***ORAL Projet IA***

Bonjour je me présente, alexandre golfier et je vais vous présenter notre projet d’intelligence artificielle avec adrian lardon. Tout d’abord je vais vous définir ce qu’est une intelligence artificielle

L'**intelligence artificielle** (IA) est « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'**intelligence** ».

Dans un premier temps nous allons définir et identifier le problème puis comprendre les données et voir à quoi elles servent. Dans la troisième partie nous allons vous expliquer notre modèle et enfin comparer les résultats de notre programme.

Je laisse la parole à adrian pour démarrer

***I - Définir et identifier le problème (AD LE BOSS)***

***II - Comprendre les données***

En **informatique**, une **donnée** est la représentation **d**'une information dans un programme : soit dans le texte du programme (code source), soit en mémoire durant l'exécution. ... Les **données** peuvent être conservées et classées sous différentes formes : textuelles (chaîne), numériques, images, sons, etc.

Document csv qui contient nos données avec une colonne texte du mail et une colonne numérique de chiffre binaire 0 ou 1 qui nous indique si un mail est spam ou non.

 Quantité de données :

o   Le format des données est en texte et du binaire 0 et 1

* o   La taille de la base de données est de 5728 lignes et 2 colonnes. La colonne texte contient des string et la colonne spam du binaire

 Qualité des données :

Plus la base de données est grande plus il y aura de choses intéressantes à comparer

Les données sont très importantes dans notre projet elles vont nous servir de base pour notre programme en fait c’est le repère du programme, il va aller chercher dans les données pour les comparer et nous donner un résultat. Sans ces données le programme n’aurait rien à analyser ou comparer donc aucun résultat à nous donner.

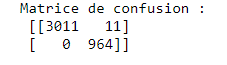
***III - Le modèle (AD LE BOSS)***

***Stopwords***

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.

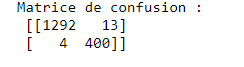
***IV - Résultat(AD LE BOSS ET MOI LOL)***



Cette matrice de confusion est appliquée uniquement avec l’échantillon entraînement 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.



|  |  |  |
| --- | --- | --- |
| 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) |