Lardon Adrian
Golfier Alexandre
Groupe 5

Projet Modélisation IA

Classification de mails spams/non spams

Etape I : Définir et Identifier le problème

Nous avons choisi de classer des mails pour savoir si ce sont des mails spams ou non spams. L'entrée du modèle sera un mail en texte(str) et la sortie sera un entier(int) qui nous indiquera si le mail est un mail spam ou non spam (0 ou 1).

Le problème sera une classification car nous allons classer les mails pour savoir s'ils sont spams ou non spams.

Etape II : Comprendre les données

Quantité de données :

- Le format des données est en texte et du binaire
- La taille de la base de données est de 5728 lignes et 2 colonne (texte et si c'est un spam ou non)

Qualité des données :

- Oui il y a du texte et un colonne spam
- La colonne texte contient des string et spam du binaire

Etape III : Le Modèle

STOPWORDS (Mot vide) :

Les stopwords sont des mots inutiles, des mots "vides", ils pourraient être « le », « la », « de », « du », « ce »…
On utilise une base de données de stopwords dans notre projet pour pouvoir trier les mots dans les mails pour n'avoir que l'essentiel. On se retrouve ensuite avec des mots évidents qui permettent au programme d'être plus efficace et plus rapide pour analyser les mails et déterminer s' ils sont spam ou non spam.

Modèle :

Nous allons trier les données du tableau en enlevant toutes les ponctuations et les stopwords. Par la suite nous allons utiliser une fonction qui permet de créer une

matrice de jetons avec chaque mot unique du tableau de données. Nous allons par la suite diviser à 70% les données en entraînement.

Une fois toutes ces étapes faites nous allons utiliser une fonctionnalité de python MultinomialNB qui permet de faire une classification avec le nombre d'apparitions de chaque mot et s'il s'agit d'un spam ou non.

On affiche par la suite les résultats obtenus et on les compare avec les vraies valeurs.

<u>En utilisant une matrice de confusion on va pouvoir déterminer la fiabilité de notre IA :</u>

```
Matrice de confusion :
 [[3011   11]
 [   0  964]]
```

Cette matrice de confusion est appliquée uniquement avec l'échantillon train qui correspond à 70% des mails analysés.

n = 3986	Prédit : NON	Prédit : OUI
Actuellement : NON	**3011** (nombre de fois que le programme à prédit non spam et que c'était bien non spam)	**11** (nombre de fois que le programme a prédit spam et que c'était non spam)
Actuellement :OUI	**0** (nombre de fois que le programme a prédit non spam alors que c'était spam)	**964** (nombre de fois que le programme à prédit spam et que c'était bien spam)

Cette autre matrice de confusion est appliquée avec le reste l'échantillon donc l'échantillon test qui correspond à 30% des mails analysés.

```
Matrice de confusion :
 [[1292   13]
 [   4  400]]
```

n = 1709	Prédit : NON	Prédit : OUI
Actuellement : NON	**1292** (nombre de fois que le programme à prédit non spam et que c'était bien non spam)	**13** (nombre de fois que le programme a prédit spam et que c'était non spam)
Actuellement :OUI	**4** (nombre de fois que le programme a prédit non spam alors que c'était spam)	**400** (nombre de fois que le programme à prédit spam et que c'était bien spam)