Université de Monastir - Institut Supérieur d'Informatique et de Mathématiques de Monastir

Département en Sciences de l'Informatique

1ér année ingénieur informatique : Cours Algorithmes de Machine Learning - 2022/2023

Enseignant(e) du cours : Mariem Gzara

Projet Algorithmes de Machine Learning

Sujet Proposé par : Mariem Gzara et Azza Chebil

1- Objectif du projet

Analyse des versets et des sourates du Coran par les algorithmes de machine learning et interprétations.

2- Description de l'ensemble de données

- Donatrice : Madame Chadia bouselama, formatrice en Sciences de récitation du Coran.
- Format du fichier : fichier CSV
- Les attributs :
  - Texte du verset du Coran
  - Numéro du verset du Coran dans la sourate
  - Numéro de la sourate ∈ 1 .. 114
  - Nom de la sourate
  - Type de la sourate
- L'ensemble de données à analyser dans votre projet : l'ensemble de données brutes données par Madame Chadia a été préparé par Mademoiselle Azza Chebil, doctorante en Science de l'informatique dans le domaine du "Text Mining ".

Les prétraitements suivant sont effectuées :

- Tokenisation

- Elimination des stops words
- Lemmatisation et stemming
- Vectorisation TF- IDF

Les attributs suivants sont retenus pour l'analyse :

- Numéro du verset dans la sourate
- Nom de la sourate
- Numéro de la sourate
- Type de la sourate
- Liste des mots "token" retenus après la préparation de chaque verset
- Le Tf-IDF de chaque "token" mot retenu dans le Coran.

$$Tf - Idf(i,j) = Tf(i,j) * Idf(i,j)$$

$$Tf(i,j) = \frac{nombre\ d'occurrences\ du\ mot\ j\ dans\ le\ document\ i}{nombre\ de\ mots\ dans\ le\ document\ i}$$

$$Idf(i,j) = \log(\frac{nombre\ de\ documents\ qui\ contiennent\ le\ mot\ j}{nombre\ total\ de\ documents\ dans\ l'ensemble\ de\ données})$$

Dans notre cas, un document i est un verset du Coran et l'ensemble de données est l'ensemble de versets du Coran.

3- Travail à réaliser

Nous nous limitons à donner quelques recommandations pour vous aider à mener votre projet.

1 Description des données :

- Dimension
- Les attributs et leurs types
- Y-a-t-il des données manquantes ?
- Nombre de sourates
- Nombre de versets
- Nombre de classes de sourates et lesquelles

- Nombre de sourates dans chaque classe
- Signification des attributs et leur nombre

2 Exploration des données

- Répartition des versets par classe
- Répartition des sourates par classe
- Distribution :
    - Nombre de versets par sourate
    - Nombre de token "mots clés" par verset
    - Nombre de token "mots clés" par sourate
- Les 100 mots clés les plus fréquents
- Les 100 mots clés les moins fréquents
- Corrélation entre les mots clés :
    - Les mots clés les plus corrélés
    - Les mots clés non corrélés
- Etude du nombre de versets par mots clés, ie, le nombre de versets qui contiennent un mot clés.
- y-a-t-il des mots clés à éliminer et pourquoi ?

4- Analyse des données

- Application des algorithmes de réduction de dimension telle que l'ACP et l'AFC.
- Interprétation des résultats et nombre des axes retenus. Justifier.
- Application des algorithmes de sélections des attributs. Interprétation des résultats et quels sont les mots clés retenus.

5. **Classification non supervisé**

- Clustering des versets du Coran
    - Quel est les nombre de clusters retenus ?
    - Visualiser les mots clés fréquents par cluster et interpréter le résultat.
    - Comparer les résultats de plusieurs algorithmes de clustering
- Clustering de sourates du Coran en deux classes et comparaison avec le clustering réel en deux classes madeni et mequi.

5- Classification supervisé

- Classification des sourates du Coran et des versets du Coran.

- Prédire la classe d’une sourate en mequi ou madeni.
- Prédire la classe d’un verset selon les classes crées dans la question précédente.

6- Analyse des associations

- Analyser les associations entre les mots clés des versets du Coran.
- Générer les itemsets fréquents et les règles d’associations et Interpréter les règles

BON TRAVAIL