


République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Akli Mohand Oulhadj de Bouira
Faculté des Sciences et des Sciences Appliquées
Département d'Informatique

Mémoire de Master

En Informatique

Spécialité : *Génie des Systèmes Informatiques*

Thème

Annotation et classification automatique des articles d'actualité

Encadrer par :

M^r Bal Kamal

Réalisé par :

Bechar Amine

Tenfir Nassim

2019/2020

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

Remerciements spéciaux à nos chers parents pour leurs Patience, Amour, Soutien et Encouragement.

Nous voudrions spécialement remercier, notre encadreur de mémoire M^r Kamal Bal pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui nous ont été précieux afin de mener notre travail à bon port.

Nous remercions également toute l'équipe pédagogique du département informatique et les intervenants professionnels responsables de notre formation.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de prés et de loin à la réalisation de ce travail.

Dédicaces

Je dédie ce projet à
Mon père, à ma mère qui n'ont jamais cessé de formuler des prières à mon égard, de me soutenir et m'épauler pour que je puisse atteindre mes objectifs.
Sans oublier mon frère et ma sœur, merci pour votre soutien et vos encouragements.
A mon binôme Nassim, merci pour son entente et sa sympathie.

Bechar Amine

Dédicaces

Je dédie ce travail à
Mes très chers parents qui m'ont soutenu toute ma vie, aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez pour tous les sacrifices suprêmes que vous n'avez cessé de me donner depuis ma naissance.
Mes chers frères, je vous souhaite un avenir plein de réussite.
Mon binôme AMINE et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible, je vous dis merci.
Mes enseignants et surtout mon encadreur KAMEL BAL.
Mes amis et me collègues qui m'encourageait toujours.

Tenfir Nassim

Résumé

Quotidiennement, plusieurs agences de presse publient des milliers d'articles contenant plusieurs événements de toutes sortes (politique, économique, culturel, etc.), la classification automatique de texte est plus qu'essentielle à l'ère d'internet et des Bigdata.

Ce travail de recherche concerne l'annotation et la classification automatique des articles. Nous avons évalué et comparé notre système avec d'autres systèmes de classification automatique et constaté que l'étape de filtrage par la reconnaissance des entités nommées augmente l'efficacité du clustering.

Nous avons tout d'abord annoté ces articles afin de pouvoir extraire les entités nommées dans le but d'augmenter la précision du clustering et réduire le taux de données, ce qui permettra par conséquent la réduction du temps de traitement. Après nous avons appliqué un algorithme de clustering (Kmeans) sur un espace de données qui contient uniquement des E.Ns. Finalement, nous avons évalué notre travail en utilisant le coefficient de silhouette pour mesurer le score du clustering.

Mots-clés : Presse, Article, Classification automatique, Bigdata, Annotation automatique, Entités nommées, Clustering, Coefficient de silhouette .

Abstract

Every day, several press agencies publish thousands of articles containing several events of all kinds (political, economic, cultural, etc.), automatic text classification is now more than essential in the age of the Internet and Bigdata.

This research concerns the automatic annotation and classification of articles, we have evaluated and compared our system with other automatic classification systems and evaluated that the approch of filtering by recognizing the named entities increases the efficiency of clustering.

We first use the automatic annotation of these articles in order to be able to extract the named entities, in the aim to increase the precision of the clustering and to reduce the data space which lead to reduce the processing time as a result. then we applird of a clustering algorithm (Kmeans) on a data space containing only E.Ns. Finally, we evaluated our work using the silhouette coefficient to measure the clustering score.

Keywords : Press, Article, Automatic classification, Bigdata, Automatic annotation, Named entities, Clustering, Silhouette coefficient.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

API Application Programming Interface

Broadcasting Le partage d'audio ou de vidéos

DBSCAN Density-based spatial clustering of applications with noise

E.N Entité Nommé

FakeNews Fausse information ou fausse rumeurs

FD Fréquence de document

GI Gain d'information

IDE Integrated Development Environment

IDF Inverse Term Frequency

JSON JavaScript Object Notation

$K - PPV$ K Plus Proche Voisin

MAP Maximum à posteriori

ML Maximum Likelihood

MVT Model View Template

NEWS Terme designe les informations

NLP Traitement de language naturel

NLTK Natural Language Toolkit

NTIC Nouvelle Technologies de l'Information et de la Communication

OMS Organisation mondiale de santé

ReelTime Les événements qui se passe en temps réels

REN Reconnaissance des Entiées Nommées

SSE Sum of Squared Errors

SVM Machine à vecteurs de support

SVR Regression à vecteurs de support

TF Term Frequency

TV Télévision

URL Uniform Resource Locator

USA Etats Unies

Introduction générale

Motivations et problématique

Le monde de la presse et de l'actualité a toujours été parmi les centres d'intérêt partagés par l'immense majorité des citoyens. Savoir ce qui se passe autour de lui comme événements politiques, économiques, sociales, culturels et sportifs est une activité quotidienne chez l'homme.

Les supports de transmission et de diffusion des actualités (les médias) ont beaucoup évolué, allant de l'oral à l'écriture pour arriver aux moyens modernes journaux, radio et télévision pour arriver de nos jours à la diffusion en utilisant les nouvelles technologies, de l'information et de la communication (internet et réseaux sociaux).

L'utilisation des NTIC (Nouvelle Technologies de l'Information et de la Communication) pour la production et la diffusion des actualités a conduit à l'émergence, sur l'internet, de nouveaux vecteurs de diffusion de l'information journalistique comme les moteurs de recherche et autres services de veille spécialisés dans le domaine de l'actualité. De plus, les utilisateurs se trouvent en face d'une prolifération croissante des sources d'actualités et une très grande quantité d'informations d'actualités produites d'une manière continue. Il devient ainsi difficile pour un utilisateur, devant ce déluge d'actualité, de cerner facilement la synthèse de l'actualité qui l'intéresse.

Comment peut-on dans ce cas là aider l'utilisateur à détecter les différents sujets d'actualités traités ? Surtout que dans les faits, des centaines d'articles d'actualités issues de sources différentes peuvent concerner le même sujet d'actualité. Comment aider l'utilisateur à s'orienter vers des sujets de son intérêt (politique, économie, sport,. . .) ?

Nous envisageons à travers ce travail de répondre à ces préoccupations à travers le développement d'un système de classification automatique des articles d'actualité. Cette classification permettra d'une part de produire automatiquement une synthèse de l'actualité en regroupant les articles similaires et d'autre part de détecter automatiquement les sujets d'actualités.

Comme c'est connu, une bonne classification suppose au préalable une bonne représentation des données de base. Pour cette raison et vu l'importance des entités nommées (noms de personnalités, noms des lieux, noms des organisations, de pays, . . .) dans les contenus d'actualité, nous estimons qu'une représentation à base d'entités nommées des contenus des articles d'actualité permettra d'avoir de meilleurs résultats en terme de qualité de classification.

Objectifs à atteindre :

Le système de classification automatique des articles d'actualité que nous envisageons de développer vise à atteindre certain nombre d'objectifs que nous citerons ici :

- ✓ Offrir un accès facile et efficace au grand flux d'actualité mise en ligne quotidiennement.
- ✓ Orienter l'utilisateur vers les sujets d'actualités de son intérêt - Avoir une vue globale et organisée des principaux sujets d'actualités.
- ✓ Explorer facilement l'espace d'actualité à travers la génération automatique d'une synthèse d'actualité (détection des sujets d'actualités).

Organisation du mémoire

Ce mémoire est organisé en quatre chapitres :

- **1er chapitre : Média et actualité** : dans cette partie nous définissons le domaine de notre étude qui est les médias, leurs types ainsi que les domaines d'utilisations, comme nous touchons l'actualité en particulier.
- **2eme chapitre : Annotation et classification automatique :** ce chapitre présente un état de l'art sur l'apprentissage automatique. Il est divisé en deux catégories (apprentissage supervisé et non supervisé), nous définissons les stratégies et techniques utilisées dans ce domaine ainsi que les algorithmes disponibles pour chacune des catégories.
- **3eme chapitre : Système de classification automatique d'articles d'actualité :**c'est la partie de conception qui permet de définir une architecture générale de l'approche proposée en apportant notre contribution.
- **4eme chapitre : Implémentation et expérimentation :** cette dernière partie est consacrée à l'implémentation de notre système, à son expérimentation et évaluation ainsi qu'à la présentation de l'environnement et les outils utilisés.

Chapitre 1

Média et actualité

1.1 Introduction

Dans toute activité de recherche ou de développement, la maitrise des concepts du domaine étudié est une étape primordiale avant toute démarche de conception ou de développement. Dans ce sens, ce premier chapitre sera consacré à la présentation de notre domaine d'étude ainsi que tous les concepts qui y sont en relation. Il sera donc principalement question de présenter le domaine des médias en général et de l'actualité (ou des news) en particulier.

1.2 Les médias

1.2.1 Définition

Depuis l'Antiquité, les médias sont révélés comme l'un des outils les plus importants de transmission des informations et de l'actualité vers un public de masse. L'évolution de ce moyen est passé de plusieurs étapes. Au début le partage de l'information s'effectuait avec les peintures et des écritures. Mais maintenant, la création et l'invention des nouveaux modes et techniques de diffusion des informations ont révolutionné le domaine du partage et la communication des messages avec les auditeurs, pour le but de les sensibiliser et affecter leurs perceptions.

On trouve cette définition du terme média dans le dictionnaire LAROUSSE : « le terme média désigne tout moyen de distribution, de diffusion ou de communication, des œuvres, de documents, ou de messages écrits, visuels, sonores ou audiovisuels (comme la radio, la télévision, le cinéma, Internet, la presse, les télécommunications, etc. ».[1]

1.2.2 Types des Médias

FIGURE 1.1 – Les types des médias [2]

Avant l'apparition des moyens technologiques les plus répandus de maintenant tel que l'internet et la télévision. La presse écrite était le seul moyen d'interaction et de partage de l'information avec l'audience.

Mais Aujourd'hui, après le développement extraordinaire de l'internet, on peut maintenant avec un simple clic de surfer dans des sites WEB et de trouver touts types d'information et d'actualité en $ReelTime$. Par exemple (les flashs news, les conférences de presse, les résultats et les statistiques des matchs de football en direct ex...)

Les médias peuvent être classés en trois grandes familles :

1.2.2.1 : Les Médias écrites (imprimés)

A. **Journaux (Newspapers)** – imprimé qui contient un ensemble d'articles et de publication, présente l'actualité liée à tous les domaines, publié sur une base quotidienne ou hebdomadaire, c'est l'un des moyens les plus importants de Print Media.

B. **Magazine** – imprimé sur une base trimestrielle ou annuelle, elle contient des informations sur l'économie, la finance, la nourriture, la mode, etc. [3]

1.2.2.2 : Les Médias de diffusion (ou de Broadcasting)

Appelé aussi un média de l'offre, c'est un ensemble de moyens les plus répandus de divertissement et de communication des informations avec le grand public dans ces dernières années, elle signifie le partage des contenus audio ou audiovisuels.

A. **Télévision** – dans le passé, il y avait quelques chaînes qui partageaient des contenues générales, comme les chaines terrestres. Mais avec l'apparition des chaines satellites, on a aperçu l'augmentation et diversification des types de contenu que chaque chaîne propose. On voit donc des chaines se spécialisent dans l'actualité, les films, les sports, la nature etc....[4]

   C'est le média de diffusion le plus important en raison de sa portée auprès du public dans le monde car au moins 1,67 milliard de foyers disposait d'au moins un téléviseur en 2020. [5]

B. **Film** – les films, les scénarios, les images animées ont une accessibilité mondiale, c'est le meilleur type de médias de masse pour envoyer des messages ciblés et influencer les grandes audiences. L'industrie du cinéma joue un rôle très important dans la diffusion de la conscience sociale. [3]

C. **Radio** – c'est un moyen d'émission des news, la météo et aussi même un outil de divertissement, utilisée aussi pour la publicité et même pour la sensibilisation. C'est un média important dans nos jours. Les informations sont transmises par des ondes radio. [6]

1.2.2.3 : Internet

L'outil Internet a évolué pour devenir un média de la demande, c-à-d l'utilisateur qui choisit le type de contenu et d'actualité que souhaite voir et consulter. Par contre les médias de l'offre comme le Broadcast Média, ils offrent le type de contenu, ce qui a permis l'augmentation de l'utilisation des moteurs de recherche. [3]

A. **Réseaux sociaux ou sites Web** – incluant Facebook, Instagram, Twitter, YouTube, Tumblr, LinkedIn, Snapchat, Quora, Reddit, Pinterest, etc.

   Ils sont largement utilisés par les gens du monde entier, donc on peut trouver des actualités exclusives dans un temps réel.

   Mais ce moyen contient des inconvénients tel que le partage des fausses rumeurs, aussi des $FakeNews$ qui peuvent aider à déstabiliser toute une société.

B. **Les Forums** – un endroit en ligne où se groupe un ensemble de gens pour le but de poster des publications, commenter, envoyer des messages ou discuter d'un sujet particulier. Les forums nous permettent de partager nos connaissances avec d'autres personnes ayant le même intérêt.

1.2.3 Types de contenus médiatiques

- **Actualités :** Ce sont les informations et news sur ce qui se passe dans le monde, on trouve plus de détails dans la section (1.3 Actualité).
- **Événement courant :** Couverture d'un grand événement comme (Tournoi Roland Garos) du sport de Tennis.
- **Divertissements :** Des histoires sur la musique, l'industrie du cinéma ou encore le théâtre, et si ce que cela vaut la peine de voir ou non.
- **Rapports détaillés :** Par exemple (documentaires, un reportage d'investigation sur une catastrophe naturelle).

- **Opinions :** Article qui présente l'opinion d'une personne de façon subjective. Ce type permet de partager un avis d'un expert sur un cas, par exemple "un politicien donne sa vision sur laquelle va gagner les élections ? ".[7]

1.2.4 Les plus grands propriétaires de médias au monde

Position	Nom	Revenu	Description
1	Alphabet	$136.81 milliards	C'est une entreprise américaine précédemment détenues par la société Google, elle domine ce classement avec une large distance. La plupart de ses revenus viennent des services de publicité comme AdSence
2	Comcast	$94.51 milliards	88 % de ses revenues vient de la diffusion des programmes de télévision
3	The Walt Disney Company	$59.43 milliards	C'est l'une les plus grandes compagnies de cinéma et de création des dessins animés, la plupart de ses revenus viennent de la publicité
4	Facebook	$55.01 milliards	La migration de la population vers l'internet et l'apparition des smart-phones ont permis à Facebook d'être l'un des médias les plus importants dans dernières années
5	21st Century Fox	$30.4 milliards	C'est une grande entreprise de divertissement dans le monde, elle est possédée par la compagnie Fox

TABLE 1.1 – Classement des organisations selon les statistiques de la société *Statista* en 2018[8]

1.3 L'actualité :

1.3.1 Définition

Dans Wikipédia on trouve qu'une actualité ou une actu, c'est un message ou une information d'un événement récent qui s'est passé dans la journée ou la semaine, délivré par les médias vers le grand public.

Dans ce moment actuel, il se passe un grand mélange d'événements dans tous les domaines de la vie, ce qui causera un chaos d'information. Heureusement, les journalistes essayent de structurer ce chaos de sorte que chaque jour, le public reçoit les nouvelles bien triées et soignées. Les nouvelles seront publiées dans les médias, tel que la télévision et les journaux. [9]

On trouve cette définition du terme actualité dans le dictionnaire LAROUSSE : « Événements **actuels** intéressant un domaine d'activité ».

1.3.2 Critères d'actualité

Il existe une priorité d'affichage des informations dans les médias. Les nouvelles importantes qui feront le buzz seront affichées en premier et avec des détails, les nouvelles moins importantes seront affichées en dernier et avec moins de détails.

La distinction de l'importance des informations se fait par le jugement des journalistes, en se basant sur le niveau d'intérêt de la société et l'importance relative de l'événement, ceci les critères d'actualité les plus importants pour juger la qualité d'un article :[9]

A. **La nouveauté de l'actualité :** le facteur de temps est très important pour déterminer l'importance d'une nouvelle, chaque heure passe, l'information ne sera pas nouvelle, donc le public ne le donnera pas le niveau d'intérêt qu'elle le mérite.

B. **L'irrégularité (Inhabituelle) :** ce sont des nouvelles qui ne se produisent pas souvent où se passe rarement. Cependant, un événement habituel varie d'une société à une autre, par exemple, dans notre société : "Un chien mord l'homme" n'est pas une nouvelle, mais "L'homme mord le chien" est une nouvelle.

   Mais dans des sociétés comme en Chine, ce n'est pas une nouvelle car c'est habituel de manger des chiens dans leur société, donc cette information ne fera pas l'actualité.

C. **Contenu intéressant :** par exemple une information comme "Les scientifiques ont trouvé de l'eau dans mars ou dans une autre planète " est une information nouvelle et inhabituelle, mais le problème elle ne trouvera pas le centre d'intérêt par le public.

   Par contre une information comme "La chute de 70% des prix de pétrole" est une information qui fera le bruit car elle touchera directement la situation financière de la population d'un pays pétrolier.

D. **L'exclusivité :** chaque journal ou chaine TV ont des sources dans chaque grande entreprise dans le monde, ces sources leur donnent des informations exclusifs. Par exemple "Le président va démissionner la prochaine semaine ".

   Ce type de nouvelle va beaucoup attirer l'attention du public de masse. Donc elle fera la Une dans les journaux ou dans les chaines d'information.

E. **La proximité :** La place ou se passe l'événement est important pour les lecteurs ou les auditeurs, par exemple "Le président des USA a fraudé les élections pour gagner la présidence" fera un bruit énorme dans le monde et l'information se propagera très rapidement. Mais si ce scandale a eu lieu dans un petit pays, il ne fera pas la Une dans les grands médias du monde.

Ceux-ci étaient quelques critères principaux qui vont définir si l'information est une information ordinaire ou une information d'actualité, et décider si cet article sera dans les premières pages ou dans les dernières.

1.3.3 Les domaines d'actualité

Ceux-ci sont les domaines les plus fréquents que le public attende des nouvelles [10] :

A. **Conflits :** les guerres, les luttes militaires et les coups d'État feront toujours un sujet important des médias, par exemple l'invasion *Nato* en Libye l'année 2011.

B. **Politique :** couvrir les élections et les luttes pour acquérir le pouvoir politique est l'une des topics les plus importants couvertes par les médias.

C. **Économie :** comporte des informations sur les crises économiques dans le monde aussi le prix des matières premières dans les bourses mondiales. Ce domaine est très important, il attire l'attention des auditeurs parce que ces crises toucheront directement leur mode du vie. Par exemple la crise économique mondiale de 2008 et la crise du pétrole en 2013 ont fait la Une dans ces périodes.

D. **Catastrophe et tragédie :** comporte deux types :.

   - Catastrophe humaine :comme l'accident d'usine de *Chernobyl* en 1986
   - Catastrophe naturelle :comme le séisme d'Haïti en 2010.

E. **Crime :** comporte tout type de crime comme le meurtre et le viol. Généralement les meurtres inhabituels feront la Une dans les médias.

   Aussi la couverture des procès des grandes personnalités politiques et économiques fera aussi la Une dans les médias.

F. **Personnes célèbres :** les célébrités sont des influenceurs, ils attirent l'attention du public lorsque par exemple ils visitent des lieux ou des pays des autres continents, faire les dons, ou aussi leur implication dans les scandales. Alors reporter ces informations augmentera le nombre de vues de ces articles.

G. **Santé :** reporter l'état et les statistiques de la propagation du coronavirus dans le monde, et aussi reporter les conférences de presse, les déclarations et décisions des chefs d'État pour combattre ce virus pour un seul but "sensibiliser" les gens.

H. **Sport :** les médias spécialisés en sport essayent toujours de couvrir les sports populaires comme le football en Europe ou le basketball en USA. Pendant la période estivale des transferts de joueurs, les médias exploitent les sentiments des lecteurs envers les clubs qu'ils aiment, en partagent beaucoup de rumeurs pour attirer leurs attentions. Par exemple le titre "Riyad Mahrez rejoint Manchester City pour 60m euro" a fait la Une des presses en 2017.

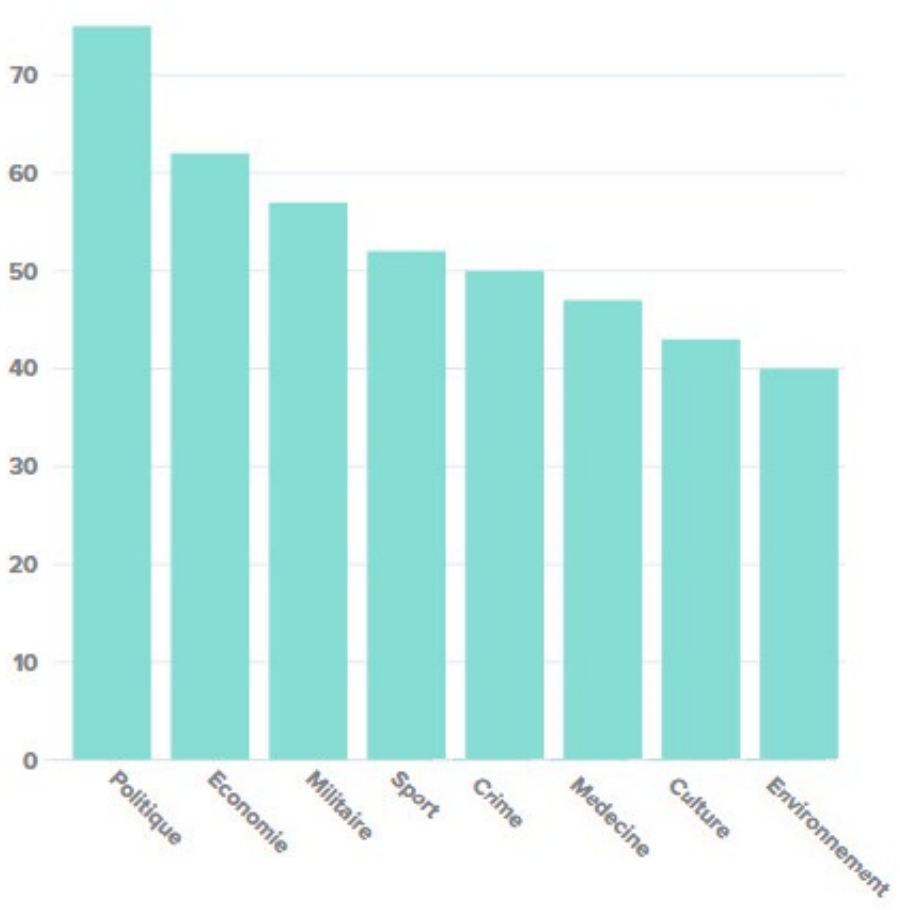

FIGURE 1.2 – Les topics le plus consulté par le public de masse [11]

1.3.4 Diffusion des actualités

Récemment, il y a eu un déplacement de la consommation de l'actualité par le public vers Internet, ce qui a baissé le nombre de vues des informations diffusées traditionnelles par l'audience. Le journalisme imprimé est donc menacé.

Une partie de la migration de l'audience a été vers les journaux en ligne associés à la presse. En outre, il existe des systèmes tels que les services de recherche de nouvelles, les moteurs de recherche et les agrégateurs de nouvelles, qui permettent d'accéder à un ensemble de sources d'information. Ce dernier ne constitue cependant pas une menace pour les organisations de médias existantes, tant que ces organisations s'adaptent à l'internet pour suivre leur public.

Les diffuseurs traditionnels tentent de se faire concurrence en fournissant des informations de dernière minute sur des développements importants, ceci mène que les nouvelles du jour ne sont pas fixes au moment des derniers tirages mais varient continuellement.[11]

1.3.5 Classement des médias de l'actualité les plus consulté au monde

Pays et régions	Moy Durée de la visite	Moyenne de visite par mois	Position
yahoo.com	00 :07 :36	1.88 milliards	1
cnn.com	00 :12 :54	587 millions	2
msn.com	00 :17 :43	321 millions	3
foxnews.com	00 :04 :41	356 millions	4
news.google.com	00 :07 :46	283 millions	5

TABLE 1.2 – Classement des sites les plus populaires selon le trafic des visiteurs, de *similarweb.com* en 2020 [12]

1.4 Conclusion

Au niveau de ce chapitre, nous avons donné une description bien détaillée sur notre domaine d'étude, en présentant le domaine des médias et de l'actualité avec tout ce qui leur concerne.

On a vu que quotidiennement, plusieurs agences de médias publient des milliers d'articles contenant toutes sortes d'événement comme (politique, économique, santé, etc...), ces articles seront consultés par des milliers d'auditeurs dans le monde. On a constaté que plusieurs critères déterminent si ces articles sont de qualité, afin d'attirer l'attention du public.

Afin de rendre toute cette actualité visible et accessible facilement aux utilisateurs, des outils basés sur la classification automatique d'articles d'actualités doivent être développés.

Chapitre 2

Annotation et classification automatique des documents

2.1 Introduction

Après avoir présenté dans le chapitre précédent le domaine des médias et de l'actualité, nous allons présenter les concepts d'annotation et de classification automatique de données textuelles. Ces techniques sont largement utilisées dans le domaine du traitement et d'analyse d'information textuelle. Nous estimons que ces deux techniques seront largement recommandés dans le développement de tous systèmes qui permet de bien organiser les flux d'actualités et de permettre un accès efficace et rapide aux grands volumes de news disponibles actuellement.

2.2 Annotation automatique de texte

Souvent, quand on veut analyser des données, on les trouve généralement sous forme numérique comme par exemple les données des ventes, les mesures physiques (comme la masse, le poids ...) et les catégories quantifiées. Les machines peuvent facilement analyser les données numériques, mais la question qui se pose, comment analyser les données de type "texte" ?

Pour cela, afin de bien analyser un texte, il faut préparer les données de sorte que la machine trouve facilement les patterns et les inférences, plusieurs techniques ont été créé pour le traitement de texte.

2.2.1 Définition :

L'annotation est une technique qui consiste à ajouter des **méta-données** aux mots d'un texte ou document pour réaliser des tâches tel que l'extraction des entités nommées et la détection des frontières des phrases. On donne ci-dessous quelques définitions de l'annotation :

Dans Le Robert :

- **Annotation** : Note critique ou explicative qui accompagne un texte
- **Annoter** : Accompagner un texte de notes critiques ; mettre sur un livre des notes personnelles.

En informatique :

Une annotation est un commentaire, une note, une explication ou tout autre remarque externe qui peut être attachée à un document ou à une partie de celui-ci. L'annotation textuelle consiste à enrichir un texte avec des informations, rattachées aux parties du texte. Le mot annotation signifie à la fois :

– Le processus dans lequel le texte est enrichi avec des informations supplémentaires.

– Les informations qui sont rajoutées au texte.

Pendant l'annotation, certains éléments textuels (mots, expressions, phrases, ...) sont étiquetés par un ensemble de catégories d'annotation, suite à une analyse des propriétés de l'élément annoté selon une méthode donnée. Les annotations expriment certaines propriétés des éléments textuels, par ex. des catégories grammaticales, le contenu sémantique, etc. en suivant un modèle.

L'annotation est une étape très essentielle du pré-traitement des données dans la classification automatique des textes.

2.2.2 Processus d'annotation :

L'annotation d'un document passe par plusieurs étapes

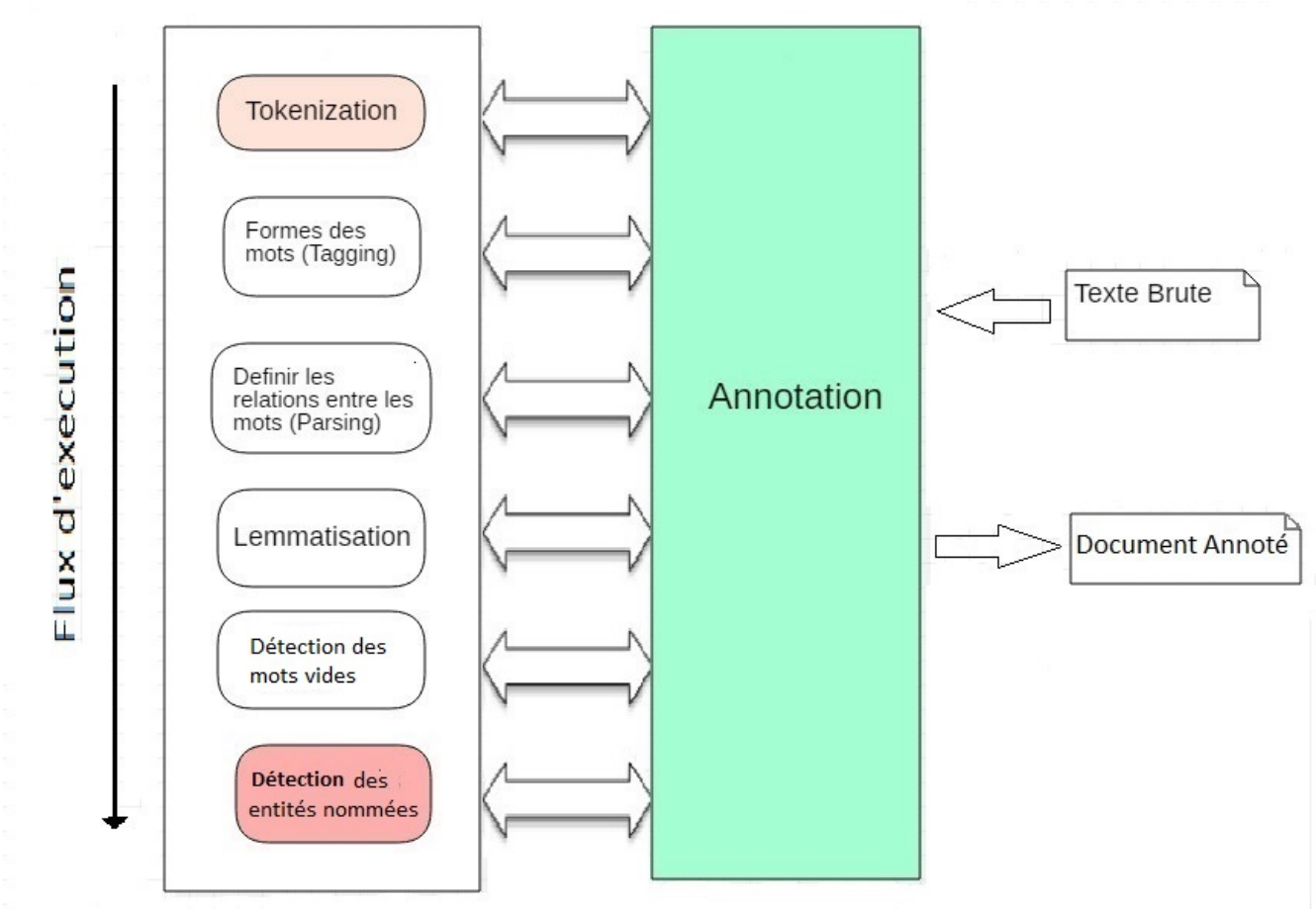

FIGURE 2.1 – Les étapes d'annotation d'un texte brute [13]

2.2.2.1 : Tokenization :

La Tokenization est une tache qui permet de diviser un texte en plusieurs composants qui s'appelle *token*. Les tokens sont des blocs de base qui constitue un document, les tokens peuvent être des mots, des nombres ou des ponctuations.[14]

La tokenization d'un texte se passe par plusieurs étapes :

A. **Suppression des espaces** : Lors de la division du texte, la tokenization ne considère pas les espaces comme token.

   Par exemple : la phrase "Bienvenue à Bouira" se compose de trois tokens : "Bienvenue", "à" et "Bouira"

B. **Détection des préfixes** : Ce sont les caractères spéciaux (ponctuations) qui viennent en début des phrases.

   Par exemple parenthèse ouvrante "(" les guillemets '"'...

C. **Détection des suffixes** : Ce sont les caractères spéciaux (ponctuations) qui viennent à la fin de chaque phrase.

   Par exemple parenthèse fermente ")" les guillemets '"'...

D. **Détection des infixes** : Les règles d'exception comme les apostrophes,

   point d'exclamation "!", d'interrogation "?" ...

Ci-dessous un exemple de division d'une phrase en plusieurs tokens :

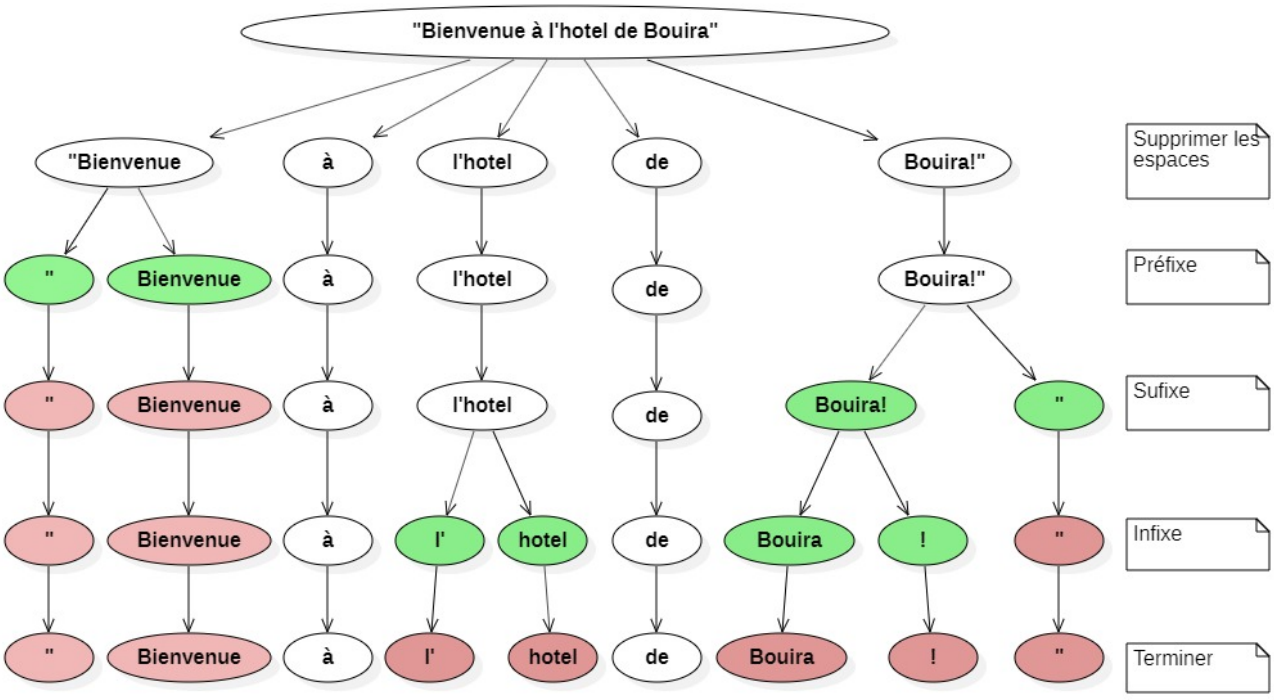

FIGURE 2.2 – Les étapes de Tokenization d'un texte

2.2.2.2 : Définir les formes grammaticales des mots : (Tagging)

Ce processus permet de correspondre chaque token à sa forme grammaticale, exemple :nom, verbe, auxiliaire.

L'annotation utilise un modèle de prédiction pour déterminer quelle est la forme du mot dans la phrase. Par exemple les mots qui suivent "le, les..." sont des noms, mais cette prédiction peut être trompeuse comme dans cet exemple : "je le trouve sympa". [15]

Ceci un exemple de la forme grammatical des mots d'une phrase :

L'	Algerie	est	le	plus	grand	pays	du	continent	africain.
DET	NOUN	AUX	DET	ADV	ADJ	NOUN	DET	NOUN	ADJ

FIGURE 2.3 – Détection des formes grammaticales de mots en utilisant python

2.2.2.3 : Dépendance grammaticale des mots dans une phrase (Parsing) :

Ce processus permet de trouver les relations entre les mots, par exemple : sujet, verbe, complément, groupe nominale...[15]

Exemple de représentation des dépendances des mots :

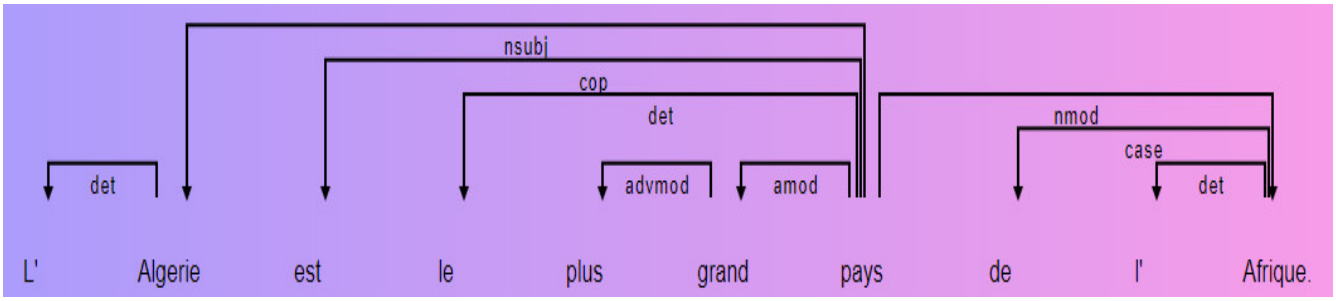

FIGURE 2.4 – Les dépendances des mots d'une phrase en utilisant python

Cette opération aide à détecter les frontières des phrase.

2.2.2.4 : Lemmatisation :

La lemmatisation consiste à analyser les termes de manière à identifier leurs formes canoniques (lemmes) afin de réduire les différentes formes (pluriel, féminin, conjugaison, etc.).[13]

Par exemple le lemme des mots "économie", économiquement", "économe", et "économétrie" est "*ÉCONOM*".

2.2.2.5 : Détection des mots vides :

Cette étape sert à détecter tous les mots qui ne sont pas porteurs de sens sur le plan sémantique et lexical, par exemple : les prépositions, les déterminants, les adverbes... comme "la, le, dans, car, pour, cependant, etc" ...

2.2.2.6 : Détection des entités nommées :

Ce sont tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (politique, économique, géographique, etc.) tel que les noms propres, des expressions de temps ou de quantité.[16] Cette opération a pour objectif de repérer et catégoriser tout contenu dans un texte, il existe plusieurs approches pour réaliser la détection des entités nommées :

A. **Symbolique** : Approche à base des règles, cette approche est très utilisée par les processus de détection des entités nommées. Les règles d'extraction sont écrites par des experts du domaine linguistiques.

B. **Statistique** : C'est une approche appelée aussi "approche par apprentissage", elle utilise des processus automatiques pour l'extraction d'information. Son principe est de mettre en points des modèles d'analyse à partir d'une grande masse de données.

C. **Hybride (Symbolique + Statique)** : Elles permettent de comprendre que dans une phrase comme « Orange n'est pas cotée en bourse », « Orange » réfère à une entreprise, alors que dans « Notre voyage à Orange s'est bien terminé », « Orange » réfère à la ville et que dans « J'ai fait de la confiture à l'orange », « Orange » réfère aux fruits et non pas à une entité nommée comme dans les deux précédents. Cette forme aide à traiter les homographes (les mots qui ont la même forme écrite, mais qui ont une signification différente).[17]

Voici un exemple d'une phrase contenant des entités nommées :

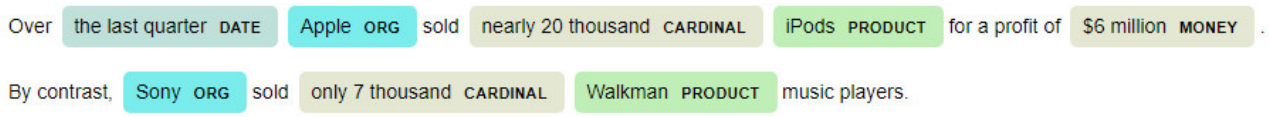

FIGURE 2.5 – Exemple de détection des entités nommés en utilisant Spacy

2.3 Classification automatique des documents

2.3.1 Objectif de la classification

Comme le nombre et le volume des documents numériques s'accroissant de façon exponentielle, on a besoin de les catégoriser afin de faciliter leur manipulation. La classification de textes a pour objectif de regrouper les textes similaires, c'est-à-dire thématiquement proches, au sein d'un même ensemble. En d'autres termes, trouver un algorithme permettant d'assigner un texte à une classe avec le plus grand taux de réussite possible. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'information efficace.

2.3.2 Classification supervisée vs non-supervisée

La classification automatique cherche à répartir un corpus en groupes de documents (catégories, classes, clusters) de façon à mettre ensemble les documents qui se ressemblent et de séparer celles qui diffèrent. Deux méthodes de classification sont utilisées, la méthode supervisée et la méthode non supervisée.

2.3.2.1 : Classification supervisée :

C'est une méthode qui consiste à attribuer une ou plusieurs classes à un document dont les classes sont connues à priori, c'est-à-dire basé sur des documents d'entrée et de sortie étiquetées.

Cette méthode permet à un système d'être capable de prédire la classe d'un nouveau document non classé (non étiqueté) à base d'un ensemble de descripteur. **La performance de la classification dépend toujours de l'efficacité de la description.**

2.3.2.2 : Classification non-supervisée (Clustering) :

Le clustering est une méthode d'apprentissage **non-supervisé** permettant de trouver des patterns dans les données. Par contre à la première approche c'est que dans cette approche les classes ne sont pas connues à priori, les documents ne sont pas étiquetés au préalable.

Ça signifie qu'un regroupement ou partitionnement des documents en fonction de leurs similarités (regroupement des documents qui se ressemblent). On dispose des documents non classés dans le but de trouver des modèles communs, il existe plusieurs stratégies de construction des clusters :[18]

A. **Clustering hiérarchique**

Dans ce cas on va faire une décomposition en arborescence des groupes. La sortie principale du clustering hiérarchique est un dendrogramme, qui montre la relation hiérarchique entre les clusters, il se divise en deux types :[19, 20]

- **Clustering agglomératif :**

  Dans le cas du clustering agglomératif (ou bottom-up), on considère que chaque cluster est indépendant, puis on cherche les deux clusters les plus proches, et on les agglomère en un seul cluster, on répète l'opération jusqu'à que tous les documents appartiennent à un seul cluster.

- **Clustering divisif :**

  C'est une approche inverse de la précédente, le clustering divisif (ou top-down), on considère que la collection des documents appartienne à un seul cluster, on la décompose jusqu'à ce que chaque document appartient à son propre cluster.

B. **Clustering non-hiérarchique**

Le clustering non hiérarchique vise à trouver un regroupement qui sert à décomposer l'ensemble des documents en K groupes. On cherche toujours à minimiser la distance de similarité entre les documents de même groupe. Cette similarité est calculée à base de certaines mesures, chaque document est représenté comme un point de l'espace ayant pour coordonnées ces mesures. On cherche toujours à regrouper les documents qui se ressemblent comme le clustering via les centres mobiles (exp :K-means). [18]

2.3.3 Processus de classification automatique des documents

La classification des documents passe par plusieurs étapes :

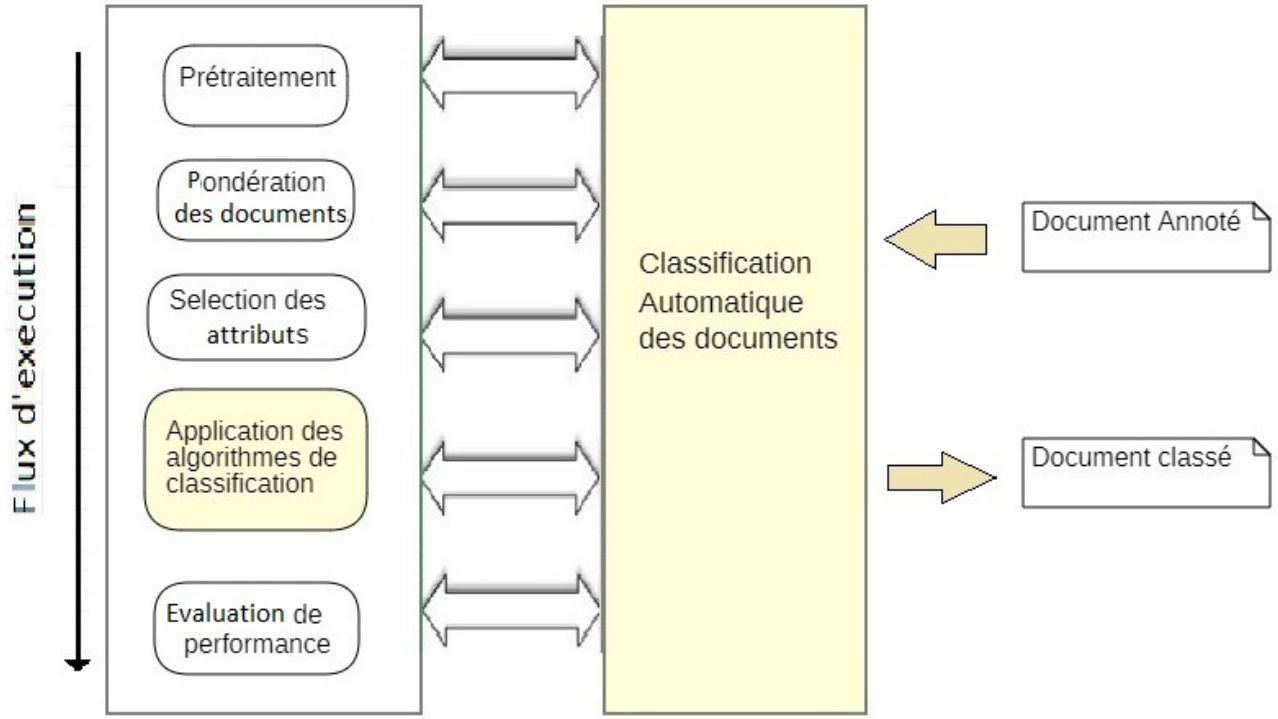

FIGURE 2.6 – Le processus d'une classification automatique de texte [21]

2.3.3.1 : Pré-traitement :

Les algorithmes d'apprentissage ne sont pas capables de traiter directement les données textuelles, c'est pourquoi qu'une étape de représentation est nécessaire. Cette étape consiste à représenter chaque document par un vecteur qui contient tous les mots présents au moins une fois dans le document. Une collection des documents (corpus) peut être représentée par une matrice.

Une fois l'annotation des documents est faite, on obtient des documents repérés en tokens. À partir de ces derniers on va construire un dictionnaire qui s'appelle le "dictionnaire des termes", qui est composé de tous les termes qui sont présents dans le corpus. La matrice document terme contient des lignes correspondant aux documents et colonnes correspondant aux termes.

Un inconvénient de cette représentation est que plus le corpus est grand plus la matrice devient grande et couteuse au niveau de l'espace mémoire et au niveau du temps de traitement, il est cependant facile de constater que certains de ces tokens sont présents dans tout le corpus mais ne sont pas porteurs de sens et n'apportent aucune influence sur la classification, ce sont **les mots vides**. Par contre, leur suppression réduit le coût du travail. [22]

Le résultat de cette première étape est une représentation matricielle du corpus.

2.3.3.2 : Pondération des articles :

C'est une étape qui permet d'assigner un poids pour chaque terme dans un document. Elle désigne le nombre de fois qu'un certain descripteur (terme) est apparu dans chacun des documents d'un corpus. À partir de ce nombre on peut dire si un descripteur est discriminant ou non par rapport à un document donné.

Cette étape permet aussi de caractériser les termes importants dans un document, l'idée est que les termes importants doivent avoir un poids important. Il existe plusieurs techniques de pondération comme le montre le tableau ci-dessous :

Technique	**Définition**	**Forme**
Pondération binaire	Comptabiliser la présence de chaque terme dans le document, sans se préoccuper du nombre d'occurrences (de la répétition- TF booléenne)	TF = 1 ou 0
Fréquence des termes-TF (Term fréquency)	TF désigne la fréquence d'un terme (descripteur) dans un texte donné	TF absolu : TF= NT NT est le nombre de fois où le terme est apparu dans le texte. TF relative : TF =NT/ST NT est le nombre de fois que le terme est apparu dans le document. et, ST est le nombre de tous les termes du document.
Fréquence documents inverses (IDF)	Elle mesure le degré de rareté d'un terme, non pas dans un document, mais dans tous les documents (l'influence d'un terme)	$IDF(T, D) = \log_{10} \frac{N}{NT}$ D : corpus N : nombre de documents dans le corpus NT : nombre de documents où le terme apparaît
Pondération TF-IDF	Relativiser l'importance d'un terme dans un document (TF) par son importance dans le corpus (IDF).	TF-IDF(T,d,D)= TF(T,d)*IDF(T, D) D : document quelconque (TF relative ou absolue)

TABLE 2.1 – Les différentes techniques de pondération [22]

2.3.3.3 : Sélection des attributs

Afin d'améliorer l'efficacité et la précision de la classification automatique, on utilise la technique de sélection des sous-ensembles d'attributs (**Feature Selection**).

La sélection des attributs consiste à trouver des sous-ensembles des termes les plus pertinents (important) en utilisant des techniques *comme fréquence de document (FD), le gain d'information (GI).* Ces techniques vont calculer un score pour chaque terme qui servira comme indice de pertinence. Les termes seront donc triés en ordre décroissant pour le but de choisir les mots les plus pertinents selon des critères prédéfinis.[23]

Cette étape permet de régler des problèmes de grandes dimensionnalités de la matrice document-terme, ci-dessous on va présenter quelques techniques de sélection des attributs :

A. **La sélection des attributs basée sur la fréquence de documents(FD) :**

Le principe de cette méthode est de calculer le nombre de fois qu'un certain terme est apparu dans chacun des documents d'un corpus, cette technique va filtrer les termes ayant une fréquence inférieure à un seuil prédéterminé. Le but de cette technique c'est d'éliminer tous les termes inutiles qui n'ont pas d'influence sur la classification des documents.

B. **La sélection des attributs basée sur le gain d'information :**

Cette méthode vise à faire la réduction du vocabulaire en se basant sur la présence ou l'absence d'un terme dans un document. Cette technique calcule la valeur de l'**entropie** (quantité d'information pour chaque document). C-à-d il existe une corrélation entre la valeur de l'entropie et la variété des informations dans un document.[23]

Le gain d'information d'un terme T dans un ensemble de documents D se calcule comme suit :

$$GAIN(D,T) = Entropie(D) - Entropie(T) \tag{2.1}$$

Avec :

$$Entropie(D) = -\sum_{i=1}^{c} d_i \log_2 d_i \tag{2.2}$$

Soit c le nombre de catégories dans lesquelles les documents seront classés.

d_i est le nombre de documents dans D appartenant à la catégorie i.

Et :

$$Entropie(T) = \sum_{v\in 0,1} \frac{|D_v|}{|D|}.Entropie(D_v) \tag{2.3}$$

Soit v la valeur de présence (v=1) ou d'absence (v=0) du terme T dans les documents.

$|D_v|$ le nombre de documents dans D à qui le terme T appartient si (v=1), ou n'appartient pas si (v=0).

et l'entropie de D_v est égale à :

$$Entropie(D_v) = -\sum_{i=1}^{c} d_{vi} \log_2 d_{vi} \tag{2.4}$$

Ou d_{vi} est le nombre de documents dans D appartenant à la catégorie i où la valeur du terme T est égale à (0 ou 1).

2.3.3.4 : Application d'un algorithme de classification :

L'application d'un algorithme de classification sur un corpus de documents est réalisée avec des algorithmes spécifiques. Elle consiste à trouver un modèle mathématique capable de représenter, puis comparer la sémantique des textes. On citera la classification par catégorie, par genre ou bien d'autres critères qui sont définis au préalable, il existe plusieurs algorithmes qui engendrent ses problématiques de classification et que l'on va les présenter dans la section 2.3.5.

2.3.3.5 : Évaluation de performance :

C'est la dernière étape de la classification de texte, c'est une technique qui permet d'évaluer la performance d'une approche ou d'un système de classification, c'est une évaluation comparative entre approches pour déterminer les mieux performant. Il existe une mesure de performance pour l'apprentissage supervisé et non-supervisé :

A. **Pour l'apprentissage supervisé** :

Ce processus permet de mesurer la performance de la classification, elle mentionne la capacité à sélectionner des documents pertinents en utilisant deux facteurs, le rappel(ou valeur prédictive positive) et la précision(ou sensibilité) :[24]

- **Rappel :**

  La capacité à sélectionner tous les documents pertinents de la collection. C'est le ratio du nombre des documents pertinents sélectionnés par le nombre total des documents pertinents :

$$Rappel = \frac{\mid documents\ pertinents \cap documents\ selectionnes \mid}{documents\ pertinents} \tag{2.5}$$

- **Précision :**

  La capacité à sélectionner que des documents pertinents. C'est le ratio du nombre des documents pertinents sélectionnés par le nombre total des documents sélectionnés. La formule de la précision :

$$Precision = \frac{\mid documents\ pertinents \cap documents\ selectionnes \mid}{documents\ selectionnes} \tag{2.6}$$

- **Score :**

  Le score se calcule avec cette formule :

$$Score = \frac{Rappel\ *\ Precision}{(Rappel\ +\ Precision)/2} \tag{2.7}$$

B. **Pour l'apprentissage non-supervisé** : La mesure de performance dans l'apprentissage non supervisé s'appuie sur deux critères :

- **La forme des clusters :** la forme des clusters qu'un clustering produit (sont-ils denses, bien séparés). On utilise ici souvent le coefficient de silhouette qui est souvent utilisé pour déterminer qu'un algorithme de clustering vérifier l'homogénéité (faible similarité inter-groupe) et la séparation (grande similarité intra-groupe). Pour un point x donné, le coefficient de silhouette s(x) permet d'évaluer si ce point appartient au « bon » cluster. il est donné par :[25]

$$S(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))} \tag{2.8}$$

Avec :

$$a(x) = \frac{1}{|C_k| - 1} \sum_{u \in C_k, u \neq v} d(u, x) \tag{2.9}$$

Tel que a(x) est la distance moyenne entre x(article) et tous les autres points du cluster C_k auquel il appartient et $d(u, x)$ la distance entre x et tout points différents de x d'un même cluster.
Et :

$$b(x) = \min_{l \neq k} \frac{1}{|C_l|} \sum_{u \in C_l} d(u, x) \tag{2.10}$$

Tel que b(x) est la plus petite distance moyenne entre *x (article)* appartient au cluster C_k et tous les autres points d'un autre cluster C_l et $d(u, x)$ la distance entre x et tout points du cluster C_l.

Si $\mathbf{a(x) < b(x)}$ alors *x (article)* a été correctement assigné à son cluster.

Le coefficient de silhouette est donc compris entre -1 et 1, et d'autant plus proche de 1 que l'assignation d'un article x à son cluster est satisfaisante.

Pour évaluer la classification (clustering), on peut calculer son coefficient de silhouette moyen :

$$S_{moy} = \frac{1}{n} \sum S(x) \tag{2.11}$$

Tel que n est le nombre de tous les points.

- **La stabilité de l'algorithme(stabilité des clusters) :** l'un des critères les plus importants pour mesurer la performance qui signifie que l'algorithme doit être déterministe autrement dit, les résultats sont les mêmes à chaque exécution (implémentation) de l'algorithme.[25]

2.3.4 Mesures de similarité utilisée dans le cadre de la classification :

Avant toute application d'un algorithme de classification, le choix d'une mesure de similarité qui permet de comparer chaque paire d'éléments du corpus est nécessaire. Dans ce cadre, plusieurs mesures de similarités sont utilisées. Nous citerons quelques une dans ce qui suit : [26]

document	Covid-19	Fièvre	Virus	Toux	Mort
1	1	1	1	0	0
2	1	0	2	1	1
3	1	0	2	0	1
4	3	0	6	0	3

FIGURE 2.7 – Exemple d'application de mesure de similarité

A. **Distance euclidienne :**

$$d(u,v) = \sqrt{\sum_{j=1}^{p}(u_j - v_j)^2} \tag{2.12}$$

p est le nombre de termes.

u_j (v_j) est la pondération du terme j pour le document u (resp.v)

$$d(3,4) = \sqrt{(1-3)^2 + ... + (1-3)^2} = 4.9$$

B. **Similarité COS :**

La formule de cette méthode est comme suit :

$$cos(u,v) = \frac{u.v}{\|u\|\|v\|} \tag{2.13}$$

Voici un exemple de cette méthode :

$$cos(3,4) = \frac{1*3 + ... + 1*3}{\sqrt{(1^2 + ... + 1^2) * (3^2 + ... + 3^2)}}$$

C. **Indice de Dice :**

Mesurer la similarité entre deux documents en se basant sur le nombre des termes communs entre eux.

$$d(u,v) = \frac{2N_c}{N_1 + N_2} \tag{2.14}$$

N_c est le nombre des termes communs entre les documents u(resp.v)

N_1 (resp. N_2) le nombre des termes de u(resp.v).

2.3.5 Algorithmes de classification :

2.3.5.1 : Les K plus proches voisins (K-PPV) :

Les K plus proches voisins (k-NN pour-Nearest-Neighbor en anglais) est un algorithme de classification supervisé qui consiste à déterminer pour chaque nouveau document non-classé la liste des **K** plus proches voisins parmi les documents déjà classés sachant que le **K** est un entier initialisé au départ.

Le document sera donc affecté à la classe qui contient le plus de document parmi ses plus proches voisins en se basant sur une distance (ex : distance euclidienne). Il a été décrit pour la première fois en 1950, mais initialement appliqué à la classification des articles de presse par Massand et al. en 1992.[27]

Le principal inconvénient de K-NN est qu'il coûte du temps pour classer les documents mais il existe des heuristiques qui ont pour but de réduire le temps d'exécution.

2.3.5.2 : Classificateur Naïf Bayes :

Naïf Bayes est un algorithme de classification supervisé fondé sur le Théorème de Bayes.

$$P(A/B) = \frac{P(B/A) * P(B)}{P(A)} \tag{2.15}$$

Cette méthode permet d'apprendre un modèle de classification à partir des données, l'ensemble d'apprentissage (A) est connue a priori dont chaque document est étiqueté par sa classe, l'objectif est de chercher à classer un nouveau document « X_{nouv} » non encore étiqueté. Le classificateur bayésien cherche à trouver la classe « y» qui a la plus grande probabilité, on parle de règle MAP (maximum à posteriori).[28]

$$Y_{MAP} = \underset{y}{\operatorname{argmax}}\, P(y|X_{nouv}) = \underset{y}{\operatorname{argmax}}\, P(X_{nouv}|y)P(y) \tag{2.16}$$

Des fois on suppose que toutes les probabilités sont égales, on parle de la règle ML (maximum likelihood)

$$Y_{ML} = \underset{y}{\operatorname{argmax}}\, P(X_{nouv}|y) \tag{2.17}$$

Si on a besoin de déterminer la classe la plus probable pour l'instance X_{nouv}, on fait le calcul suivant pour un exemple de test :

$$Y = \underset{y}{\operatorname{argmax}}\, P(y) \prod_{i=a}^{|A|} P(X_{nouv}|y) \tag{2.18}$$

2.3.5.3 : K-Means :

K-Means est un algorithme d'apprentissage non-supervisé, il est de loin l'algorithme le plus populaire et le plus simple des algorithmes de clustering.

L'algorithme est utilisé pour trouver des groupes et ensembles qui n'ont pas été étiquetés, trouver les patterns afin de choisir les meilleures décisions. Cette méthode suit une procédure très simple pour classer un ensemble de données à travers un certain nombre(**K**) de cluster.[29]

K-Means est défini par la fonction objective qui minimise les distances entre les données dans un cluster, elle sera appliquée dans tous les clusters.

A. **La fonction objective de K-means :**

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \Big(\sum_{x_j \in S_i} \big\| x_j - \mu_i \big\|^2 \Big) \tag{2.19}$$

- k est le nombre total des clusters.
- S_i contient l'ensemble des objets x dans un cluster i.
- μ_i est le centroïde du cluster S_i.[30]

B. **Fonctionnement de l'algorithme K-means :**

- L'algorithme divise le data en N dimensions (dans cet exemple deux dimensions).
- Choisir la valeur **K**, le nombre de clusters qui sera généré.
- Initialiser k points dans le dataset comme étant le centroïde initial du cluster.
- Pour chaque donné dans le dataset faire
  - Calculer la distance entre les données du dataset et le centroïde du cluster.
  - Affecter les données vers le cluster du centroïde le plus proche.
- Déplacer les centroïdes vers la moyenne(*mean*) de l'emplacement des données du cluster.
- Répéter les étapes 4 et 5 jusqu'au nombre maximal d'itération, ou que les centroïdes arrêtent de déplacer. [29, 30]

C. **Le pseudo-algorithme de K-means :**

Algorithme 1 : K-means

Input : D Data-set
m //Le nombre des données dans le dataset
$(x_1, x_2, x_3...x_m)$ //Les documents du dataset
K //Initialiser le nombre des clusters.

Initialiser de façon aléatoires les centroides des clusters $(\mu_1, \mu_2...\mu_k \in R^n)$
répéter
  pour *i = 1 jusqu'au m* **faire**
    //assigner chaque document à un cluster
    c_i = l'index de centeroide de cluster le plus proche d'un document x_i
  fin
  pour *j = 1 jusqu'au k* **faire**
    μ_j= La moyenne(mean) de points (données) qui sont dans le cluster
  fin
jusqu'à μ_j *arrête de changer*;

D. **Exemple du Fonctionnement de l'algorithme :**

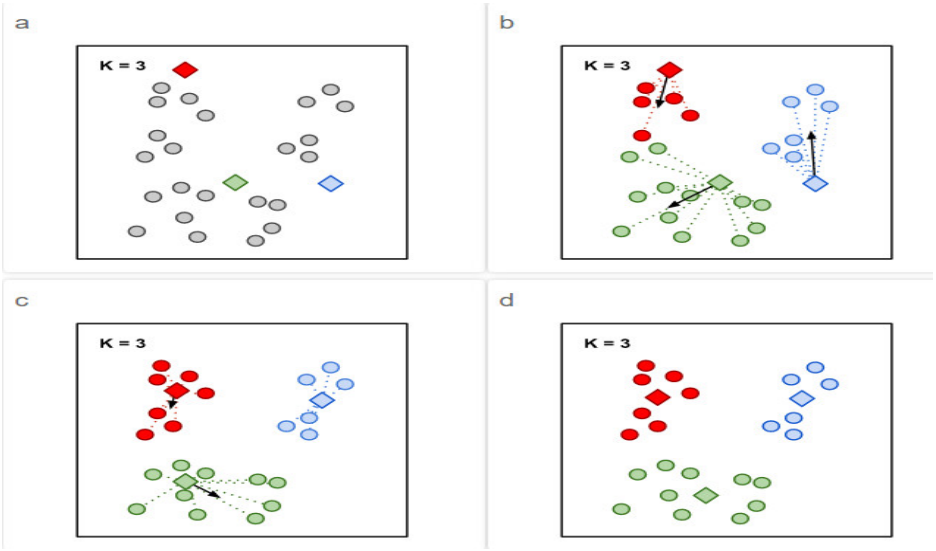

FIGURE 2.8 – Fonctionnement de l'algorithme K-means avec K=3

E. **Comment trouver le meilleur K :**

Pour utiliser l'algorithme de k-means, les utilisateurs sont censés de trouver le meilleur K (nombre de groupes) afin d'acquérir des bons résultats.

Cependant, il existe une méthode pour trouver le meilleur K, c'est la **méthode du coude** (Elbow method).

L'idée de cette technique, est d'exécuter l'algorithme K-means sur une plage de valeur de K et calculer le SSE (somme carré de la distance entre un document du cluster et son centroide).

Après, on trace le graphe (qui correspond à un bras), la valeur qui se situe dans le coude du bras c'est la valeur du meilleur K de l'algorithme.[30]

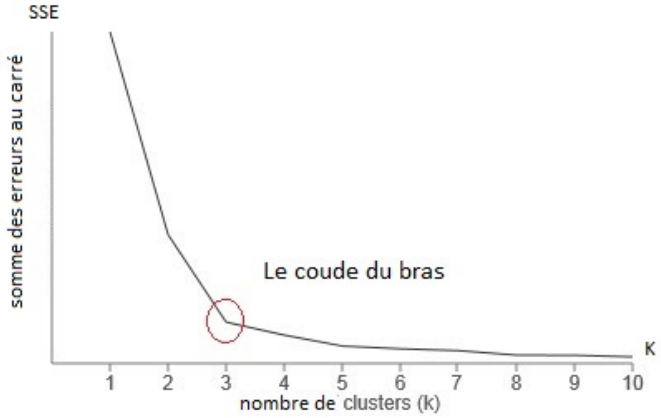

FIGURE 2.9 – Fonctionnement de la méthode du coude

2.3.5.4 : DBSCAN :

DBSCAN (aussi appelé algorithme de clustering basé sur la densité) est un algorithme d'apprentissage non-supervisé, le but de cet algorithme est de séparer les clusters qui ont une petite et une grande densité. Cette méthode est utilisée pour construire les modèles et les algorithmes de machine learning [31]. Le fonctionnement de l'algorithme se fait avec les étapes suivantes :

- L'algorithme divise le data en N dimensions (dans cet exemple deux dimensions).
- Sélection de deux paramètres
  - Le nombre de documents minimum dans un cluster (**minPts**).
  - La distance minimale entre deux documents dans un cluster appelé **epsilon**.
- Choisir un document aléatoire et voir s'il est un document noyau d'un cluster ou non selon deux critères
  - La distance minimale entre le document noyau et un voisin est inférieur à epsilon.
  - Le nombre de document voisin supérieur ou égal au nombre de document minimum (**minPts**)
- **Important :** Les données qui sont à l'intérieur de l'epsilon des données voisines et qui contienne le nombre minimal de nœuds dans leur epsilon, sont aussi intégré dans le cluster.
- Les données qui n'appartient à aucun cluster sont appelées valeurs aberrantes (*Noise*). [31]

2.4 Conclusion

Dans ce chapitre on a étudié la conception des méthodes d'apprentissage automatique (discrimination et clustering) de manière générale.

Pour faire la classification des données de type texte, il faut ajouter des méta-données afin de passer d'un texte brute vers un texte annoté pour que la machine comprend et apprend de façon efficace, comme on l'a montré dans la première partie.

Dans la première partie, on a vu comment une machine peut apprendre dans la première approche (la discrimination ou la classification) à partir d'un échantillon d'exemples classés (étiquetés) de classer un nouvel exemple non étiqueté.

On va passer à la deuxième approche (le clustering) si la première ne peut pas traiter un certain problème c'est-à-dire, d'apprendre à partir d'une base sans aucune connaissance préalable.

Chapitre 3

Système de classification automatique des articles d'actualités

3.1 Introduction

La forte croissance des articles disponibles en format numériques et la nécessité de les organiser nous oblige à concevoir **un système de classification automatique des articles d'actualité** qui peut gérer la masse considérable des articles et permettra de bien organiser ce grand espace de données selon les thèmes (sujets) abordés. Après avoir présenté dans la partie précédente de ce mémoire, un état de l'art sur les concepts et techniques en relation avec notre cas d'étude, nous allons entamer dans ce chapitre l'étape de conception de notre système. Nous allons rappeler les objectifs de notre travail, présenter les défis que nous avons rencontré ainsi que notre contribution.

Nous rappelons ici que l'objectif principal de notre travail, c'est de développer un système de classification des articles d'actualité qui permettra de :

- Offrir un accès facile et efficace au grand flux d'actualité mise en ligne quotidiennement
- Orienter l'utilisateur vers les sujets d'actualités de son intérêt
- Avoir une vue globale et organiser des principaux sujets d'actualités.
- Explorer facilement l'espace d'actualité.
- Augmenter la visibilité sur le web.

En effet, avec le développement des NTIC, de gros volumes de données d'actualités sont publiées continuellement, l'utilisateur se trouve submerger par ce grand flux d'actualité ininterrompu et il devient difficile pour lui d'avoir une vue globale des sujets d'actualité traités. Il est incapable de trouver facilement les sujets de son intérêt. De plus on trouve souvent des milliers d'articles qui traite le même sujet d'actualité ce qui crée une redondance chez l'utilisateur

3.2 Défis liés à la classification des news :

Le développement d'un système de classification automatique d'articles d'actualité présente naturellement un certain nombre de défis. Ces défis sont liés essentiellement au très grand volume de données à classer, ce qui engendre un problème de complexité. En plus ces données sont

de nature textuelle, toutes les problématiques liées à leurs représentations, à la pondération de leurs contenus, à leurs sémantiques rendent difficile le processus de détection des sujets traités. Ci-dessous une liste de défis que nous pouvons rencontrer dans le cadre de ce travail :

- Comment choisir les sources d'actualité
- Défis liés au gros volumes de données.
- Problème lié à la représentation des données (de nature textuelle)
    - o Comment choisir les descripteurs ?
    - o Comment les pondérer afin de mieux refléter les contenus des articles ?
- Défis liés à la détection de sujets d'actualité traités

3.3 Approche et contribution :

Pour surmonter ces défis et pour développer notre système, nous basons sur deux choix de conception (ou constats) qui nous semblent importants :

- Tout d'abord, **opter pour une représentation à base d'entité nommée (et non pas à base de termes)** des contenus des articles d'actualité.
- Ensuite, **exploiter l'importance de la rubrique « titre »** de l'article dans la représentation des contenus des articles.

Ces choix de conception se trouvent amplement justifiés dans le cadre de la classification des news. En effet les noms des personnes (personnalités politiques, sportives, scientifiques, célébrités...), les noms des lieux (ville, pays, région), des organisations (politiques, sociales, économiques, locales, internationales,...) et les dates sont des éléments de contenus très représentatifs. Ces noms sont porteurs de sens qui véhiculent une grande richesse d'information.

Nous devons pour cela mettre en évidence l'ensemble des entités nommées (EN) dans un corpus de news pour en faire une représentation à base d'E.N de chaque article. Ce choix nous permet ensuite de réduire considérablement le volume de notre dictionnaire (les descripteurs), Nous devons bien sur mesurer l'importance de chaque E.N dans un article via une opération de pondération.

Il est clair aussi que le titre d'un article d'actualité reflète grandement sa thématique et son sujet. Le titre ne peut être choisi au hasard, il représente le sujet d'un article et décrit de manière générale son contenu. Pour cela, nous envisageons dans notre système d'exploiter ce constat dans la phase de pondération des descripteurs. Un descripteur (E.N dans notre cas) apparaissant dans le titre d'un article doit avoir un poids plus important que celui apparaissant dans le corps de l'article.

3.4 Formulation du problème

On formuler le problème comme suit :

2 : Formulation du problème

Input :

- **S** =S_1, S_2,S_3,S_4,S_5,. . . ,Sn. *l'ensemble des sources d'actualités*
- **C_News** =N_1, N_2,N_3,N_4,N_5,. . . ,Nm. *le corpus de news (l'ensemble des articles).*
- **E** = E_1,E_2,E_3,E_4,. . . . . En. *l'ensemble des entités nommées existantes dans tout le corpus.*
- *Chaque article* Ni *sera représenté par un vecteur pondéré.*
  - $Ni = (WiE_1,WiE_2,WiE_3,\ldots..,WiEn)$.
  - $WiEj$ *est le poids (pondération) de l'entité* Ej *dans l'article* Ni.

Output :

- **Cluster** = ($C1$, $C2$, $C3$,..., Ck) *un ensemble de clusters d'articles. Chaque cluster (Ci) regroupe un certain nombre d'articles* ***qui traitent le même sujets d'actualité*** .

Contrainte :

- **Densité du cluster (Intra-cluster) :**Minimiser la distance moyenne entre l'article Ni et tous les autres points du cluster C_k auquel il appartient.
- **Diversité (inter-cluster) :** Maximiser la distance moyenne entre l'article Ni qui appartient au cluster C_k et tous les autres points d'un autre cluster C_l

3.5 Architecture du système

Notre système repose sur l'analyse textuelle (annotation et pré-traitement) des contenus d'articles d'actualité, on dispose des articles comme des documents bruts non compréhensibles par la machine, il faut suivre quelques étapes pour les rendre compréhensibles et atteindre notre objectif – leurs classifications –. La figure ci-dessous montre l'architecture générale de notre système :

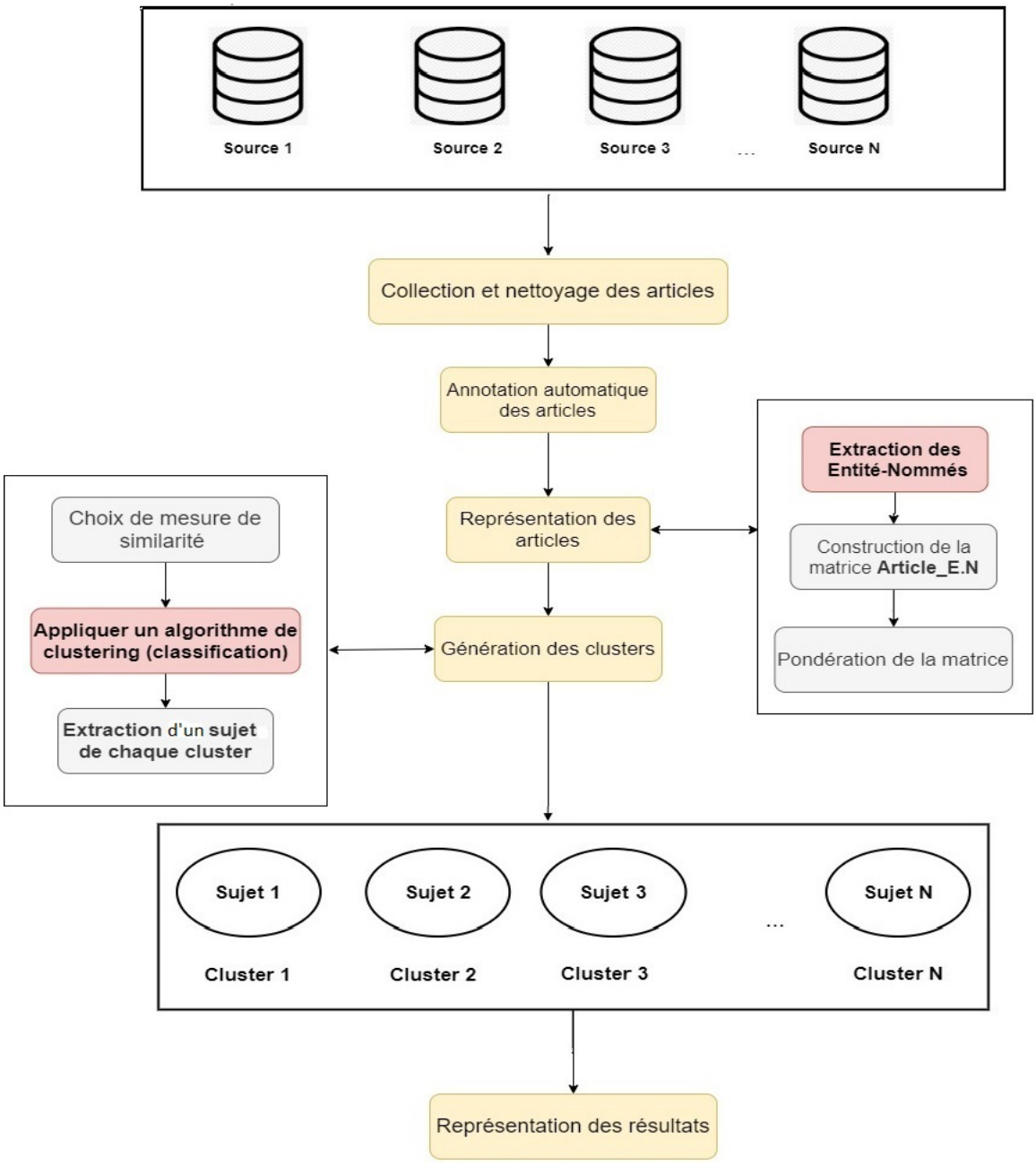

FIGURE 3.1 – Architecture générale du système

Nous détaillons dans ce qui suit les différentes phases du notre système :

3.5.1 Collection et nettoyage des articles

Notre corpus (ensemble d'articles) est collecté à partir des ressources disponibles sur internet. Certains sont entièrement gratuits et disponibles à 100%, d'autres nécessitent des étapes d'authentifications ainsi que des renseignements personnels et professionnels (tel que l'organisme de recherche par exemple), tandis que d'autres exigent un coût pour certaines fonctionnalités.

Pour notre approche, cette collection est constituée principalement d'extraits de journaux mondiaux, des magazines et généralement de presse, qui sont des articles hétérogènes à base des thèmes.

Le schéma ci-dessous résume le processus du pré-traitement des articles et la construction du dataset.

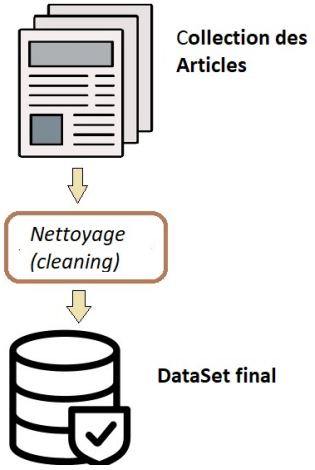

FIGURE 3.2 – Collection et nettoyage des articles

Pour collecter les articles de façon quotidienne, notre système envoie une requête *API* pour avoir l'accès à tous les articles de multiples sources qui sont ajoutés récemment. La réponse du service de fournisseurs des articles est généralement de forme *JSON* (non-structuré), ce qui oblige notre système à structurer ces articles afin de faciliter leur utilisation.

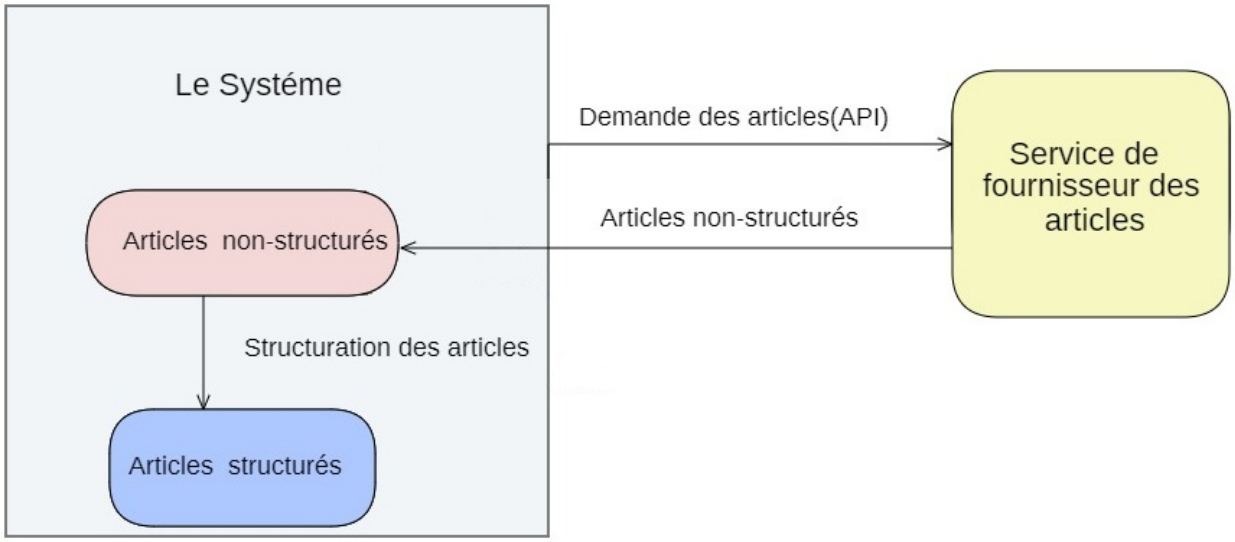

FIGURE 3.3 – Processus de collection des articles

Ces articles nécessitent un nettoyage ou un pré-traitement, car il existe des articles qui ne possèdent pas des données ou des informations qu'on juge importante, tel que un titre, contenu ou date...

L'absence du titre ou du contenu dans un article affectera la précision de la classification, donc il faut les filtrer, aussi il existe des articles du type multimédia (vidéo, son) qui ne contiennent pas un contenu textuel, il faut les supprimer aussi. Et enfin il faut vérifier s'il existe des articles dupliqués (en vérifiant les ID dupliqués) afin de les supprimer et minimiser la taille du *Dataset*.

Après la fin de la tache du pré-traitement des articles, on passe à l'étape construction du dataset, ce dernier contient des informations suivantes pour chaque article :

- L'identifiant de l'article.
- La source d'ou l'article a été collecté.
- Le titre de l'article.
- L'URL qui va orienter vers la source de l'article.
- La date et l'heure de la création de l' article.
- L'image qui illustre l'article.
- Le contenu de l'article.

On note que **si le dataset est bien traité et nettoyé, la précision de la classification augmentera** ce qui est très important dans notre système.

3.5.2 Annotation automatique des articles :

Dans notre système, l'annotation des articles est très importante. Elle permet de joindre pour chaque article un ensemble des méta-données pour le but d'augmenter la précision de la classification et de faciliter la tâche du pré-traitement, afin de construire les descripteurs de la matrice de pondération. Le schéma ci dessous explique le résultat du processus d'annotation automatique d'un article :

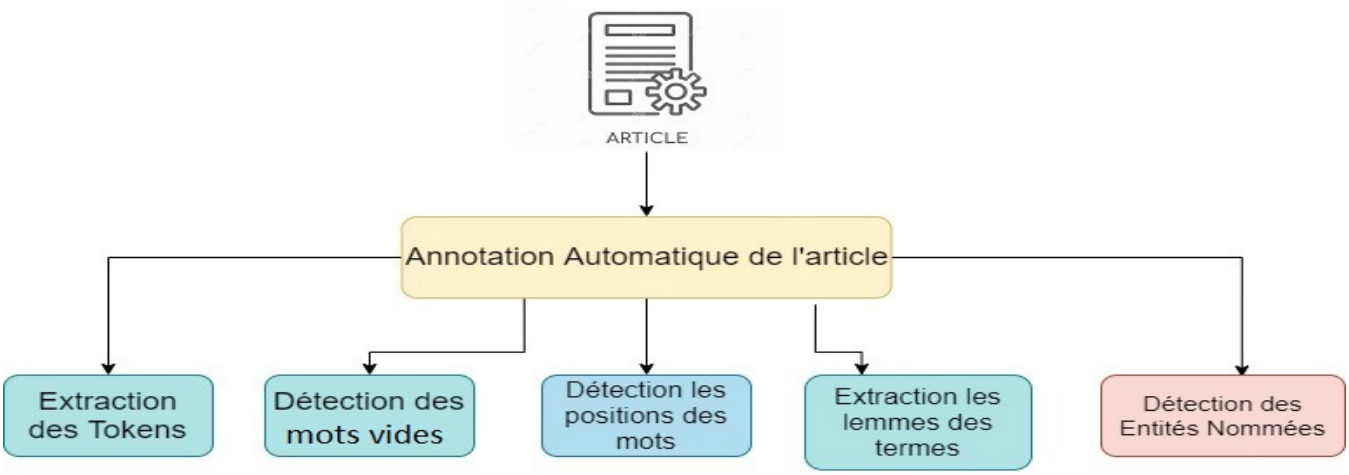

FIGURE 3.4 – Processus d'annotation d'un article

Le processus d'annotation des articles passe par plusieurs étapes (voir section (2.2)). A la fin de ce processus, on aura des méta-données pour chaque token qui existe dans les articles.

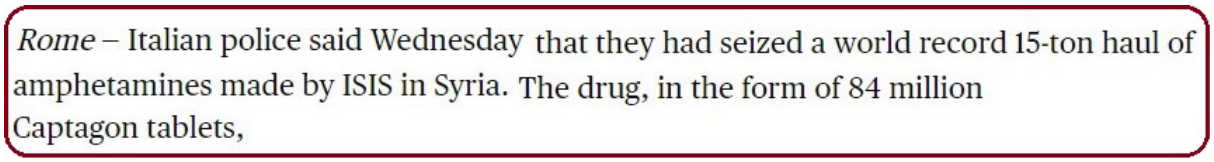
Rome – Italian police said Wednesday that they had seized a world record 15-ton haul of amphetamines made by ISIS in Syria. The drug, in the form of 84 million Captagon tablets,

FIGURE 3.5 – Exemple d'article de presse

Le tableau ci-dessous contient toutes les méta-données de cet article après le processus d'annotation automatique :

Terme	ID Token	Mot vide	POS	Lemma	Entité-Nommé
Rome	1	Faux	Nom PROPRE	Rome	Vrai
—	2	/	/	/	/
italian	3	Faux	ADJ	italy	Vrai
police	4	Faux	Nom	police	Faux
said	5	Faux	Verbe	say	Faux
wednesday	6	Faux	Nom PROPRE	Wednesday	Vrai
. . .	. . .	. . .	. . .	. . .	. . .
amphetamines	17	Faux	Nom	amphetamine	Vrai
made	18	Faux	Verbe	make	Faux
by	19	Vrai	Adverbe	by	Faux
ISIS	20	Faux	Nom PROPRE	ISIS	Vrai
in	21	Vrai	Adverbe	in	Faux
Syria	22	Faux	Nom PROPRE	Rome	Vrai
.	23	/	/	/	/

TABLE 3.1 – Exemple de méta-donnés que contient un article

Lors de la détection des entités nommées, on suggère d'utiliser la méthode hybride (statistique et symbolique présentée au chapitre précédent) afin d'augmenter le nombre des ENs existantes dans un article, par exemple :

- La compagnie Orange lance de nouvelles offres mobiles. *Orange est une E.N*
- L'orange est un fruit. *orange ici n'est pas une E.N*

3.5.3 Représentation des articles

Le but de cette étape est d'arriver à produire une représentation à base d'entités nommées du corpus de news sous forme d'une matrice où les lignes correspondent aux articles et les colonnes aux E.Ns présentes dans le corpus de news. Chaque article (ligne) sera représenté par un vecteur pondéré, chaque coordonné correspondra au poids de chaque EN dans cet article. Comme le montre la figure suivante cela passera par les étapes suivantes :

- Extractions des E.Ns
- Génération de la matrice Article/EN
- Pondération des descripteurs

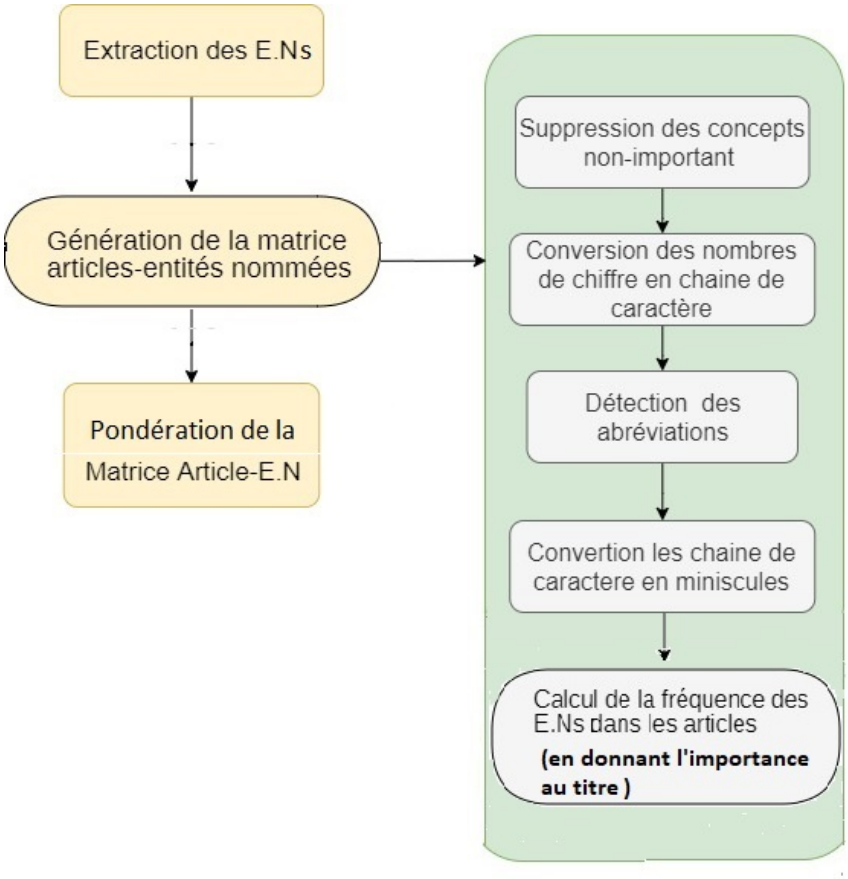

FIGURE 3.6 – Représentation des articles

3.5.3.1 : Extraction des entités nommées :

Les entités nommés sont des noms de lieux, de personnes, d' organisations... etc, l'un des avantages de l'annotation automatique c'est qu'elle permet d'affecter une catégorie pour chaque E.N, dans ce tableau on va citer les multiples catégories des (E.N) :

Concept	**Exemple**
Les Personnes ou personnalité publique	Trump, Putin...
Partie politique, Nationalité	Algerien, anglais, F.L.N ...
Nom d'établissement comme les écoles, aéroports	Aéroport Houari Boumedienne, Ecole Ibn Badis
Les compagnies	Djezzy, Oreedoo, Orange...
Nom des pays et location	Algerie, Maroc, Alger, Bouira...
Événement	Coronavirus, Tsunami de Haiti...
Domaine artistique	Nom des livres, des musique...
Date et nom des jours	Mercredi, 21 octobre 2019
Mesure de distance ou de poids	20kg, 50km
Nombre ordinal	premier, deuxième, dernier...

TABLE 3.2 – Les différents catégories des entités nommés

Il existe des outils et des techniques qui permettent l'extraction des E.Ns comme *OpenNLP* [1] et *Spacy* [2].

On utilisera donc les E.Ns pour sélectionner les descripteurs de la matrice de pondération des articles, ceci un exemple de l'extraction des entités-nommées à partir d'un article :

1. Bibliothèque OpenSource de Java
2. Bibliothèque OpenSource de Python

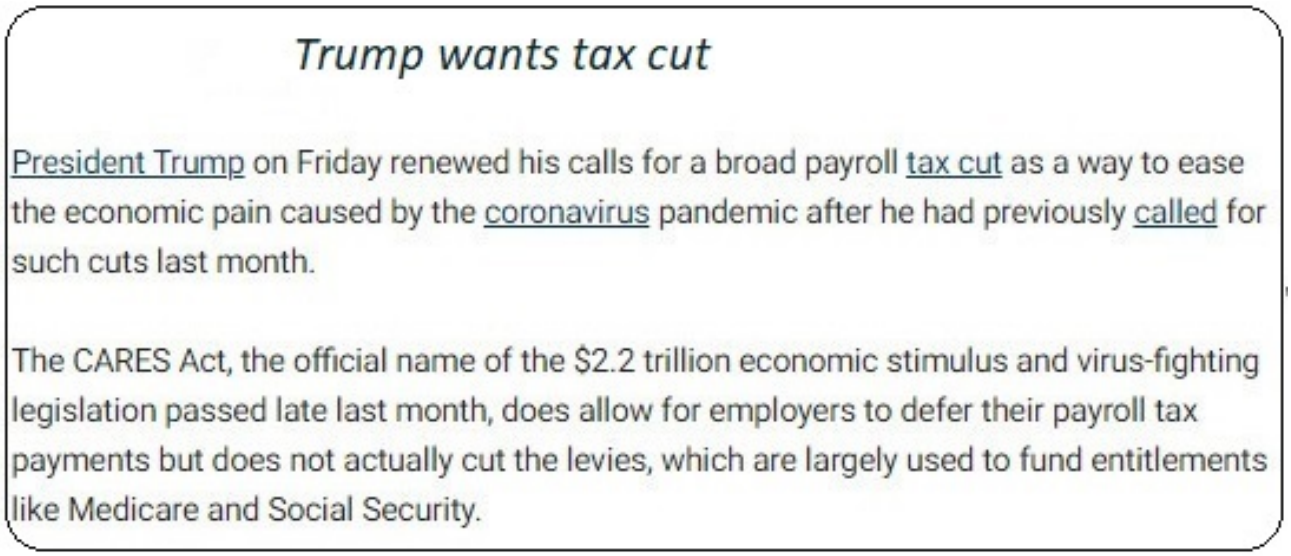

FIGURE 3.7 – Articles contenant des E.N

Après avoir terminer le processus d'annotation de cet article, on extraira les E.Ns avec leurs types. On représente les résultats dans ce tableau :

ID E.N	Texte de E.N	Concept	ID E.N	Texte de E.N	Concept
1	Trump	Personne	8	trillion	Argent
2	Friday	Date	9	late	Date
3	last	Date	10	last	Date
4	month	Date	11	month	Date
5	Cares	Nom d'organisation	12	Mediacare	Nom d'organisation
6	Act	Nom d'organisation	13	Social	Nom d'organisation
7	2.2$	Argent	14	Security	Nom d'organisation

TABLE 3.3 – Extraction des E.Ns

3.5.3.2 : Génération de la matrice articles-entités nommées

L'indexation à base d'entités nommées permet non-seulement de réduire l'espace de dimensionnalité (jusqu'à 70%), aussi elle permet d'améliorer la précision de la classification car l'indexation sera basée sur des concepts que nous jugeons très importants pour les actualités.

Elle permet aussi de passer certaines étapes du pré-traitement tel que **la lemmatisation** (les entités nommées ne contiennent pas les verbes ou les adverbes) **et suppression des mots vides.**

Après avoir extrait les entités nommées d'un article, on passe à l'étape de construction de la matrice article-EN. Pour des raisons d'efficacité, il est important d'effectuer un certain nombre de traitements et ne pas prendre toutes les E.Ns détectées :

A. **Suppression des concepts non-importants :** On rappelle que l'annotation automatique affecte à chaque E.N son type, Il existe des concepts des E.Ns qui ne sont pas importants dans les articles, tel **les dates** comme (l'année passée, demain, dans le prochain mois ...) aussi **les nombres ordinales** (premier, deuxième, ...), **les noms des journées** (Mardi, Mercredi ...), Ces concepts peuvent avoir un impact sur notre classification ce qui nous oblige à les filtrer.

B. **Conversion des chiffres en chaine de caractère :** Cette étape permet de convertir les chiffres en chaine de caractères pour éviter les entités dupliquées. Par exemple la conversion du chiffre "**155**" vers une chaine de caractère nous donnera "**cent cinquante cinq**".

   Il existe beaucoup de librairies et d'outils qui réalise cette conversion comme les bibliothèques *Spacy* et *NLTK* [3] de Python.

C. **Détection des abréviations :** Comme dans l'étape précédente, on peut trouver dans un article des E.Ns écrites sous forme normale ou sous forme d'abréviation ce qui provoquera des descripteurs dupliqués, il existe des méthodes qui permet de convertir **les abréviations** comme (O.M.S ou N.H.S) **en chaine de caractère** (Organisation Mondiale Santé) ce qui va supprimer les descripteurs qui ont le même sens mais qui sont dupliqués. Le point faible de cette méthode c'est qu'elle ne peut pas convertir les abréviations si leur sens n'ont pas été cités dans l'article.

D. **Conversion les chaine de caractère en minuscules :** Cette étape permet de convertir tout les chaines de caractère en minuscule, ce qui va supprimer les mots dupliqués écrits en majuscule, par exemple (Alger -> alger, COVID -> covid ...)

3. Bibliothèque OpenSource de Python

E. **Calcul de la fréquence des E.Ns dans les articles :** Dans notre contribution on a suggéré de donner l'importance au titre par rapport au contenu, afin de réaliser cette suggestion on doit fusionner le titre de l'article avec son contenu avec un taux de fusion lambda (c-à-d si lambda = 2 on doit fusionner le titre 2 fois avec le contenu).

Afin de calculer la fréquence des E.Ns dans un document, on suit cette formule :

$$TF(D, E.N) = \Big\{ Frqu\_Terme\_Titre * lambda + Frqu\_Terme\_Contenu \tag{3.1}$$

On utilise l'exemple de l'article précédent (figure 3.7) et on calcule la fréquence du terme "Trump" qui se trouve dans le titre est le contenu :

Ce terme apparait **une fois** dans le titre et **une fois** dans le contenu, en utilisant la formule précédente on aura pour lambda = 2 :

$$TD(\text{"}Trump\text{"}) = 2 * 1 + 1 = 3$$

Donc la fréquence du terme "Trump" dans l'article est **3**.

F. Après la fin du processus de construction de la matrice Article-E.N, au aura une matrice qui ressemble à celle-ci :

	beijing	bernardino	california	cedar	celsius	center	china	chinese	cities	coronavirus	county	degrees	department	england	feet	francisco	gavin
0	1	0	0	0	1	0	3	1	0	0	0	1	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	4	0	0	0
2	0	0	5	0	0	0	0	0	1	0	0	0	0	0	0	1	1
3	0	1	3	1	0	0	0	0	0	1	1	0	1	0	0	0	1

FIGURE 3.8 – Exemple de matrice article-EN

3.5.3.3 : Pondération de la matrice :

Il existe de nombreuses techniques dédiées à la pondération des termes dans un document comme déjà mentionné dans la partie état de l'art. La pondération TF-IDF est de loin la plus utilisée et c'est celle qui donne de meilleurs résultats.

On va utiliser l'exemple précédent (figure 3.8) afin de calculer TF-IDF du l'E.N ("california") ? On voit que la fréquence du ce terme dans l'article 2 est 5 et le nombre total de termes dans cet article est 8 termes, on rappelle la formule pour calculer TF :

$$TF(T, D) = \frac{NT}{NTm} \tag{3.2}$$

avec NT est la nombre d'apparition du terme dans l'article et NTm le nombre de termes totales dans l'article, donc on aura :

$$TF(T, D) = \frac{NT}{NTm} = \frac{5}{8} = \mathbf{0.625} \tag{3.3}$$

Ce terme apparait dans 2 articles sur 4, afin de calculer le IDF du terme on rappelle la formule de IDF :

$$IDF(T, D) = \log_{10} \frac{ND}{NDt} \tag{3.4}$$

avec ND est le nombre total des articles et NDt le nombre de document à qui le terme "california" appartient, donc on aura :

$$IDF(T, D) = \log_{10} \frac{4}{2} = \log_{10} 2 = \mathbf{0.301} \tag{3.5}$$

De (3.3) et (3.5) on déduit :

$$Poids = TF(T, d) * IDF(T, D) = 0.625 * 0.301 = 0.19 \tag{3.6}$$

Donc le poids du terme "california" dans l'article 2 est : **0.19**.

De la même façon, on calcule le poids des entités nommées dans les articles. On aura la matrice suivante :

	beijing	bernardino	california	cedar	celsius	center	china	chinese	cities	coronavirus	county	degrees	department	england	feet	francisco	gavin
0	0.1	0.00	0.00	0.0	0.1	0.00	0.1	0.1	0.00	0.0	0.0	0.1	0.0	0.00	0.1	0.00	0.00
1	0.0	0.00	0.00	0.0	0.0	0.12	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.48	0.0	0.00	0.00
2	0.0	0.00	0.19	0.0	0.0	0.00	0.0	0.0	0.08	0.0	0.0	0.0	0.0	0.00	0.0	0.08	0.04
3	0.0	0.04	0.11	0.1	0.0	0.00	0.0	0.0	0.00	0.1	0.1	0.0	0.1	0.00	0.0	0.00	0.04

FIGURE 3.9 – Exemple de pondération de la matrice article-EN

3.5.4 Génération des clusters :

Cette étape nécessite un certain nombre de traitements. Tout d'abord on doit choisir une mesure de similarité à utiliser puis appliquer un algorithme de clustering, et une fois les clusters générés, trouver un moyen de caractériser chaque cluster par un titre ou une description.

3.5.4.1 : Choix d'une mesure de similarité :

La mesure de similarité permet de comparer deux articles en se basant sur les entités nommées qui les composent, afin de détecter les articles similaires et dissimilaires. Il existe plusieurs mesures qui permettent de calculer cette similarité.

Dans notre système, on a choisi **similarité cosinus** comme mesure de similarité, c'est un paramètre nécessaire de l'algorithme de clustering qui permet de calculer la ressemblance entre deux articles A1 et A2.

On justifie ce choix de similarité car un article peut être représenté par une centaine de descripteurs. Cette mesure de similarité calcule la ressemblance entre deux articles sans prendre en considération la taille des vecteurs des articles, **elle mesure l'angle qui sépare les deux articles** (chaque fois l'angle est petit, la similarité sera grande). [32]

On prend la matrice de pondération (figure 3.9) et on calcule la mesure de similarité cosinus entre l'article 3 et les autres articles :

On rappelle les formules de mesures de cosinus :

$$cos(u, v) = \frac{u.v}{\|u\|\|v\|} \tag{3.7}$$

En appliquant cette formule sur les vecteurs des articles 3 et 4 on aura :

$$cos(A3, A4) = \frac{0.0 * 0.0 + ...0.19 * 0.11 + ... + 0.04 * 0.04}{\sqrt{(0.19^2 + ...0.04^2) * (0.11^2 + ... + 0.04^2)}} = \frac{0.0225}{0.0526} = \mathbf{0.427}$$

De la même façon, On calcule la similarité entre article 3 et les articles 1 et 2 et on obtient :

$$cos(A3, A1) = 0 \qquad et \qquad cos(A3, A2) = 0$$

Un autre avantage important de la **mesure cosinus** par rapport à la distance euclidienne si un poids du terme "coronavirus" dans un article est 0.7, et le poids du même terme dans un autre article égale à 0.2, ils seront considéré dissimilaire par la mesure euclidienne, mais la mesure cosinus traitera ce problème.

3.5.4.2 : Appliquer un algorithme de clustering :

Maintenant que les articles sont représentés dans un format qui peut être interprété par des algorithmes d'apprentissage non supervisé. Ces algorithmes nécessitent une connaissance a priori du nombre optimal du cluster. Dans notre système on a choisi l'algorithme **K-means** pour la classification des articles cet algorithme prend **la mesure de similarité cosinus** comme entré, on juge ce choix par plusieurs avantages et critères :

- Algorithme rapide pour le clustering des données.
- Facilité de l'implémentation.
- K-means produit des clusters plus serrés que le clustering hiérarchique.

Afin de trouver le nombre optimal K (nombre cluster)on a choisi la méthode du coude (Elbow Method) qu'on a expliquée dans la partie (2.3.2).

L'algorithme prend donc en entrée la matrice de pondération et le K optimal, K-Means fait plusieurs itérations jusqu'à ce que les centres des clusters s'arrêtent, enfin chaque article sera affecté à son cluster comme le montre la figure ci-dessous :

	miami	michigan	mineiro	mississippi	missouri	monterrey	mx	news	nyra	ohio	sec	sports	id_cluster
Article 0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	7
Article 1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.4	0.000000	0.000000	0.000000	0.000000	0
Article 2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	2
Article 3	0.144338	0.144338	0.144338	0.144338	0.144338	0.144338	0.144338	0.0	0.144338	0.288675	0.144338	0.144338	9

FIGURE 3.10 – Exemple de résultat de clustering des articles

A la fin de la classification, nous aurons des articles similaires qui parlent sur les mêmes topics groupés dans les mêmes clusters, dans la prochaine étape nous expliquerons comment on détecte les sujets pour chaque cluster.

3.5.4.3 : Extraction des sujets pertinents :

Après avoir classifié les articles, on passe à la dernière étape de la détection d'un sujet pour chaque cluster, l'idée c'est de trouver l'article le plus proche du centre du cluster en calculant la mesure cosinus entre le vecteur du centroide et les vecteurs des articles qui sont dans le cluster, puis considérer une E.N du titre de cet article comme le sujet abordé par le cluster en utilisant le **Topic Modeling** (sélection d'une E.N dans le titre qui apparait le plus dans touts les autres articles). Ci dessous, on va utiliser un autre exemple d'une matrice qui contient les vecteurs des articles issus du même cluster avec le vecteur du centroide :

	bernardino	census	central	coronavirus	georgia	health	hypothermia	jekyll	league	long	meters	news	nursing	orlando
Centroide	0.23	0.20	0.31	0.14	0.07	0.17	0.05	0.41	0.11	0.14	0.04	0.04	0.22	0.31
Article 1	0.12	0.33	0.04	0.24	0.52	0.00	0.13	0.13	0.00	0.00	0.00	0.14	0.51	0.21
Article 2	0.04	0.04	0.04	0.00	0.24	0.24	0.00	0.04	0.00	0.00	0.04	0.00	0.00	0.00
Article 3	0.16	0.08	0.03	0.00	0.00	0.00	0.20	0.04	0.07	0.00	0.04	0.32	0.00	0.00

FIGURE 3.11 – Exemple d'une matrice de pondération des articles d'un même cluster

Afin de trouver l'article le plus proche du centre du cluster, on utilise la mesure de **similarité cosinus** donc on trouve :

$$cos(C, A1) = 0.604 \qquad et \qquad cos(C, A2) = 0.389 \qquad et \qquad cos(C, A3) = 0.335$$

Avec C le vecteur du centre du cluster (centroide) et A_i les vecteurs des articles.

On remarque que l' Article 1 est le plus proche du centroide, donc on prend l'E.N la plus fréquente qui apparait dans le titre, **et on la considère comme sujet ou topic traité par ce cluster.**

3.5.5 Représentation des résultats

Nous allons présenter à cette étape comment représenter notre travail à l'utilisateur. Tellement elle est intéressante il faut prendre quelques considérations. Au point vu de l'utilisateur la représentation des résultats obtenus est le seul moyen qui reflète la performance de notre travail. Pour cela on est obligé d'augmenter la visibilité de l'utilisateur en facilitant la recherche et l'exploration de l'espace des résultats.

La phase du Clustering nous permet d'avoir un certain nombre de Cluster. Chaque Cluster est représenté par son sujet. On rappelle que le sujet d'un Cluster est constitué d'une entité nommée la plus fréquente qui apparait dans l'article le plus proche du centre du cluster.

La synthèse de l'actualité est représentée de la façon suivante : Trois zones de visualisation sont prévues. Dans la première zone, **zone des grands titres**, on affiche les dix sujets d'actualité (les grands titres de l'actualité) les plus pertinents, c-à-d les sujets des clusters qui contiennent le plus grand nombre d'articles, le tableau ci-dessous explique comment le classement des sujets a été fait :

Cluster sujet Numéro :	ID Cluster	Sujet traité	Nombre d'article
1	7	Washington	140
2	26	CoronaVirus	78
3	14	U.S	50
. . .	. . .	. . .	. . .
8	20	Japan	35
9	15	N.C	31
10	19	New Mexico	20

TABLE 3.4 – Classement des sujets de clusters selon le nombre des articles

Dans la deuxième zone qui est la **zone de visualisation principale d'un sujet**, les trois articles les plus récents de sujet choisi sont présentés avec des illustrations (photo, image, graphique. . .) éventuelles (voir figure ci-dessous).

Dans la troisième zone de visualisation, **zone de visualisation auxiliaire**, les autres articles appartenant au sujet choisi sont présentés.

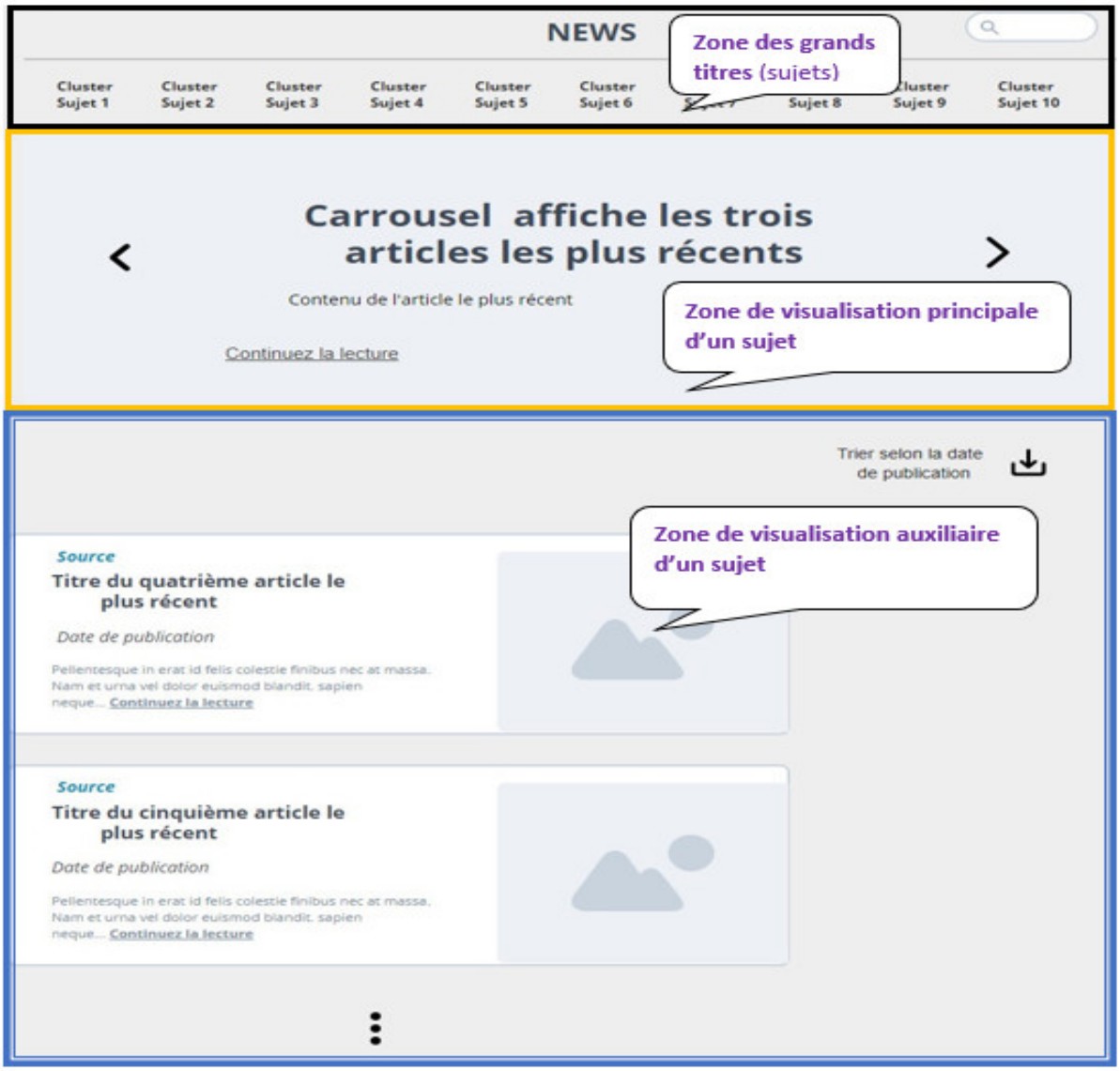

FIGURE 3.12 – Schéma de l'interface Utilisateur

Cette façon de présenter les résultats permet d'avoir une synthèse globale de l'actualité traitée (sous forme d'un ensemble de sujets ou grands titres). Elle permet aussi une exploration facile de l'actualité en allant d'un sujet à un autre pour pouvoir visualiser les articles de ce dernier.

3.6 Conclusion

Nous avons présenté dans ce chapitre l'approche que nous proposons pour effectuer la classification automatique des articles d'actualité par sujets ainsi que notre contribution.

Dans le début nous avons rappelé les objectives et les défis que nous avons rencontrés, puis nous avons ainsi exposé la conception générale de notre système d'annotation et la classification des articles à base des entités nommés, ainsi que les différentes parties qui la constituent.

Aussi nous avons donné le déroulement du processus de l'extraction des sujets à partir de chaque cluster, et enfin on a présenté les résultats sous forme d'une interface homme-machine.

Le chapitre suivant, on va présenter notre implémentation du système et l'expérimentation. On va montrer les résultats des tests sur notre système.

Chapitre 4

Implémentation et Expérimentation

4.1 Introduction

Nous avons présenté dans le chapitre précédant l'architecture de notre système qui représente le côté abstrait de l'approche. Durant ce chapitre nous allons viser l'objectif de toute étude que nous avons faite précédemment et entamer le côté concret, c'est la cible du travail qui est l'expérimentation et l'évaluation des résultats obtenues. Nous allons définir l'environnement et les outils utilisés dans notre expérimentation, définir l'interface du système et enfin évaluer l'expérimentation pour détecter si le système est précis ou non.

4.2 Résultats des évaluations :

Nous avons effectué plusieurs expérimentations afin de pouvoir faire une évaluation précise. Nous avons procédé par l'évaluation de plusieurs variantes (stratégies). À travers ces évaluations nous voulions répondre aux questions suivantes :

- Q1. Est-ce que la représentation à base d'entités nommées considérée dans notre solution permet d'améliorer la qualité du Clustering, par rapport à une représentation à base de texte ?
- Q2. Est-ce que la prise en compte de l'importance du titre dans la pondération des contenus des articles d'actualités permet d'améliorer la qualité du Clustering ?

Ces différentes variantes sont définies selon les paramètres suivants :

– L'utilisation ou non des ENs dans la représentation des contenus des articles.

– La considération ou non du titre dans la représentation de l'article

– Pondération préférentielle ou non des mots apparus dans le titre

Les différentes configurations évaluées sont comme suit :

- **Avec reconnaissance des entités nommées (REN+) :** ce type permet d'exécuter une expérimentation en utilisant uniquement des entités nommées. Le filtrage de texte est par extraction des E.Ns, on a négligé tout autre type de mots.

- **Sans reconnaissance des entités nommées (REN-) :** cette expérimentation est exécutée avec deux méthodes, une par lemmatisation des mots et suppression des mots vides (lemm stop+) et l'autre sans lemmatisation et sans suppression des mots vides (lemm stop-), autrement dit tous les mots d'un texte sont utilisés.

Le clustering des articles dans notre système est fait de différentes manières qui sont diffèrent selon le vocabulaire utilisé, le clustering est appliqué à base :

A. **Contenu des articles :** les mots présents dans le contenu seuls sont utilisés, c'est-à-dire la construction de la matrice Articles-Termes est faite à base des mots présents dans le contenu.

B. **Titre des articles :** les mots présents dans le titre seuls sont utilisés pour la construction de la matrice Articles-Termes.

C. **Titre + Contenu des articles :** le contenu et titre des articles avec une importance égale ; l'égalité ici a touché la pondération de la matrice, les mots fréquemment utilisé ont un poids plus élevé par rapport aux autres selon la technique TF-IDF.

D. **Titre + Contenu des articles :** sachant qu'on a utilisé ici un paramètre de coefficient lambda qui est égale à 2 pour le titre afin de donner un plus d'importance au titre que le contenu dans le but d'étudier l'utilité du titre dans un clustering à base de sujet. Le poids d'un mot contenu dans le titre doit être doublé pour faire une projection des descripteurs du titre dans le but de pouvoir augmenter leurs importances par rapport au descripteurs de contenu.

Dans chacun de ces types de clustering la construction de la matrice est basée sur les deux configurations décrites précédemment (REN+, REN-).

Le tableau suivant permet de montrer les résultats des évaluations des différentes variantes d'expérimentation. Pour estimer la précision du clustreing, nous avons utilisé le coefficient de silhouette moyen comme mesure de performance (voir chap 2). Le calcul de ce paramètre a été fait en utilisant la bibliothèque Scikit-learn de Python via **metrics.silhouette_score** :

	Coef silhouette moyen (en %)		
	REN^{+}	REN^{-}	
		$Lemme\_Stop^{+}$	$Lemme\_Stop^{-}$
Contenu	68	46	30
Titre	70	44	28
Contenu + Titre	54	19	16
Contenu + 2 * Titre	73	51	28

TABLE 4.1 – Résultats des évaluations des variantes d'expérimentation

Les graphes suivants montrent les différentes évaluations :

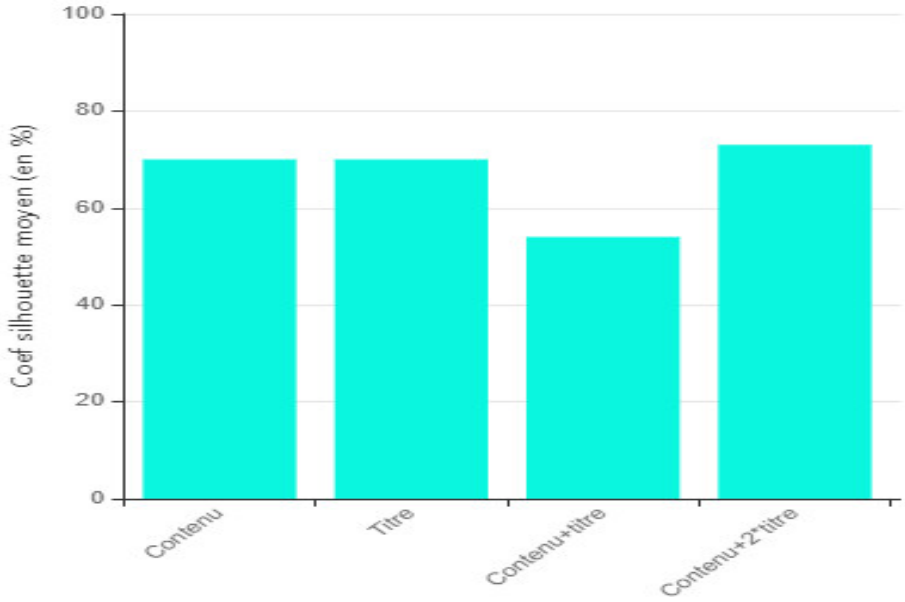

FIGURE 4.1 – Expérimentation avec reconnaissance des ENs (REN+)

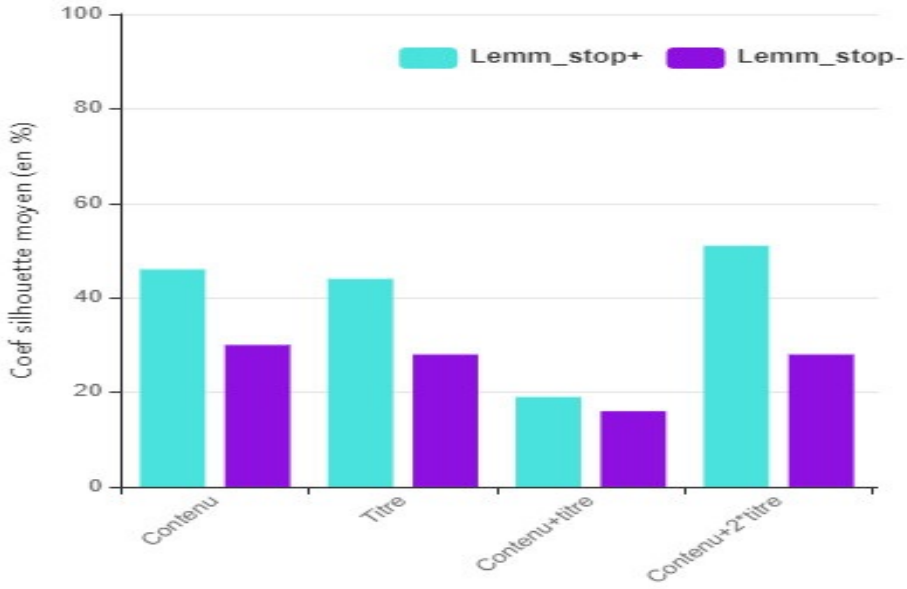

FIGURE 4.2 – Expérimentation sans reconnaissance des ENs (REN-)

4.3 Discussion des résultats

D'après les résultats, nous avons remarqué que la méthodologie que nous avons développée (clustring à base des ENs en donnant une importance au titre plus que le contenu) est la plus précise, avec une précision de **73%.**

Un autre critère qui s'avère ainsi important est la stabilité des clusters. pour prouver ce critère nous avons utilisé une méthode qui détermine dynamiquement le nombre de clusters optimal

(Elbow Method), puis exécuter plusieurs fois notre algorithme de clustering, le résultat est toujours le même, le nombre de clusters reste inchangeable ainsi que le nombre d'articles dans chaque cluster.

4.4 Interface de l'application

Afin d'afficher les articles d'actualité, nous avons réalisé une application WEB qui intègre le système de classification automatique des articles avec le framework **Django**.

Cette application contient une page d'accueil qui permet à l'utilisateur d'interagir avec le système. Dans cette page, on a choisi d'afficher les articles les plus récents (par exemple ceux d'aujourd'hui), après on a affiché dix sujets représentant les titres(sujets plus général) des clusters (Washington, Coronavirus, U.S ...) dans le navbar comme le montre cette figure :

FIGURE 4.3 – Interface d'accueil

On affiche aussi les articles avec leur source et l'URL qui va mener vers l'article complet. L'ordre d'apparition des articles est basée sur la contrainte de nouveauté des articles — c-à-d articles plus récents — comme représenté dans la figure suivante :

FIGURE 4.4 – Les articles de l'actualité les plus récents

Si l'utilisateur choisit un sujet d'intérêt (par exemple Japon) l'application va afficher tous les articles similaires qui traitent le sujet "**Japon**" comme le montre la figure suivante :

FIGURE 4.5 – Interface des articles qui traitent le sujet **"Japon"**

Voici un extrait des articles similaires qui traitent le sujet "**Japon**". Ils sont ordonnés selon leur date de création comme l'indique l'image suivante :

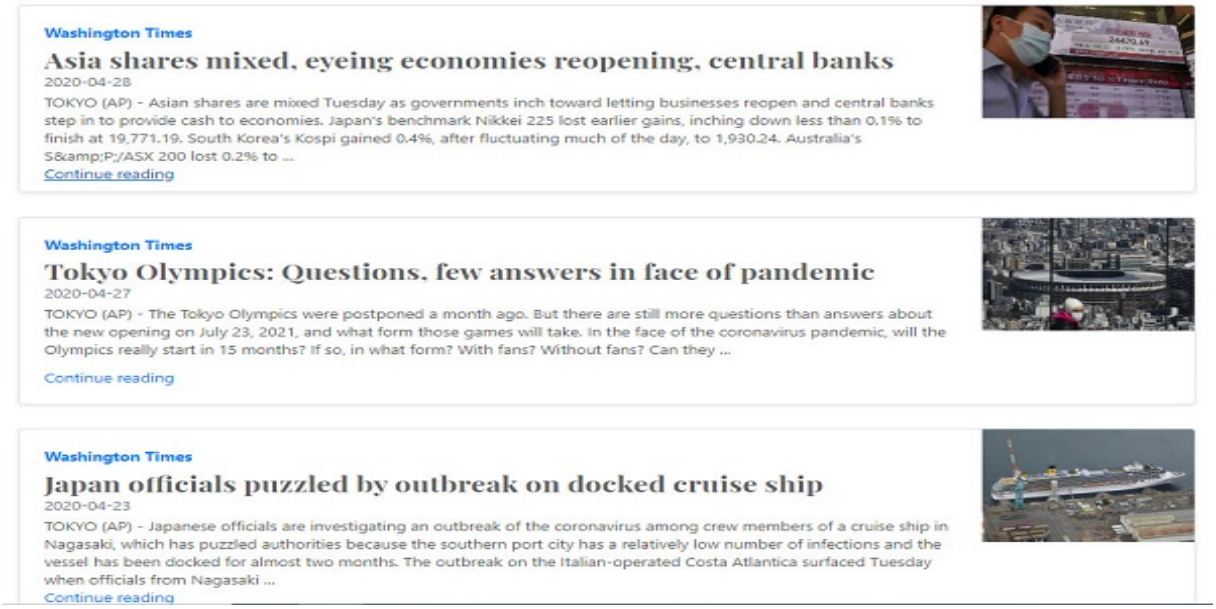

FIGURE 4.6 – Les articles les plus récent du sujet Japon

Si l'utilisateur clique sur "continuez la lecture ", le système va l'orienter vers l'article complet dans le site source de l'article de presse, ceci un exemple d'un article complet du sujet "Japon" :

FIGURE 4.7 – Exemple d'un article complet

4.5 Outils et environnements de développement

4.5.1 Anaconda :

C'est une distribution libre Python et R qui possède une collection de librairie de science des donnés tel que (Pandas, Scikit-Learn, Numpy...) ainsi que plus de 100 packages les plus populaires des langages Python, R et Scala. Ceci nous permet d'éviter d'installer les librairies et les modules de façon indépendante. Aussi Jupyter Notebook est installé dans la distribution ce qui permet d'éviter de l'installer séparément [35]

Cet environnement de travail concentre sur les domaines de l'analyse des données, l'apprentissage automatique et les sciences de données.

4.5.2 Spacy :

Spacy est une librairie open source réalisée en Python et Cython en 2015, elle est spécialisée pour le traitement avancé de langage naturel. Cette librairie est conçue pour réaliser les tâches de traitement de texte tel que l'annotation automatique d'un texte (extraction des entités nommées, tokenisation, la lemmatisation ...) Avec une implémentation efficace des modèles communs tels que les réseaux de neurones convolutifs. Ces modèles ont été créé en utilisant la bibliothèque Thinc.

Spacy utilise des modèles de réseaux de neurones statistiques afin de réaliser la détection des entités nommées, c-à-d pour chaque décision que le modèle prend —(*si un mot dans un texte est une E.N ou non*)— est **une prédiction**.[33]

4.5.3 Scikit-Learn :

C'est la librairie libre de python le plus populaire destinée à l'apprentissage automatique créé par David Cournapeau, elle construite en utilisant les libraires NumPy, SciPy, et matplotlib de python.

Scikit-Learn contient beaucoup d'algorithme intégré de l'apprentissage automatique comme de **classification** (Naïve Bayes, SVM), **régression** (plus proche voisin, SVR) et de **clustering** (KMEANS, DBSCAN), on peut utiliser ces algorithmes en faisant appel à une simple instruction, ce qui assure la simplicité de l'utilisation de cette bibliothèque.

Scikit-Learn permet aussi l'extraction extrait la structure de données complexes (bases de données, textes, images) pour les classifier en utilisant des techniques (méthode de pondération TF-IDF), afin que ces données seront supportées par les algorithmes de machine learning.[34]

4.5.4 Django :

Un framework est un ensemble d'outils qui simplifie le travail d'un développeur. Il s'agit donc d'un ensemble de bibliothèques coordonnées, qui permettent à un développeur d'éviter de réécrire plusieurs fois une même fonctionnalité, et donc d'éviter de réinventer constamment la roue. Inutile de dire que le gain en énergie et en temps est considérable.

Django est donc un framework Python destiné au web, fournit une interface graphique à l'utilisateur sous l'architecture ou patron Modèle-Vue-Template(MVT) est composé de trois entités distinctes, chacune ayant son propre rôle à remplir :

- Le modèle représente une information enregistrée quelque part, le plus souvent dans une base de données. Il s'agit d'une interface supplémentaire entre votre code et la base de données.

- La vue prend en charge tous les événements de l'utilisateur (accès à une page, soumission d'un formulaire, etc.). Il se charge, en fonction de la requête de l'utilisateur, de récupérer les données voulues dans les modèles.

- Le template qui est, comme son nom l'indique, la visualisation de l'information. C'est la seule chose que l'utilisateur peut voir. Typiquement, un exemple de template est une page web.

4.5.5 Pycharm :

PyCharm est un environnement de développement intégré (IDE) utilisé dans la programmation informatique, spécifiquement pour le langage Python.Il permet de faciliter le développement web avec le framework Django. Il est développé par la société tchèque JetBrains.

4.5.6 API :

Civicfeed est un service de fournisseurs d'article qui permet d'accéder à l'archive des articles de plusieurs sources de presse dans le monde comme New York Times et Washington Post. Il s'agit d'un grand ensemble de données, plus de 1000 articles de tous types de catégories avec toute leur information (titre, contenu, date, URL ....), ce service envoie la réponse aux requêtes (API) sous forme de JSON, ce qui permet de faciliter l'utilisation et de la collection des articles de l'actualité.

4.6 Conclusion :

Nous avons consacré ce chapitre pour l'implémentation de notre système. Nous avons configuré plusieurs variantes d'expérimentations afin de pouvoir réaliser la meilleure solution pour notre cas. Nous avons fait une évaluation pour chacune des variantes en utilisant des techniques d'évaluation de performance. Ensuite on a conçu une interface qui est développée avec une représentation des différents sujets. Et enfin nous avons défini l'environnement et l'ensemble d'outils utilisés.

Conclusion générale

Il était question dans ce mémoire de fin d'étude Master de concevoir et développer un système d'annotation et de classification automatique des articles d'actualité. Les motivations et l'intérêt de développer un tel système viennent du fait qu'aujourd'hui, les internautes se retrouvent quotidiennement face à un flux d'informations massif, diversifié et ininterrompu de flux d'actualité véhiculé par les organes de presse, les réseaux sociaux et les sites spécialisés.

Dans ce contexte, l'utilisateur lambda ne dispose toujours pas de moyens satisfaisants pour gérer ce flux. Le besoin d'outil lui permettant d'avoir une synthèse des sujets d'actualité traités bien organisée et facile à explorer se fait de plus en plus ressentir.

Les techniques d'apprentissage automatique, notamment la classification automatique sont des outils très puissants qui permettent de faciliter l'exploration d'une grande masse de données et de détecter automatiquement des populations similaires et ou dissimilaires. Elles permettent aussi d'organiser un ensemble de données selon des classes ou des hiérarchies de classes.

Nous nous somme basé sur ces techniques pour concevoir et développer un système de classification automatique des articles d'actualité dans le but de générer automatiquement une synthèse d'actualité en détectant d'une manière automatique les différents sujets d'actualités traités. Dans le but d'améliorer les performances et surtout la précision de la classification, nous nous somme basé sur une représentation à base d'entités nommées des contenus des articles au lieu d'aller sur une représentation en texte brute. Nous avons aussi essayé d'exploiter l'importance de la rubrique TITRE des articles dans la pondération des contenus des articles étant donné que le titre d'un article donne une idée globale sur le sujet traité dans l'article d'actualité.

Nous avons enfin implémenté notre système et nous avons évalué plusieurs variantes afin de confirmer la justesse de nos hypothèses de départ. Toutefois, et bien qu'étant arrivés à des résultats que nous jugeons satisfaisants, il existe des perspectives futures pour notre travail comme par exemple, le passage à l'échelle c'est-à-dire permettre toujours un temps de réponse et une précision satisfaisante quelques soit le nombre d'articles et de sources d'actualité considérées.

Bibliographie

[1] Yves Citton,*"Études de média comparés"* .Cours université Paris Vincennes-Saint Denis, 2018

[2] https ://www.acuite.fr/dossiers/les-medias-parlent-de-vous
Consulté le : 02/10/2020-20 :56

[3] Iqbal, Haider. *"Selecting an Appropriate Source of Media as an Effective Source of Promotion and Communication From ATL and BTL Modes of Advertising (A Study of FMCGs in Peshawar).*" CITY UNIVERSITY RESEARCH JOURNAL 3.2, 2013.

[4] De Rooij Laurens & Hoover Stewart. *"Television. Encyclopedia of Economics and Society"*, 2015.

[5] statista.com/statistics/268695/number-of-tv-households-worldwide/
Consulté le : 15/10/2020-10 :12

[6] Julie Yonneau .*"Le titre de l'article attire le regard"*.Cours Université Fontenay-le-Fleury, Yvelines, 2007

[7] Cours de l'Organisation des Nations unies pour l'alimentation et l'agriculture, *"Approcher les médias"*, 2012

[8] https ://www.zenithusa.com/top-30-global-media-owners-2017/
Consulté le : 05/11/2020-11 :52

[9] A. BÈguec, H Coste, *"Qu'est-ce que l'actualité ? "*, Mémoire recherche, 2005

[10] Head, Alison, Wihbey, John, Metaxas, P. Takis, et al. *"How students engage with news : Five takeaways for educators, journalists, and librarians"*..Page 18. Project Information Literacy Research Institute, 2018

[11] Thelwall, M., Byrne, A. & Goody, M."*Which types of news story attract bloggers ?*" .Page 327.Information Research, 12(4),2007

[12] similarweb.com/fr/top-websites/category/news-and-media
Consulté le : 17/09/2020-19 :06

[13] Ahmad Mazyad. *"Contributions to Automatic Text Classification : Metrics and Evolutionary Algorithms"*, Université Littoral Côte d'Opale, Thèse de docotrat. PhD thesis, 2018.

[14] Yahaya Sanoussi. *"Amélioration du système de recueils d'information de l'entreprise Semantic Group Company grâce à la constitution de ressources sémantiques"*, Thèse de docotrat. PhD thesis, 2017.

[15] Gries, Stefan Th, Andrea L. Berez. "*Linguistic annotation in/for corpus linguistics.*" Handbook of linguistic annotation. Springer, Dordrecht.Livre Pages 383-385. 2017

[16] Claude Martineau, Elsa Tolone, Stavroula Voyatzi, *"Les Entités Nommées :usage et degrés de précision et de désambiguïsation"*, 26ème Colloque international sur le Lexique et la Grammaire (LGC'07),2007

[17] Flitti Sarah, *"Identification Automatique d'Entités Nommées"*, mémoire Master, 2017

[18] Toucherifte Samira. *"Étude comparative en classification non-supervisé "*.Thèse de doctorat, 2011.

[19] Guénaël Cabanes.*"Classification non supervisée à deux niveaux guidée par le voisinage et la densité"*.Thèse de doctorat, 2010

[20] Soufiane Khedairia.*"Contribution à la classification non supervisée : Application aux données environnementales"*.Thèse de doctorat, 2014.

[21] Korde Vandana. *"Text Classification and Classifiers :A Survey."* International Journal of Artificial Intelligence & Applications. 3. Pages 85-99, 2012

[22] Ricco Rakotomalala, *"Construction de la matrice documents termes "*, Cours université Lyon 2,2016.

[23] Raheel Saeed *"L'Apprentissage Artificiel pour la Fouille de Données Multilingues"* . PhD thesis, 2010.

[24] Mohamed Boughanem .*"Évaluation des performances dans les systèmes de recherche d'information"* .Page 5,6.

[25] M.-J. Huguet.*"Apprentissage non supervisé :Méthodes de Clustering"*. Cours de l'institut des sciences appliquées de Toulouse, page 27,2019.

[26] Matthieu Constant.*"Traitement Automatique des Langues"*.Master Informatique, Université Paris-Est Marne-la-Vallée.

[27] Massand.B, Linoff. G, Waltz. D*" Classifier les actualités Histoires utilisant le raisonnement basé sur la mémoire "* .Page. 59-65, 1992

[28] Jamal Atif . *"Data Mining/ML Kppv et Naive Bayes"*.Cours Master 2 Université Paris-Dauphine .2015

[29] Nesrine Masmoudi. *"Modèle bio-inspiré pour le clustering de graphes : application à la fouille de données et à la distribution de simulations"*. Intelligence artificielle [cs.AI].Université de Normandie ; Université de Sfax (Tunisie).Thèse doctorat, 2017 .

[30] Gildas Tagny Ngompe. *"Méthodes D'Analyse Sémantique De Corpus De Décisions Jurisprudentielles"*, Mines Alès Ecole Mines - Télécom, These doctorat, 2020.

[31] Abderraouf Boukhatem, Alexandre Duhamel et David-Alexandre Eklo.*"Étude de méthodes de Clustering pour la segmentation d'images faciales, Université Paris-Dauphine"*.page 6, 2017

[32] Adèle Désoyer. *" Appariement de contenus textuels dans le domaine de la presse en ligne : Développement et adaptation d'un système de recherche d'information "*, Université Paris Nanterre, Thèse de docotrat. 2017.

[33] https ://spacy.io/

Consulté le : 03/10/2020-11 :30

[34] Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, et al. *"Scikit-learn : Machine learning in Python"*, The Journal of machine Learning research, vol. 12, p. 2825-2830,2011

[35] Kadiyala Akhil et Kumar Ashok. *"Applications of python to evaluate environmental data science problems"* Environmental Progress & Sustainable Energy, vol. 36, no 6, p. 1580-1586, 2017