزيون	هندسة مدنية
النجاح	مدنية
العام	مات
متخافوش	فيزياء
خايف	فيزيا
خايفين	رياضيات
تلاميذ	علوم تجريبية
سنة	الدروس
لغات	التمرين
كراريس	التمسخير
قسم	تمسخير
كراستي	نقرا
كتاب	نقرو
ستيلو	نقراو
كراس	تقرا
متقريش	قراو
النصوص الأدبية	يقري
مراجعة	قريتو
قناة	ملخصات
الثانوية	قروب
ثانوية	تريفيزو

مراقب	بليز
مراقبين	مسنجر
تكنيك	نورمالموا
الدروس الخصوصية	مواضيع
المحاسبة	باكات
دروس	باك
يوتيوب	الباك
فيسبوك	اجتماعيات
فليكس	دعم
كونيكسيون	دالة
سنة بيضاء	داله
شيخ	دالة تألفية
صحابي	دالة تآلفية
أصحابي	أسئلة
اصحابي	أسئلة
نورمالمو	اختبار
مالا نهاية	الحل
نهاية	حل
الكلاصة	العدد المشتق
التلاميذ	النهايات
العطلة	حالات عدم التعيين
عطلة	ثانوي
الحب	المرجح
حب	الزوايا
سنة بيضاء	سوجي
وزير التربية	

Tableau 4 : dictionnaire des lycéens.

B-Dictionnaire des universitaires :

Les profils des universitaires	Les profils des universitaires
اللسانيات	التعليم
احترافية	إشكالية
معايير	المنظمة
إستراتيجية	اضرابات

أمسية علمية	واضرابات
سيميولوجية	مموار
ناشر	كمنظمات
مقاييس	الطلبة
مقياس	جامعة
انثروبولوجيا	جامعية
مسابقة	جامعية
تسويق	جامعي
طور	ماستر
كلاسيك	منظمات
متخرج	حقوق الطلبة
ارشاد	طلبة
كيميائية	مموار
موارد	إضرابات
بشرية	الاضراب
الكتروني	اضراب
احترازية	إضراب
وقاية	وقفة
توعية	منظمة
عطلة أكاديمية	المصلحة العامة
الارطفونيا	الحقوق
صهيوني	حقوق
الموساد	الاقامات
يهود الفلاشا	الإقامة
كهروضوئية	الخدمات
توربينات	خرافة
الطاقة الشمسية	تحليل
الطاقة	سوسيولوجي
افتراضي	عمق
الطب	المجتمع
قرصنة	المحلل
جمعيات	الاجتماعي
مادي	السياسي

تقنيات	الدكتور
تدوين	الأستاذ
أنتروبولوجية	البروفسور
عصر	اقامة
مهندس	هوية
لاتيني	قانون
عثماني	الحرية
كفاءات	الوزارة
مبادر	ترجم
تكوين	أطروحات
دورة	أطروحة
أسدل الستار	الإنجليزية
فعاليات	إشكالية
مفاهيم	قرار سياسي
مناهج	سياسي
نموذج	اللغوي
مثالي	التعليم
تعريب	التعليم العالي
فهرس	الفيروسات
معجم	الالكترونية
عميد	البيولوجية
شيتور	الرداءة
مداخلة	لا فض فوك
تنمية	المطبوعات
استشار	فريد من نوعه
المجتمع المدني	بلقاسم حبة
ديمقراطية	إنسانية
تشاركية	الدروس على الخط
أطروحة	الوزارة الوصية
اطروح	متابعة قضائية
أطروح	الأساتذة
الجامعيين	طالب
إحصائية	برميل نفط

إحصائيات	برميل النفط
الجامعة الصيفية	دولار
البنكية	مدى جدية
الخدمات	مدى قانونية
الاستحقاقات	التعليمة
ديمقراطية	المصدر
الاتفاف	قطاع
الاقتصادية	العالم
رجال الأعمال	دول
الاستثمارية	مبتكر
تدريجي	درجة
الهيدروكلورين	معلوماتية
الدكتور	ابتكر
محكمة	براءة اختراع
البيرقراطيين	اختراع
المدونين	العالم العربي
نظرية المؤامرة	جائحة
الماسونية	الضرائب
الشنتو	حافظ السيوطي
الاسلاموية	الشريعة
العلمانية	برفسور
الاحزاب	وزير التعليم العالي
اجندات	بحث
السلطة	علمي
سلطة	اداري
الهيكل	تعليمة
منظمة	المرسوم
المنظمة	وزارية
المنظمات	الشفافية
سنن	الرقابة
الحق	العرف
الباطل	الفكرة
القومية	عقل

الوطنية	عقول
الدستور	أفكار
الكونية	الأمم
العلماء	هويتنا
النخب	شبكات
الفكرية	كلمة السر
المجتمع	كلمة سر
الدولة	اسم المستخدم
مداخلة	حقوق الملكية
إسلامية	الفكرية
الشرائية	القضاء
البيروقراطيون	برامج
الضرائب	رقمية
الضرائب	مؤلفات
الحكومة	علمية
حكومات	ترخيص
النواميس	منصة
الاستحقاقات	دولي
الانتخابية	نسختة
الراسمالية	الورش
اشتراكية	دورات
الليثيوم	رابط
البيروقراطية	روابط
الرئيس	الأمين العام
السلم	مراسلة
الاجتماعي	النشر
السياسية	بيداغوجية
اقتصاديا	بحث
إشكاليات	بحوث
الإشكاليات	مخبر
ابن خلدون	مخابر
الأغلبية	كلية
الاحصائيات	كليات

اقتصادي	معهد
نظام	معاهد
النظام	دراسات
الأنظمة	مقال
الظواهر	مقالة
الظاهرة	منشور
العلمي	مجلة
الاعجاز	ملتقى
الفيزيائية	ندوة
المنطقية	تربص
إزدواجية	بيوطبية
الخطاب	مرجع
لمالك بن نبي	مراجع
زغماتي	الإحصائية
يتناقشون	الإحصاء
مناقشة	عين الإعتبار
الازلي	باحث
الصناعية	بروفيسور
التطور	مجلات
العلم	خبراء
النمو	اللجنة الوطنية
الصناعي	اللجنة
الغذائي	اللجان
الاكتفاء	الصوديوم
محاكات	غاز النيتروجين
الفكر	جائزة
العالمية	القانوني
بوكروح	سلبيات
الحضيض	اجابيات
التجاذبات	عالمية
الإنسانية	طلابية
PDF	الدكتوراه
إنشائي	بيان

تعبير	ليسانس
العدالة	لسانس
الأيديولوجية	الاميل
براغماتية	الايميل
البراغماتية	الايمايل
الاستقلال	محاضرة
هوية	محاضرات
الهوية	الاستدراك
الأجيال	المنح
المنظومة	المنحة
منظومة	بريد الكتروني
مصدر	البريد الالكتروني
الجهوية	وزاري
الوعي	رسالة
الحضاري	ماجستير
البيئة	مقالات
كيسنجر	ثقافي
الاحتباس	اقتصادي
لمنظمه	المؤتمر
الط؟ب	مؤتمر
المصلحة	تربوية
العامة....	علاجية
ماستر	علم النفس
الاضراب	جداريات
حقوق	دراسية
الطلبة	تكوينية
لطلبة	صنف
الوضع	سكوبس
الكارثي	مذكرة
الاقامات	تظاهرة
الخدمات	مجلس
درجة	نظرية
الجامعة	نظريات

للطلبة	مناقشة
الإقامات	مناقشات
الجامعية	الأساتذة
الخرافة	الباحثين
خرافة	الاجتماعية
مجرد	مسابقات
الاستفتاء	مقدمة
النتيجة	خاتمة
سوسيولوجي	إمضاء
المجتمع	أخلاقية
السياسي	ارتجالية
اقامة	ديبلوم
هوية	الإنترنت
قانون	الانترنت
الحرية	علم
الوزارة	تأهيل
أطروحات	اللجنة العلمية
سياسي	المجلس العلمي
اللغوي	اكاديمي
للتدريس	مكتبات
القرار	قواعد البيانات
الماستر	مشرف
الصميم	مبتكر
المصائب	ابتكر
والجرائم	حائز
المزرية	العربي
البيولوجية	عالميا
الفيروسات	المدمج
الجهاز	التعليم
ارتقاء	التقنية
المحتوى	الكفاءة
الرداءة	المعلومات
دكتور	الكلاسيكي

للتعليم	الطالب
راقي	الجائحة
بلقاسم	السيوطي
حبة	العبرة
مبادرة	الشريعة
إنسانية	باستور
قانونية	استعمارية
قضائية	المستنقعات
الأساتذة	السذاجة
الطلبة.	الاحتباس
الطلبة	الصهيونية
برميل	مدلول
نفط	تحليل
بالدولار	النظام
المصدر	الإشكال
الرداءة	الاشكال
جامعاتنا	أخلاق
الصحة	الغاية
العالم	الحراك
دول	القطب
إستثناء	

Tableau 5 : dictionnaire des universitaires

3. Expériences et résultats :

Tant que nous utilisons la méthode d'apprentissage supervisé, nous divisons le corpus en deux parties, 80% pour la formation et 20% pour les tests.

3.1 Analyse des résultats:

Nous avons réalisé plusieurs tests. Le tableau 4.1 présente les résultats de la classification et une discussion partielle de chaque test.

Test	Type de classification et nombre de commentaires	Classificateur	Résultats	Discussion
1	Le dataset est équilibrée, la première classe contient le niveau secondaire et le reste contient le niveau universitaire et le nombre de commentaires dans les deux classes est388, 391	**SVM**	**0.794**	Le meilleur résultat de ce test est de 80% donné par le DT et le pire résultat de ce test est de 74% donné par le
		DT	**0.801**	
		RF	**0.778**	
		NB	**0.743**	
2	Le dataset est équilibrée, la première classe contient le niveau universitaire et le reste contient le niveau secondaire et le nombre de commentaires dans les deux classes est391, 388	**SVM**	**0.807**	Le meilleur résultat de ce test est de 82% donné par le RF et le pire résultat de ce test est de 80% donné par le
		DT	**0.814**	
		RF	**0.820**	
		NB	**0.807**	
3	Le dataset n'est pas équilibrée, la première classe contient le niveau secondaire et le reste contient le niveau universitaire et le nombre de commentaires dans les deux classes est388, 558	**SVM**	**0.820**	Le meilleur résultat de ce test est de 82% donné par le SVM et le pire résultat de ce test est de 78% donné par le
		DT	**0.807**	
		RF	**0.807**	
		NB	**0.814**	
4	Le dataset n'est pas équilibrée, la première classe contient le niveau universitaire et le reste contient le niveau secondaire et le nombre de commentaires dans les deux classes est 558,388	**SVM**	**0.803**	Le meilleur résultat de ce test est de 80% donné par le SVM et le pire résultat de ce test est de 73% donné par le
		DT	**0.787**	
		RF	**0.797**	
		NB	**0.734**	

Tableau 6 : Analyse des resultats.

3.2 Résultats (analyse et discussion)

3.2.1 Analyse des résultats

Dans cette section, nous analyserons et discuterons les différents résultats de classification obtenus après avoir appliqué les quatre algorithmes que sont les machines vectorielles de support (SVM), l'arbre de décision (DT), la forêt aléatoire (RF) et Naive Bayes

(NB) sur l'ensemble des classifications précédemment proposées avec un dictionnaire de fonctionnalités.

Tant que nous utilisons la méthode d'apprentissage supervisé, nous divisons le corpus en 2 parties, 80% pour la formation et 20% pour les tests. Nous avons compté quatre cas dans quatre modèles; nous expliquerons ce que nous avons fait à ce qui suit:

Cas 1: Un ensemble de données équilibré et dictionnaire de niveau pré-supérieur, la première classe contient les commentaires de niveau pré-supérieur et l'autre contient les commentaires de niveau supérieur, et le nombre de commentaires est proche.

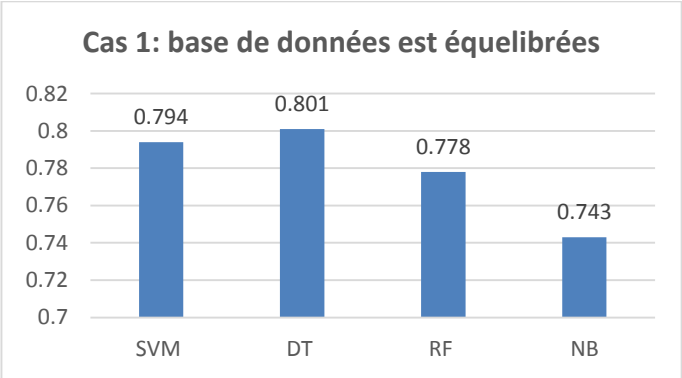


Figure 11: Cas 1-Résultat: ensemble de données équilibré et dictionnaire de niveau pré-supérieur

Le meilleur résultat de ce test est de 80% donné par le DT et le pire résultat de ce test est de 74% donné par le NB.

Cas 2: Un ensemble de données équilibré et dictionnaire de niveau supérieur, la première classe contient les commentaires de niveau supérieur et l'autre contient les commentaires de niveau pré-supérieur, et le nombre de commentaires est proche.

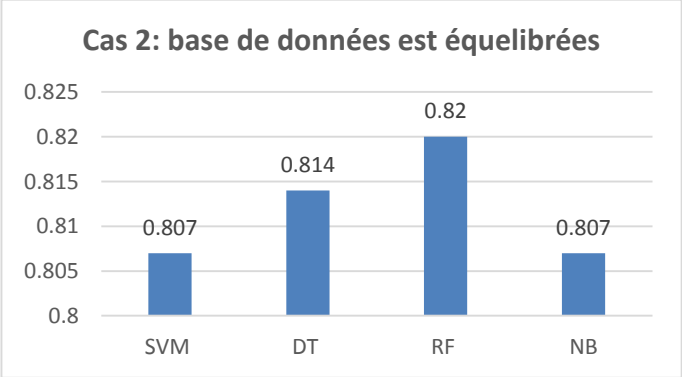


Figure 12. Cas 2-Résultat: ensemble de données équilibré et dictionnaire de niveau supérieur

Le meilleur résultat de ce test est de 82% donné par le RF et le pire résultat de ce test est de 80% donné par le SVM et NB; mais ils sont généralement tous proches.

Cas 3: ensemble de données non équilibré et dictionnaire de niveau pré-supérieur, la première classe contient les commentaires de niveau supérieur et l'autre contient les commentaires de niveau pré-supérieur, avec une grande différence dans le nombre de commentaires.

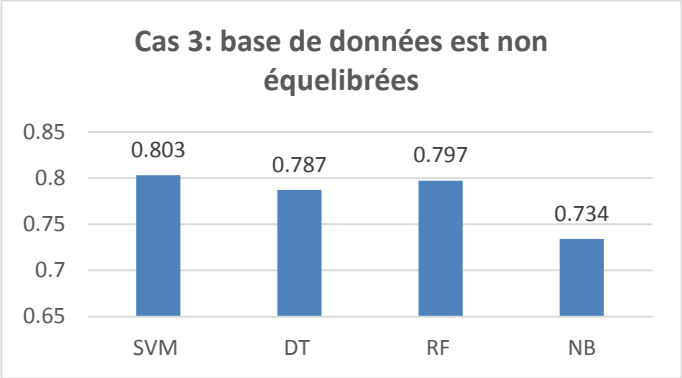


Figure 13. Cas 3-Résultat: ensemble de données déséquilibré et dictionnaire de niveau pré-supérieur

Le meilleur résultat de ce test est de 80% donné par le SVM et le pire résultat de ce test est de 73% donné par le NB.

Cas 4: Un ensemble de données non équilibré et dictionnaire de niveau supérieur, la première classe contient les commentaires de niveau supérieur et l'autre contient les commentaires de niveau pré-supérieur, avec une grande différence dans le nombre de commentaires.

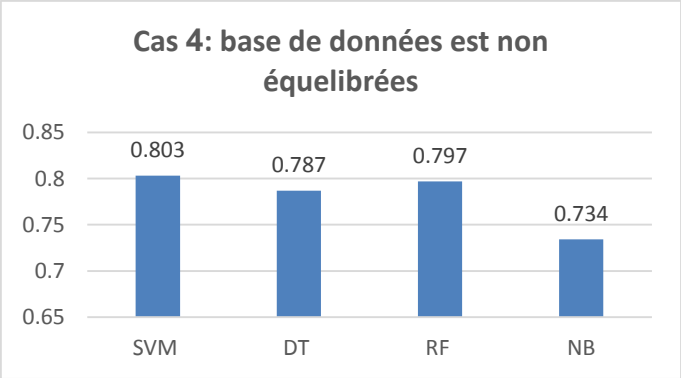


Figure 14. Cas 4-Résultat: ensemble de données déséquilibré et dictionnaire de niveau supérieur.

Le meilleur résultat de ce test est de 80% donné par le SVM et le pire résultat de ce test est de 73% donné par le NB.

3.2.2 Discussion sur les résultats:

D'après les résultats, nous concluons que:

- ✓ Le meilleur résultat dans tous les tests était le (82%) que RF a donné sur le cas 2 s'applique à l'ensemble de données équilibré et inclut les commentaires de niveau pré-supérieur et supérieur.
- ✓ Les meilleurs résultats (80,3%), donnés par SVM dans les cas 3 et 4;
- ✓ Le pire résultat des quatre cas (73%), donné par le NB dans le cas 4 où il a été appliqué à la classe 4 contenant un ensemble de données déséquilibré comprenant des commentaires de niveau pré-supérieur et supérieur;
- ✓ Nous concluons que le nombre de commentaires d'entraînement joue un grand rôle dans l'amélioration du résultat, et la sélection la plus appropriée de l'algorithme de Machine Learning augmente la précision du résultat.

3.3 Discussion sur les résultats :

Le meilleur résultat dans tous les tests était le 82%que RF a donné avec le test 2

La classe 2 s'applique au DATASET équilibré et comprend les commentaires de niveau secondaire et universitaire.

Et SVM l'a donné avec le test 3. La classe 3 s'applique au dataset déséquilibré et comprend les commentaires de niveau secondaire et universitaire.

Le pire résultat des quatre tests 73% a été donné par le NB au test 4 où il a été appliqué à la classe 4 contenant une dataset non équilibrée comprend des commentaires du secondaire et du universitaire.

Nous concluons que l'amélioration des résultats n'est pas due à un déséquilibre mais au grand nombre de DATASET

Conclusion

Dans ce chapitre, nous avons identifié des profils généraux qui définissent chaque niveau d'éducation et analysé un ensemble de données une fois équilibrées et une autre fois déséquilibrées en termes de nombre de commentaires. Le déséquilibré contenait près de 946 textes et le déséquilibré en contenait 779 comprenant deux niveaux d'enseignement (secondaire / universitaires).

Enfin, nous avons créé deux dictionnaires, un dictionnaire de niveau secondaire de 206 mots et un dictionnaire de niveau universitaires de 513 mots.

Chapitre IV :

L'implémentation

Introduction

Dans ce chapitre, nous parlerons de Python et des libraires les plus importants, en particulier les libraires d'apprentissage automatique, puis les étapes du programme d'application, les plus importants sont :

- Télécharger les bibliothèques requises
- Téléchargez les données
- Calculez les mots dans les données et séparez-les
- Créer une table de poutres pour les commentaires
- Séparation du contenu de la table
- Le programme scanne le contenu du tableau et le compare à 0/1
- Une fois qu'une correspondance existe, les résultats sont placés dans la colonne de comptage univ
- Ensuite, nous appliquons les quatre algorithmes d'apprentissage automatique supervisé.

1. Python

Est un langage de programmation interprété, orienté objet et de haut niveau avec une sémantique dynamique. Il a été créé par Guido van Rossum et sorti en 1991. Python prend en charge les modules et les packages, ce qui encourage la modularité du programme et la réutilisation du code. L'interpréteur Python et la bibliothèque standard étendue sont disponibles sous forme source ou binaire sans frais pour toutes les principales plates-formes, et peuvent être librement distribués.

2. A quoi sert Python؟

2.1 Développement web

Récemment, les frameworks qui utilisent Python pour le développement Web sont devenus populaires, tels que Django et Flask.

Ces progiciels aident à développer des serveurs Web backend basés sur Python. Django et Flask sont les logiciels les plus populaires et les meilleurs pour les débutants.

2.2 Scripting

Les scripts écrivent de petits programmes destinés à automatiser des tâches simples. En fait, Python est le langage parfait pour des tâches aussi simples car la construction de sa syntaxe est très simple et facile. Python est également le mieux adapté pour développer et tester rapidement de petits programmes

2.3 Science des données (apprentissage automatique, analyse et présentation du Big Data)

Il existe de nombreuses bibliothèques et progiciels d'apprentissage automatique Python. Les meilleures et les plus populaires de ces bibliothèques est **scikit-learn**.

3. Notebook Jupyter

Le notebook Jupyter est un environement HTML pour Python. Il fournit un environement organisé en cellules interactives qui peuvent être exécutées, permettant l'organisation et la documentation de calculs de façon structurée.

4. Editeur Spyder

Spyder est un IDE MATLAB-like pour le calcul scientifique avec Python. Il permet d'éditer, d'exécuter et de déboguer du code dans un seul environement.

Spyder fournit entre autres :

- Un editeur de code efficace avec coloration syntaxique, permettant de l'introspection dynamique de code, intégré au débogueur de Python.
- Un exploreur de variables, une invite de commande IPython

- Variable exploré, IPython command prompt.
- Une documentation intégrée.

5. Les bibliothèques de Python :

La bibliothèque est une ensemble de fichiers de code et de fichiers d'aide que vous pouvez inclure dans votre code qu'un groupe de développeurs a écrit et publié pour le rendre disponible à tout le monde et le but des bibliothèques de programmation est de faciliter la mise en œuvre de certaines choses dans n'importe quelle application en premier lieu en plus de donner à l'application des capacités supplémentaires. Il ne peut le faire qu'après avoir utilisé ces bibliothèques

5.1. Bibliothèque Requests :

Si vous souhaitez établir une connexion via le protocole HTTP, alors cette bibliothèque est votre choix approprié pour le faire automatiquement.

5.2. Bibliothèque Pluggy :

Il est considéré comme l'une des bibliothèques relativement nouvelles du langage Python et vous donne la flexibilité de gérer les plugins (plugins) d'ajout, de modification et de gestion, car il vous permet de gérer plus de 500 plugins .

5.3. Bibliothèque SQLAlchemy :

Cette bibliothèque donne aux programmeurs Python la puissance et la flexibilité nécessaires pour gérer SQL, et cette bibliothèque vous permet d'effectuer les principales opérations relatives aux bases de données de type SQL.

5.4. Bibliothèque Numpy :

Cette bibliothèque permet d'effectuer des calculs numériques avec Python. Elle introduit une gestion facilitée des tableaux de nombres.

5.5. Bibliothèque SciPy :

Est une bibliothèque Python a usage mathématiques elle dépend de NumPy, qui fournit une manipulation des tableaux à N dimensions pratique et rapide.

5.6. Bibliothèque Skit-learn

Scikit-learn ou **Sklearn** est une bibliothèque basée sur Python permettant de créer des modèles d'apprentissage automatique.

- Il fournit de nombreux algorithmes d'apprentissage pour la régression, le regroupement et la classification.
- Sklearn est compatible avec NumPy et SciPy. Cela signifie que vous pourrez interagir facilement avec différentes bibliothèques pour Python.

5.7. Bibliothèque Pandas :

Pandas est une **bibliothèque Python open source** permettant de manipuler des structures de doe'nnées hautes performances et faciles à utiliser ainsi que des outils d'analyse de données pour le langage de programmation Python.

5.8. Bibliothèque CSV :

La bibliothèque csv fournit des fonctionnalités de lecture et d'écriture dans des fichiers CSV. Conçu pour fonctionner directement avec les fichiers CSV générés par Excel, il est facilement adapté pour fonctionner avec une variété de formats CSV. La bibliothèque csv contient des objets et d'autres codes pour lire, écrire et traiter des données depuis et vers des fichiers CSV.

6. Matériel de configuration :

J'ai utilisé mon pc DELL et leurs caractéristiques sont les suivants :

- **Processeur** : processeur Intel® Core ™ i3-2377M à 1,50 GHz.
- **RAM** : 4 Go.
- **Type de système** : système d'exploitation Windows.

7. Les étapes d'exécution :

Dans la partie implémentation, nous avons appliqué les algorithmes à deux types de dataset équilibrées et non équilibrées en termes de nombre de commentaires, L'exemple ci-dessous concerne le dataset non équilibrées

« Modèle DataSet non équilibré »

7.1 La déclaration et l'importation des bibliothèques nécessaires :

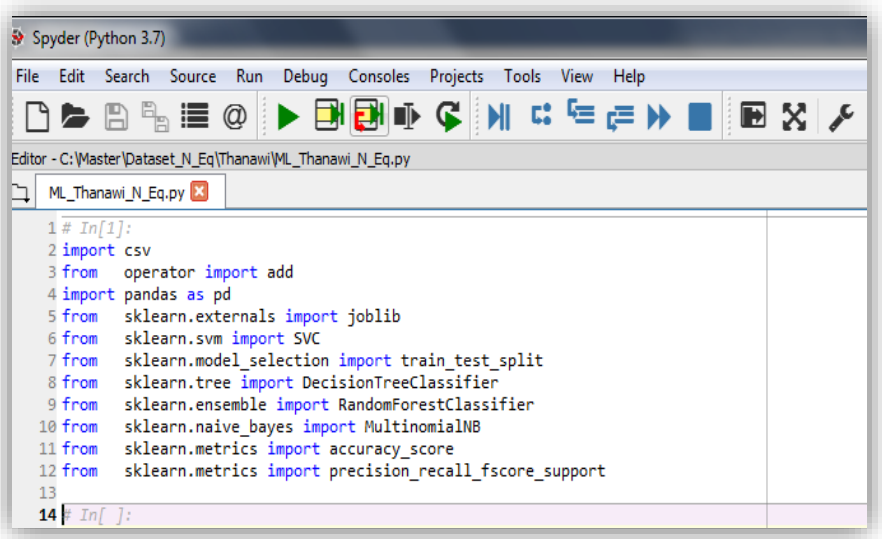


Figure .11: La déclaration et l'importation des bibliothèques nécessaires

7.2 Importation des données (les commentaires et les dictionnaires) sont lus et répertoriés dans des tableaux pour être traités ultérieurement

```
15
16 Corpus = pd.read_csv("C:/Master/Dataset_N_Eq/Thanawi/Dataset.csv", encoding = "ANSI" )
17
18 # In[]
19 Thanawi = pd.read_csv(r"C:/Master/Dataset_N_Eq/Thanawi/D-Thanawi.csv", header=0, sep='delim
20
```

Figure .12: Importation des données

7.1.3 On compte les mots dans chaque commentaire du Dataset .la séparation entre le mot et le mot est l'espace et la fonction Split () qui effectuent cette opération

```
22 # In[3]
23 # Calcule le nbr des mot dans les commentaires
24 word_count = []
25 for i in Corpus.Tweets:
26     word_count.append(len(i.split()))
27
28 Corpus["word_count"] = word_count
29
30 # In[ 4]:
```

Figure .13: compte les mots

le résultat est une table Word_count qui affiche le nombre de mots pour chaque commentaire, puis répertorie le résultat dans une colonne appelée Word_count dans la table Corpus.

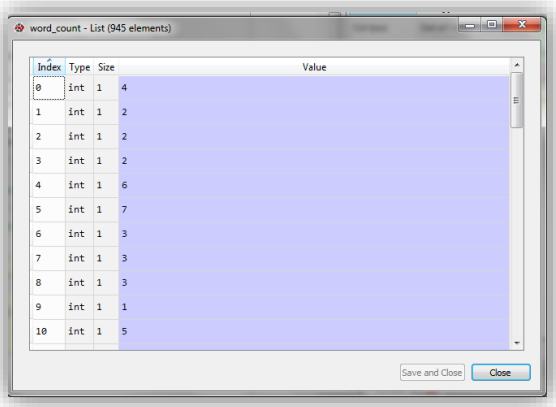


Figure .14: Word count list

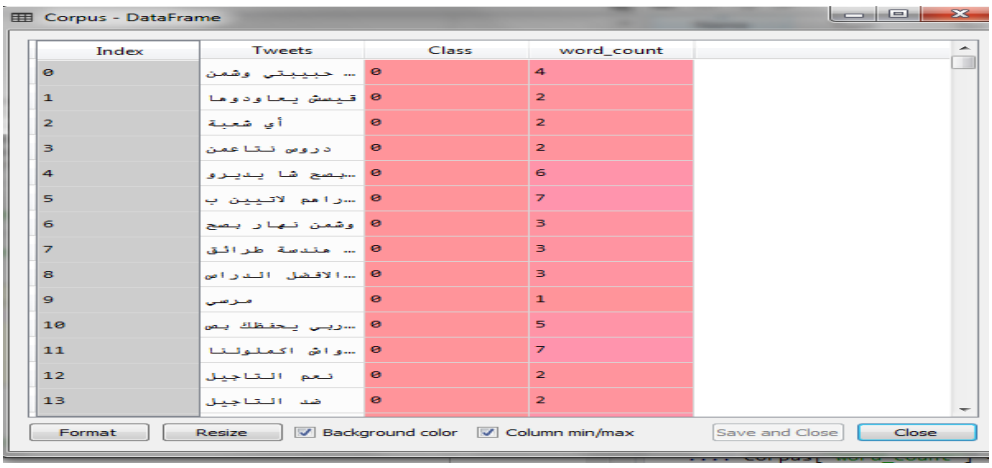


Figure .15: Aprés l'ajout d'une colonne COUNT WORD.

7.1.4 Deuxièmement:

Le programme crée un tableau avec des vecteurs pour les commentaires afin qu'il sépare chaque commentaire pour une unité et ici il représente un vecteur et sépare chaque mot du commentaire séparément pour devenir un composant des vecteurs

```
# In[ 4]:

Corpus_words       =[]

csv_reader = csv.reader(open(r'C:/Master/Dataset_N_Eq/Thanawi/Tweets.csv'  ), delimiter=','
arr = list(csv_reader)
for line in arr:
        Corpus_words.append(line[0].split(" "))

# In[ 4]:
```

Figure .16: tableau des vecteurs

Le résultat :

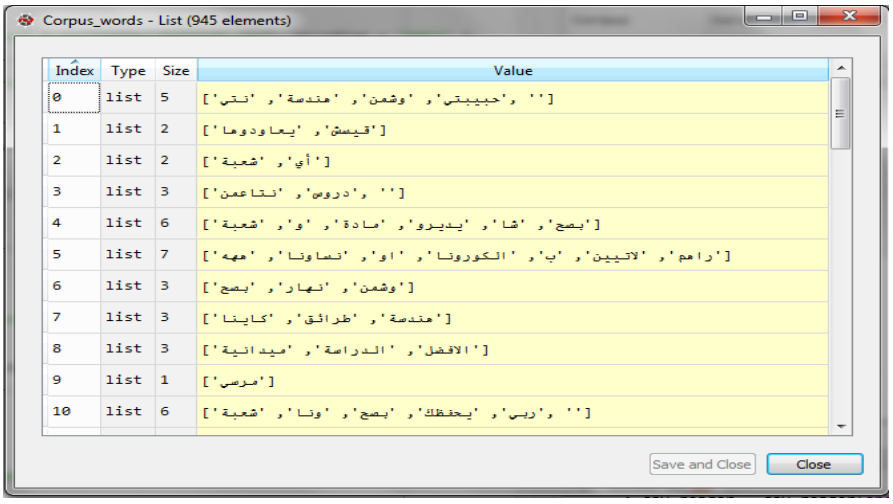


Figure .17: Le résultat

Nous effectuons le processus avec notre table de caractéristiques et le divisons en rayons de composants séparés

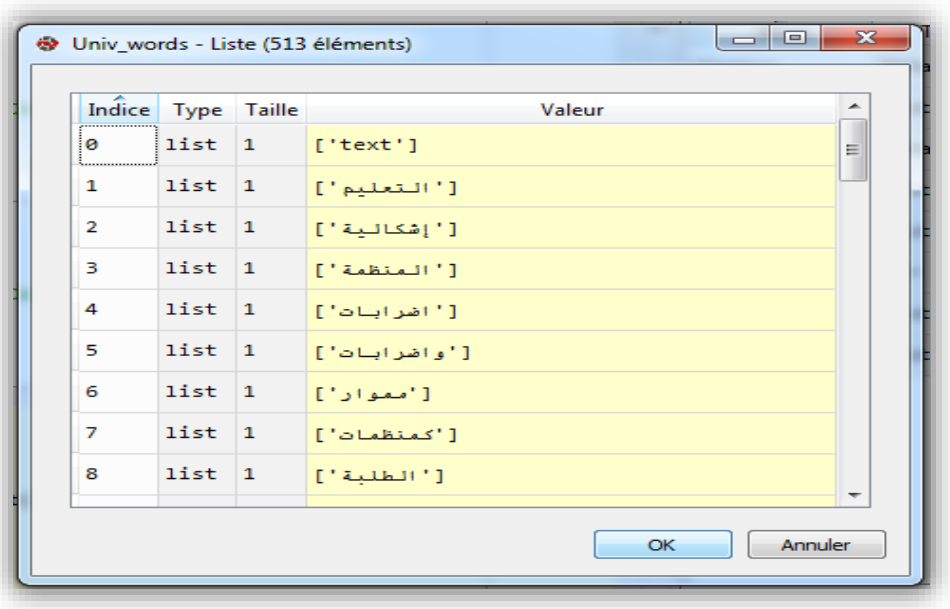


Figure .18: table de caractéristiques

```
48 # In[8]
49
50 F_Univ    = []
51
52 for t in Corpus.Tweets:
53     if any(w in t.split() for w in Univ.text):
54         F_Univ.append(1)
55     else :
56         F_Univ.append(0)
57
58 Corpus["Futurs_Univ"] = F_Univ
59
```

Figure .19: verification de correspondance

Troisièmement: le programme scanne tous les mots du datasat sur les tables de la table des entités dans une table pour chaque vecteur séparément (commentaire) S'il y a une correspondance qui donne la valeur 1 ou la valeur 0, insérez ces valeurs dans un colonne dans la table de l'ensemble de données (Corpus).

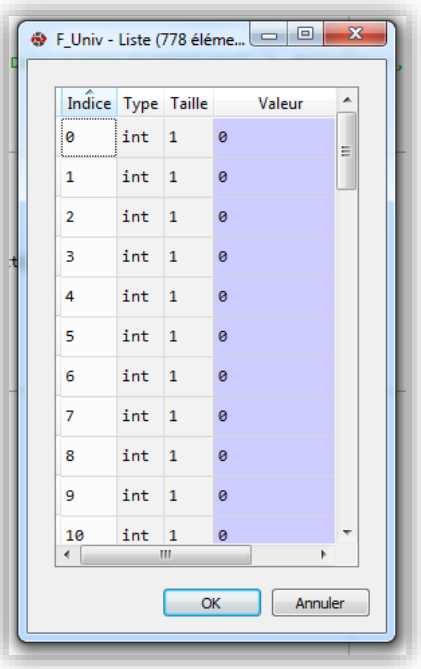


Figure .20: resultat de verification

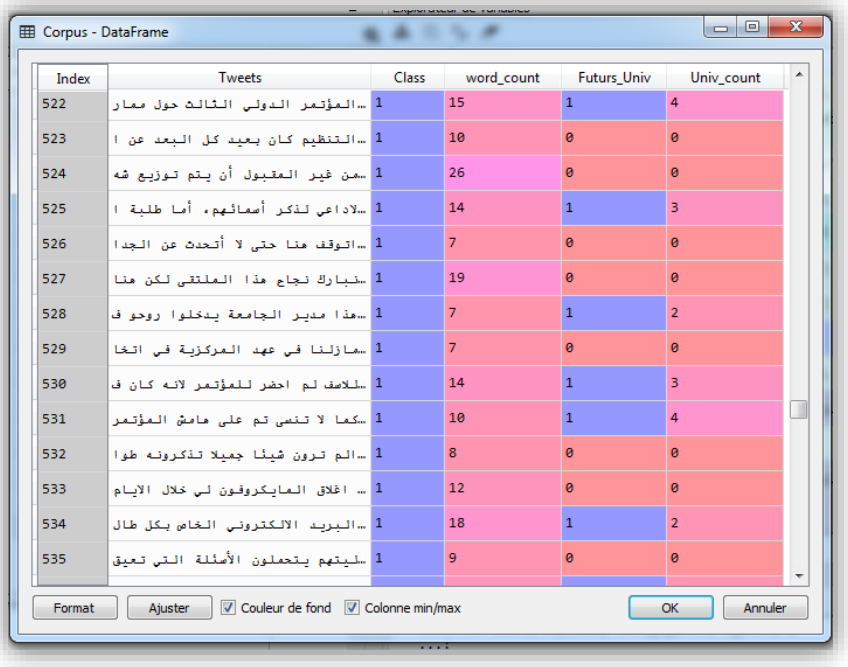


Figure .21: resultat

Les resultats :

Le numéro Un en bleu est le résultat de la correspondance de cette ligne avec les entités nommées dans l'en-tête de colonne.

Quatrièmement: Après avoir trouvé des correspondances avec les caractéristiques , nous les suivrons: (Autrement dit, collectez-les pour chaque caractéristique séparément), et ce processus avec toutes les tables de fonctionnalités.

Et nous insérons le résultat dans les colonnes de la table data set, pour notre exemple ici nous l'avons inclus sous le nom « univ count ».

Remarque: les fonctionnalités de notre site sont souvent des phrases complètes.

```
60 # In[9]:
61
62 Cont_W_Univ      = []
63 count = 0
64 for t in Corpus.Tweets:
65     for word in t.split():
66         for word1 in Univ.text:
67             if word == word1:
68                count +=1
69     Cont_W_Univ.append(count)
70     count = 0
71 fn = []
72 fn= list( map(add, Cont_W_Univ, F_Univ) )
73
74 Corpus["Univ_count"] = fn
75
```

Figure .22: insertion des resultat dans la table de dataset

Les resultats

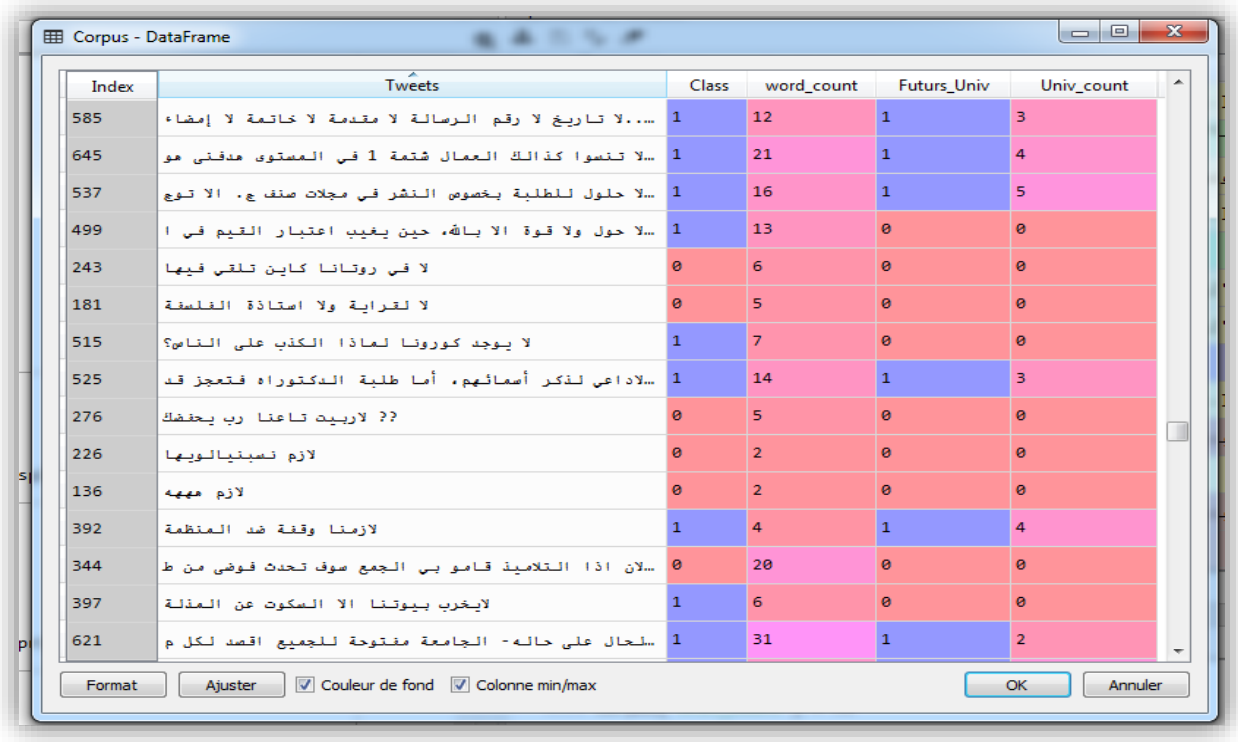


Figure .23: resultat

Phase d'entraînement machine et de tirage des résultats :

Premièrement: nous prenons les valeurs sur lesquelles le système va s'entraîner et passer le test

```
588 # In[21]:
589
590 Z = Corpus.get(Corpus.columns[1:])
591 Y = Corpus.get(Z.columns[0])
592 X = Corpus.get(Z.columns[1:])
593 x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=4)
594
```

Figure .24: entrainnement du programme

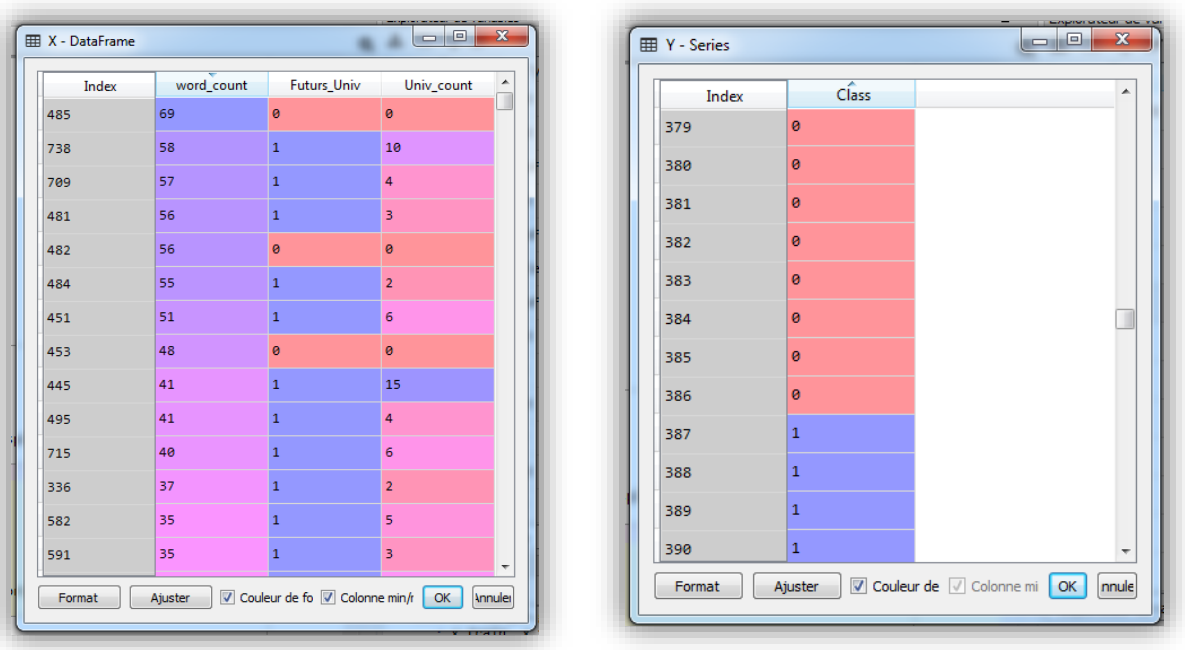


Figure .25: les correspondances avec des caractéristiques

La fonction get ne donne que les valeurs numériques de ces copus, ce sont des valeurs X et les valeurs Y sont les valeurs de classe.

Ensuite, nous les ajoutons à la valeur de Z.

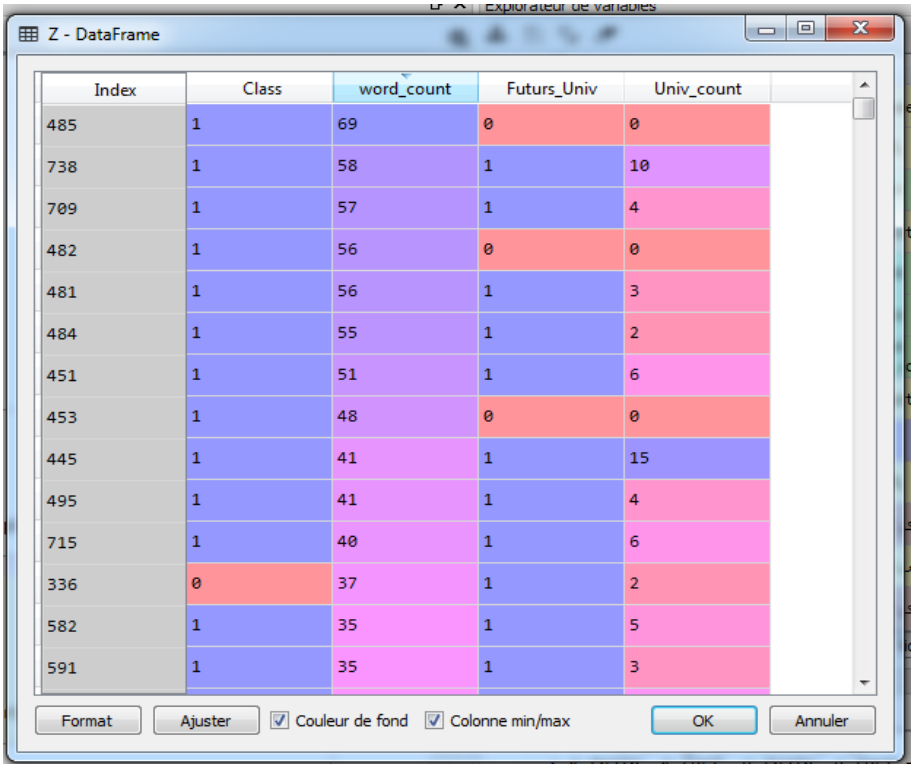


Figure .26: resultat

Et puis nous passons les données sur les quatre algorithmes comme suit pour créer les fonctions requises de chaque algorithme

```
# In[22]:

svm = SVC()
svm.fit(x_train, y_train)
acc1_svm = svm.score(x_test, y_test)
pr1_svm = svm.predict(x_test)
ac = accuracy_score(y_test, pr1_svm)
ww1 = precision_recall_fscore_support(y_test, pr1_svm, average='macro')


# In[23]:

dt = DecisionTreeClassifier()
dt.fit(x_train, y_train)
acc2_dt = dt.score(x_test, y_test)
pr2_dt = dt.predict(x_test)
ww2 = precision_recall_fscore_support(y_test, pr2_dt, average='macro')
ac2 = accuracy_score(y_test, pr2_dt)

# In[24]:

rf = RandomForestClassifier(n_estimators=40)
rf.fit(x_train, y_train)
acc3_rf = rf.score(x_test, y_test)
pr3_rf = rf.predict(x_test)
ww3 = precision_recall_fscore_support(y_test, pr3_rf, average='macro')
ac3 = accuracy_score(y_test, pr3_rf)

# In[25]:

NB = MultinomialNB()
NB.fit(x_train, y_train)
acc4_NB = NB.score(x_test, y_test)
pr4_NB = NB.predict(x_test)
ac4 = accuracy_score(y_test, pr4_NB)
ww4 = precision_recall_fscore_support(y_test, pr4_NB, average='macro')
```

Figure .27: partie d'application des algorithmes

Enfin, obtenons la précision de chaque algorithme

ac	float64	1	0.8076923076923077
ac2	float64	1	0.8141025641025641
ac3	float64	1	0.8205128205128205
ac4	float64	1	0.8076923076923077

Figure .28: resultat

Notez que SVM est le meilleur algorithme, nous gardons donc le modèle .

```
# In[15]:
    # Output a pickle file for the model

joblib.dump(svm, 'Model_of_Niveau_Univ.pkl')
```

Figure .29: SVM meilleur resultat

Le résultat :

```
In [14]: joblib.dump(svm, 'Model_of_Niveau_Univ.pkl')
Out[14]: ['Model_of_Niveau_Univ.pkl']
```

Figure .30 : Resultat

Remarque :

En ce qui concerne l'application d'algorithmes d'apprentissage automatique au dataset équilibrée, ce sont les mêmes étapes et les différences sont les suivant :

La précision de chaque algorithme

ac	float64	1	0.8076923076923077
ac2	float64	1	0.8141025641025641
ac3	float64	1	0.8205128205128205
ac4	float64	1	0.8076923076923077

Figure .31 : Resultat

Notez que Random Forest est le meilleur algorithme, nous gardons donc le modèle

Conclusion :

Dans ce chapitre, nous avons découvert Python et ses bibliothèques les plus importantes, et nous avons appliqué 4 algorithmes d'apprentissage supervisé

Ces algorithmes sont : Support Vector Machines (SVM), Tree Decision (DT), Random Forest (RF) et Naïve Bayes (NB).. sur les jeux de données et affichent leurs résultats pour obtenir un modèle.

Conclusion générale

Conclusion générale

Le but de cette mémoire est de réaliser une approche intelligente pour déterminer le niveau d'éducation des commentateurs dans les réseaux sociaux.

Pour appliquer cette approche, nous avons utilisé le programme Python parce qu'il contient des bibliothèques d'apprentissage automatique et il traite divers fichiers.

Dans le premier chapitre, nous nous sommes familiarisés avec l'intelligence artificielle et ses domaines les plus importants, y compris l'apprentissage automatique, les types d'apprentissage automatique sous supervision et sans supervision, et ses algorithmes les plus importants.

Où, dans la deuxième chapitre, nous avons défini le terme de réseaux sociaux, puis analysé les réseaux sociaux et les méthodes utilisées et nous avons parlé du data mining, et en particulier du text mining, de ses méthodes, objectifs et applications, avec un exemple appliqué en Python qui montre l'intérêt d'utiliser le concept de word bag en machine learning, qui transforme les textes en nombres pour faciliter le traitement des données.

Aussi, dans le troisième chapitre, nous avons commencé à mettre en œuvre l'approche intelligente, où nous avons commencé à collecter des données dans des fichiers csv, et nous avons utilisé deux types de dataset qui sont équilibrées et déséquilibrées en termes de nombre de commentaires et nous avons analysé et discuté des résultats de l'application d'algorithmes aux deux types de dataset.

Et puis définir les profils pour chaque niveau, ce qui nous a permis de créer facilement deux dictionnaires, un dictionnaire secondaire et un dictionnaire universitaire

Dans la partie implémentation utilisant Python, nous avons téléchargé les bibliothèques nécessaires, en particulier les bibliothèques d'apprentissage automatique, et appliqué 4 algorithmes sur les données, et nous avons constaté que le SVM est celui qui nous donne le meilleur résultat.

Références bibliographiques

[1] analyse des réseaux sociaux et communautés en ligne : quelles applications en marketing ? Maria Mercanti-Guérin Management Prospective Ed. | « Management & Avenir » 2010/2 n° 32 | pages 132 à 153 ISSN 1768-5958 Article disponible en ligne à l'adresse : ----------
---

https://www.cairn.info/revue-management-et-avenir-2010-2-page-132.htm Analyse des réseaux sociaux et communautés en ligne : quelles applications en marketing ? Maria Mercanti-Guérin Dans Management & Avenir 2010/2 (n° 32), pages 132 à 153 Mis en ligne sur Cairn.info le 17/05/2010
https://doi.org/10.3917/mav.032.0132

---

[2] Anil Jain and Douglas Zongker, "Feature selection : Evaluation, application, and small sample performance," IEEE transactions on pattern analysis and machine intelligence, vol. 19, no. 2, pp. 153–158, 1997.

[3] Antoine Cornuéjols, Laurent Miclet, Yves Kodratoff, *Apprentissage Artificiel : Concepts et algorithmes*, Eyrolles, 2002

[4] Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers par Rachid MIFDAL

[5] D. Boyd and E. Nicole. "Social Network Sites Definition, History," .Journal Of Computer—Mediated communication. 2008.

[6] Data Mining/ML Apprentissage non-supervisé Jamal Atif jamal.atif@dauphine.fr M2 ID Université Paris-Dauphine.

[7] Gilbert Saporta, Probabilités, Analyse des données et Statistiques, Paris, Éditions Technip, 2006, 622 p. [détail des éditions] (ISBN 978-2-7108-0814-5, présentation en ligne [archive])

[8] http://edutechwiki.unige.ch/fr/Communaut%C3%A9_virtuelle.

[9] http://www.intelligenceartificielle.fr/domaines_IA.php

[10] http://www.unitheque.com/Livre/springer/Regression_avec_R-37175.html.
http://cedric.cnam.fr/~saporta/multicolinearite.pdf
http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/tp_reg-penal_ukcomp.pdf
http://apiacoa.org/publications/teaching/data-mining/m2p6/supervised-slides.pdf

[11] https ://digitiz.fr/blog/reseaux-sociaux-definition-liste/ ,consulté le 16/03/20. " (Larousse, 1867)."

[12] https://fr.wikipedia.org/wiki/Apprentissage_non_supervis%C3%A9

[13] https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[14]https://www.24pm.com/117-guide-de-l-intelligence-artificielle/definitions/479-apprentissage-par-renforcement

[15] https://www.agence-web-cvmh.fr/reseaux-sociaux/

[16] https://www.mygreatlearning.com/blog/bag-of-words/

[17] https://www.piloter.org/business-intelligence/datamining.htm

[18] https://www.wizaly.fr/blog/post/machine-learning-quelles-applications"

[19] I "PLS-DA Classification Supervisée Julien JACQUES 26/09/2018"

[20] L. Lebart, A. Morineau, M. Piron, Statistique Exploratoire Multidimensionnelle, Dunod, 2000.

[21] M. Bardos, Analyse Discriminante - Application au risque et scoring financier, Dunod, 2001.

[22] M. Tenenhaus, Méthodes Statistiques en Gestion, Dunod, 1996.

[23] M.DJELLOULI,A.TABTI, Mémoire Analyse de Contexte des Personnes Sur Les Réseaux Sociaux : Analyse des sentiments sur Twitter 2017/2018.

[24] MÉMOIRE DE FIN D'ÉTUDES pour obtenir le grade de MASTER EN GÉNIE BIOMÉDICAL Spécialité : Informatique Biomédicale présenté et soutenu publiquement par Lakhdari Salsabil & Saidi Amaria

[25] Michel Volle, Analyse des données [archive], Economica, 4e édition, 1997,

[26] Mohamed-Rafik Bouguelia. Classification et apprentissage actif à partir d'un flux de données évolutif en présence d'étiquetage incertain. Intelligence artificielle [cs.AI]. Université de Lorraine, 2015. Français"

[27] O Chapelle, B Schölkopf, and A Zien, "Semi-supervised learning, ser. adaptive computation and machine learning," 2006.

[28] P. TORLOTING, enjeux et perspectives des réseaux sociaux paris ,2006. » (Divjak et Peharda, 2010 : 69)

[29] Romain Rissoan. Les réseaux sociaux Facebook. Twitter. Linkedin. Viadeo, Google+ Comprendre et maitriser ces nouveaux outils de communication, ENI. 2011

[30] support de cour pour etudiant en master en Intelligence artificielle