Mini Projet : Machine Learning

Génie informatique 4

Détection du niveau de lisibilité des textes arabes

Réalisé par :

Kessou Kaouthar

Hmama Mohammed

Harmaze Aymane

Jaouane Taha

Encadré par:

Bouchentouf Toumi

Haja Zakariae

I. Introduction

Les systèmes de mesure de lisibilité sont utilisés pour mettre en évidence le niveau de rédaction des textes. Ils peuvent aider les auteurs à rédiger des textes dans un style facile à lire. De plus, ils affichent souvent un score global de lisibilité qui est dérivé d'une formule de lisibilité ou d'un modèle de prédiction.
Dans le contexte de notre projet , les textes arabes seront classifiés selon cinq niveaux :

- Niveau 1 : Très facile
- Niveau 1+ : Facile
- Niveau 2 : Moyen
- Niveau 2+ : Difficile
- Niveau 3 : Très difficile

II. Données et Caractéristiques

1. Données

Pour développer un modèle de prédiction du niveau de difficulté des textes arabes, nous avons besoin d'une collection de textes accompagnés de leurs niveaux de difficulté afin de suivre une approche de classification supervisé. Pour ce faire, nous avons collecté un ensemble de 271 textes à partir de la plateforme ''Global Language Online Support System (GLOSS[1])''. La plateforme offre des milliers de leçons dans des dizaines de langues pour les apprenants indépendants dans le but de renforcer leurs compétences en écoute et en lecture.
Les textes arabes dans GLOSS sont annotés avec cinq niveaux de difficulté en utilisant l'échelle de l'organisation "Interagency Language Roundtable (ILR)". L'échelle permet de mesurer les compétences linguistiques au sein du gouvernement fédéral américain. Cette échelle évalue les compétences linguistiques des personnes sur des niveaux allant de 0 à 5. Les niveaux 0+, 1+, 2+, 3+ ou 4+ sont utilisés lorsque les compétences d'une personne dépassent largement celles d'un niveau donné, mais sont insuffisantes pour atteindre le niveau suivant. Le Tableau 1 présente la répartitions des textes que nous avons collecté sur cinq niveaux (1, 1+, 2, 2+, et 3).

Niveau de difficulté	Nombre de textes
1	34
1+	29
2	95
2+	68
3	45

Tableau 1:Distribution des textes par niveaux

2. Caractéristiques

Pour représenter vectoriellement nos textes, nous avons collecter un ensemble de six caractéristiques très utiliser dans le domaine de la prédiction de la difficulté des textes à savoir :

[1] https://gloss.dliflc.edu/

- SL (Sentence Length): un attribut qui représente le nombre de phrases dans un texte. Il est claire que plus le nombre de phrases augmente plus le texte devient difficile à lire et à comprendre par les étudiants/lecteurs ;
- WL1 (Word Length 1): représente le nombre de mots dans un texte. Come pour SL, il est claire que le nombre de mots dans un texte influence son niveau de difficulté ;
- WL2 (Word Length 2): cette variable calcule le nombre de caractères dans un texte. Elle permet a l'algorithme d'apprentissage de connaitre la longueur moyenne des mots dans un texte en ce basant sur une combinaison entre WL1 et WL2. Et donc un texte ayant des mots avec une longueur moyenne énorme est difficile à lire ;
- SC (Stems Count) : représente le nombre de stems dans un texte. Un stem représente le mot sans clitique. C'est le résultat de la combinaison d'une racine avec des affixes infléchissant pour indiquer des caractéristiques grammaticales telles que le nombre, la personne, le temps, etc. Afin de calculer le nombre de stems dans l'ensemble de nos textes, nous avons utilisé Farasa[2]. Ce dernier est un progiciel complet à la pointe de la technologie pour le traitement de la langue arabe. Il permet d'avoir différent informations morphologiques pour chaque mot du texte analysé à savoir : le stem, le lemme, etc.
- LC1 (Lemmas Count 1) : un attribut qui calcule le nombre de lemmes dans un texte. Un lemme représente l'entrée du dictionnaire. Il s'agit d'une forme spécifique qui représente le mot. Dans la langue arabe, le lemme des verbes est défini comme leur forme conjuguée à l'accompli à la 3ème personne du singulier. Alors que le lemme nominal est représenté par le singulier et masculin (si possible). Encore une fois nous avons utiliser l'analyseur Farasa pour le calcul des lemmes.
- LC2 (Lemmas Count 2) : représente le nombre de lemmes fréquents dans un texte. Un lemme est dit fréquent s'il se répète plus qu'une fois dans le text. Il est claire que si un texte contient un nombre de lemmes redondants c'est que le lecture il est entrain de pratiquer le même vocabulaire avec des conjugaison différents et donc le texte est facilement compréhensible.

III. Démarche Poursuivie

1. Extraction des caractéristiques

a. Importation des bibliothèques nécessaires

```
import os
import csv
from farasa.segmenter import FarasaSegmenter
from farasa.stemmer import FarasaStemmer
```

Pour extraire nos caractéristique nous avons commencer par importer les bibliothèque nécessaires, nous avons utilisé la bibliothèque **Farasa** pour la

[2] https://farasa.qcri.org/

lemmatisation et la stemetisation des textes. La bibliothèque **csv** a été utiliser pour l'écriture du fichier csv finale contenant l'ensemble des vecteurs caractérisant nos textes. Et finalement, la bibliothèque **OS** a été utiliser pour le parcourt de l'arborescence des dossier et fichiers des données.

Pour les caractéristiques nous avons développé six fonctions de base (une pour chaque variable) en plus de quatre fonctions secondaires utiliser par les six premier pour aboutir à leur buts.

Le résultat de cette première phase est un fichier CSV contenant 271 vecteurs qui feront l'objectif d'une deuxième étape de génération du modèle.

2. Génération du model et évaluation sur l'ensemble de teste :

a. Importation des bibliothèques nécessaires

```
from sklearn.svm import SVC
from sklearn import svm
from sklearn.metrics import f1_score, precision_score, accuracy_score
from sklearn.model_selection import train_test_split
import pandas as pd
```

Pour générer et tester le model de prédiction du niveau de difficulté des textes, nous avons utilisé différents fonctionnalités de la bibliothèque Sklearn a savoir : le modèle SVM, les différents métriques de mesure (accuracy, precision, etc.), et la fonctionnalité de séparation des données en des ensemble d'apprentissage et de test. Nous avons en plus utiliser la bibliothèque pandas pour la manipulation des fichiers de vecteurs.

Le processus que nous avons adopter pour générer et tester notre modèles consiste à :

1. ***Préparation des données*** : une étape de chargement des donnes et de séparation des caractéristiques (features) de la classe (target) ;
2. ***Partition apprentissage-test*** : nous avons divisé les données en 80% pour l'apprentissage et 20% pour le test ;
3. ***Créer un classificateur SVM*** : a ce stade nous avons tester les quatre noyaux de l'algorithme SVM {linear, sigmoid, poly,rbf}. Les meilleurs résultats ont été obtenus par le noyau linéaire ;
4. Tester le modèle généré sur l'ensemble de teste ;
5. Finalement afficher les valeurs des différents métriques afin d'évaluer les performance de notre modèle.

IV. Conclusion

La réalisation d'un système qui mesure les cinq niveaux de lisibilité pour les textes en arabe a vraiment enrichi nos connaissances dans plusieurs domaines tels que la linguistique et les statistiques.

Aujourd'hui, les modèles de prédiction de la lisibilité sont plus populaires que jamais et il existe des modèles de lisibilité pour l'espagnol, le français, l'allemand, le néerlandais, le suédois, le russe, l'hébreu, l'hindi, le chinois, le vietnamien et le coréen . Ainsi , les variables communs entre ces modèles sont celle de base que nous avons adopter dans ce travail. Donc pour chaque langue, nous pouvons incorporé d'autres variables dédié pour augmenter les performances que nous avons obtenus dans ce premier travail.