Yann Bernard
Nicolas Gauville

TP Filtre Anti-Spam

Répartition des tâches
Fonctions charger_dictionnaire / lire message / charger_base_app : Yann
Fonctions calculer_bj / calculer_probabilite / predire : Nicolas
Interface utilisateur : Les deux

Implémentation
Nous avons choisit d'implémenter le filtre anti-spam en python.
Nb : python 2 est nécessaire pour faire fonctionner le programme.

Utilisation : python filter.py

Difficultés rencontrées :
- L'utilisation de python3 provoque de nombreux problèmes de lecture à cause de l'encodage. Le problème ne survient pas avec python 2

Choix réalisés :
- Nous avons choisit d'utiliser Python parce qu'il est adapté à ce type de problème et que nous ne l'avions jamais utilisé.
- Nous avons modifié les formules pour optimiser le programme de la façon suivante :
  - On a appliqué ln() sur l'ensemble des calculs de probabilités
  - Les xi / 1-xi sont convertis en binaire (0 ou 1) au lieu d'entiers

Tests d'exécution

```
Nombre de spam de la base d'apprentissage ? (max 500) 200
Nombre de ham de la base d'apprentissage ? (max 2500) 200
Nombre de spam à tester ? (max 500) 20
Nombre de ham à tester ? (max 500) 20

Apprentissage realise sur 200 SPAMs et 200 HAMs

SPAM numero 0 : P(Y=SPAM | X=x) = -281.7232475575908, P(Y=HAM | X=x) =
-284.7473355145567 => identifie comme un SPAM
SPAM numero 1 : P(Y=SPAM | X=x) = -250.2545374084485, P(Y=HAM | X=x) =
-250.29974962911518  => identifie comme un SPAM
SPAM numero 2 : P(Y=SPAM | X=x) = -270.37221600126463, P(Y=HAM | X=x) =
-294.2756843917197  => identifie comme un SPAM
SPAM numero 3 : P(Y=SPAM | X=x) = -270.37221600126463, P(Y=HAM | X=x) =
-294.2756843917197  => identifie comme un SPAM
[…]

Resultats :
Erreur de test sur les 20 SPAM       :  5 %
Erreur de test sur les 20 HAM        :  10 %
Erreur de test globale sur 40 mails  :  7 %
```

Exemple complet dans le fichier test_execution.txt

Test sur la totalité de la base

```
Nombre de spam de la base d'apprentissage ? (max 500)500
```

```
Nombre de ham de la base d'apprentissage ? (max 2500)2500
Nombre de spam a tester ? (max 500)500
Nombre de ham a tester ? (max 500)500

Apprentissage realise sur 500 SPAMs et 2500 HAMs
[…]
Resultats :
Erreur de test sur les 500 SPAM       :  26 %
Erreur de test sur les 500 HAM        :  1 %
Erreur de test globale sur 1000 mails  :  13 %
```