الجزائرية الديمقراطية الشعبية الجمهورية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي - 08 ماي 1945 – بسيدي بلعباس
Ecole Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbes

Mémoire de Fin d'étude

Pour l'obtention du diplôme d'**ingénieur d'état**

Filière : **Informatique**

Spécialité :**Ingénierie des Systèmes Informatiques (ISI)**

Thème

Classification des textes en arabe basée sur les représentations en graphes

---

Présenté par **:**

- Mlle LACEFAR Iness Chaima

Soutenu le : **29/09/2020** Devant le jury composé de :

- M. KESKES Nabil — Président
- Mme BENNABI Sakina Rim — Encadreur
- M. SAIDI Fatih — Examinateur

Année Universitaire : 2019 / 2020

Résumé

Le traitement automatique du langage naturel est très difficile à maîtriser à cause de la complexité du langage humain et des particularités qu'il contient, ce qui empêche d'appliquer des règles fixes qui couvrent tous les aspects. Les recherches dans le domaine du traitement de la langue arabe n'ont vu le jour que ces dernières années où différentes tentatives ont présenté des performances plus ou moins satisfaisantes. Ces recherches se basent généralement sur des représentations de base qui utilisent des notions simples pour pondérer et classer les termes. Cependant, ces représentations ne sont pas assez précises pour refléter les contenus des documents arabes dans lesquels les dépendances entre les mots sont assez fortes,et aussi du fait de l'ambiguïté, d'agglutination et de divers problèmes de cette langue. Dans ce travail, nous présentons quelques approches qui montrent quelques représentations en graphes et en embedding des mots proposées pour la classification des textes, et nous étudions aussi les connaissances linguistiques et morphologiques de la langue arabe ce qui permet d'améliorer les performances, en exploitant pleinement sa richesse.Nous présentons également une comparaison entre ces travaux et une synthèse pour extraire suffisamment d'informations dans ce domaine de recherche.

En se basant sur cette étude bibliographique, une application web est proposée pour la classification des textes arabes en se basant sur une représentation des graphes, en utilisant l'algorithme **TextRank** et des méthodes d'apprentissage automatique supervisé.

Mots-clés : La classification textuelle, représentations en graphes, embeddings des mots, TextRank.

Abstract

Automatic natural language processing mastering is complicated due to the complexity of human language and its charasteristic, which avoids the application of fixed rules to the totality of aspects. Research in the field of language processing for the Arabic language only emerged a few years after the first work, various attempts have yielded more or less satisfactory performance. These searches are generally based on traditional representations ; use simple concepts to weight and classify the terms. However, these representations are not precise enough to represent the contents of Arabic documents in which the dependencies between the words are strong, and because of the ambiguity, agglutination, and various problems of this language. In this work, we present some approaches which show some representations in graphs and embedding of the words proposed for the classification of texts, and also we study the linguistic and morphological knowledge of the Arabic language, which allows improving performances, in fully exploiting its wealth. This allows at the end to make a comparison between these works and then a synthesis to obtain enough information in this area of research.

Based on this bibliographic study, a web application is proposed for the classification of Arabic texts based on a representation of graphs, using the textbf TextRank algorithm and supervised

machine learning methods.

Keywords : Textual classification, Graph representations, Words embedding, TextRank.