TP info : Filtre de spam - Durée : 2 séances de 1h

Date

1 Objectif

Objectif L'objectif de ce TP est de créer un programme qui lorsqu'on lui donne un mail (sous forme fichier texte .txt), est capable de dire si c'est un spam ou non. Pour un être humain, il est en général assez facile de distinguer un spam d'un mail normal. Mais ce n'est pas le cas pour un ordinateur, et il faudra trouver des critères fiables pour distinguer les deux. Malgré cela, le programme peut se tromper, et il faudra tenir compte du facteur erreur.

Question 1.1. *Le programme ne sera pas parfait, et il pourra se tromper. Que vaut-il mieux :*

- *Dire qu'un mail est un spam alors que ça en est pas un*
- *Dire qu'un mail n'est pas un spam alors que ça en est un.*

On peut résumer les choix dans un tableau :

vérité/prévision	Oui (Spam)	Non-spam
Oui	OK	Erreur
Non	Erreur	OK

2 Méthode

Première étape Nous allons faire une analyse des mots utilisés dans les spams et dans les e-mails non spams. L'objectif est de montrer que certains mots reviennent beaucoup plus souvent dans les spams (comme le mot "gagner", ou "argent", "félicitations", etc).
`http://vadeker.net/reponses/arnaques/spammers_scammers_arnaques.html`
`http://www.exemplede.fr/exemple-de-spam/`

Deuxième étape En ce basant sur les données statistiques de la fréquence d'apparition des mots, on regarde un mail : si les mots "félicitations", "argent", etc apparaissent beaucoup, alors on conclura que c'est un spam.

3 Etude de la fréquence d'apparition des mots dans un spam et dans un mail classique

Première étape On récupère un fichier texte contenant une centaine de spams (*exemple_spam.txt*) et un autre fichier contentant une centaine de mails classique (*exemple_mails.txt*).

On donne un programme (fait par moi-même).
En entrée : un fichier .txt (contenant le/les mail(s) à analyser)
En sortie : Un tableau. Première colonne les mots apparaissant dans les mails; deuxième colonne le nombre d'occurrence du mot

Analyse des spams

Question 3.1. *Lancer le programme donné, en mettant le fichier des mails spams en entrée. Mettre les résultats dans un tableur.*

Question 3.2. *Calculer la fréquence d'apparition de chaque mot.*

Analyse des mails non-spams

Question 3.3. *Recommencer les questions de la partie précédente avec les mails non-spams.*

Comparaison spams et non-spams

Question 3.4. *Comparer les mots les plus fréquents dans les spams et les non-spams.*

4 Filtre spam

On arrive maintenant à l'étape cruciale de l'activité. On donne un mail. Je veux un programme qui me dise : "oui c'est un spam" ou "non ce n'est pas un spam", en se basant sur les mots présent dans ce mail.

Le mail se trouve dans *mail.txt*.

Question 4.1. *Analyser la fréquence d'apparition des mots dans ce mail.*

Question 4.2. *A vu d'oeuil, cela ressemble-t-il à un spam ?*

Il faut donc faire comprendre à l'ordinateur que les résultats de fréquence des mots ressemblent à ceux d'un spam.