Faculté
des Sciences
Aix*Marseille Université

Sémantique et langue naturelle

Réalisé par :

- ***Waël KOUJUK***
- ***Lucien KERISIT***
- ***Patty RANDRIAMBOLOLONA***

Encadrant :

- ***Line JAKUBIEC-JAMET***

Introduction

Le Traitement Automatique des Langues (TAL) est un domaine pluridisciplinaire ou 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.

Sujet détaillé

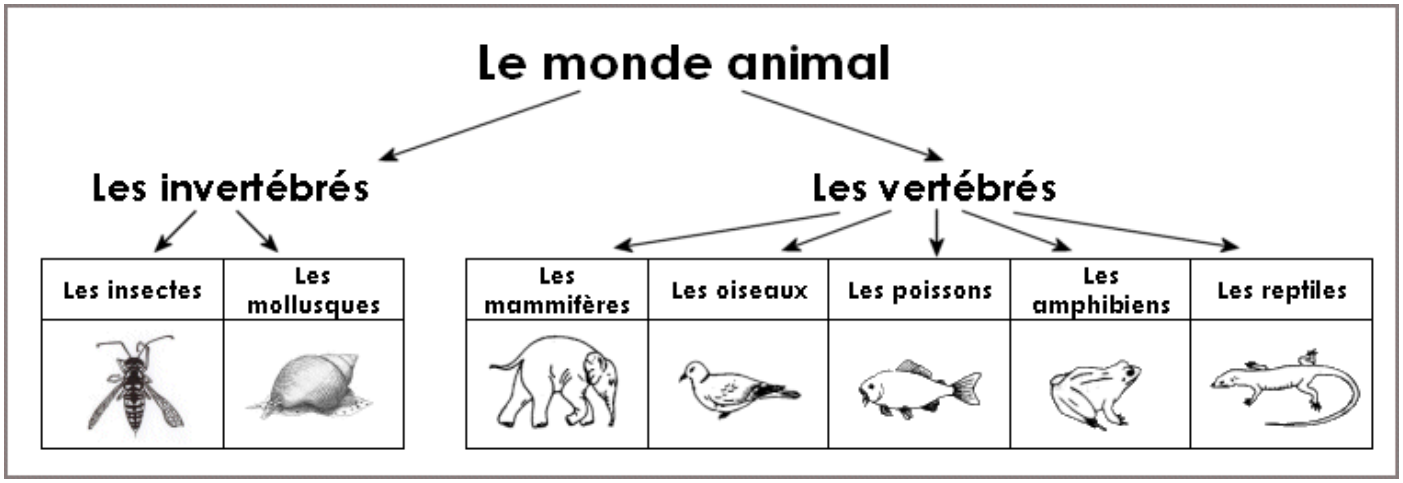
1) Objectifs

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) 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 utile et intéressant pour les futurs utilisateurs.



3) 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

4) Mise en œuvre et répartition des tâches

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.

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.

5) Architecture de l'application et technologies utilisées

- 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.

- ## 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), le Framework CSS Bootstrap

Pourquoi Bootstrap ?

Tout d'abord, Bootstrap est un Framework HTML, CSS & JavaScript fonctionnant sur n'importe quelle technologie serveur ou environnement serveur avec une dizaine de

composants, et plugin JavaScript. Son concept est la création et maintenance rapide d'un site internet avec rendu correct, et interface complète.

Les avantages de Bootstrap sont sa facilité, accessibilité, sa structure, sa méthodologie, sa vitesse de développement accrue, porté vers le futur, grille fixe ou fluide, personnalisable et modulable.

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 ?

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

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 proposé 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 test du 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.

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 restreinte.

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.

PROLEX – PROLEXBASE 0.1

Adresse URL : http://www.cnrtl.fr/lexiques/prolex/recherche.php?page=tex&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 limité pour pouvoir reconnaitre certaine pharse voire basic.

7) Schéma de la base de données