

SEMANTIQUE et LANGUE NATURELLE

PROJET : TRAVAUX D'ETUDE ET DE RECHERCHE

Réalisé par :

Waël KOUJUK
Lucien KERISIT
Patty RANDRIAMBOLOLONA

Encadré par :

Line JAKUBIEC-JAMET

Année 2015

SOMMAIRE

REMERCIEMENTS	2
1 INTRODUCTION.....	3
2 SUJET DETAILLE	5
2.1 OBJECTIFS	5
2.2 DOMAINE DE CLASSIFICATION	5
3 CONCEPTION	7
3.1 L'EQUIPE DE DEVELOPPEMENT.....	7
3.2 MISE EN ŒUVRE ET REPARTITION DES TACHES	7
3.2.1 Rédaction de rapport :	7
3.2.2 Centralisation du projet avec gestionnaire de version :	8
3.2.3 Contraintes techniques et solutions	8
3.2.4 Justification des choix technologiques.....	9
3.3 ETAT DE L'ART	11
4 DEVELOPPEMENT	16
4.1 SCHEMA DU DOMAINE ANIMALE	16
4.1.1 La Classification du vivant	16
4.1.2 Le règne animal.....	16
4.1.3 Adaptation du domaine à l'application.....	18
4.1.4 Le schéma de l'arborescence.....	18
4.2 DONNEES ET SCHEMA DE LA BASE DE DONNEES.....	19
4.2.1 Les données	19
4.2.2 Modèle conceptuel de données.....	20
4.3 DEVELOPPEMENT DE L'APPLICATION WEB.....	22
4.3.1 Interface graphique ; lancement de l'analyse	22
4.3.2 Interface graphique : résultat de l'analyse.....	23
4.3.3 Amélioration du projet	24
5 CONCLUSION	25

Remerciements

Nous remercions tout particulièrement notre encadrant Line JAKUBIEC-JAMET de nous avoir proposé ce sujet. Toute notre gratitude pour ses conseils tout au long de la réalisation du Projet, ainsi que pour son écoute et sa disponibilité.

1 Introduction

Le Traitement Automatique des Langues (TAL) est un domaine pluridisciplinaire où travaillent en collaboration étroite des linguistes et des informaticiens dans le but de développer des applications capables de traiter des données linguistiques.

Ce projet s'inscrit dans le cadre du TER de première année de Master Informatique à l'Université Aix-Marseille. Le sujet proposé par Madame Line Jakubiec-Jamet s'insère dans le domaine de la formalisation et du traitement des langues naturelles.

Il nous est demandé de développer une interface Web qui permettra d'attribuer à chaque mot d'un texte une étiquette sémantique, en fonction d'une description déjà définie au préalable, dans le but d'évaluer ensuite la sémantique de phrases entières voire du texte complet.

L'application, évaluant la sémantique de mots rencontrés dans un texte, vise un public qui par exemple veut apprendre le français.

Dans ce rapport nous allons vous présenter notre projet et détailler sa conception de manière explicite de la phase d'analyse au produit final.

Pour cela, nous ferons dans un premier temps un examen attentif du sujet et nous évoquerons les méthodes et outils pouvant être utilisés pour mener à terme le projet.

Dans une deuxième partie, il s'agira de détailler le travail réalisé.

Enfin, une troisième partie fera office de conclusion dans laquelle on s'efforcera de faire le point sur la réalisation du projet par rapport aux objectifs fixés par le cahier des charges.

ANALYSE DU PROJET

2 Sujet détaillé

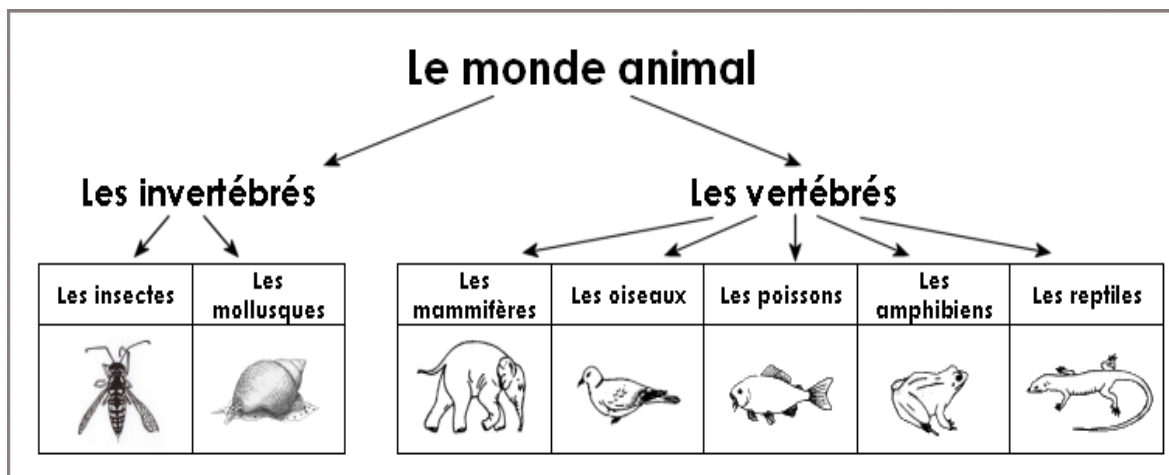
2.1 Objectifs

L'application doit récupérer en entrée un texte pour ensuite l'analyser et afficher la classification de chaque mot en fonction d'une description déjà définie dans un domaine choisi. L'analyse des mots permettra ensuite de vérifier la sémantique des phrases.

L'interface Web de l'application doit être conviviale, ergonomique et facile d'utilisation.

2.2 Domaine de classification

Le choix d'un domaine de classification des mots nous a demandé un certain temps de réflexion de la part de notre équipe afin de choisir un domaine pouvant être utiles et intéressant pour les futurs utilisateurs.



Le monde animal est un grand domaine intéressant qui, jusqu'aujourd'hui reste un sujet de recherche importante et vaste, c'est la raison pour laquelle on a choisi de faire d'explorer ce monde. Dans le TAL, plusieurs projets de recherche de plusieurs universités ont été réalisés dans d'autres domaines, mais pas dans le domaine du règne animal et encore moins en français.

Notre application va se concentrer sur la reconnaissance de l'entité animale et de résoudre la sémantique dans une phrase en entrée.

Méthode :

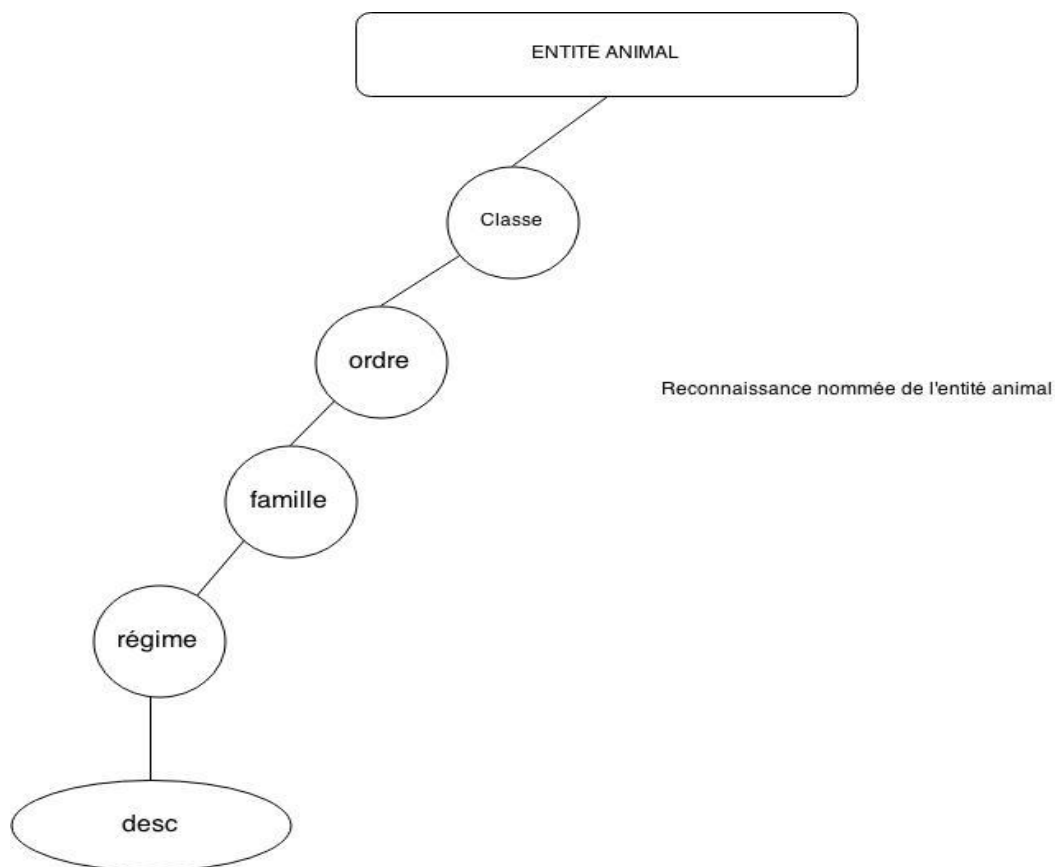
Le mécanisme de fonctionnement de l'application est le parcours de parcourir l'arbre depuis le lemme en remontant chaque sommet jusqu'à atteindre son hyperonyme dans le but d'avoir tous les caractéristique commune de l'animale et son brève description.

Analyse d'un lemme:

Un lemme d'une phrase correspond plusieurs définition et de caractéristique et sera représenté comme suit. Admettons la phrase :

« Le **chien** est l'animal domestique le plus célèbre et le plus varié ».

Le mot 'chien ' est représenté comme suit :



Pour résoudre la sémantique dans un phrase, cette étiquetage va avoir une importance capitale, d'où le faite de classer chaque mot. Ici, le chien esthétiqué par plusieurs définition dont sa classe, ordre, famille, régime, et sa brève définition. Ce qui permettra par la suite de prévoir le prochain mot utilisé pour ce bon lemme tout en évitant la sémantique.

3 Conception

3.1 L'équipe de développement

L'équipe est formée de 3 étudiants encadrés par Madame Line JAKUBIEC-JAMET.

Nom	Rôle	Contact
Line JAKUBIEC-JAMET	Encadrant	line.jakubiec@univ-amu.fr
Lucien KERISIT	Etudiant / Développeur	lucien.kerisit@etu.univ-amu.fr
Waël KOUJUK	Etudiant / Développeur	wael.koujuk@etu.univ-amu.fr
Patty RANDRIAMBOLOLONA	Etudiant / Développeur	patty.randriambololona@etu.univ-amu.fr

3.2 Mise en œuvre et répartition des tâches

3.2.1 Rédaction de rapport :

La définition et la rédaction de chaque document sera réalisée par un ou plusieurs membres du groupe. Une fois la rédaction terminée, le document sera vérifié par l'ensemble de l'équipe. Ces derniers devront s'assurer de la bonne qualité du travail réalisé, puis émettre d'éventuelles remarques et suggestions de modifications.

Enfin, les auteurs du document devront prendre en compte les retours apportés par les approbateurs et seront chargés d'effectuer les corrections nécessaires.

3.2.2 Centralisation du projet avec gestionnaire de version :

Afin de centraliser les documents et sources, nous utiliserons le gestionnaire de version Git pour pouvoir mener à bien la réalisation de ce projet informatique.

Un repository GitHub a été créé spécialement pour ce projet, l'ensemble des participants ont donc accès aux sources du projet pour y apporter leur modifications.

Le gestionnaire de version Git nous permettra de suivre scrupuleusement les modifications apportées au projet, la sauvegarde des sources et de beaucoup d'autre fonctionnalités pour pouvoir travailler en équipe de façon efficace.

Architecture de l'application et technologies utilisées

3.2.3 Contraintes techniques et solutions

La première question qui s'est posée pour la réalisation de la partie technique du projet a été de choisir une architecture pour l'application.

En effet, le site requiert une structure qui puisse stocker des informations, les utiliser et les traiter. L'application doit pouvoir gérer une base de données de façon optimale et nécessitera d'avoir des pages fluides et réactives.

Les types de périphériques qui accèderont à l'application seront variés, ce qui imposera un développement de l'application de manière « responsive » pour s'adapter à tous ces médias.



3.2.4 Justification des choix technologiques

Le développement de l'application nécessite une phase d'étude et de conception préalable à la mise en œuvre technique proprement dite. Cette phase requiert des compétences pointues et une connaissance approfondie des technologies web afin d'être en mesure de faire les choix pertinents desquels dépendront la pérennité et les performances de l'application.

Dans ce cadre, nous avons recensé les spécifications fonctionnelles de notre projet, ce qui nous amène à la rédaction du dossier d'analyse et des spécifications détaillées, pour entamer par la suite la conception d'une architecture fiable, couvrant les exigences fonctionnelles et techniques du site.

Afin d'en assurer la mise en œuvre, ce projet sera réalisé à travers plusieurs technologies récentes, telles que le langage de programmation JavaScript (jQuery, Ajax), un design HTML5 ultra-Lean.

Pourquoi HTML kickstart ?

L'innovation dans le web design nous a poussés à opter un Template ultra Lean par rapport à Bootstrap tweeter qui n'est pas mal non plus. HTML KICKSTART est un Template ultra-Lean HTML5, CSS & JavaScript fonctionnant sur n'importe quelle technologie serveur ou environnement serveur avec une dizaine de composants, et plugin JavaScript. HTML KickStart a été testé et fonctionne sur IE 8+, Safari, Chrome, Firefox, Opera, Safari IOS, et le navigateur Chrome sur Android qui permet d'avoir un design innovant et complet.

HTML kickstart est open source créée par le professeur Joshua Gatcke dans le but de faciliter la création d'un site web rapidement et facilement.



Pourquoi JQuery ?

Il existe beaucoup de bibliothèque JavaScript alors pourquoi utiliser jQuery ?

Car jQuery est le plus complet et pour être le plus complet jQuery se base sur certaines stratégies.

Exploiter le CSS

jQuery base son mécanisme de localisation des éléments de la page sur les sélecteurs CSS, ce qui permet au développeur de pouvoir sélectionner très facilement ce qu'il souhaite.

Accepter les extensions

jQuery accepte très bien et très facilement les extensions.

La création de plugin jQuery est simple et très bien documentée ce qui permet à jQuery de posséder une multitude d'extensions.

Autoriser plusieurs actions sur une ligne.

Le résultat de la plupart des opérations sur un objet est l'objet lui-même, ce qui permet de ne pas dupliquer l'objet mais de travailler tout le temps sur le même.

Pourquoi MySql ?

MySQL est la base de données open source la plus répandue au monde, elle est facile d'utilisation, le coût d'exploitation est faible, elle offre de bonnes performances et possède une véritable communauté de développeurs pouvant répondre et aider en cas de soucis quelconque.

Pourquoi PHP ?

Par rapport à Java le langage de programmation crée par Sun Microsystems qui est aussi un langage de programmation performant. Java est langage de programmation orienté objet et pas mal utilisé mais reste lourd au point de vue ressource. Pour des raisons de performance, design, rapidité et d'interaction avec la base de données, on a opté pour PHP.

PHP a été créé en 1994 par Rasmus Lerdorf, un langage purement destiné au web. La dernière version, PHP5, introduit la programmation orientée objet. Contrairement à Java et à Dot Net dont les codes sont d'abord précompilés puis s'exécutent sur une machine virtuelle, le PHP est interprété par le serveur sur lequel il s'exécute. Un même code PHP est donc capable de s'exécuter sur n'importe quel serveur, mais il n'est pas « optimisé ». Le serveur réalise cette opération à la première utilisation de l'application. Concrètement, le PHP cumule 2 avantages majeurs :

- Performance et rapidité
- LAMP (Linux Apache Mysql PHP) open source

3.3 Etat de l'art

Une partie du temps a été consacré à la recherche des projets et des recherches existants dans le domaine du traitement automatique des langues. Nous nous sommes focalisés sur les projets prenant en compte l'analyse sémantique du texte écrit.

Trois d'entre eux ont retenus notre attention de par leur avancée dans l'analyse sémantique du texte écrit et des fonctions qui y sont proposées.

Une description détaillée des avantages et inconvénients de ces projets a donc été réalisés afin de pouvoir apporter une évolution dans l'annotation du texte écrit et dans l'apport d'une nouvelle approche structurée de données.

Projets existants :

NOM	DESCRIPTION	URL
DBPEDIA spotlight	Outil permettant d'annoter automatiquement un texte (multilangages)	http://dbpedia-spotlight.github.io/demo/
Stanford NLP Tools	Etiquette un mot et renvoie sa classe grammaticale possible.	https://github.com/agentile/PHP-Stanford-NLP
ProlexBase	Reconnaissances des mots possibles en français	http://www.cnrtl.fr/lexiques/prolex/

3.3.1.1 DBPEDIA-SPOTLIGHT

Adresse URL : <http://dbpedia-spotlight.github.io/demo/>

Catégorie d'usage : Outil d'analyse de texte écrit.

Cible principale : grand public.

Droits d'inscription: gratuit.

Description du projet :

Le projet Dbpedia a été lancé en juin 2010 par des chercheurs de l'université libre de Berlin du groupe System Web dans le but de proposer un annotateur sémantique intégré avec le dépôt sémantique Dbpedia.

Dbpedia-spotlight est l'outil qui utilise la base de données Dbpedia-spotlight en tant que lien documentaires. Dbpedia-spotlight est un projet open source de type web sémantique également utilisable afin de tester le service web.

Dbpedia Spotlight est accessible librement via un service web pour des besoins de tests. Pour des usages plus intensifs, le code source en langage Java et Scala est disponible sous licence Apache. La distribution de The DBpedia Spotlight inclut également un plugin jQuery3 qui permet aux développeurs d'annoter des pages web à la volée lors de leur consultation. Divers clients sont proposés en Java et en PHP pour simplifier l'intégration de Spotlight dans un programme.

Notre avis :

Bien que la base de données dbpedia reste conséquente, l'annotation sémantique n'est malheureusement disponible qu'en anglais et propose très peu d'annotation en français voire pas du tout.

3.3.1.2 STANFORD NAMED ENTITY RECOGNIZER (NER)

Adresse URL : <http://nlp.stanford.edu/index.shtml>

Catégorie d'usage : Outil d'analyse grammatical.

Cible principale : grand public.

Droits d'inscription: gratuit.

Description du projet :

L'application Stanford de reconnaissance d'entité nommée est implémentée en langage Java. Cette application annote une séquence de mots dans un texte telle que les noms des choses, les noms de personnes et de sociétés, ou des noms de gènes et des protéines.

L'application est très efficace pour l'annotation des mots anglais notamment pour les mots appartenant à ces 3 classes (personne, organisation, lieu). Les domaines de classification sont donc assez restreints.

De type licence GNU disponible au téléchargement, la disponibilité des sources a permis l'émergence de projets utilisant Stanford NER avec l'utilisation d'autres langages.

Notre avis :

L'application ne peut pas détecter les erreurs et comportent souvent des bugs. En effet, en pratique peu de phrases sont reconnues. La reconnaissance se fait uniquement en anglais et l'ajout mineur du support d'autres langues reste très superficiel.

3.3.1.3 PROLEX – PROLEXBASE 0.1

Adresse URL : <http://www.cnrtl.fr/lexiques/prolex/recherche.php?page=lex&lang=fra>

Catégorie d'usage : Outil d'analyse de mot.

Cible principale : grand public.

Droits d'inscription: gratuit.

Description du projet :

Le projet Prolex, piloté par le Laboratoire d'informatique (LI) de l'université François-Rabelais de Tours, a pour but de fournir, à la communauté du traitement automatique des langues (Tal), des connaissances sur les noms propres, qui constituent, à eux seuls, 10% des textes journalistiques. C'est une plate-forme comprenant un dictionnaire électronique relationnel multilingue de noms propres, des systèmes d'identification des noms propres et de leurs dérivés, des grammaires locales, etc.

Prolexbase est un projet TAL du laboratoire informatique regroupant les ressources de base de l'application. Cette ressource est maintenue en étroite collaboration avec :

- le laboratoire ligérien de linguistique ;
- l'université de Belgrade;
- l'académie des sciences de Varsovie.

Notre avis :

L'application ne peut pas détecter les erreurs et comportent de nombreux bugs. En effet, la base de données semble trop limitées pour pouvoir reconnaître certaines phrases même basique.

DEVELOPPEMENT

4 DEVELOPPEMENT

4.1 Schéma du domaine animale

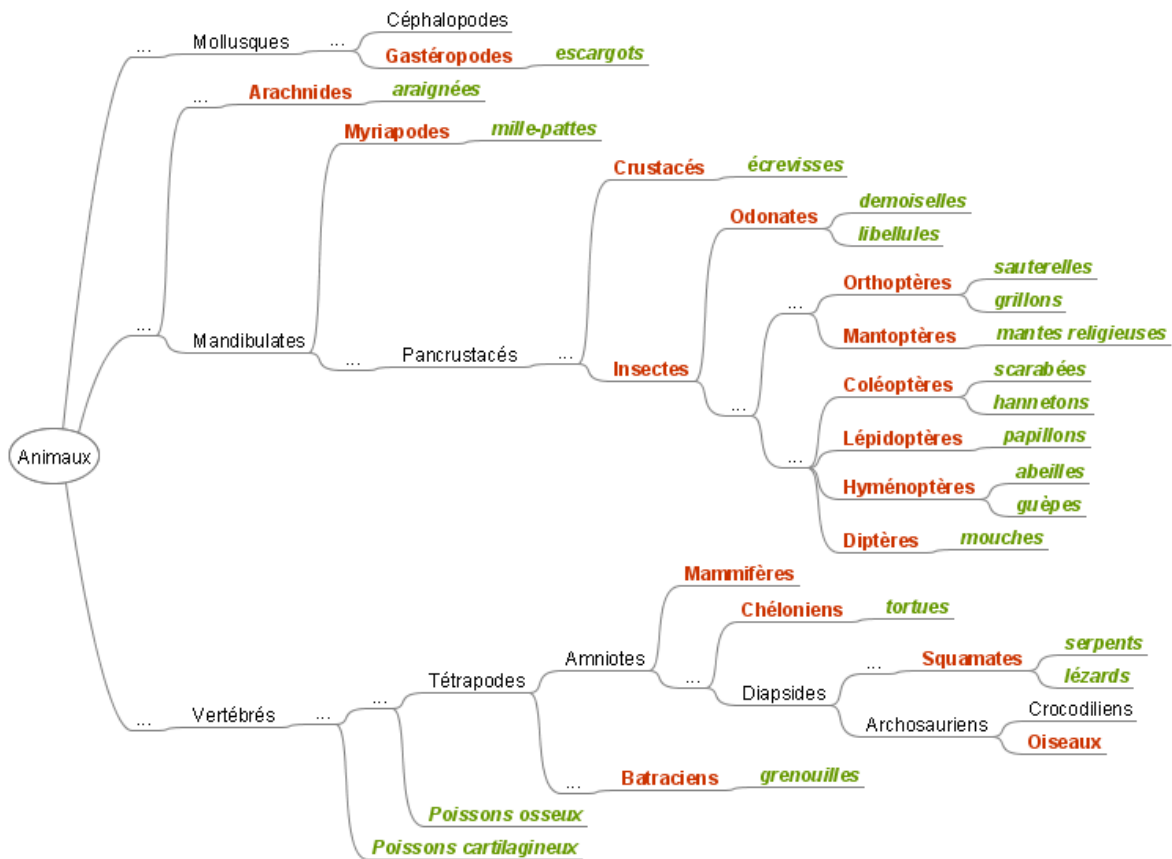
4.1.1 La Classification du vivant

Différentes classifications des êtres vivants ont été élaborées au cours des époques. La classification actuellement utilisée est dite "phylogénétique". La classification phylogénétique est un système de classification des êtres vivants qui a pour objectif de rendre compte des degrés de parenté entre les espèces, et qui permet donc de comprendre leur histoire évolutive (ou phylogénie). Elle ne reconnaît pas certains groupes comme les reptiles ou les poissons contrairement à la classification classique.

Dans notre projet, on opte pour la classification classique pour des raisons de simplicité et d'utilisation.

4.1.2 Le règne animal

Les principales familles animales se décomposent (très grossièrement) d'une manière illustrée ci-dessous. Les noms en brun sont ceux que nous avons utilisés pour classer les observations, tandis que les noms en vert donnent des exemples d'animaux dans les différentes classes.



Bien que la classification phylogénétique ne les reconnaisse pas, par soucis de simplicité, nous parlerons sur ce site :

- de reptiles pour englober les squamates (serpents et lézards) et les chéloniens (tortues)
- de poissons quelque soit leur nature (osseux ou cartilagineux).

4.1.3 Adaptation du domaine à l'application

La représentation du règne animale reprend le schéma au-dessus (voir domaine de classification), qui sera le modèle dans notre application.

Un animal est représenté comme suit dans notre base de données :

- Identité (id)
- Nom
- Classe
- Description

L'ID est la clé primaire de la table animale, de type numérique. Pour optimiser la recherche d'un animal, on peut indexer l'ID afin réduire la complexité de la recherche d'une information concernant un animal.

Le Nom comme son nom l'indique fait l'objet de représentation du nom commun d'un animal. La classe, on a aussi ajouté des informations sur la classe des animaux. Ces classes sont d'une importance capitale pour relever la sémantique plus tard. Les classes qu'on a définies sont les mammifères,

La description est la définition propre de l'animale.

4.1.4 Le schéma de l'arborescence



4.2 Données et schéma de la base de données

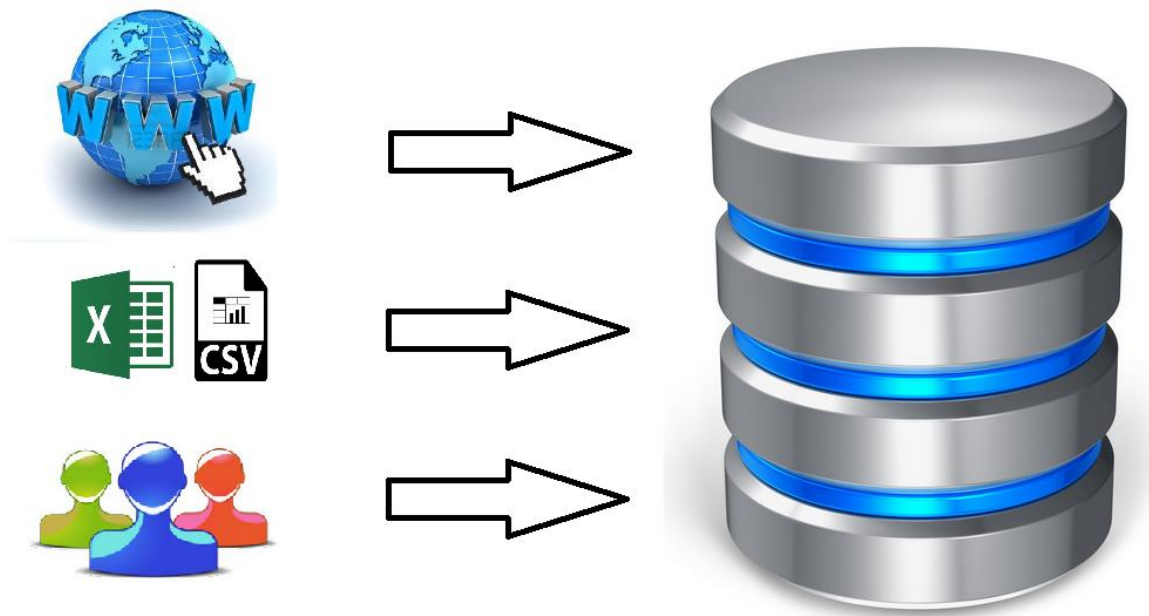
4.2.1 Les données

La conception de la base de données s'est avéré être la partie du projet qui a nécessité le plus de temps en terme de recherche. Le domaine animal est un domaine assez vaste et l'élaboration d'une base de données la plus complète possible s'est avérée complexe.

En effet, la base de données a pu être alimentée à partir de 3 sources principales :

- Des sites spécialisés sur les animaux (diconimoz.com, animaux.org, ..)
- Des sources déjà existantes de travaux de laboratoire de recherche notamment des dictionnaires de mots.
- De choix arbitraire de notre part en insérant des données pertinentes.

La difficulté a été la réalisation de scripts permettant de récupérer les données spécifiques choisies de plusieurs centaines de pages.



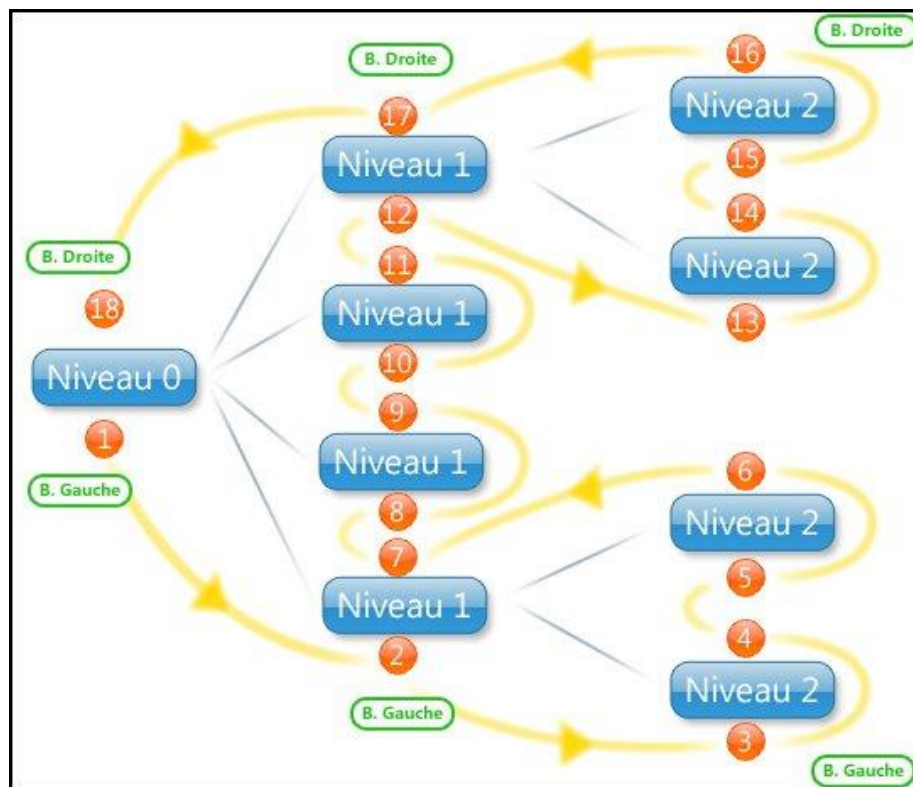
4.2.2 Modèle conceptuel de données

Le modèle de la base de données est extrêmement simple. Elle comporte 2 tables principales qui modélisent le domaine animal.

Cependant, la quantité d'informations contenues dans chaque table, notamment la table « animal_tree » est énorme. Pour pouvoir palier au problème de lenteurs de recherches possible au vues des requêtes qui y seront effectuées, nous avons mis en place une représentation sous forme d'arbre ou appelée « représentation intervallaire » afin d'y optimiser grandement les requêtes.

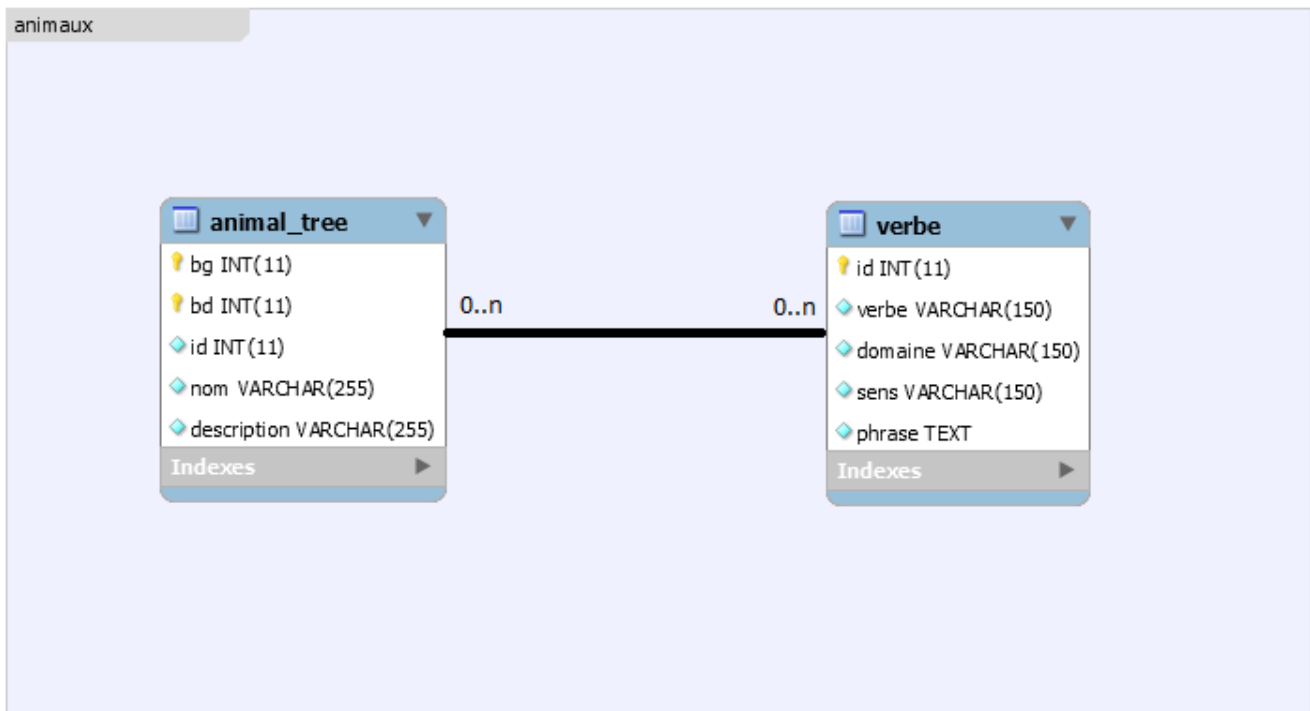
Cette technique peut permettre de situer un élément dans une hiérarchie. Elle introduit la notion de borne gauche (bg) et borne droite (bd). Chaque borne de chaque élément se voit attribuer un nombre. Ce nombre permet de déterminer la position de l'élément dans l'ensemble.

Voici un schéma qui résume le principe de représentation intervallaire :



Si l'insertion ou la suppression dans une telle représentation est un peu coûteuse, l'interrogation et notamment la recherche s'exprime la plupart du temps en une seule requête.

- Avantages : Les interrogations de la base de données sont plus rapides !
- Inconvénients : les insertions dans la base de données sont plus coûteuses, ce qui dans notre cas nous convient car les insertions dans la base sont rares.

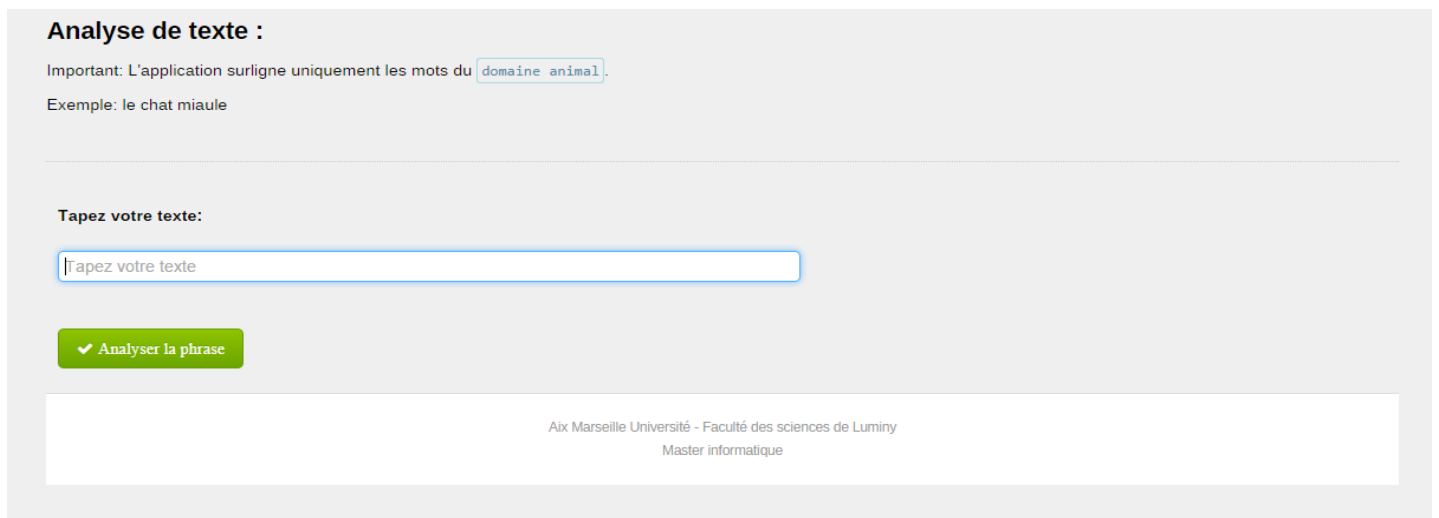


4.3 Développement de l'application web

4.3.1 Interface graphique ; lancement de l'analyse

L'utilisation de l'application a été conçue de manière à être simple d'utilisation. Les différentes pages apportent à l'utilisateur diverses informations au sujet de ce projet ainsi que les auteurs qui l'ont réalisé.

L'analyse sémantique se fait de façon intuitive. L'utilisateur peut entrer une phrase ou un texte plus long et lancer l'analyse. L'algorithme de l'application traite donc le texte avec sa base de données.



The screenshot shows a web application interface for text analysis. At the top, it says "Analyse de texte :". Below this, there is a note: "Important: L'application surligne uniquement les mots du domaine animal." and an example: "Exemple: le chat miaule". A horizontal line separates this section from the input area. The input area has the label "Tapez votre texte:" followed by a text input field containing the placeholder "Tapez votre texte". Below the input field is a green button with a checkmark icon and the text "Analyser la phrase". At the bottom of the interface, there is a footer area with the text "Aix Marseille Université - Faculté des sciences de Luminy" and "Master informatique".

Image : C'est l'interface permettant d'entrer le texte à analyser.

4.3.2 Interface graphique : résultat de l'analyse

Pour réaliser un affichage agréable de l'analyse du texte, nous avons disposé l'écran des résultats sur 2 sections distinctes :

- A gauche, le texte d'origine avec en plus les mots reconnus correspondant au domaine animal qui seront surlignés en vert.
- A droite, l'annotation sémantique du mot sur lequel l'utilisateur a cliqué.

L'affichage des annotations sémantiques de chaque mot se fait de façon dynamique, c'est-à-dire que la page ne se recharge pas. A chaque clique que le mot choisi, une requête AJAX interroge la base de données et affiche grâce à une fonction JavaScript l'annotation de ce mot.

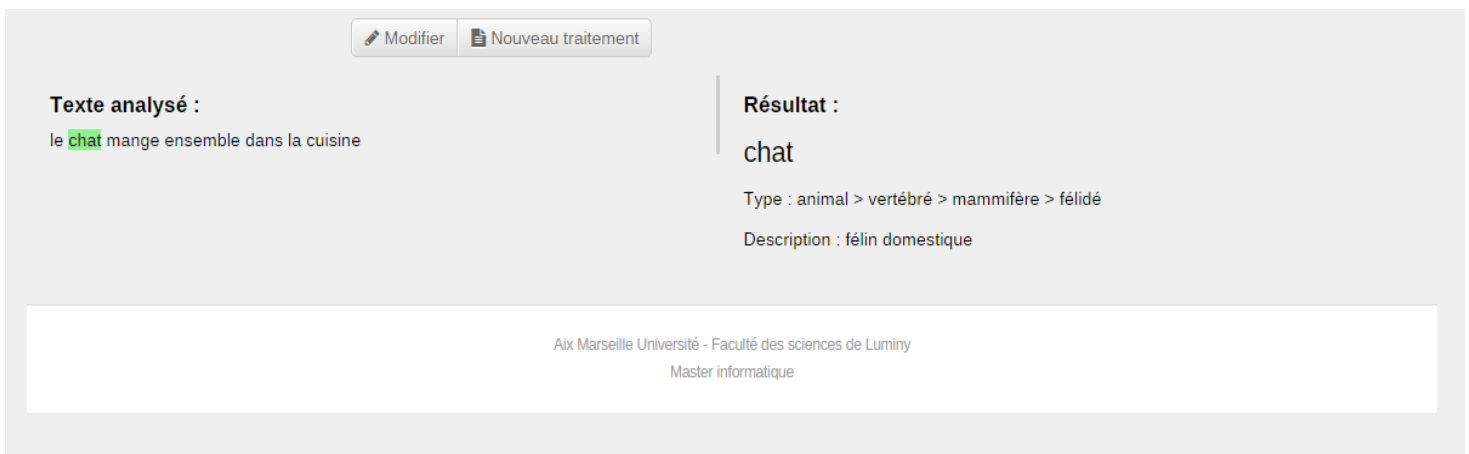


Image : l'affichage résultat de l'analyse du texte.

4.3.3 Amélioration du projet

Plusieurs perspectives d'améliorations du projet sont envisagées pour faire évoluer l'application.

La première est de continuer à alimenter la base de données avec encore plus d'informations sur les mots. Plus le nombre de mots reconnus et leurs informations seront complets, plus la précision de l'application en sera grande.

Une analyse du lemme d'un mot permettrait grandement de pouvoir reconnaître des mots sous différentes formes, c'est-à-dire que si l'application reconnaît un mot du domaine animal, il pourrait ainsi à partir de celui-ci retrouver les pluriels de ce mot ou sous des formes différentes (conjugaisons, etc...).

5 Conclusion

Le projet TER réalisé a été pour notre part une expérience universitaire exceptionnelle. Il nous a permis de mettre en pratique tout l'éventail d'enseignements dont nous avons bénéficiés lors de cette année de Master 1 Informatique.

Ces semaines de recherches, de réflexion et de développement ont été pour nous l'occasion d'évaluer nos compétences dans le domaine de la gestion de projet, et du développement ainsi que la mise en pratique de nombre d'outils et méthodes de travail enseignés tout le long de cette année.

Nous avons ainsi pu démontrer nos capacités à pouvoir travailler en équipe en mettant en place des méthodes de travail modernes apprises durant cette année universitaire pour mener à bien ce projet.

Ce projet a été une expérience enrichissante sur tous les niveaux et particulièrement la phase de recherches et de réflexions sur un thème qui nous était au départ parfaitement inconnu.

Nous avons été très satisfait de la manière dont le projet s'est déroulé et des conseils précis qu'a pu apporter notre encadrant.