Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté d'électronique et d'informatique
Département d'informatique

Rapport

Module : Traitement automatique du langage naturel

Master 1 SII

Mini projet

Réalisation d'unn outil d'aide au développement d'un dictionnaire de la langue arabe sur des bases historiques

- Réalisé par :

  BENHADDAD Wissam
  BOURAHLA Yasser
  MOHAMEDI Haroune
  LAHBIB Abdelghani

10-12-2018

Table des matières

Table des figures

CHAPITRE 1

MOTIVATIONS ET PROBL[PLEASEINSERT\PRERENDERUNICODE{ÃĽ}INTOPREAMBLE]MATIQUE

1.1 Introduction

Depuis son apparition (au IIe siècle), la langue arabe n'a cessé d'évoluer, donnant naissance à de nouveaux mots, ou modifiant le sens de mots existants. Cette évolution a particulièrement enrichi le vocabulaire de la langue de l'islam, en conséquent et au fil du temps, plusieurs ouvrages destiné a recenser les différentes définitions et sens d'un mot ont vu le jour, chacun durant sa période. néanmoins, il est primordial de garder une trace des différents changement qui ont eu lieu sur ces mots, et cela depuis leurs émergence. C'est avec cette idée en tête que les lexicographes des temps modernes en eu l'initiative d'entamer la construction de dictionnaires historiques afin de regrouper toutes les nuances des mots a travers les ages.

La création d'un tel ouvrage n'est pas chose facile, en effet elle demande d'une part une grande connaissance sur les différentes périodes historiques de la langue, ainsi que sur la langue en elle même durant ces périodes. Chercher et regrouper des écrits, documents et ouvrages des différentes auteurs durant ces périodes est une tâche qui est en elle même très ardue, cela peut prendre plusieurs décennies pour créer une collection de documents asses représentative de chaque période. Analyser le contenue de tout ces documents est la phase qui dure le plus de temps, une vérification minutieuse de chaque information ajoutée au dictionnaire finale doit être faite, puis sujette à l'approbation de plusieurs experts du domaine.

C'est avec l'avènement de l'informatique, de l'intelligence artificielle et plus récemment avec l'explosion du volume de données présent sur internet, que l'idée d'utiliser ces technologies pour faciliter et accélérer le processus de création d'un dictionnaire historique de la langue arabe à émergé. En effet la grande quantité et diversité de documents présente sur internet pourrait être exploité par un lexicographe pour ne pas s'attarder sur fastidieuse tâche de collecte des données, et cela en utilisant des techniques de traitement automatique du langage (TALN), de recherche d'information(RI) et d'intelligence artificielle.

Ce besoin d'un outillage informatique est la principale motivation derrière ce mini-projet, avec suffisamment de données et une bonne conception, la réalisation d'un tel outillage pourrait faire gagner énormément de temps aux lexicographes du monde arabe.

Le but de projet étant maintenant établi, nous allons maintenant passer à la schématisation de ce rapport. Nous commencerons d'abord par de petites définitions pour se situer dans la suite du rapport, nous enchaînerons ensuite sur la conception du système pour expliquer le travail réalisé, viendra ensuite la présentation de notre application, enfin nous finirons par une conclusion générale comportant un bilan du projet, des critiques sur notre système ainsi que les perspectives envisagées.

1.2 Définitions

1.2.1 DataSet et Corpus

Un jeu de données (DataSet) est un ensemble de données traité et organisé dans un schéma spécifique aux besoins d'un système, un dataset peut être une base de données relationnelle, un ensemble de fichiers texte, une banque d'images/videos ...

Dans notre cas nous nous intéresserons plus particulièrement à un type de dataset appelé Corpus, informellement un corpus est un dataset principalement utilisé dans le domaine du TALN, il est constitué d'un ensemble de fichier texte (annotés ou pas) qui représentent un domaine, une thématique, un(ou des) type(s) d'ouvrages ...

Un corpus est un composant essentiel pour la l'application des techniques de TALN, la taille et la qualité d'un corpus est donc un facteur primordial pour assurer une bonne performance d'un système.

1.2.2 TALN

Le Traitement Automatique du Langage Naturel (TALN) est un sous domaine de l'intelligence artificielle qui vise à analyser et à modéliser les composants du langage humain, que ce soit du point de vue syntaxique, sémantique ou pragmatique. l'aspect principal du TALN est le fait de permettre aux machine de traiter les séquence de texte non plus comme une simple suite de symboles, mais comme des entités informationnelles. Des connaissances sur la langue sont un prérequis essentiel pour le développement d'un système utilisant le TALN, ainsi que la disponibilité d'un grand ensemble de données pour faire de l'apprentissage automatique.

Le TALN est découpé en un ensemble de techniques et opérations à appliquer sur du texte, une multitude de domaine d'application existent pour l'utilisation de ces derniers. Dans ce projet nous nous intéresserons principalement aux techniques suivantes :

Lemmatisation

La lemmatisation est un terme désignant l'analyse lexicale d'un texte dans le but de regrouper les mots d'une même famille. Les mots d'une même famille sont donc réduits en une unique entité appelée « **lemme** ». Ainsi la lemmatisation consiste à regrouper les différentes flexions d'un mot unique.[1]

Segmentation

La segmentation d'un texte est l'opération de découpage de ce dernier en composantes linguistiques plus petites(des phrases, des groupes nominaux, des mots ...), c'est un processus non-trivial car chaque langue dispose de règles spécifiques en ce qui concerne les marqueurs de fin de phrases.

Étiquetage morphosyntaxique (PoS-Tagging)

il consiste à identifier pour chaque mot sa classe morphosyntaxique(Nom,Verbe,Nom pluriel, ...) à partir de son contexte dans un corpus ou texte ,ainsi que de connaissances lexicales de la langue.

1.2.3 Dictionnaire historique

Informellement, un dictionnaire historique est un ouvrage qui rassemble, sous forme d'un liste d'entrées, un ensemble de mot d'une langue donnée avec leurs définitions et/ou des exemples d'utilisation selon des périodes historiques prédéfinies (en rapport avec la langue ou pas).

1.3 Conclusion

Au terme de ce chapitre, nous avons une idée plus claire sur le travail qui doit être réaliser, nous allons donc attaquer l'aspect conceptualisation, il s'agira principalement de définir les composants de notre système.

CHAPITRE 2

CONCEPTION DU SYSTÈME

2.1 Introduction

Comme mentionné précédemment, nous allons nous intéresser dans ce chapitre à la conception que nous avons réalisé, nous présenterons un schéma global du système, puis nous nous détaillerons le rôle de chaque composant, en donnant un exemple d'utilisation et/ou du flux de donnés qui entre/sort de ce dernier, nous parlerons ensuite du déploiement du système dans une plateforme serverless(dans le cloud), principalement car c'est un aspect important de l'expérience d'utilisation(UX).

2.2 Schéma global du système

Notre système se compose essentiellement de deux parties(elles même subdivisées en plusieurs modules) :

- **Récupération et pré-traitement des données** : principalement, c'est depuis des sites web que le système cherche des données, puis il se charge d'organiser les fichiers téléchargés dans un espace de stockage.
- **Exploitation et mise à jour des données récupérées** : c'est la partie où les données qui sont maintenant structurées et organisées seront utilisées par l'application, qui dans notre cas se trouve être une application web hébergé dans le cloud.

Le schémas suivant explicite un peu plus l'explication précédente :

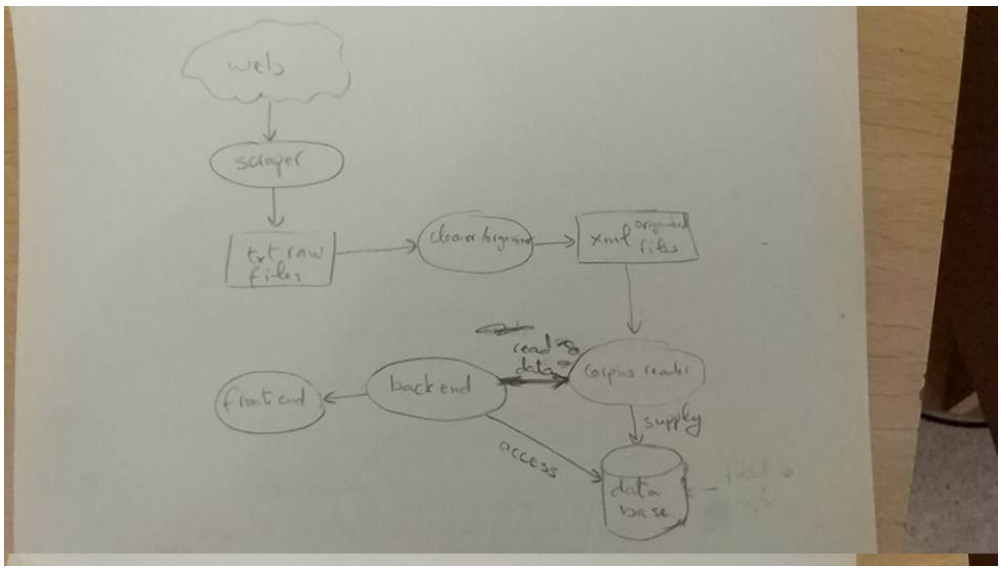

FIGURE 2.1 – Schéma global du système

2.3 Les modules du système

2.3.1 Aspirateur de sites web

2.3.2 Organisateur de corpus

2.3.3 Corpus reader

2.3.4 Base de données

2.3.5 Application (Front-end et Back-end)

2.4 Déploiement des modules dans le cloud

2.5 Conclusion

CHAPITRE 3

RÉALISATION DE L'APPLICATION

3.1 Introduction

3.2 Environnement de travail et outils utilisés

3.2.1 Python

3.2.2 JavaScript

3.2.3 NLTK

3.2.4 VueJS

3.2.5 Django

3.2.6 PostgreSQL

3.2.7 Google Cloud Platform

3.3 Présentation de l'application

3.3.1 Interface principale

Insert for each subsubsection an exemple of usage

3.3.2 Corpus

Upload new corpus into the server

3.3.3 Corpus Browser

Explore the corporas

3.3.4 Add Entry

Add new entries into the historical dico

3.3.5 Dicos

All the words from the historical dico

3.3.6 Graphs

Display different statistics for some words

3.4 Fonctions supplémentaires

3.5 Conclusion

CHAPITRE 4

CONCLUSION GÉNÉRALE

4.1 Objectifs atteints

4.2 Limites du système

4.3 Perspectives futures

BIBLIOGRAPHIE

[1] “La lemmatisation : Optimiser le seo avec la lemmatisation.” https ://www.yakaferci.com/lemmatisation-seo/. (Accessed on 12/13/2018).