MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DELA RECHERCHESCIENTIFIQUE

UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master en Informatique

Option : **Ingénierie des Systèmes d'Information**

Présenté par :

Mortet Hadjer

Encadrant(e) :

Kenniche Ahlem

THÈME :

Classification des entités nommées

Année Universitaire 2022-2023

Résumé

La reconnaissance des entités nommées fait aujourd'hui figure d'incontournable en Traitement Automatique des Langues est basé sur des règles linguistiques qui exploitent l'étiquetage syntaxique, des déclencheurs et des dictionnaires de noms propres. La tache de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. Notre rapport présente les différentes étapes à utiliser de la conception d'un Système de recherche d'entités nommées sur Wikipédia. Ce système s'appuie sur la classification des entités nommées par année, catégorie et par langue est il est basé sur la méthode et l'entropie de Shannon.

Mots-clés:

Reconnaissance d'entités nommées, classifications, notoriété, Wikipédia, l'entropiede Shannon.

Abstract

The recognition of named entities is now essential in Natural Language Processing is based on linguistic rules that exploit syntactic labeling, triggers and dictionaries of proper names. The task of recognizing and categorizing the names of people, places, organizations, etc.Our report presents the different steps to use in the design of a Named EntitySearch System on Wikipedia. This system is based on the classification of named entities by year, category, language based on Shannon's entropy method.

Keywords:

Named entity recognition, classifications, notoriety, Wikipedia, Shannon entropy.

Dédicaces

C’est avec grand respect et gratitude que je tiens à exprimer toute ma reconnaissance et ma sympathie et dédier ce travail modeste à mes parents MOHAMED et KHEIRA , les mots ne sauraient exprimer l’immense et profonde gratitude que je leur témoigne ici pour leurs précieux soutien, pour leurs patience, pour avoir crus en moi, pour leurs sourires réconfortants et pour leurs sacrifices qui m’ont permis d’atteindre cette étape dans ma vie et qu’ils n’ont jamais cessé de consentir pour mon instruction et mon bien être dieu me les gardes et les protèges.

- A mes deux sœurs HAKIMA et SARA et mon frère ABDERAHMAN.
- Toute ma famille et surtout mes grand-mère HOURIA et FATMA.
- À mes meilleures amies MERIEM, MAROUA, NIHEL, NESSRIN, ZOHRA, SOUHILA.
- Tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.
- Tous mes enseignants tout au long des cycles de mes études.

Remerciements

Je remercie ALLAH de me avoir donné la force et la capacité qui ma menées à ce niveau.

Je tiens à remercier notre encadrent Mme KENNICHE AHLEM pour son encadrement son écoute, ses élucidations, ses conseils, ses directives et encouragements qu'elle ma a afflué.

Je remercie aussi les membres de jury pour avoir bien voulu donner de leur temps pour lire et juger notre projet.

Merci aussi à tous mes amis(es), mes collègues, et à tous ceux qui m'ont aidé de près ou de loin. Je leur exprime ma profonde sympathie et leur souhaite beaucoup de bien.

Jen'oublierai pas non plus de remercier toutes les personnes que j'ai pu côtoyer pendant ces cinq ans à la faculté des sciences exactes et d'informatique pour leur soutien moral et amical.

Liste des figures

Liste des tableaux

Liste des abréviations

Abréviation	Expression Complète	Page
REN	Reconnaissance des Entités Nommées	6
EN	Entités Nommées	6
TAL	Traitement Automatique des Langue	8
L	Wikipédia	30
HPI	l'indice de popularité historique	30

Liste des équations

Numero d'équations	Expression
1	Calcul du poids de chaque critère
2	Calcul l'entropie Ej
3	Calcul le poids Wj de chaque critère
4	Calcul des scores de la méthode SAW
5	La répartition entre trios valeurs de notoriété
6	Attribuons la notoriété

Sommaire

Introduction Générale

Chaque fois que nous entendons un mot ou lisons un texte, nous avons la capacité naturelle d'identifier et de catégoriser le mot en personnes, lieu, emplacement, valeurs, etc.

En philosophie du langage, un nom propre désigne un être linguistique, comme étant une expression, un signe ou une combinaison de signes qui désigne un objet déterminé (au sens large, privé des concepts et des relations). Et pour cette linguistique a pu profiter de la puissance des ordinateurs pour acquérir une nouvelle dimension et ouvrir la voie à de nouveaux domaines de recherche, parmi ces domaines la recherche d'entité nommées (REN).

Dans notre travail Nous utiliserons comme corpus l'encyclopédie Wikipédia qui est une ressource libre. La Wikipédia nécessite un développement et un enrichissement permanents via l'exploitation des ressources libres et riches en textes du web sémantique, entre autres. La notoriété d'une personne est sa renommée publique, le fait qu'il soit connu (ou non), et l'adjectif peut signifier que l'entité est connu d'un grand nombre de personnes, qui fait partie de la sphère publique, qui est célèbre. Dans ce mémoire nous proposons une méthode pour la classificationdes entités nommées en utilisant l'entropie de Shannonpour mesuré la notoriété de celle- ci. Nous utilisant comme corpus les éditions de la wikipedia dans plusieurs langues. Notre but et de classifier les entités par notoriété type par type et voire l'évolution de celle-ci selon l'années et la langue.

Dans le premier chapitre nous avons abordé la définition de la reconnaissance des entités nommées. Puis nous allons décrire les catégorisations. Enfin, nous allons aborder les trois Approches de reconnaissance des EN (linguistique, statistique et hybride).

Dans le deuxième chapitre, nous allons rappeler la définition de la Wikipédia. Puis, pour chaque approche nous allons présenter les systèmes de reconnaissance des entités nommées. De plus, nous allons définir la Wikimédia. Finalement, nous allons citer les trois méthodes qui existe pour calculer la notoriété d'un nom propre ; l'entropie de Shannon, Wikipédia Index et Panthéon.

Dans le troisième chapitre, nous allons décrire les différentes étapes à suivre pour construire notre travail (choisi les url, choisi la catégorie, …), calculer la notoriété on a utilisé l'entropie de Shannon et la modélisation de notre système à travers le langage UML.

Dans le quatrième chapitre, qui est un chapitre d'implémentation nous représentons les résultats, présentation de l'application et les différentes technologies et les outils (JavaScript, html, CSS, etc…) que nous pouvons utiliser lors du processus de création de notre système.

Ce rapport se termine par une conclusion générale.

Chapitre 1

Reconnaissance d'entités nommées

1.1 Introduction

La reconnaissance d'entité nommée (REN) est une tâche importante du traitement automatique des langues (TAL), et qui sert généralement de point de départ à d'autres taches telles que l'extraction d'informations, etc.

Le système de la reconnaissance des entités nommées est basé sur des règles linguistiques qui exploitent l'étiquetage syntaxique, des déclencheurs et des dictionnaires denoms propres est de traiter des données structurées etnon structurées.

La plupart des systèmes d'étiquetages utilisent des grammaires formelles associées à des modèles statistiques, éventuellement complétées par des bases de données (listes de prénoms, de noms de villes ou de pays par exemple Algérie, Paris).

Dans ce chapitre on s'intéresse à l'étude et à la reconnaissance des entités nommées (EN) et les approches utilisées pour cela.

1.2 La reconnaissance des entités nommées

Le concept d'entité nommée est apparu dans les années 90 à l'occasion de conférence d'évaluation MUC (Message Under standing Conférence) (Grishman et Sundheim 1996).

Ces conférences avaient pour but de promouvoir la recherche en extraction d'information. Les tâches proposées consistaient à remplir de façon automatique des formulaires concernant des événements. Dans ce cadre, certains objets textuels,

ayant une importance applicative particulière dans plusieurs domaines du TAL ont été regroupés sous le nom d'entités nommées.

La reconnaissance de ces dernières est donc considérée comme une sous-tâche à part entière de l'extraction d'information.[4]

1.3 Les catégories des entités nommées

1.3.1 Noms propres

Le nom propre est donc une dénomination attachée à une personne, un peuple, un lieu, une marque, une institution, un animal.
Ils tirent leur origine du langage courant comme les noms géographiques et les noms donnés aux personnes ou aux divinités païennes.
Par exemple : François, Chien, reine Elisabeth 2.[10]

1.3.2 Noms de lieux

Les noms de lieux désignent les villes, les pays, les villages, les montagnes et les fleuves.[10]
Par exemple : le Nil, Paris, l'Everest.

1.3.3 Noms d'organisations

Les noms d'organisations sont assez nombreux et sont difficilement quantifiables puisque leur apparition et leur disparition dépendent de la situation dans le monde.[10]
Par exemple : l'ONU, FMI, l'ONG.

1.3.4 Entités numériques

Les entités numériques sont définies comme un lien technologique entre une entité réelle (personne, organisme ou entreprise) et des entités virtuelles (sa ou ses représentations numériques).

Elles sont divisées en deux catégories : les expressions de temps et les nombres.

Les expressions de temps incluent les dates, la période et toute autre expression exprimant le temps.[10]

Par exemple : 28 Juillet 1914 (Début de la première guerre mondiale).

Les nombres incluent principalement les systèmes de mesures (poids, distance, volume, vitesse), les pourcentages, ainsi que les devises.

Par exemple : 384 400 km (Distance entre la terre et la lune), 72%(Pourcentage d'eau qui recouvre la surface du globe).

1.4 Approches de reconnaissance des EN

Les approches de reconnaissance d'EN trouvent leurs origines dans le domaine de la linguistique computationnelle. Il connut un fort développement depuis la fin des années 80 sous l'impulsion des conférences MUC.

Aujourd'hui, toutes les approches offrent des taux de reconnaissance (repérage et catégorisation élémentaire) au-dessus de 90%. Cependant, l'attribution des catégories reste une tâche assez complexe.

Il existe trois approches fondamentales de REN qui peuvent être fondées sur des démarches linguistiques ou non et qui sont nommées linguistique, symbolique, statistique ou à base d'apprentissage et hybride.

Les trois approches permettent de réaliser les mêmes objectifs définis tandis qu'elles admettent des principes différents.[10]

D’autres facteurs peuvent distinguer aussi ces approches comme l’acquisition de données et leur manipulation. La section courante se divise en trois sous-sections dont chacune est dédiée à présenter une approche de REN.

1.4.1 Approche linguistique

L’approche linguistique repose sur la construction manuelle des règles à formaliser via des grammaires et à appliquer sur le corpus étudié.

Les règles construites ont la forme des patrons dont ils reposent sur les caractéristiques ; morphologique, syntaxique et sémantique.

Les règles peuvent être alimentées à l’aide des lexiques spécifiques. Etant donné que ces lexiques sont parfois non exhaustifs et ouverts alors la création des règles peut se baser sur des indices contextuels (des preuves externes déclenchant les EN) [38].

Dans le cas de la reconnaissance d’un nom de personne, une règle linguistique peut se composer d’un mot déclencheur (السيد
/Monsieur) plus deux mots se trouvant respectivement dans le lexique des prénoms et celui des noms de famille.

L’approche linguistique repose sur des règles lisibles ce qui permet de cerner les erreurs rencontrées lors de leur application sur un texte.

1.4.2 Approche statique

L’approche statistique repose sur diverses techniques d’apprentissage qui se diffèrent au niveau du degré de supervision exigé. La supervision concerne l’intervention humaine pour l’étiquetage de l’ensemble de données dontl’objectif est de guider un modèle d’apprentissage déjà conçu.

Au sein de l’approche statistique, nous distinguons trois types d’apprentissage ; supervisé, semi-supervisé et non supervisé. L’apprentissage supervisé consiste à

créer un modèle à la base d'un ensemble de données annotées (nom de catégories) afin de classifier des nouvelles données.

L'apprentissage semi-supervisé utilise deux ensembles de données ; annotées et non annotées visant à améliorer la qualité d'apprentissage. Le dernier type est l'apprentissage non supervisé visant à créer des classes de données à partir d'un ensemble de données non annotées. Chaque type déjà cité fait recours à des algorithmes d'apprentissage qui seront appliqués pour entraîner le système élaboré [36].

1.4.3 Approche hybride

L'approche hybride est la combinaison des approches symbolique et statistique. La direction du flux de traitement peut être du symbolique vers le statistique ou vice versa. Autrement dit, les règles sont, soit écrites manuellement puis corrigées et améliorées automatiquement, soit apprises automatiquement puis révisées manuellement [37].

La combinaison de ces deux approches augmente la puissance descriptive des règles linguistiques d'une part et remède aux faiblesses d'apprentissage d'autre part. Cette approche aide à atteindre des améliorations importantes pour la performance des systèmes de REN.

1.5 Conclusion

Dans ce chapitre, nous avons abordé le concept de base de la reconnaissance des entités nommées, les catégories des entités nommées et les approches fondamentales de la reconnaissance des entités nommées.

Le chapitre suivant sera consacré à la description de la Wikipédia, sa structure générale, la Wikimédia, l'accès au contenu de Wikipédia, et le calcul de notoriété d'un nom propre via Wikipédia.

Chapitre 2 Wikipédia

2.1 Introduction

La Wikipédia est une encyclopédie collaborative universelle et multilingue en ligne qui fonctionne sur le principe de wiki, c'est-à-dire une application web permettant la modification des pages web écrites en utilisant un langage de balisage par ses visiteurs via un navigateur web.

La richesse de la Wikipédia en termes d'EN et son aspect multilingue ont joué un rôle important pour la proposition des systèmes de reconnaissance des entités nommées. Elle est visitée chaque mois par près de 500 millions de visiteurs et propose plus de 30 millions d'articles dans plus de 280 langues. Plus de 25 000 articles sont créés par jour sur les différentes versions linguistiques de la Wikipédia et on compte plus de 10 millions de modifications par mois ; à l'heure actuelle, les éditions de la Wikipédia les plus importantes en nombre d'articles sont la Wikipédia en anglais, en allemand, en français, en néerlandais et en suédois.

La Wikipédia constitue une ressource pour les chercheurs et les développeurs travaillant sur des problèmes de bases de données, d'indexation ou de classification de documents. [1]

2.2 Wikipédia

2.2.1 Définition

Le terme Wikipédia est étymologiquement issu de la fusion de deux termes : wiki-, issu de l'hawaïen wiki, qui signifie rapide, se référant au fait que l'encyclopédie ait toujours vocation à s'améliorer rapidement et à être constamment active par son

mode de fonctionnement, et -pédia, lui-même dérivé du mot grec paideia, instruction et éducation.[11]

2.2.2 Structure générale d'une page Wikipédia

Les articles de la Wikipédia sont constitués dans les différentes versions linguistiques avec généralement une structure quasi identique ; ils se composent de textes écrits en langage naturel, d'images et aussi d'autres informations structurées et de plusieurs types de liens. Ci- dessous, nous détaillons certainsde ces composants.[1]

2.2.2.1 Les info boxes

Elles représentent les caractéristiques d'une entité donnée, correspondent aux tableaux reprenant des informations factuelles et structurées, et sont placées en général en haut à droite de certains articles. Le contenu de ces Info boxes est une base pour l'alimentation de la base de donnéesDBpedia, cependant, leur présence est limitée ; dans le cas des articles biographiques, moins d'un article sur trois propose ainsi une Info box ; les biographies font pourtant partie d'un des types d'articles les plus fréquents sur la Wikipédia.[15]

Les Info boxes ou les boîtes d'information affichent des informations pertinentes pour le sujet de l'article en utilisant la fonctionnalité du logiciel modèle considérant le type d'entrée ; ces informations peuvent être des clés pour les recherches d'informations.[1]

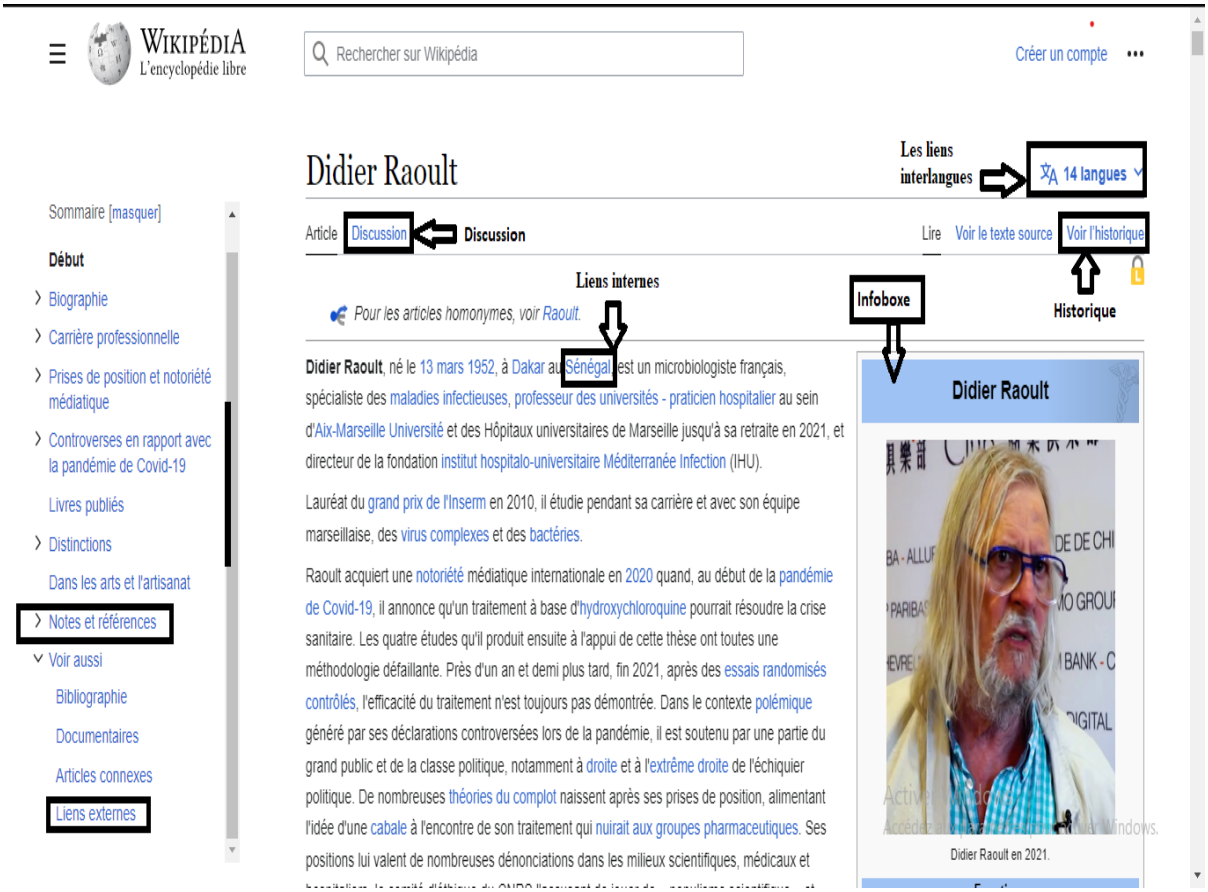


Figure 1 : Une capture d'une partie de la page de Didier Raoult (Version française) qui montre la page entière (l'info boxe, discussion, les liens inter langues, historique, les liens internes, les liens externes, notes et références)

2.2.2.2 Les catégories

Elles indexent chaque page de la Wikipédia oùun ensemble de catégories mères visibles etcliquables par l'utilisateur est placé en bas de chaque page.[1]

2.2.2.3 L'historique

Il désigne un lien nommé « Historique » dans laversion française, placé en haut à droite, près du moteur de recherche ; via ce lien on peut accéder à la page de

l'historique conservant l'ensemble des modifications qui ont été effectuées à la page cible depuis sa création. La page de l'historique permet de connaître la date, l'auteur et la teneur exacte de chaque modification ; elle contient des outils externes et statiques relatifs à la page cible : Auteurs et statistiques, Recherche de l'auteur d'un passage de l'article, Statistiques de consultation, Contributeurs suivant cette page et Modifications par utilisateur.[1]

2.2.2.4 La discussion

Il existe un lien appelé « Discussion » (en français) en haut à gauche de la page, qui conduit vers la page de discussion où se trouvent les différents points de vue des contributeurs et les résultats du système d'évaluation fourni par le projet Wikipédia sur le contenu de la page cible.[1]

Figure 2 : Capture d'écran de la page Didier Raoult qui montre la Discussion [8]

2.2.2.5 Pages liées

C'est un lien vers une page d'outil via lequel on peut connaître la liste des pages liées à la page cible ; cette page contient un outil externe pour lenombre de pages

liées, les inclusions, les liensinternes et les redirections contenus dans la page cible.[1]

Figure 3 : Capture d'écran d'une partie de la page Didier Raoult (Version française) qui montre les pages liées, les informations sur la page et l'aide [5]

2.2.2.6 Informations sur la page

C'est un lien vers une page contenant des informations de base sur la page cible comme le titre, la taille, le nombre de contributeurs, le nombre de redirections vers cette page et d'autres informations.[1]

2.2.2.7 Les liens inter langues

Ce sont des liens vers les articles correspondants dans les autres langues ; ces liens sont situés dans un cadre à gauche de la page.

Ainsi, le lecteur ou le contributeur peut trouver l'article équivalent dans les autres langues.[1]

2.2.2.8 Les liens inter wiki

Ils sont appelés aussi liens inter-projet car ce sont des liens entre les différents projets de la fondation Wikimédia ; ce sont des liens intégrésdans le texte comme les liens internes ordinaires, à utiliser principalement dans les discussions, en dehors donc des articles.[9]

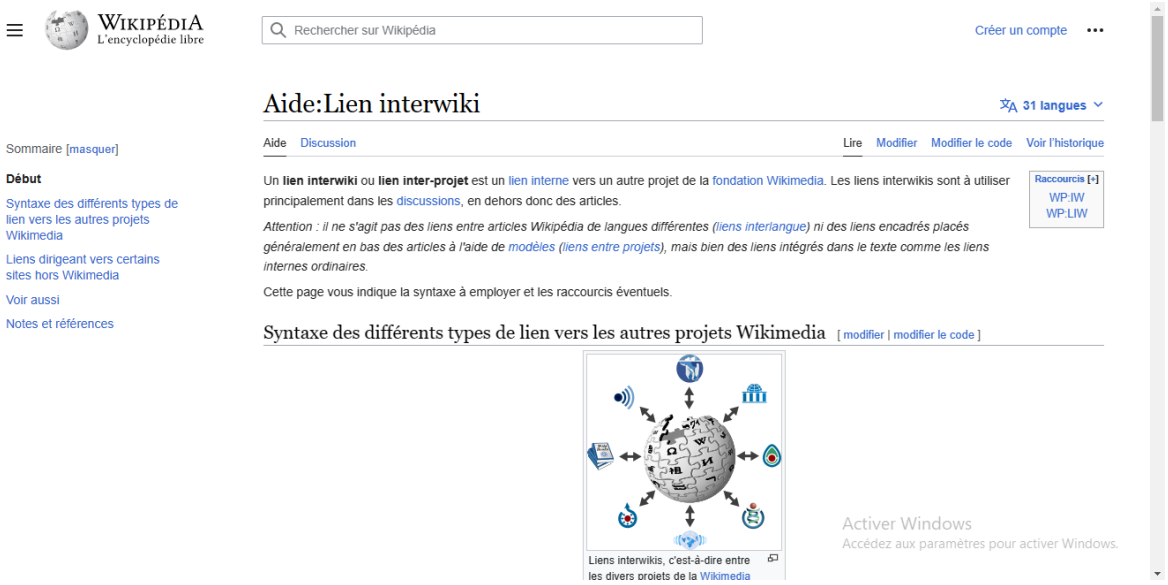


Figure 4 : Capture d'écran de Didier Raoult qui montre un Aide :lien interwiki

2.2.2.9 Liens externes

Ce sont des hyperliens qui mènent vers d'autres sites web que la Wikipédia. Dans lesarticles de la Wikipédia, on peut en trouver à deux endroits différents. Toutd'abord, dans la liste des sources permettant de vérifier ce qui est écrit dansl'article. Ce type de lien externe, aussi appelé source ou référence, est généralement regroupé dans une section intitulée Références ou bien Notes et références. Undeuxième endroit possible pour ces liens est une section tout simplement appelée Liens externes en fin d'article.[7]

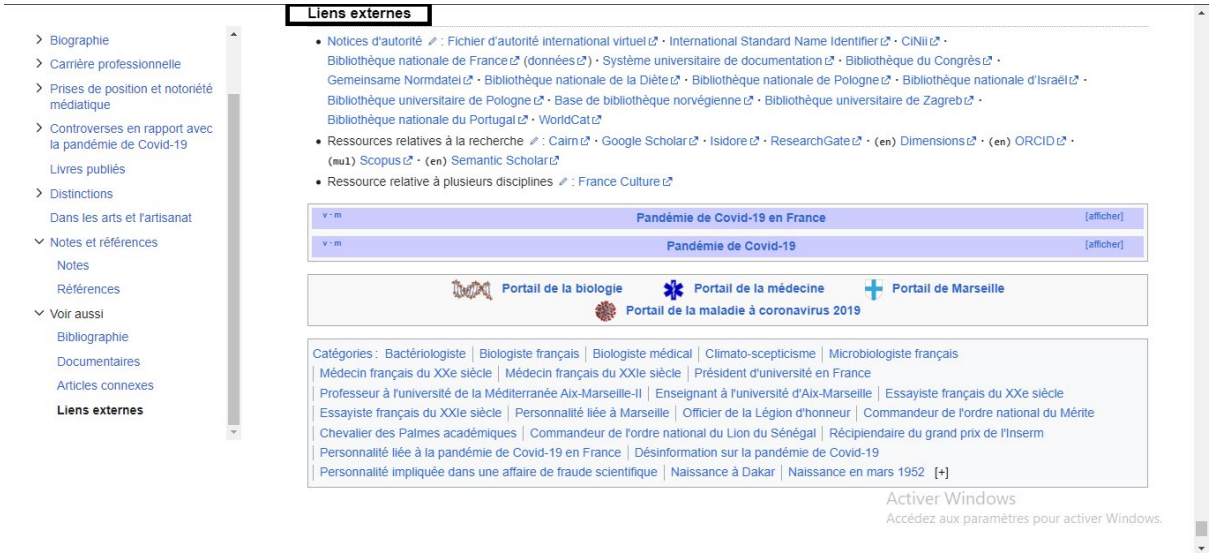


Figure 5 : Capture d'écran de la page Didier Raoult qui montre les liens externes [7]

2.2.2.10 Liens internes

Ce sont des liens internes à la Wikipédia ou wikiliens pointant vers d'autres articles de la Wikipédia ; ils se mettent dans le corps de l'article. Leur utilisation peut parfois pécher dans leur pertinence (le lien doit apporter une information utile), leur efficacité (l'article correspondant à ce lien doit exister et le lien ne doit pas être répété) ou leur esthétique. Les liens internes connexes à un article sont regroupés en fin d'article dans unesous-rubrique Articles connexes de la rubriqueVoir aussi, un lien interne s'affiche par défaut enbleu et quand il pointe vers un article qui n'existe pas, il s'affiche en rouge.

2.2.2.11 Notes et Références

Elles se trouvent à la fin d'un article Wikipédia et elles sont des sources qui sont insérées dans le texte d'un article en les précédantpar «↑» pour les distinguer des

autres types[12]. Pour terminer cette section, nous illustrons en images des exemples clarifiant certains composants qui sont mentionnés plus haut.[5]

nous illustrons en images des exemples clarifiant certains composants qui sont mentionnés plus haut ; la Figure 1 représente une partie de la page Didier Raoult dans l'édition Wikipédia française en entourant L'info boxe , discussion , les liens inter langues , historique, les liens internes , les liens externes , notes et références. La Figure 2 et 3 indique le résultat de l'évaluation de cette page dans la page de Discussion, les pages liées,les informations sur la page, l'aide ; la page Wikipédia française de Didier Raoult reçoit le stade A pour l'avancement ce qui signifie « Article Avancé » ; finalement, les Figures4 ,5 et 6 montrent respectivement la forme des références, liens externes et lien interwiki se trouvant dans la page « DidierRaoult ».

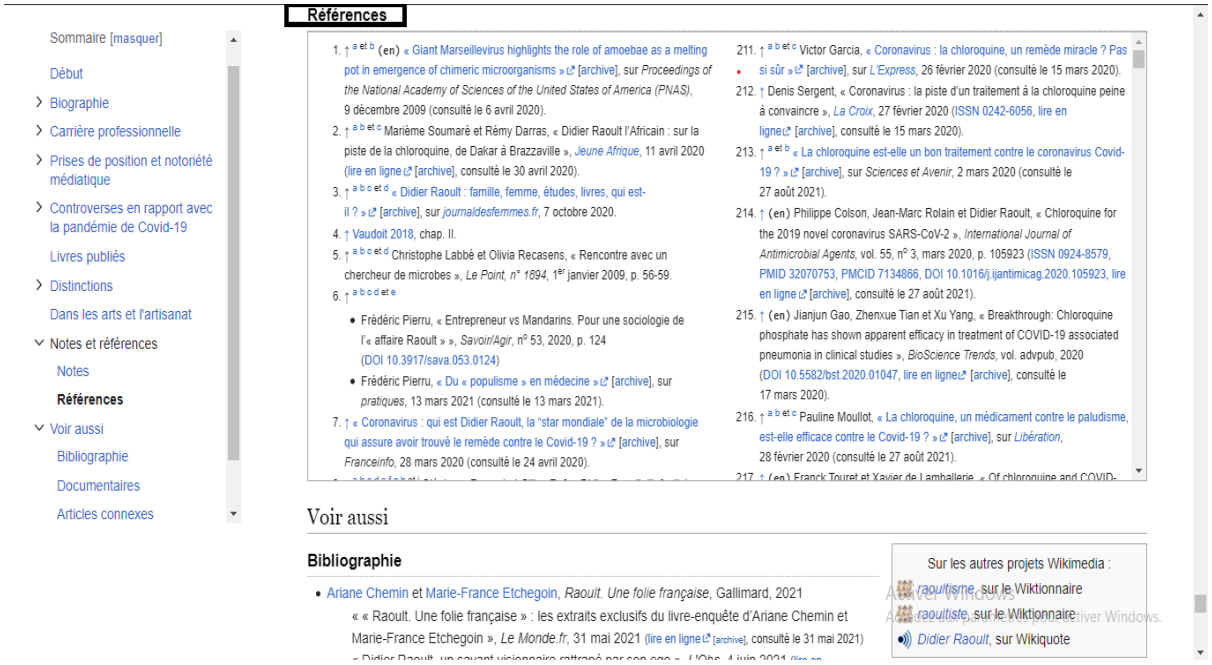


Figure 6 : Capture d'écran de la page de Didier Raoult qui montre les Références [6]

2.3 Wikimédia

2.3.1 Définition

Wikimédia est un mouvement international qui apporte un contenu éducatif gratuit, et ça à travers différents projets, chapitres, et une structure adéquate.[19]
Ainsi elle va faire fonctionner l'un des plus importants projets de l'édition collaborative, qui est Wikipédia.

2.4 L'accès au contenu Wikipédia

Les outils qui permettent d'accéder au contenu de l'encyclopédie Wikipédia sont les suivants :

2.4.1 Les dumps

Le terme dump désigne la copie brute de données des projets Wikimédia.

Ils contiennent les publications, les historiques, les métadonnées, les liens externes qui sont des fichiers de grandes tailles en format XML ou SQL.
On peut télécharger les sauvegardes après avoir installé le logiciel Média Wiki.[13]
Néanmoins, cette méthode peut contenir quelques problèmes parce qu'elle utilise beaucoup de mémoire vive.

2.4.2 API Médiawiki

C'est un service web qui permet d'accéder à certaines fonctionnalités du Wiki comme accéder à ses bases de données, ses métadonnées en utilisant des requêtes HTTP.
La requête doit commencer par une URL principal, par exemple l'URL pour l'API Wikipédia néerlandais https://nl.wikipedia.org/w/api.php
Via ses requêtes, les clients demandent des actions en définissant un paramètre « query »[16], ça permet de récupérer les informations sémantiquement concernant l'historique, la liste des contributeurs, les modifications récentes, etc.
Les réponses à ces questions sont en format JSON [17] ou XML.[1]
Par exemple : Spécification de pages par ID de page.
api.php ?action=query&pageids=123|456|75915&format=json&formatversion=2

2.4.3 DBPEDIA

DBpedia est un projet universitaire et communautaire d'exploration et extraction automatiques de données dérivées de Wikipédia. Son principe est de proposer une version structurée et normalisée au format du web sémantique des contenus de Wikipédia. [14]
DBpedia vise aussi à interconnecter Wikipédia avec d'autres ensembles de données ouvertes provenant du Web des données.

L'accès aux données se fait à l'aide d'un langage d'interrogation de type SQL pour RDF, appelé SPARQL[18].

Par exemple, on va s'intéresser à la série shōjo japonaise Tokyo Mew Mewet on veut rechercher d'autres œuvres réalisées par son illustrateur Mia Ikumi.

DBpedia combine des informations provenant des entrées de Wikipédia sur Tokyo Mew Mew, Mia Ikumi et sur des œuvres telles que Super Doll Licca-chan et Koi Cupid.

```
PREFIX dbprop: <http://dbpedia.org/ontology/>
PREFIX db: <http://dbpedia.org/resource/>
SELECT ?who, ?WORK, ?genre WHERE {
 db:Tokyo_Mew_Mew dbprop:author ?who .
 ?WORK  dbprop:author ?who .
 OPTIONAL { ?WORK dbprop:genre ?genre } .
}
```

Figure 2 : Figure qui représente la recherche des genres d'œuvres ressemblantes à celle de la série Tokyo Mew Mew

2.5 Estimation de la notoriété d'un nom propre via Wikipédia

La notoriété est un terme signifiant célébrité, réputation, renommée et l'objectif de calculer la notoriété est conté le nombre de consultations. Nous allons maintenant présenter les trois méthodes pour l'estimation de la notoriété d'un nom propre qui sera basée sur des liens vers Wikipédia.

2.5.1 L'entropie de Shannon [1]

2.5.1.1 Le choix des critères

L'ensemble des critères est déduit de la Wikipédia.

Il y a cinq critères ; les trois premiers critères concernent l'article, et les deux derniers concernent les liens de cet article avec les autres articles et l'ensemble de Wikipédia.[1]

Voici les cinq critères :

1. Le nombre de consultations de l'article
2. Le nombre de contributeurs à l'article
3. La taille de l'article
4. Le nombre de liens internes à la Wikipédia pointant vers l'article
5. Le nombre de liens externes à la Wikipédia contenus dans l'article

Les calculs ont été faits à l'intérieur d'une même édition linguistique, et le nombre de consultations de l'article a été fait sur l'ensemble des données disponibles sur le nombre de consultations mensuelles de l'article depuis 2008.[1]

2.5.1.2 Le calcul des indices

Pour le calcul du nombre de consultations il a été calculé d'après Mouna Elshter selon la période de 2008-2015 qui couvre 96 mois via Wiki média.

Vu que la notoriété d'un nom propre peut varier dans le temps, les auteurs ont attribué pour chaque un, un coefficient d'oubli. Ainsi janvier 2008 sera associé au coefficient 1/96 etc.

Donc pour chaque nom propre et pour chaque mois, les auteurs ont calculés le produit du nombre de consultations de l' article par le coefficient d'oubli du mois en question. Et en dernier lieu, la somme de ces valeurs sera faite. Cette somme représentera le premier indice.

Pour le calcul des autres indices, le nombre de contributeurs à l'article, la taille de l'article, le nombre de liens internes et le nombre de liens externes ;sont extrait en utilisant le service Web l'action API de Média wiki, l'un des outils d'exploitation du contenu de l'encyclopédie Wikipédia.

Ils ont illustré ça avec trois exemples de la version Wikipédia française avec les valeurs de ces cinq indices récupérés dans le tableau ci-dessous. Cette illustration a pour objectif de mettre au clair les différentes valeurs possibles qu'un article de la Wikipédia peut avoir selon la réputation et la célébrité de son sujet.[1]

Le tableau 1 représente 3 trois noms propres avec le Nombre de consultations, le Nombre des contributeurs,la Taille de l'article, le Nombre de liens internes et externes

Nom propre	Nombre de consultations	Nombre des contributeurs	Taille de l'article	Nombre de liens internes	Nombre de liens externes
Platon	1 481 827	513	1 546 77	3 120	43
David Beckham	1 398 370	499	82 251	518	63
Stefan Niesiolowski	1 298	11	487	4	0

Tableau 1 : Ensemble de trois noms propres et leurs valeurs de cinq indices de notoriété récupérées via le programme dans l'édition française de la Wikipédia.[1]

2.5.1.3 Le calcul de la notoriété

Après l'obtention des cinq indices de notoriété, il est maintenant tant de calculer la valeur finale qui est égale à 1, 2 ou 3.

Pour cela, la méthode SAW (simple additive weighting), nécessite d'attribuer un poids à chaque critère.[1]

Ce poids sera calculé en utilisant l'entropie de Shannon.[1]

Pour le calcul du poids de chaque critère, y aura pour chaque nom propre *i* et chaque critère *j*, nous normalisons les valeurs *xij* obtenues précédemment en une valeur *cij* comprise entre 0 et 1 ; si *m* est le nombre total de prolexèmes considérés dans une langue donnée :

$$cij = \frac{xij}{\sum_{i=1}^{i=m} xij} \text{pour } i = 1..m,\ j = 1..5 \qquad ...1$$

Puis, nous calculons l'entropie *Ej* (comprise entre 0 et 1) :

$$Ej = \left(\frac{-1}{ln\ (m)}\right) \sum_{i=1}^{i=m} [c\ ij\ ln(c\ ij)] \text{pour } j=1..5 \qquad ...2$$

avec, par convention, $cij \ln(cij) = 0$ pour $cij = 0$

Et le poids *Wj* de chaque critère qui est calculé via la formule suivante:

$$Wj = \left(\frac{1-Ej}{\sum_{j=1}^{j=5}(1-Ej)}\right) \quad \text{pour } j=1..5 \qquad ...3$$

Maintenant, c'est au tour du calcul des scores de la méthode SAW,

On multiplie chaque valeur normalisée *cij* par le poids *Wj* du critère correspondant et qui a été calculé juste avant et nous obtenons le score *Si* d'un nom propre en additionnant ces cinq valeurs :

$$Si = \sum_{j=1}^{j=5} cij * Wj \text{pour } i=1..m \quad ...4$$

Maintenant place à la répartition entre les trois valeurs de notoriété.

1 pour la plus forte notoriété, 2 pour une notoriété moyenne et 3 pour une notoriété faible.

Pour cela, nous attribuons tout d'abord la notoriété 1 aux prolexèmes de scores supérieurs à la moyenne plus l'écart-type de l'ensemble des scores :

$$\begin{cases} \bar{M} = \text{Moyenne(Si)} \text{ pour} i = 1..m \\ \bar{E} = \text{Ecart\_type(Si)} \text{ pour} i = 1..m \\ \quad \text{Si } Si > \bar{M} + \bar{E}, Ni = 1 \end{cases} \quad \ldots 5$$

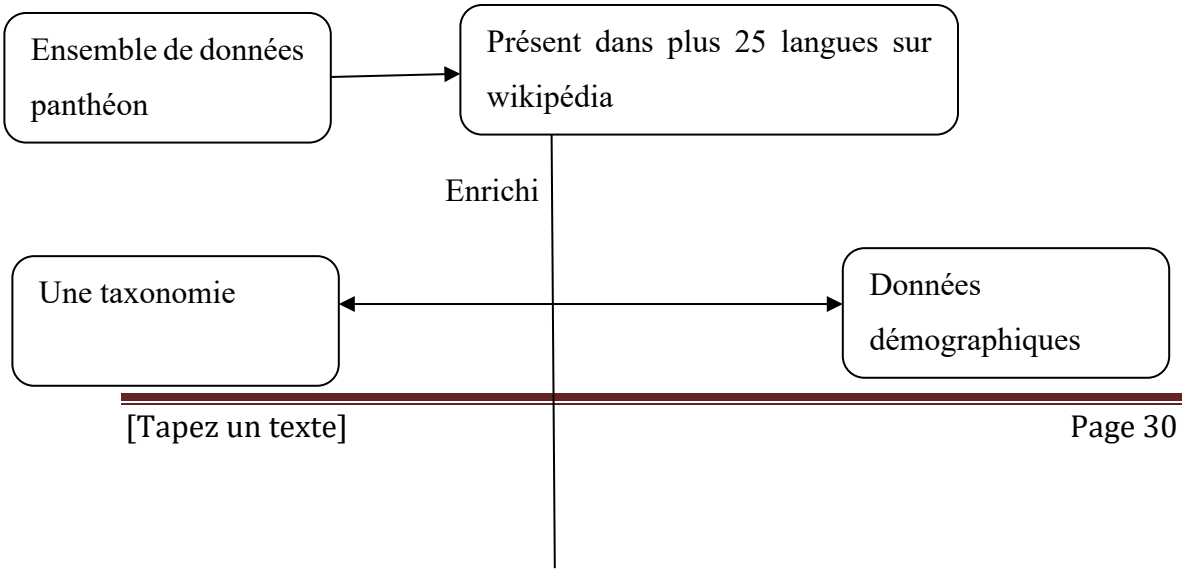
Ensuite, nous attribuons la notoriété 2 aux prolexèmes de scores supérieurs à la moyenne plus la moitié de l'écart-type de l'ensemble des scores restants et la notoriété 3 aux autres prolexèmes :

$$\begin{cases} \bar{\bar{M}} = Moyenne(Si) \text{ pour } \bar{S}i \leq \bar{M} + \bar{E} \\ \bar{\bar{E}} = Ecart\_type(Si) \quad \bar{S}i \leq \bar{M} + \bar{E} \\ Si \bar{M} + \bar{E} \geq Si > \bar{\bar{M}} + 1/2\, \bar{\bar{E}}, Ni = 2 \\ \quad \text{Si} Si \leq \bar{\bar{M}} + 1/2 \bar{\bar{E}}, Ni = 3 \end{cases} \quad \ldots 6$$

2.5.2 Panthéon [2]

2.5.2.1 Définition

La méthode Panthéon représente comme suite :



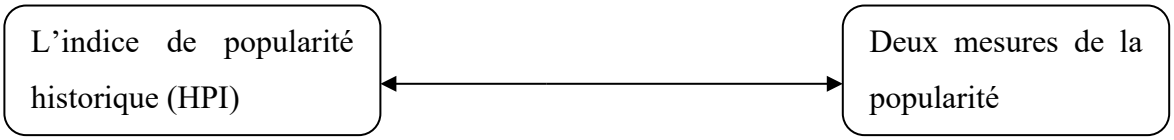


Figure 8 un diagramme qui définit panthéon

2.5.2.2 Méthodes

On utilisera une méthode pour appliquer Panthéon :

2.5.2.2.1 Collection des données

Puisque aucun ensemble de données de ce type n`existe, pour crée un ensemble de données plus simple concentré uniquement sur les informations biographiques en utilisant les données de Freebase et 277 éditions linguistiques de Wikipédia. Ensuite, Les auteurs ont lié les individus à leur page Wikipédia en anglais en utilisant leur identifiant d`article Wikipédia unique, et à partir de là, obtenu des informations sur les éditions de langues supplémentaires en utilisant l`API Wikipédia à partir de mai 2013. L’ensemble de données Pantheon 1.0 est limité aux 11 341 biographies présentes dans plus de 25 différentes langues dans Wikipédia (L>25).[2]

La figure 9 résume les principaux composants du flux de travail utilisé pour créer le jeu de données Panthéon.

On mettre sous forme de schéma suivent :

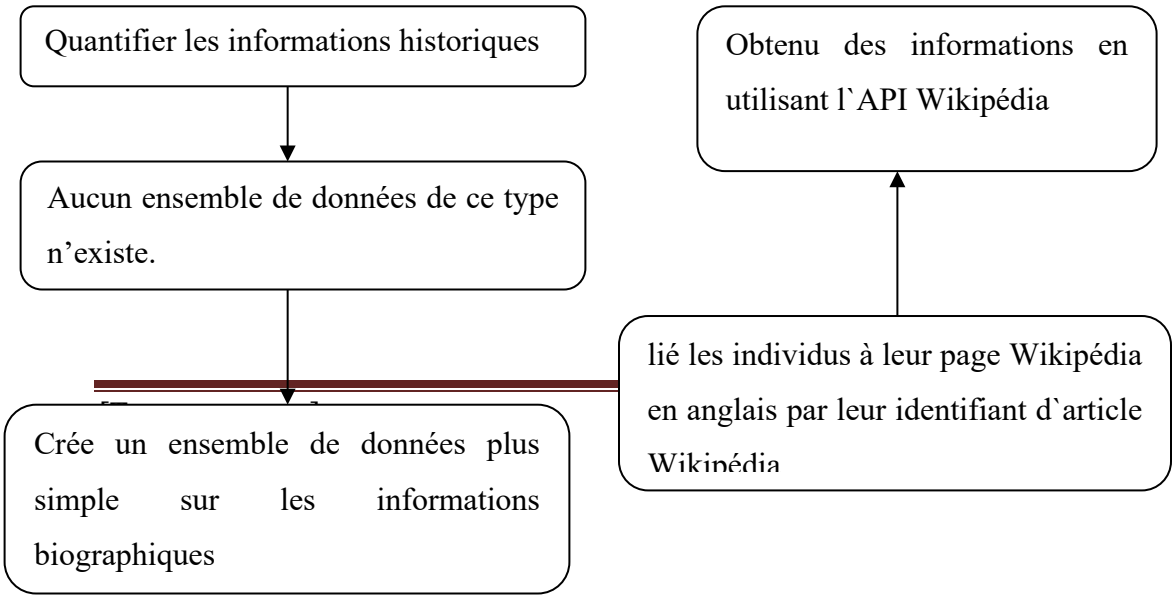


Figure 9 les flux de travail utilisé pour créer le jeu de données Panthéon

2.5.2.2.2Taxonomie de conception

Un système de classification normalisera l'échelle mondiale pour classer les biographies en fonction des professions, pour introduise une nouvelle taxonomie reliant les biographies aux professions. Suivant les meilleures pratiques de création de taxonomie à partir des sciences de l'information, dérivent un vocabulaire contrôlé à partir des données brutes et concevoir une hiérarchie de classification permettant trois niveaux d'agrégation.[2]

2.5.3 Index de Wikipédia [3]

2.5.3.1 Définition

Le calcul de l'indice est basé surla distribution de citations du travail de chercheur. Selon Hirsch s'il a l'index h, si h de ses articles (Np) cités au moins h fois chacun, tandis que les deux articles restant (Np - h) cité pas plus de h fois chacun. Cet indice a gagné le soutien et est utilisé dans des systèmes.[3]

2.5.3.2 La règle de l'index de Wikipédia

Les auteurs ont proposé les règles suivantes pour calculer le Wiki-index de popularité de l'auteur. On suppose que les références sur l'auteur se trouvent dans N articles de Wikipédia. Classés par nombre décroissant de paramètres qui déterminent le nombre de fois que le nom de l'auteur apparaît dans les références bibliographiques de ces articles.[3] Le nom de l'auteur est présent dans les références bibliographiques de ces articles, nous noterons comme :

R_1 , R_2 , ……, R_N

Wiki-index de popularité de l'auteur (*WI*) correspond au maximum nombre d'articles (*WH*) de Wikipédia, dans lequel le nombre de références ne dépasse pas le *WH* qui est multipliée par une certaine fonction intégrale, qui n'est pas décroissante (par exemple, la racine carrée est considérée ci-dessous) le N, c'est-à-dire :

$WI = WH * \sqrt{N} = \max (i : R_i > i) * N$

Le Wiki-index de popularité des auteurs est idéologiquement proche de l'indice de Hirsch ; Cependant, il ne prend pas en compte le nombre d'articles qui font référence à l'article de l'auteur et les citations à l'article de l'auteur. L'article de l'auteur et les citations de son travail et le nombre d'articles de Wikipedia qui contiennent ces liens de données. Une autre différence par rapport à l'indice de Hirsch est la multiplication par une fonction de N, reflétant la considération qu'ild'une plus grande popularité et la plus grande dispersion des valeurs de l'indice pour différentsauteurs.

Il faut noter que le niveau de popularité de l'auteur doit être rattaché à son domaine thématique d'une part afin d'éviter un comptage erroné pour les homonymes, et d'autre part pour assurer l'exhaustivité sur le domaine thématique.[3]

2.5.3.3 Exemple

Supposons que l'article de Wikipédia comportant le plus grand nombre de références à l'auteur George Smith (dans un domaine donné) contient 100 références. Le site deuxième - 20 documents, un troisième - 10, un quatrième - 5, un cinquième - 5, 4 autres - un seul lien. [3]

Ainsinous avons un certain nombre de valeurs :

R_1= 100, R_2 = 20, R_3 = 10, R_4 = 5, R_5 = 5, R_6 = 1, R_7 = 1, R_8 = 1, R_9 = 1

1article contient le nombre de références le moins élevéR_1= 100

2articles contiennent le nombre de références le plus faibleR_2 = 20

3articles contiennent le nombre de références le plus faibleR_3 = 10

4articles contiennent le nombre de références le plus faibleR_4 = 5

5articles contiennent le nombre de références le plus faibleR_5 = 5

Il n'y a pas 6 articles qui contiennent le nombre de références d'au moins 6.

Dans ce cas :

N=9, *WH* = 5

Donc :

WI = 5*$\sqrt{9}$=15

2.6 Conclusion

Nous avons présenté dans ce chapitre les deux ressources principales pour la réalisation de notretravail : l'encyclopédie Wikipédia, et Wiki média. Pour la Wikipédia, nous avons exposé certains composants de la structure d'une page donnée, en particulier les liens internes, les liens externes, l'information sur la page et les liens inter langues, etc. Nous avons expliqué également les différentes méthodes d'accès au contenu d'un article de la Wikipédia en se basant plus sur l'action API média wiki que nous allons utiliser afin d'extraire les informations concernées.et l'estimation de la notoriété d'un nom propre via Wikipédia avec la méthode : l'entropie de Shannon.

Chapitre 3 conception

3.1 Introduction

Dans le présent chapitre, nous allons présenter les différentes étapes de l'estimation de la notoriété d'un nom propre via Wikipédiaen utilisant l'entropie de shannon. Notre tâche étant d'entamé la classification de chaque entité nommée sur wikipedia type par type. Les critères de classement que nous avons choisi sont l'année et la langue .

3.2 Les Etapes de conception de notre système de classification des entités

Nous présentons notre système de classification des entités nommées par les étapes suivantes :

3.2.1 Choisir la catégorie

Les catégories sont un système de classement thématique des articles de Wikipédia, présent en bas de chaque page. Chaque article est classé dans unecatégorie. C'est-à-dire chaqueentité nommée a un type. Alors on choisit la catégorie de l'entité nommée (figure 10)après on calcul sa notoriété et on l'enregistre dans la table Type Prolexbase[39].

1

[39] Prolexbase est une base de données lexicale qui contient toutes les informations syntaxiques, morphologiques et sémantiques concernant

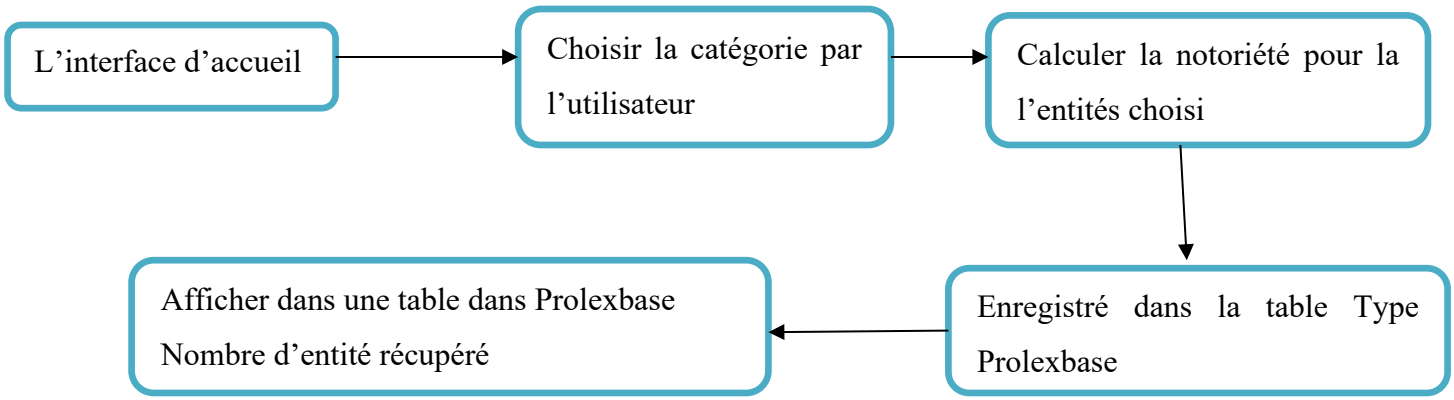


Figure 10choix de la catégorie pour chaque entité

3.2.2 Choisir l'URL

Une URL est un lien hypertexte ou hyperlien, sur lequel un internaute peut juste cliquer sur un lien (un mot, ou un groupe de mots) pour être redirigé vers une autre ressource sur le Web. Notre point de départ dans ce travail est l'url car une entité nommée est définie par son url et lecalcul de la notoriété est basée sur les liens vers la Wikipédia. La figure 11 présente l'algorithmepour la récupération des entités nommes à partir de leurs url.

les noms propres.

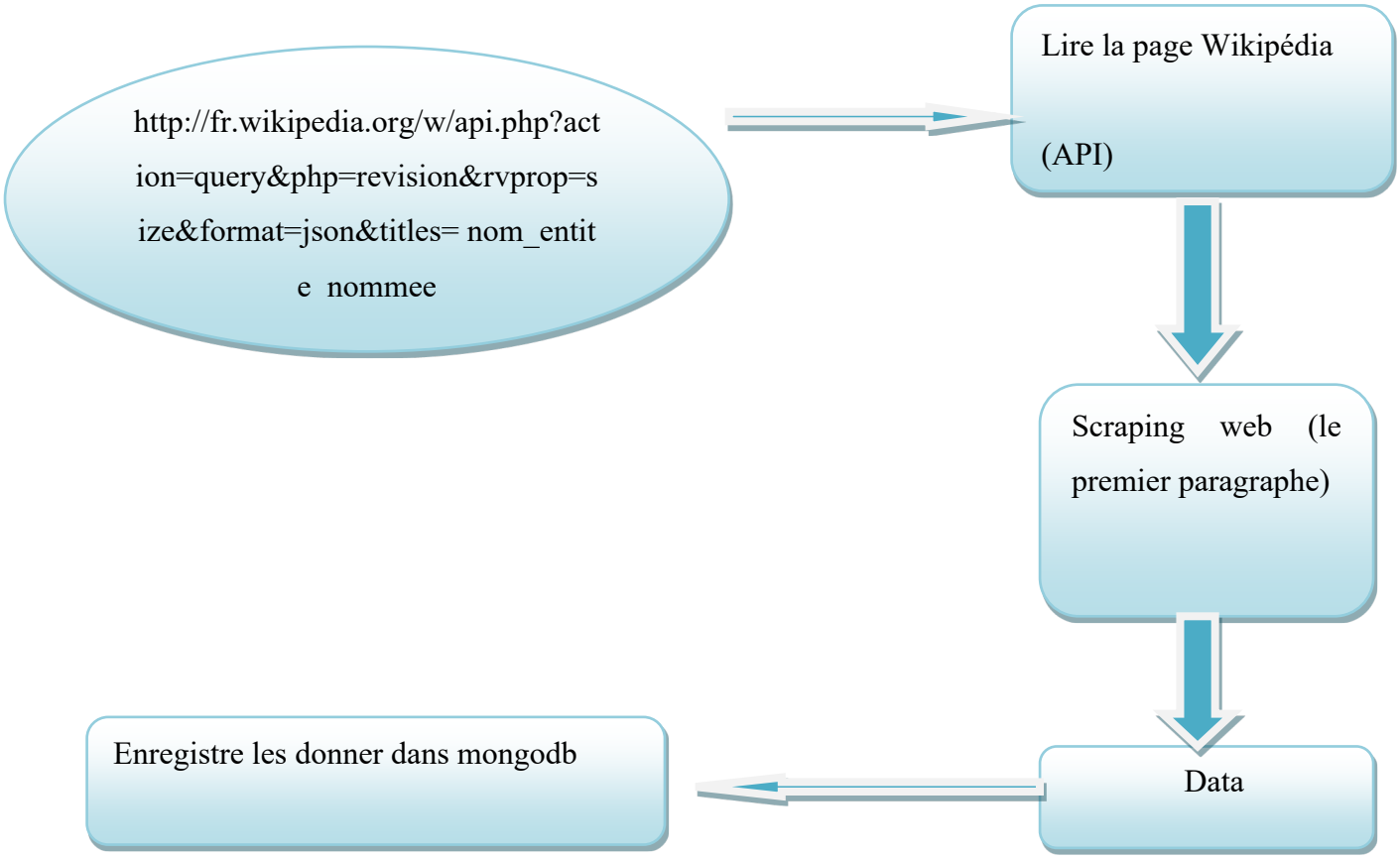


Figure 11 Choix de l'URL pour une entité

3.2.3 Calculer la notoriété

Nous avons calculéla notoriété des entitésnommées à partir ducorpus Wikipédia via la technique de l'entropie de Shannon précédemment évoquée dans le chapitre 2. Les étapes de calcul sont présentées ci-dessous. La figure 12 montre L'algorithme de calculde la notoriété (N= Notoriété).

1. **Calcul du nombre de consultations [1]**

Pour le calcul du premier indice, via des services webWikimédia[34] du nombre de consultations, Nous avons choisi de prendre comme point de départ le premier mois de l'année et comme fin le dernier mois complet. Dans notre travaille nous considérerons la période 2016-2022.Tout d'abord, l'agrégation des statistiques de consultations pour un article cible a été effectuée mois par mois à partir de 2016 et jusqu'à 2022 via l'outil «Statistiques de consultation »[1] de la Wikipédia. Cette partie du travail a été réalisée langue par langue et type par type. A la fin de cette étape, nous avons créé une table SQL pour chaque type dans une langue donnée.

- **Les étapes :**
  - Ona un lien Wikipédia qui est associé à un type
  - Envoyer le lien vers l'outil de Wikipédia Statistiques de consultation
  - Ouvrir la page de lien de l'outil et récupérer le nombre après (has been viewed)
  - Stocker le nombre dans un tableau
  - Après tous ça on a deux boucles externe « tableau d'années » et interne « tableau de mois »
  - Stocker le nombre dans fichiers SQL pour chaque type de chaque langue
  - Envoyer le a la base de données Prolexbase.

2

[1] Mouna Elshter

3

2. Calcul des quatre indices [1]

La Wikipédia fournit un service web permettant d'obtenir la liste de tous ses contributeurs, et donc, leur nombre. Le même service permet d'obtenir la taille de l'article, le nombre de liens entrants et le nombre de liens externes. En effet, le principe général effectué pour récupérer ces informations d'une page de la Wikipédia est l'action API de Media Wiki[13](chapitre 2), l'un des outils d'exploitation du contenu de la Wikipédia qui se base sur une requête soumise via une URL ; elle permet d'interroger la Wikipédia, et de fournir une réponse dépendant des paramètres utilisés et de leurs valeurs associées

- **Les étapes :**
  - Ona un lien Wikipédia qui est associé un type
  - Pour chaque lien construire l'url de l'action API Media Wiki selon « l'indiceà extraire »
  - Ouvrir la page de la réponse de ce web service
  - [4]Extraire la valeur de l'indice cible
  - Stocker la valeur dans un tableau typique à l'indice
  - Et on a un tableau pour les quatre indices, tableau des pivots, tableau de la langue et tableau de type.

[34] http://www.cnrtl.fr/lexiques/prolex/
[13] https://fr.wikipedia.org/wiki/Aide:MediaWiki

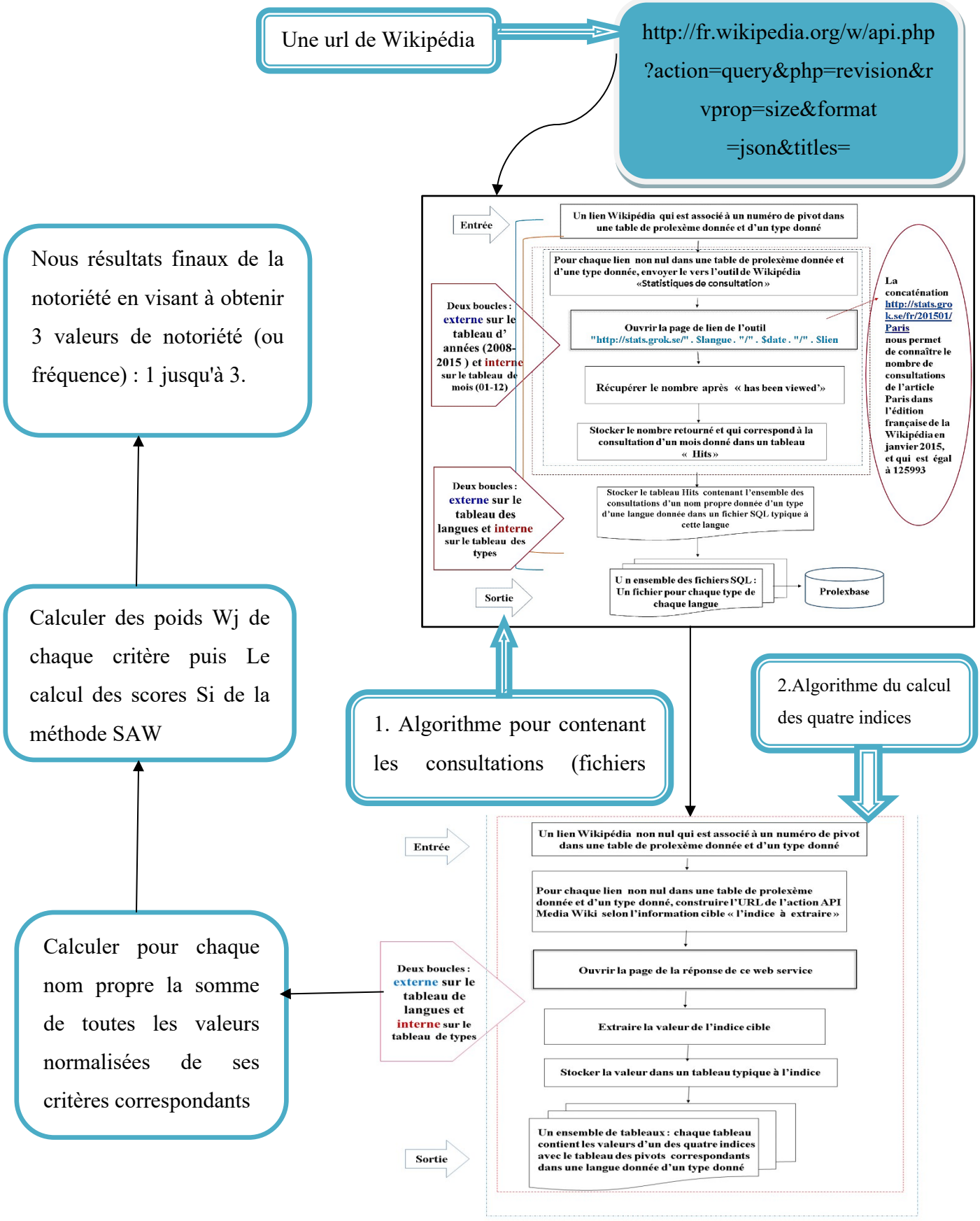


Figure 12 : L'algorithme de calcule de la Notoriété

3.3 Classification des entités nommées

Nous pouvons classer les entités nommées de trois manières :

3.3.3 Classification par langue

On calcule pour chaque entité sa notoriété dans une langue donné, car la notoriété d'une entité nommée dépend de la langue. Nous effectuons ce calcule depuis un lien dans une langue donnée, et via le concept inter langue relatif à la Wikipédia. Ce calcule de notoriété d'une langue a une autre permet de mesurer la notoriété que représente l'entité d'une région ou d'un pays a un autre ce qui permettra de classifier l'entité selon sa disposition géographique.

Dans notre travail on a choisi de travaillé sur 4 langues qui sont : l'anglais, l'arabe, le français et le polonais.

Prenons l'exemple de l'entité **EL Emir Abdelkader**. Cette entité est plus populaire en Algérie et en moyen orient par rapport aux autres régions, donc sa notoriété devrait être plus élevée en langue arabe.

La figure 13 résume comment on a classifié l'entité nommée par langue (évolution géographique).

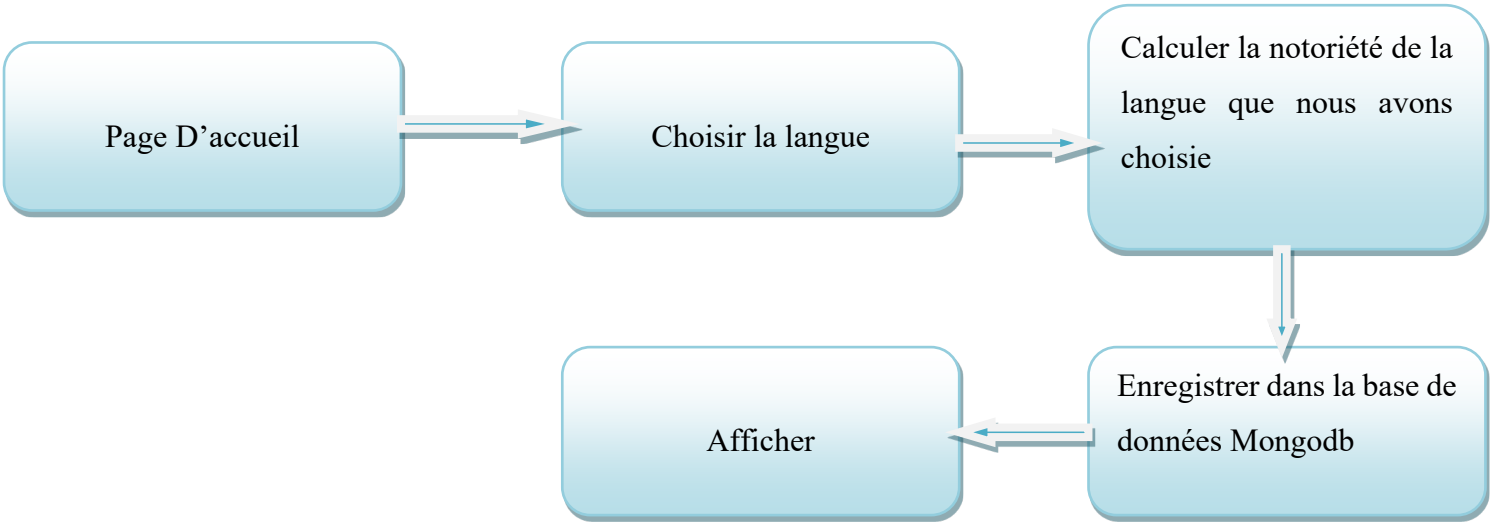


Figure13 : L'algorithme qui classifiel'entité nommée par langue (évolution géographique)

3.3.4 Classification par année

Certaines entités nommées sont classés par année, Si on prend l'exemple de l'entité Didier Raoult le microbiologiste, sa notoriété devient plus élevée pendant la pandémie du COVID 19.La figure 14résume comment on a procédé à la classification de l'entité nommée par année,et la figure 15 montre l'évolution de la notoriété sur plusieurs années (2016 jusqu'à 2022)

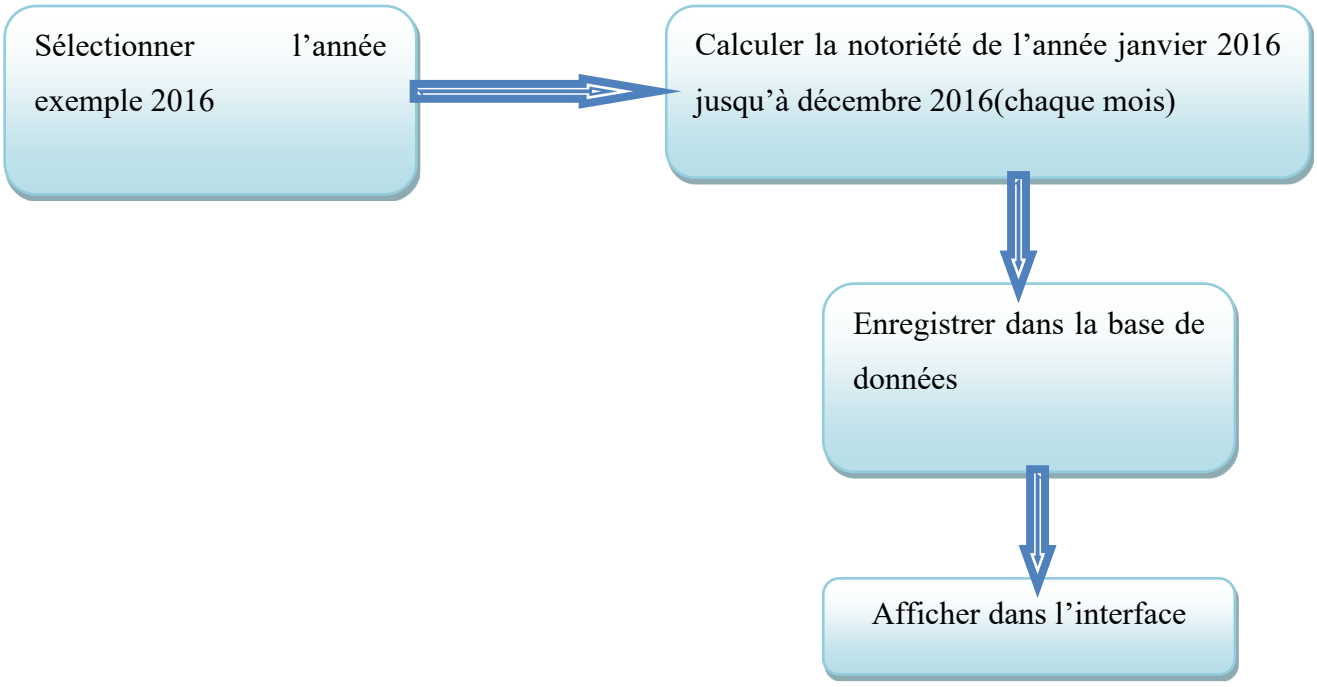


Figure 14 : L'algorithme de classification des entités nommées par année

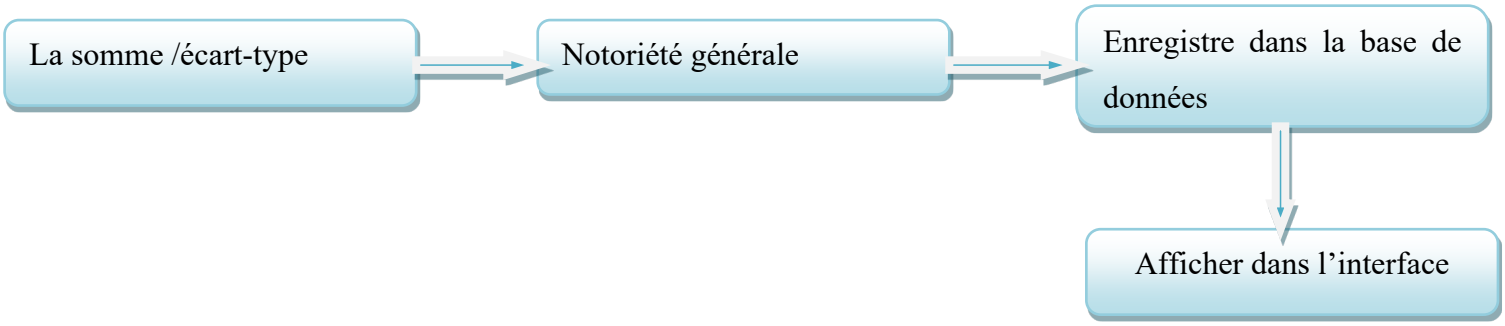


Figure 15 : L'algorithmequi classifiel'entité nommée générale en calculant l'écart type

3.4 Les diagrammes UML

3.4.1 Diagramme de cas d'utilisation

Les diagrammes de cas d'utilisation capturent le comportement d'un système, d'un sous-système, d'une classe ou d'un composant tel qu'il est vu par les utilisateurs externes. La figure ci-dessous représente le diagramme de cas d'utilisation de notre système.

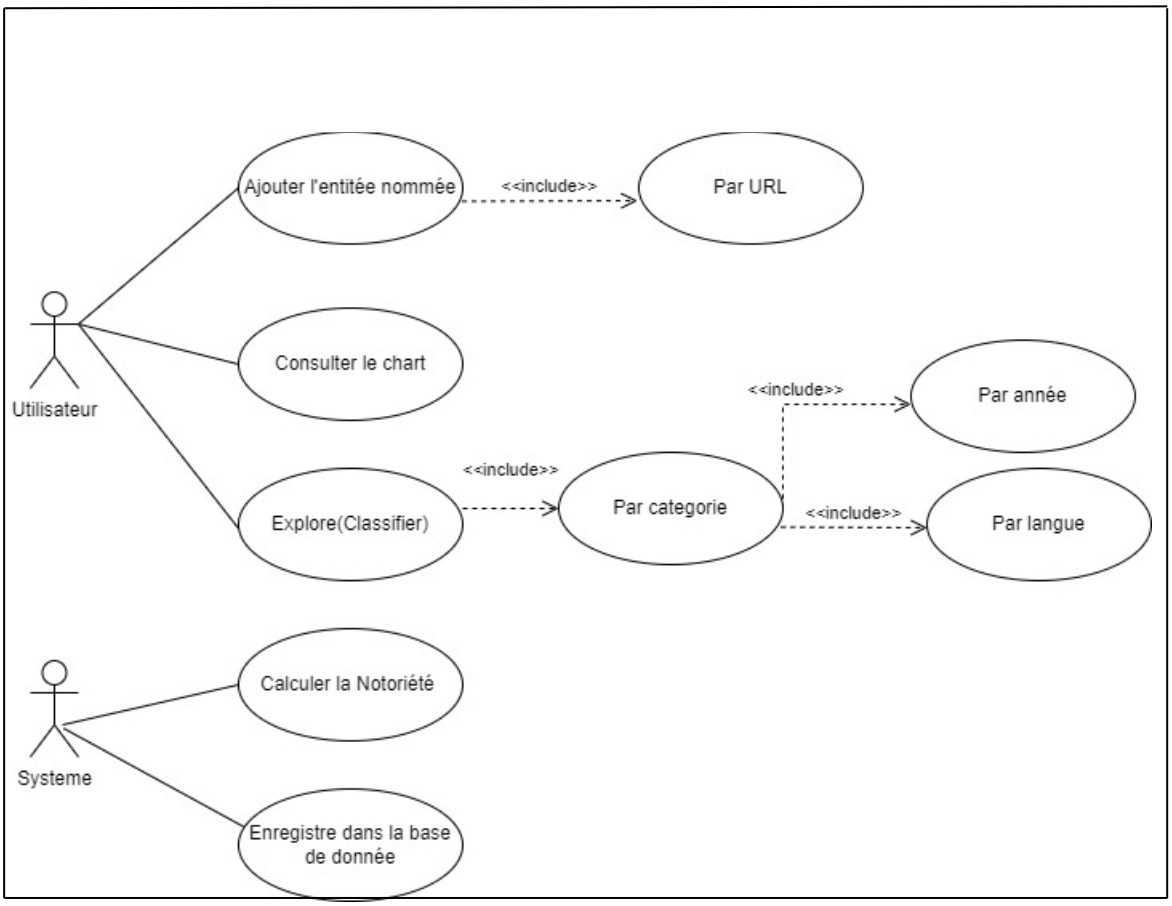


Figure 16 : diagramme de cas d'utilisation

Description textuelle de cas d'utilisation :

Scénario 1	L'accueil du site
Objectif	**C**onnecter dans le site
Acteurs principaux	Utilisateur
Pré conditions	-Avoir le site -la base de données est disponible
Post conditions	Accès dans le site
Scénario nominal	1. L'utilisateur accède au site 2. Le système affiche l'interface 3. L'utilisateur ajouter l'entité nommée, l'années, la langue et catégorie. 4. Le système vérifie les champs introduits par l'utilisateur 5. Le système vérifie l'existence de l'url 6. Le système enregistre les informations de l'utilisateur dans la base des données Prolexbase. - Si les informations introduites sont correctes, le système calcule la notoriété. 7. Le système affiche la classification des entités nommes. 8. Afficher le graphe (chart).

Scénario alternatif	-A1 : l'utilisateur possède déjà une entité nommes avec sa notoriété. L'enchainement de A1 démarre au point 7 du Scénario nominal : 7. Le système affiché la classification des entités nommées. 8. Afficher le graphe (chart).

3.4.3 Diagramme de classe

En génie logiciel, un diagramme de classes dans le langage de modélisation unifié (UML) est un type de diagramme de structure statique qui décrit la structure d'un système en montrant les classes du système, leurs attributs, opérations (ou méthodes) et les relations entre les objets.

Il est utilisé pour la modélisation conceptuelle générale de la structure de l'application et pour la modélisation détaillée, traduisant les modèles en code de programmation.

La figure ci-dessous représente le diagramme de classe de notre système :

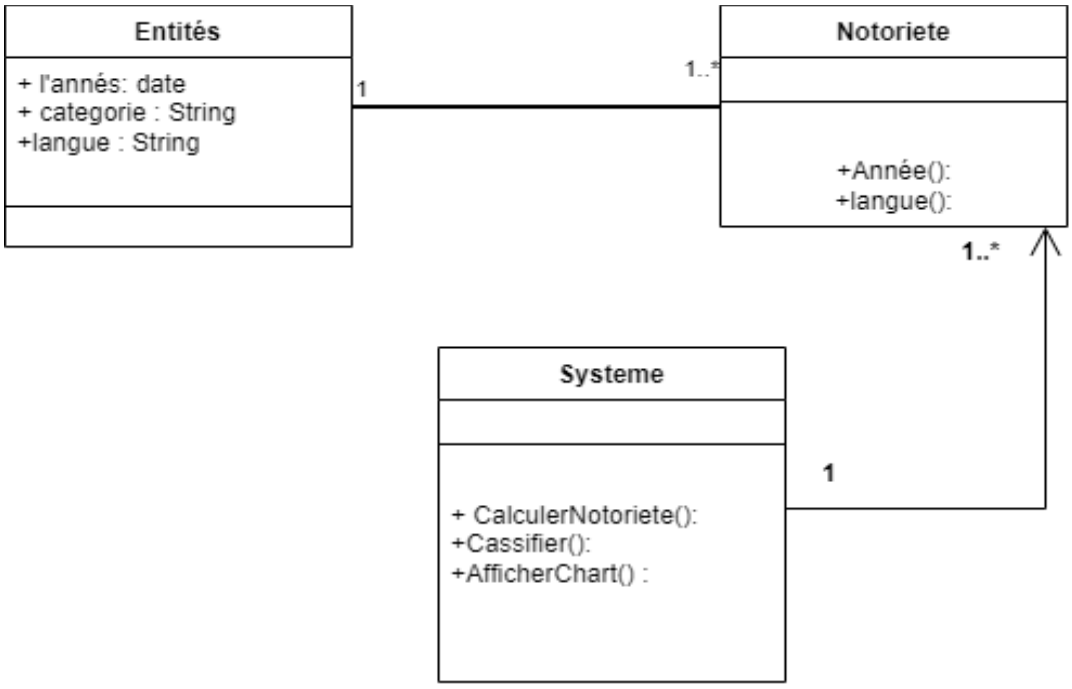


Figure 17 : diagramme de classe

3.4.4 Diagramme de séquence

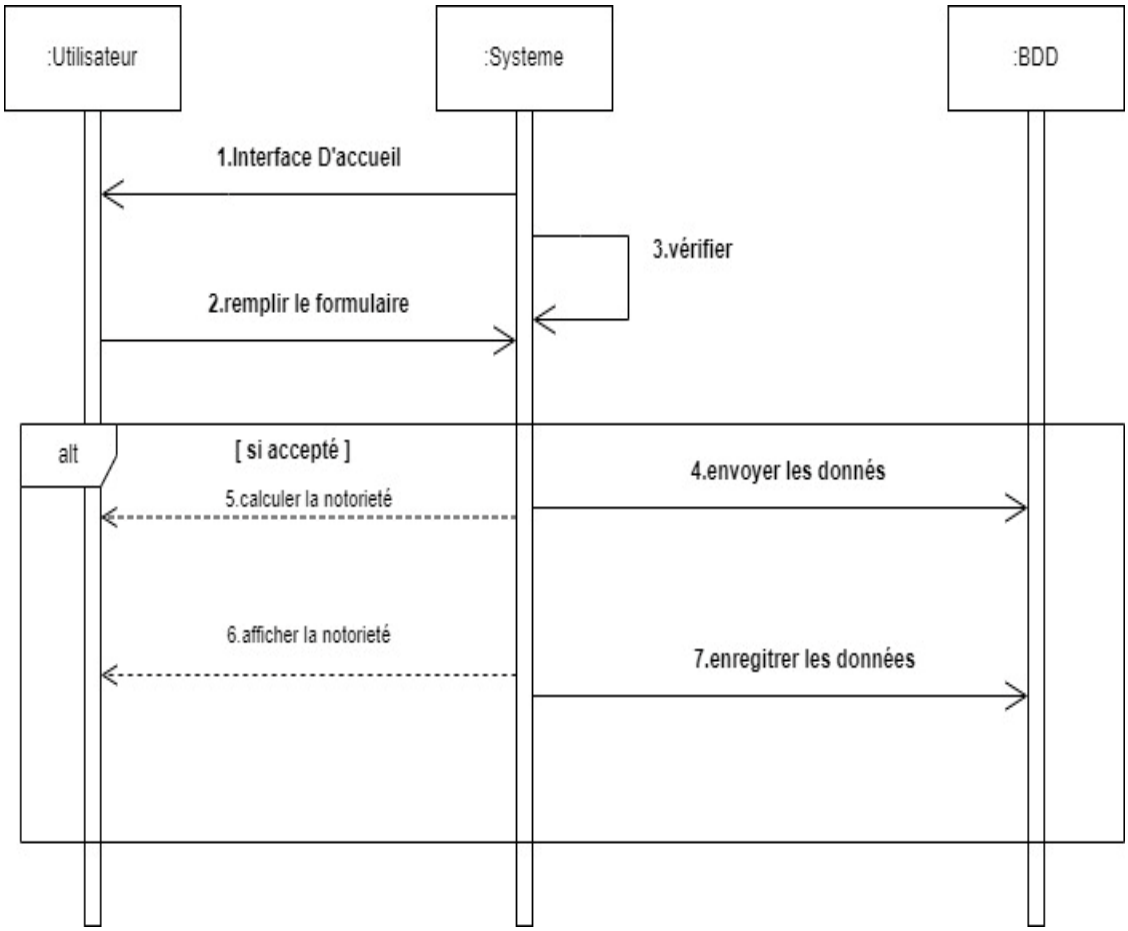


Figure 18 : diagramme de séquence

Chaque utilisateur doit remplir les champs tels que l'entité nommées, l'année, la langue et la catégorie, le système vérifiera si l'url existe. Si c'est le cas alors les informations seront automatiquement stockées dans la base de données Prolexbase, puis calculer la notoriété de l'entité nommées et afficher dans l'interface en même temps enregistré dans la base de données Prolexbase.

3.5 Conclusion

La phase de conception consiste à obtenir une représentation claire, explicite, cohérente et condensée du système et de son fonctionnement, elle facilite grandement le travail des développeurs ce qui peut nous aider à réaliser le projet et fournir les outils dont nous avons besoin dans le processus de création de notre système. Ce score de conception peut ensuite être utilisé par les algorithmes que nous avons déjà mentionnés pour avoir une meilleure recommandation pour l'utilisateur.

Chapitre 4 Implémentation

4.1 Introduction

Dans le présent chapitre, nous allons présenter le processus de classification des entités nommées tel qui l'a été présenté dans le chapitre précédent. Dans ce dernier chapitre, nous présentons l'ensemble d'outils et d'environnements de programmation utilisés pour l'implémentation de notre projet ainsi que les langages de programmation. Nous terminons par quelques interfaces qui présentent notre projet.

4.2 Environnement de développement

4.2.1 Visual studio code

Visual Studio Code est un éditeur de code open-source développé par Microsoft supportant un très grand nombre de langages grâce à des extensions. Il supporte l'auto complétion, la coloration syntaxique, le débogage, et les commandes git.[40]

4.2.2 Mongo DB

Mongo DB (de l'anglais humongous qui peut être traduit par « énorme ») est un système de gestion de base de données orienté documents, répartissable sur un nombre quelconque d'ordinateurs et ne nécessitant pas de schéma prédéfini des données. Il est écrit en C++. Le serveur et les outils sont distribués sous licence SSPL, les pilotes sous licence Apache et la documentation sous licence Créative Commons2. Il fait partie de la mouvance No SQL. [33]

4.2.3 Xampp

XAMPP est un ensemble de logiciels permettant de mettre en place un serveur Web local, un serveur FTP et un serveur de messagerie électronique. Il s'agit d'une distribution de logiciels libres (X (cross) Apache MariaDB Perl PHP) offrant une bonne souplesse d'utilisation, réputée pour son installation simple et rapide.[41]

4.2.4 Draw.io

Une application gratuite en ligne, accessible via son navigateur (protocole https) qui permet de dessiner des diagrammes ou des organigrammes. Cet outil vous propose de concevoir toutes sortes de diagrammes, de dessins vectoriels, de les enregistrer au format XML puis de les exporter.[42]

4.3 Langages de programmation

4.3.1 JavaScript

JavaScript est le principal langage de script des navigateurs Web et est essentiel aux applications Web modernes. Les programmeurs ont commencé à l'utiliser pour écrire des applications complexes, mais il y a encore peu de support d'outils pendant le développement.

4.3.2 Next.js

Next.js est un framework gratuit et open source s'appuyant sur la bibliothèque JavaScript React et sur la technologie Node.js , qui prend en charge les techniques de rendu des pages web côté serveur (SSR : Server Side Rendering), le rendu statique de pages web (SSG: Static Site Generation). Il prend également en charge la génération hybride de pages web et / ou incrémentale des pages (ISR: Incremental Static Generation.

4.3.3 Html

Le langage de balisage hypertexte, souvent abrégé en HTML, est un langage de balisage destiné à représenter des pages Web. C'est un langage d'écriture hypertexte, D'où le nom. HTML peut également être sémantiquement et logiquement structuré, et mettre en forme le contenu de la page, y comprendre les ressources multimédias, y comprendre les images, les formulaires de saisie et les programmes informatiques.

4.3.4 Css

Les feuilles de style en cascade peuvent être traduites par "feuille de style en cascade". CSS est Langage informatique utilisé sur le Web pour formater des documents HTML ou XML. Les fichiers CSS contiennent le code qui gère la conception des pages HTML.

4.3.5 REChart.js

Chart.js est une bibliothèque JavaScript open source gratuite pour la visualisation de données. Créée par le développeur Web Nick Downie en 2013, la bibliothèque est maintenant maintenue par la communauté et est la deuxième bibliothèque de graphiques JS la plus populaire sur Git Hub par le nombre d'étoiles après D3.js. Chart.js est rendu dans un canevas HTML5. Elle est disponible sous la licence MIT.Recharts est une bibliothèque de graphiques redéfinie construite avec React et D3.Le but principal de cette bibliothèque est de vous aider à écrire des graphiques dans les applications React sans aucune douleur. [32].

Dans la section suivante, nous présentons les résultats obtenus de notre classification via la wikipedia :

4.4 Résultats de Classification d'entité nommée

Le tableau 2 présente le résultat des nombres d'entités récupérés par type.

Type d'entité	Nombre d'entité récupéré
Marque	8
Groupe	2
Célébrité	32
Culture	10
Philosophie	4
Politicienne	10
Région	88
Scientifique	2
Série	6
Chanteuse / Chanteur	4
Sport	11
Technologie	4
Université	2
Unknow	7

Tableau 2 : Nombre d'entité récupéré

Le tableau 3 montre le nombre d'entités récupérés par type dans les trois langues : anglais arabe et polonais.

Type d'entité	Nombre d'entité récupéré anglais	Nombre d'entité récupéré arabe	Nombre d'entité récupéré polonais
Marque	3	3	2
Groupe	0	0	2
Célébrité	14	15	3
Culture	4	4	2
Philosophie	1	1	2
Politicienne	2	2	6
Région	27	30	31
Scientifique	1	1	0
Série	1	1	4
Chanteuse / Chanteur	0	3	1
Sport	3	7	1
Technologie	1	2	1
Université	1	1	0
Unknow	1	2	4

Tableau 3 : Nombre d'entité récupéré pour chaque langue (arabe, anglais, polonais)

- **Exemple d'entité avec sa notoriété dans plusieurs langues (évolution géographique)**

Exemple d'entité	Notoriété en générale	Notoriété en anglais	Notoriété en arabe
Paris	1	1	1
Google	2	1	2
هواري_بومدين	2	2	2

Tableau 4 : Exemple d'entité avec sa notoriété anglais arabe

- **Exemple d'entité avec sa notoriété sur plusieurs années (évolution de la notoriété de l'entité)**

Dans le tableau 5, nous remarquons que la notoriété de paris en 2016 est plus élevé que les autres années et ceux dans les différentes langues ; cela peux s'expliqué par les différents évènements qui ont lieu en 2016 a paris tél que la COP 21, l'euro 2016, ainsi que plusieurs attentats terroriste qui ont lieu cette année-là.

Année pour l'entité Paris	Notoriété générale	Notoriété en anglais	Notoriété en arabe
2016	1	1	1
2018	2	2	2
2020	2	2	2
2022	2	1	2

Tableau 5 : Exemple d'entité avec sa notoriété sur années

- **Afficher les statistiques sur un diagramme.**

LaFigure 19 représente de la notoriété d'entité Algérie en langues arabe la notoriété est élevée en 2020 jusqu'à 2022 vu les événements de cette période comme EL HIRAK, les élections présidentielles, les jeux méditerranéen. La figure 20 représente la notoriété de l'entité Algérie en anglais (2016 jusqu'à 2022)

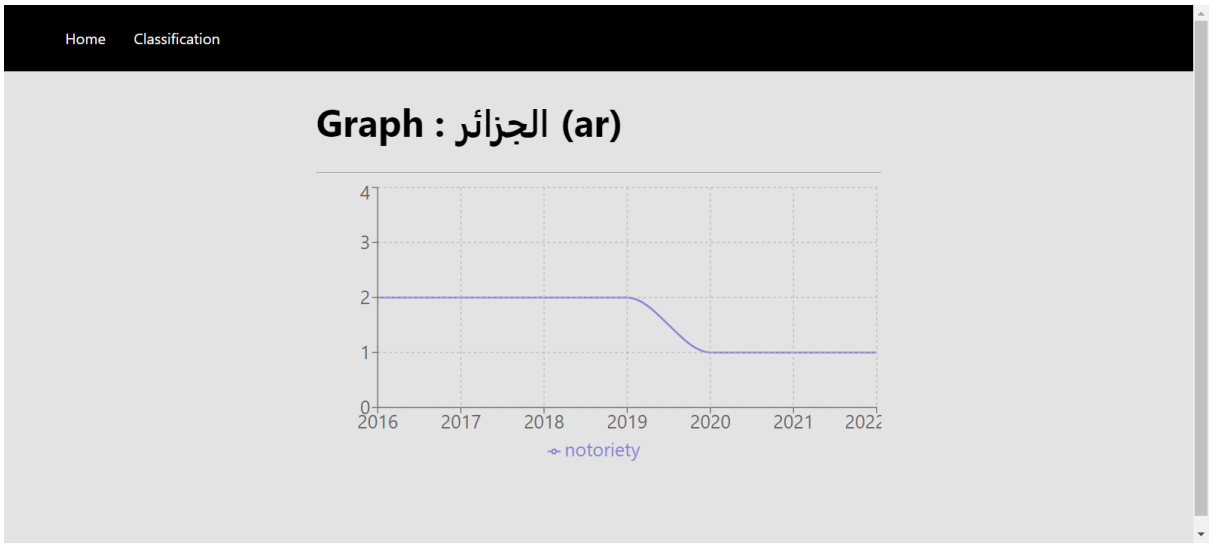


Figure 19 statistiques de la notoriété d'entité en langues arabe.

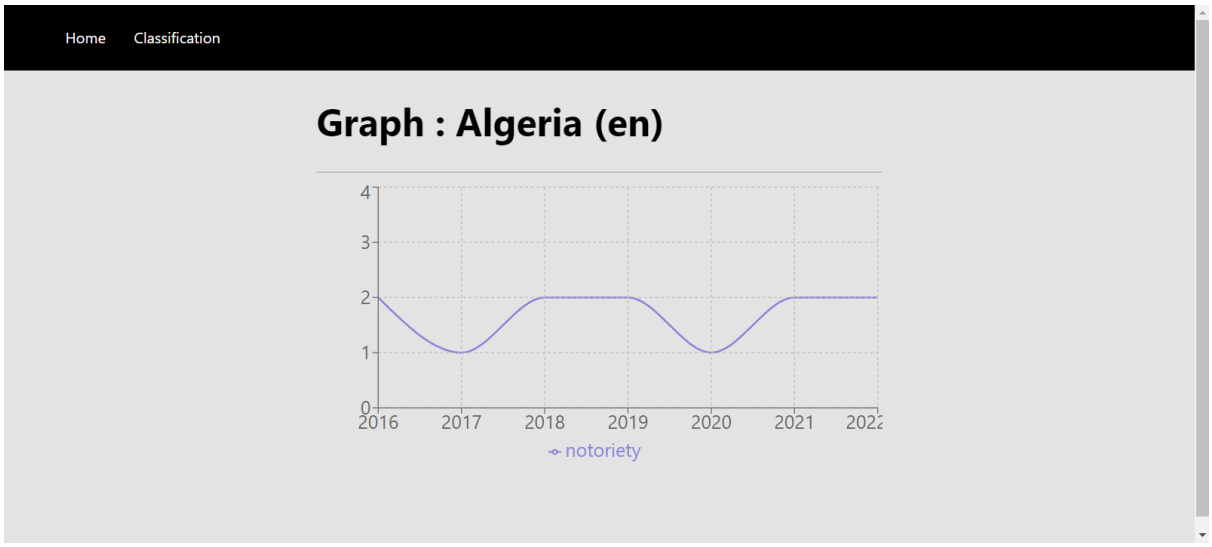


Figure 20statistiques de la notoriété d'entité en langues anglaise

4.5 Présentation du travail

Dans notre travail nous avons utilisée deux bases de données . La base de données prolexbase_3_1 (SQL) et la base de donnée chanonepro (no SQL).

Dans la base de données prolexbase_3_1 il existe quatre base de données 2_prolexbase_3_1_other_data avec la table prolexeme_arabe qui contient les noms propres arabe , la base de données 2_prolexbase_3_1_pol_data avec la table prolexeme_pol qui contient les noms propres polonais , la base de données 2_prolexbase_3_1_eng_data avec la table prolexeme_eng qui contient les noms propres anglais , la table 2_prolexbase_3_1_fra_data avec la table prolexeme_fra qui contient les noms propres francais .

- **La base de données 2_prolexbase_3_1_eng_data**

La figure 21représente une capture d'écran de la base de données 2_prolexbase_3_1_eng_data et la table prolexeme_eng (SQL). Dans cette table il existe les label prolexèmes (les noms propres) avec numéro de pivot, url de wikipedia (wikepedia link) et la notoriété (num-frequency) et num-prolexeme.

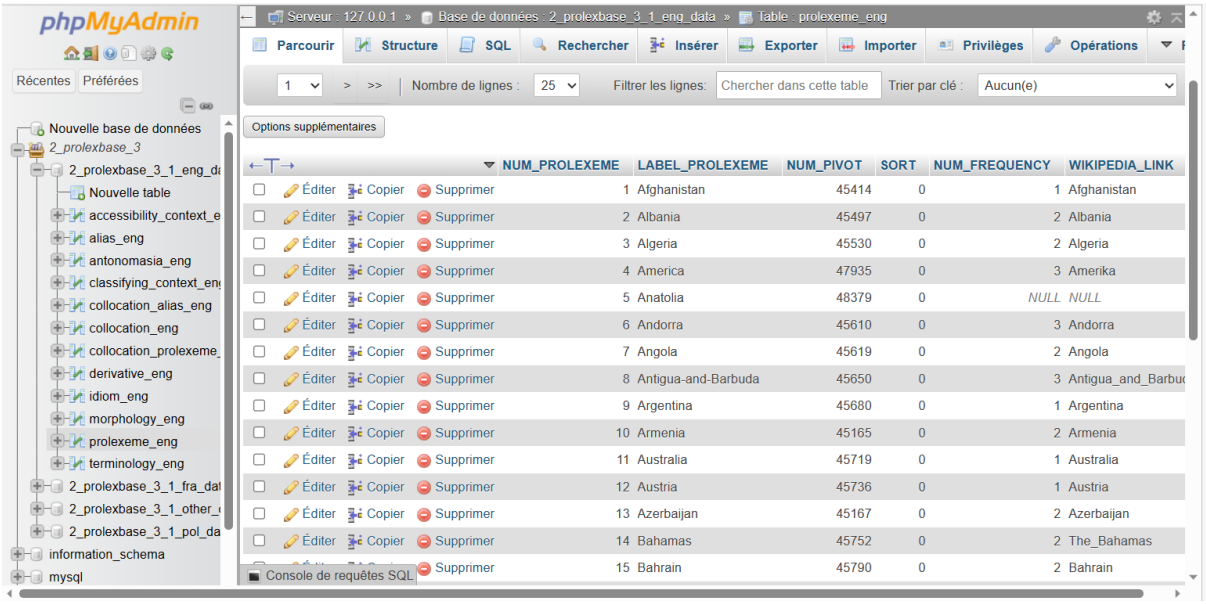


Figure 21 : la base de données 2_prolexbase_3_1_eng_data et la table prolexeme_eng

- **La base de donnée chanonepro**

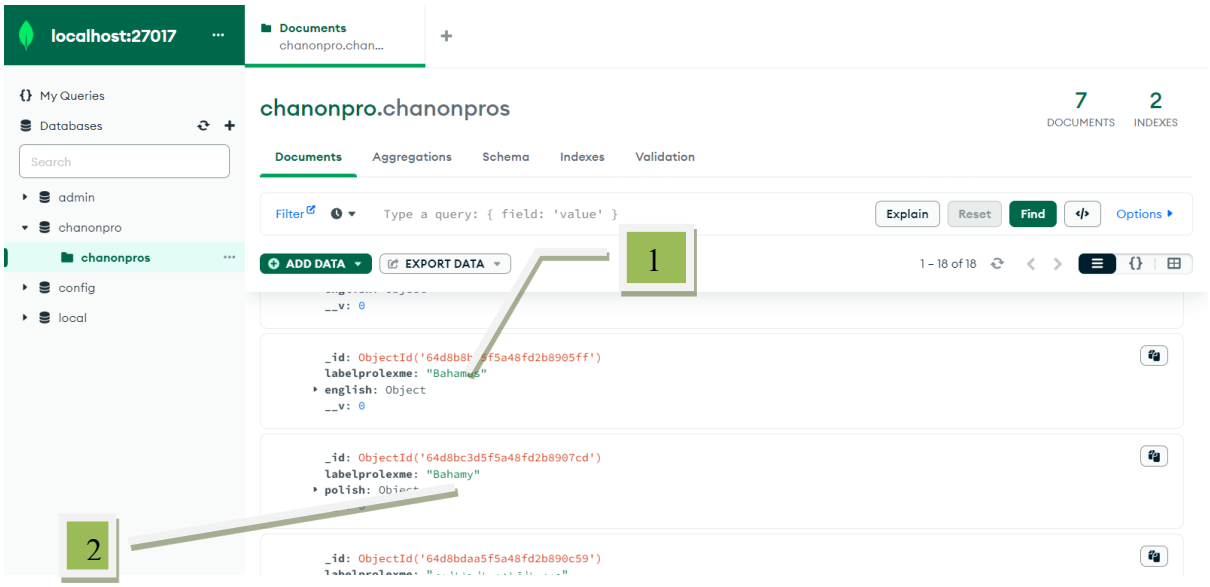
Nous avons créé une base de données chanonepros qui est une base de données no SQL pour enregistrer l'entité nommées , l'années , la notoriété et la langue dans la Table chanonepros . La figure 22représente

la base de données chanonepro d'une entité nommées par deux langues différentes anglais et polonais. Dans l'exemple de la figure 22 :

1. Représente l'entité nommes Bahamas en anglais et labelprolexme : Bahamas .

2. Représente l'entité nommes Bahamas en polonais et labelproleme : Bahamy.

3. Le nom propre Bahamas en anglais et l'années 2022 on a label prolexme : Bahamas , la notoriété : 2 et wikipédia Link (extlink) : The_Bahamas .



3

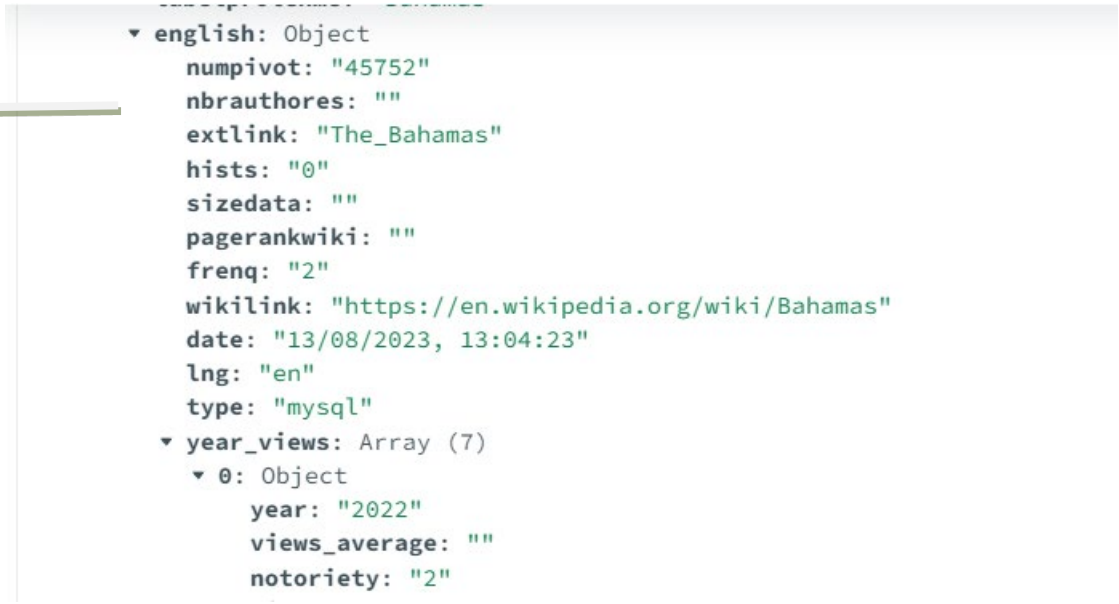


Figure 22 Base de données chanonepro

- **Interface Home**

  Sur cette interface, l'utilisateur doit cliquer sur LOAD ou saisie le nom propre dans la barre de recherche et choisi la langue et l'année. La partie Historique afficher l'historique de recherche. La figure 23 représente l'interface Classification des entités nommées .

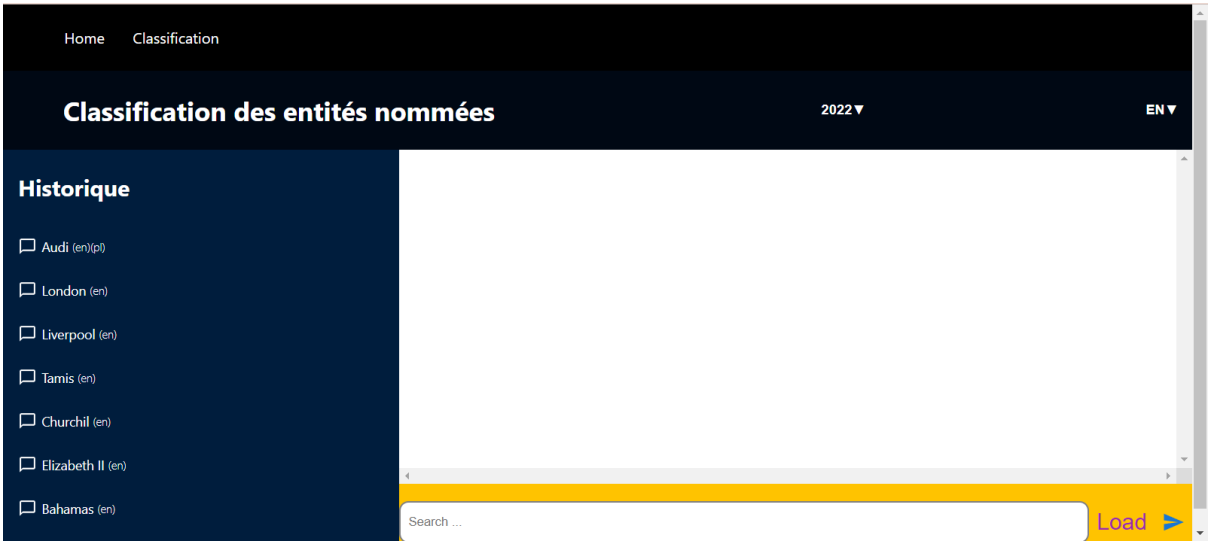
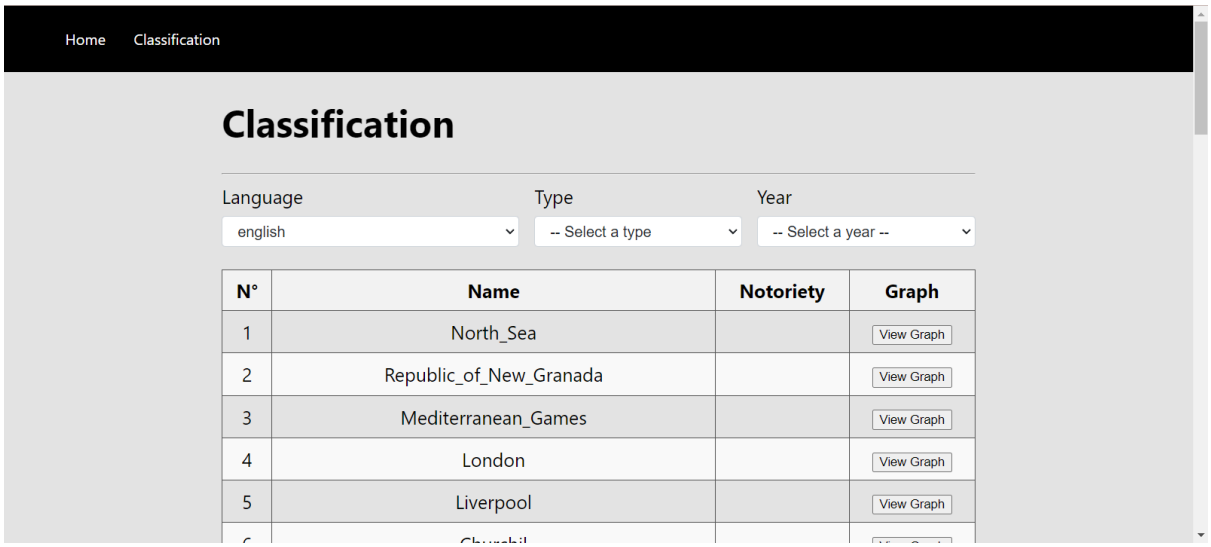


Figure 23 Interface Home

- **Interface Classification**

Cette interface représente les résultats notoriétés de chaque entité nommée d'âpres la langue, l'année et type.La figure 24 représente un exemple de classification avec la langue polonais et l'année 2016. La figure 24 représente l'interface Classification avec les résultats



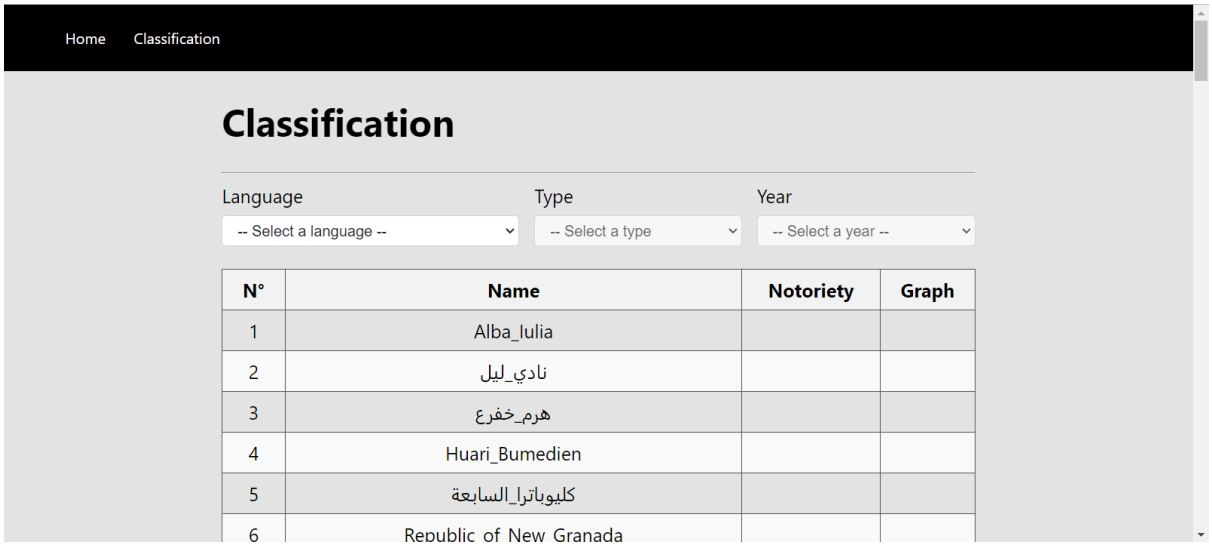


Figure 24 Interface Classification

- Dans le cas de classification par graph la figure 25 représente la notoriété de l'exemple Abdelhamid ben badis en arabe et les années 2016 jusqu'à 2021 .on remarque que la notoriété dans les années2018, 2019 et 2021 est élevée.

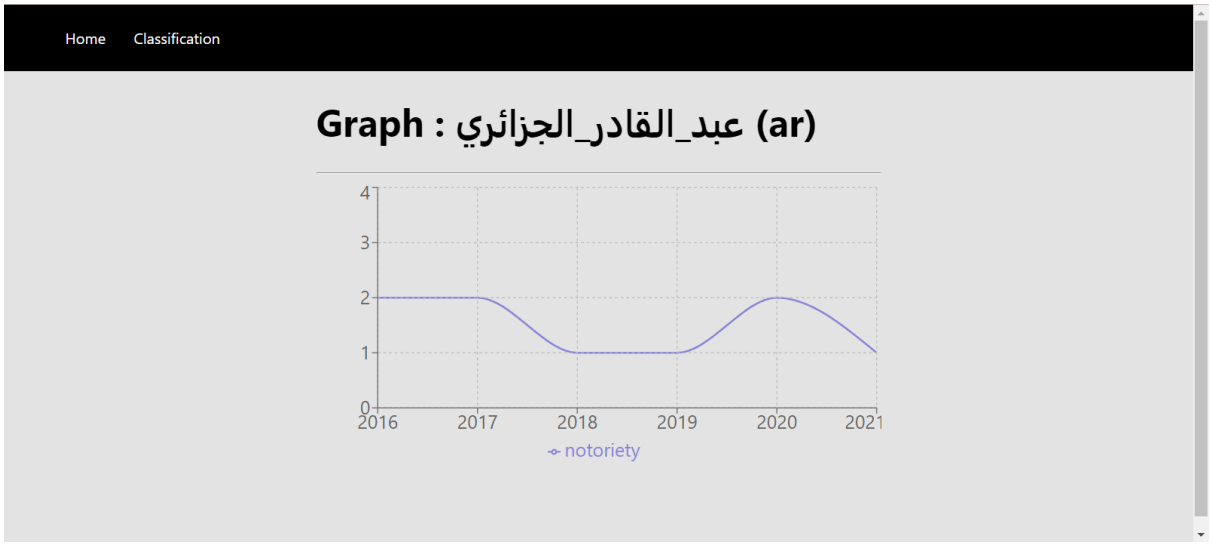


Figure 25classifications d'entité par graphe

- Dans le cas classification type par type on a fait entrer 197 entités nommes.

  La figure 26 représente les entités nommes en anglais et le type célébrité

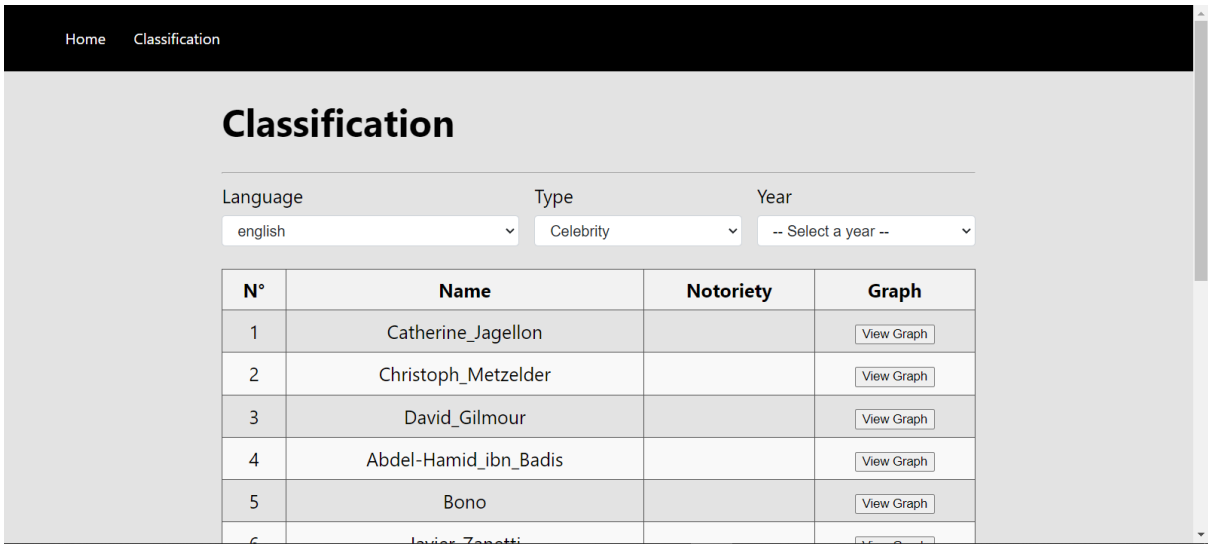


Figure 26 classifications d'entité en anglais et type célébrité .

- la figure 27 représente les entités nommes en anglais et le type région.

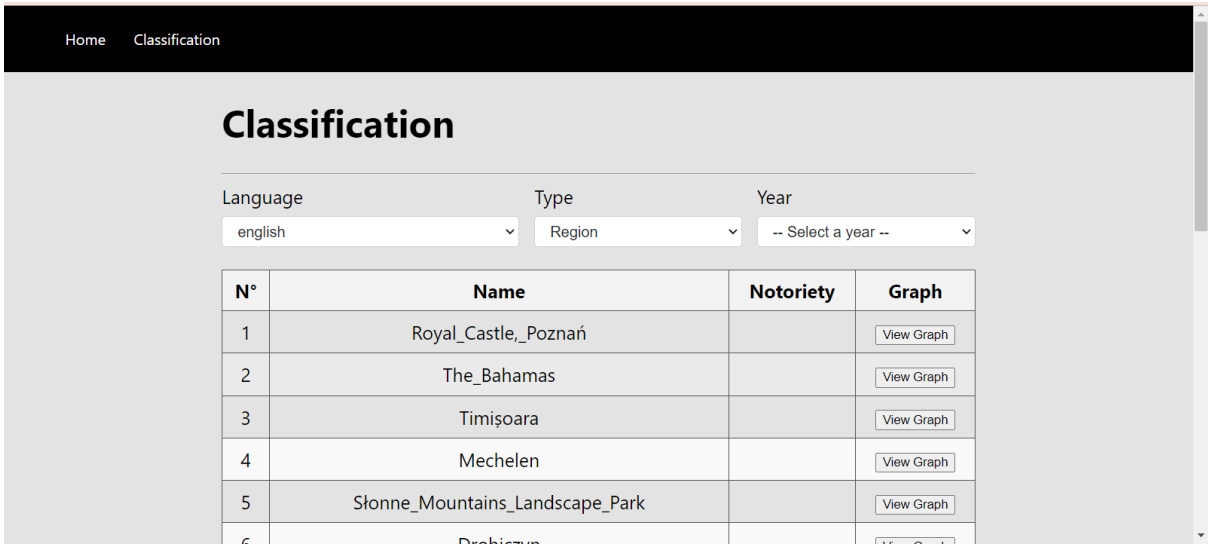


Figure 27 classifications d'entité en anglais et type région.

- **Test page view**

  On a utiliser l'outil Postman pour tester le calcule de page view de 5 critère . La figure 28 représente le calcul de l'url page view par mois.

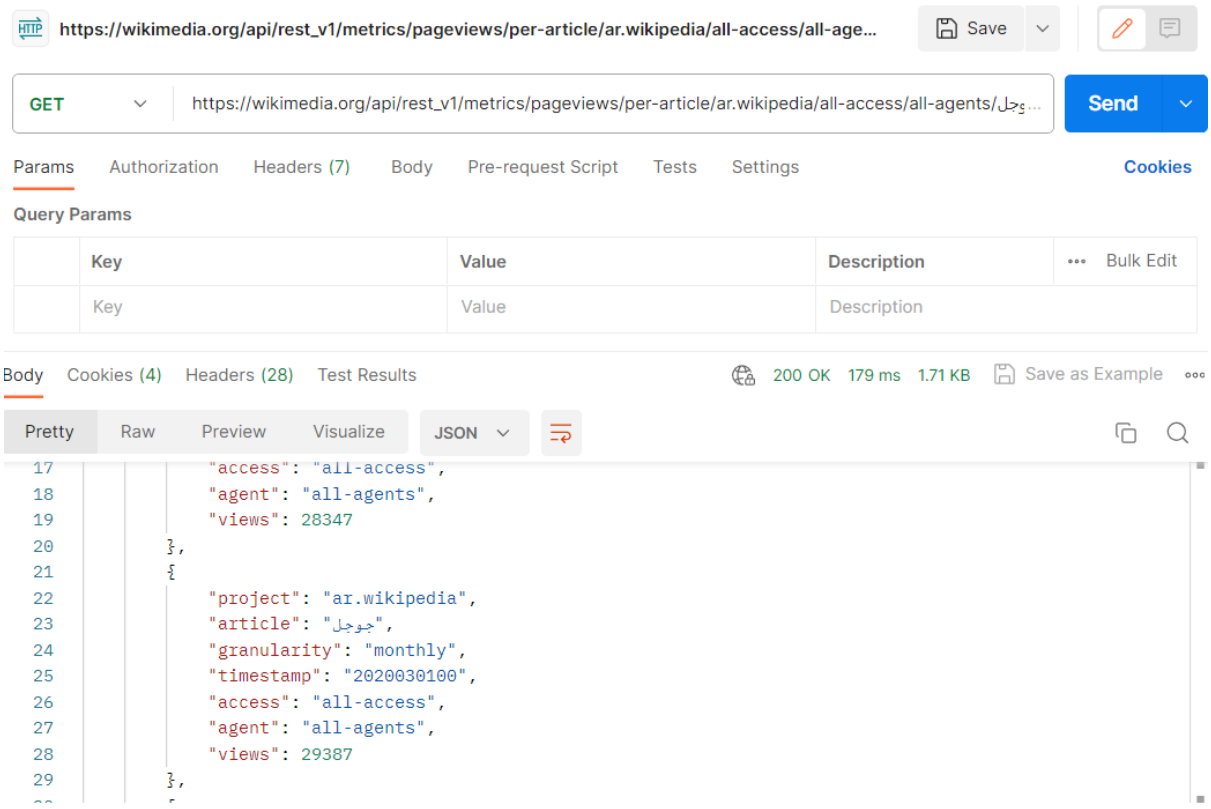


Figure 26 Tester l'url de page view

4.6 Conclusion

Dans ce chapitre, nous avons présenté les environnements de développement et les différents langages de programmation et logiciels utilisés dans l'implémentation de notre projet, et les interfaces de notre application.

Conclusion Générale

Calculer la notoriété d'une entités nommées est une tâche complexe qui requiert beaucoup d'efforts et de ressources et qui combine plusieurs techniques et méthodes. Nous nous sommes basé sur l'encyclopédie Wikipédia comme principale ressource.

Dans ce travail nous avons présenté quatre chapitres : Le premier chapitre, représente la reconnaissance d'entité nommée basé sur des règles linguistiques qui exploitent l'étiquetage syntaxique, des déclencheurs et des dictionnaires de noms propres ;Le deuxième chapitre, est consacré pour la wikipedia ; Le troisième chapitre, est consacré pour la conception et la modélisation du notre système ; dans le quatrième chapitre, nous avons présenté les outils de l'implémentation et les résultats de classification des entités nommées .

Finalement, toutes ces recherches nous ont servie à réaliser notre projet de fin d'études, la classification des entités nommées par année et langue basé sur le calcul de la notoriété dans un corpus Wikipédia. La plus importante des taches est l'utilisation de la méthode l'entropie de Shannon dans les différentes langues.

Bien sûre ce travail n'est pas complètement achevé, car nous avons plusieurs perspectives, nous citons comme exemple :

1. Ajouter d'autres langues pour voir l'évolution pour s'élargir géographiquement.
2. L'ajoute des catégories automatiquement pour rendre le systèmede classification type par type plus performants.

Bibliographie

Armel FOTSOH TAWOFAING

[29] Thèse Recherche d'entit´es nomm´ees complexes sur le Web - propositions pour l'extraction et pour le calcul de similarit´e , Doctorat de l'Universit´e de Pau et des Pays de l'Adour le 27 f´evrier 2018

Damien NOUVEL

[28] Thèse Reconnaissance des entités nommées par exploration de régles d'annotation, l'université François – Rabelais de Tours le 20 novembre 2012

Fatma Ben Mesmia Chaabouni

[23] Thèse Reconnaissance des entités nommées à partir de Wikipédia arabe : Application à la découverte des relations sémantiques

Hela Fehri

[10] Reconnaissance automatique des entités nommées arabes et leur

traduction vers le français » Université de Franche-Comté ; Université de Sfax. Faculté des sciences, (2012). Français

Mouna Elashter

[1] Gestion et extension automatiques du dictionnaire relationnel

multilingues de noms propres Prolexbase», thèse doctorat 'université François Rabelais deTours (2017).

Mohamed HATMI

[22] Reconnaissance des entités nommées dans des documents multimodaux , THÈSE DE DOCTORAT UNIVERSITÉ DE NANTES Le 20 janvier 2014

Maud Ehrmann

[30] Thèse les entites nommes de la linguistique au tal UNIVERSITE PARIS 7 - DENIS DIDEROT , DOCTORAT le 2 Juin 2008

Nadeau et Sekine ; Abdelrahman et al; Belainine

[36] Nadeau et Sekine, 2007 ; Abdelrahman et al., 2010 ; Belainine, 2017

Oudah et Shaalan; Abuleil

[37] Oudah et Shaalan 2012 ; Abuleil, 2006

Pantheon 1.0

[2] a manually verified dataset of globally famous biographies .

Shaalan

[38] Shaalan , 2010

WIKI-INDEX OF AUTHORS' POPULARITY

[3] D.V. Lande, V.B. Andrushchenko, I.V.

Balagura Institute for Information Recording of NAS of Ukraine, Kiev .

Webographie

[4]https://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es Consulte le 06 novembre 2022

[5] https://fr.wikipedia.org/wiki/Didier_Raoult Consulte le 25 décembre 2022

[6] https://fr.wikipedia.org/wiki/Didier_Raoult#R%C3%A9f%C3%A9rences

[7] https://fr.wikipedia.org/wiki/Didier_Raoult#Liens_externes

[8] https://fr.wikipedia.org/wiki/Discussion:Didier_Raoult

[9] https://fr.wikipedia.org/wiki/Aide:Lien_interwiki Consulte le 29 décembre 2022

[11] https://igm.univ-mlv.fr/~dr/XPOSE2011/Wikipedia/presentation_wikipedia.html

[12]https://fr.wikipedia.org/wiki/Aide:Ins%C3%A9rer_une_r%C3%A9f%C3%A9rence

[13] https://fr.wikipedia.org/wiki/Aide:MediaWiki

[14] https://fr.wikipedia.org/wiki/DBpedia

[15] https://www.cetic.be/Exploiter-le-contenu-de-Wikipedia

[16] https://www.mediawiki.org/wiki/API:Query

[17] https://fr.wikipedia.org/wiki/JavaScriptJavaScript Object Notation, est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l'information structurée .

[18]https://fr.wikipedia.org/wiki/SPARQL**]**« SPARQL Protocol and RDF Query Language » est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet. »

[19] https://www.wikimedia.fr/le-mouvement-wikimedia/

[20] https://fr.wikipedia.org/wiki/Notori%C3%A9t%C3%A9

[21] https://fr.wikipedia.org/wiki/Entropie_de_Shannon 10 Décembre 2022

[24] https://www.mediawiki.org/wiki/API:Data_formats

[25] https://dumps.wikimedia.org/ Wikimedia dumps

[26]https://coop-ist.cirad.fr/evaluer/le-h-index-d-un-chercheur/1-qu-est-ce-que-le-h-indexHirsch-index

[27] https://fr.wikipedia.org/wiki/Indice_h

[31]https://docwiki.embarcadero.com/RADStudio/Rio/fr/D%C3%A9finition_des_diagrammes_de_classes_UML_1.5 Définition des diagrammes de classes UML 1.5 disponible à l'adresse

[32] https://fr.wikipedia.org/wiki/Chart.js Définition de Chart.js

[33] https://fr.wikipedia.org/wiki/MongoDB Définition de MongoDB

[34]http://www.cnrtl.fr/lexiques/prolex/

[35] https://www.wikimedia.org/

[40] https://fr.wikipedia.org/wiki/Microsoft_Visual_Studio consulte le 20 aout 2023

[41]https://fr.wikipedia.org/wiki/XAMPP consulte le 20 aout 2023

[42] https://www.tice-education.fr/tous-les-articles-er-ressources/articles-internet/819-draw-io-un-outil-pour-dessiner-des-diagrammes-en-ligne consulte le 20 aout 2023