



Rapport du projet anti spam

Réalisé par:

Tarik Boulaajoul
Sara TOUZANI

Encadré par:

Mme Houda BENBRAHIM
Mme. Lamia BENHIBA
Mme Hanane El Bakkali
Mr Ismail Kassou

Introduction

L'anti-spam, est un ensemble de systèmes et moyens techniques et juridiques de lutte contre le courriers électroniques publicitaires non sollicités. Vers les années 2000, ce genre semblait être inoffensif vue que la plupart des spameur l'utilisaient afin de valoriser leurs produits. Or, avec le volume important de courriel surtout non désiré ,il est devenu nécessaire de se prémunir contre cette nuisance.

Les solutions de lutte anti-spam mettent sensiblement des techniques de plus en plus innovantes pour distinguer le pourriel du courrier légitime.ou les 'Hams' Ces techniques peuvent être mises en œuvre soit au niveau des fournisseurs de service Internet qui protègent leur messagerie, soit du côté de l'utilisateur ,du serveur (webmails, FAI et entreprises) ou des filtres du côté du routeur(Limitation de la vitesse d'envoi pour les comptes récents,Vérification des taux de plainte reçues depuis les feedback loops, etc)

. Plusieurs techniques de lutte contre les spams sont possibles et peuvent être combinée: basée sur des notions statistiques (méthode bayésienne), filtrage par contacts, *Black listes* (désignation de personnes, domaines ou de machines auxquelles il est interdit de publier dans certains lieux)

Problematique

Chaque jours, nous recevons des mails de divers expéditeurs.il y'a des emails qui sont importants, d'autres moins importants, d'autres sont moins importants, et certains sont pires des Spams ; c'est pourquoi il faut filtrer ces messages inutiles qui nous ont submergés.

Le probleme c'est qu'un facteur important mesurant la qualité d'un filtre est le taux d'erreurs ainnsi que le F_score, il vaut mieu que le filtre laisse passer quelques spams que de rejeter des emails importantsqui pourrait représenter une opportunité d'embauche ou un cour qu'on attendu sa promotion.

Solution proposée

La solution mise en place dans le cadre de notre travail s'appuie sur deux techniques de filtrage qui sont le filtrage par contacts selon le choix de l'utilisateur de l'application et le filtrage bayésien.

Filtrage par Contacts

Cette méthode est appliquée selon le choix de l'utilisateur et se base sur le rejet du courrier si l'email de l'expéditeur n'apparaît pas sur la liste que l'utilisateur va charger de sa base de contacts Gmail ou n'importe quel opérateur

Filtrage bayésien

Le filtrage bayésien du spam (du mathématicien Thomas Bayes) est un système basé sur une théorie de probabilités basée sur grande quantité de "spam" et de "ham" afin de déterminer si un courriel est à rejeter ou non.

Le filtre calcule des probabilités conditionnelles en fonction des informations observées (situées dans une base de données); chaque mot d'un message est évalué suivant les probabilités qu'il se trouve dans un message défini comme indésirable ou non.

Ce filtre est donc basé sur le théorème de Bayes, dont la formule mathématique est:

$$\Pr(\text{pourriel}|\text{mots}) = \frac{\Pr(\text{mots}|\text{pourriel}) \Pr(\text{pourriel})}{\Pr(\text{mots})}$$

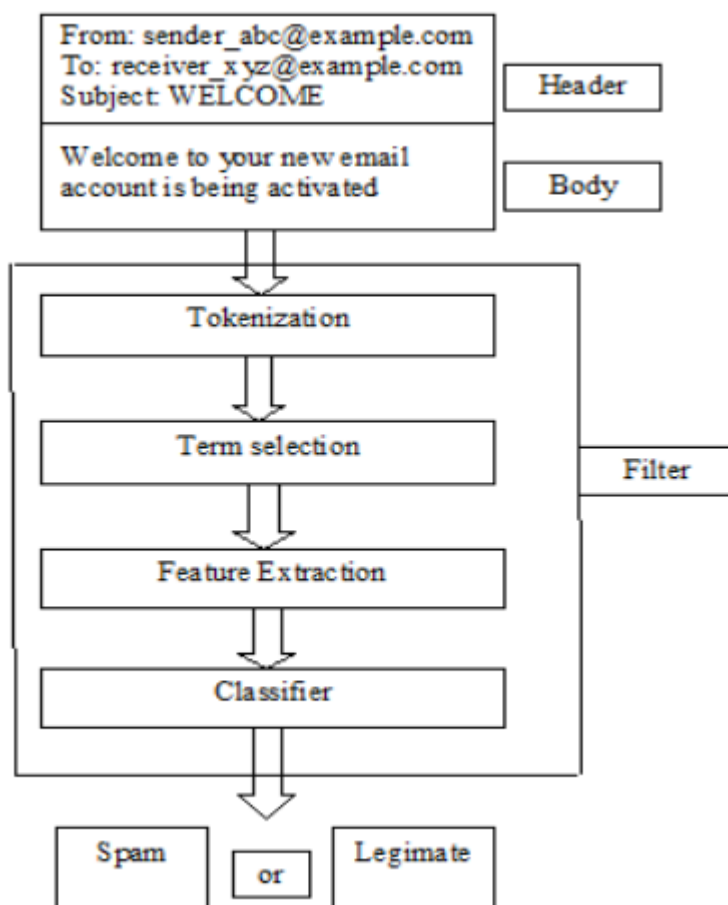
La probabilité qu'un courrier soit un pourriel, compte tenu qu'il contienne certains mots, est égale à la probabilité de trouver ces mots dans un pourriel multipliée par la probabilité qu'un courrier soit un pourriel, divisé par la probabilité de trouver ces mots dans un courrier.

[1]

[1] http://monge.univ-mlv.fr/~dr/XPOSE2007/fthomann_sujet/filtrage_stat.html

Afin de garantir une bonne qualité de notre filtre, le corpus de spam et de ham doit être riche en terme de data. Le filtre Bayésien est fondé sur le principe que la plupart des événements sont dépendants et que la probabilité qu'un événement se répète dans le futur peut être déduite des précédents de ce même événement.

Le message à identifier est découpé en morceaux, puis traité suivant une fonction qu'on a écrite preprocess() une pour la langue anglaise et une pour française suivant la langue de notre mail de test utilisant des word_tokenizer pour une séparation des mots plus efficace , enlevant les stop words, ainsi que des stemmer qui nous remplace chaque mot par son radical en minuscule de cette façon que 'PYthoning', 'PYTHON', 'pythoner' deviennent tous 'python' tant que ceci revient au même en terme de sens de mot. Enfin en remplaçant tout les caractères spéciales par '#' (Ici je parle pas de signe de ponctuation, car les les linguistique considèrent les signes de *ponctuation* comme de véritables signes *linguistiques* où même des mots car elles finalisent le sens de la phrase). Ce processus est monté dans le but d'améliorer le score de notre algorithmes qui évalue la qualité de notre algorithme . Ses données elles sont comparés à tout le corpus de courriels (*Training data*) qui a subit aussi le même processus de preprocess() , pour déterminer la fréquence des différents morceaux dans les deux catégories. Une formule statistique est utilisée afin de calculer la probabilité que le message soit un pourriel ou non. Lorsque la probabilité est suffisamment élevée, le système bayésien catégorise le message comme un pourriel. Sinon, il le laisse passer.



(1)Tokenization: la separation du text en des phrases ou des mots selon le besoin

(2)Term Selection: suivant laquelle nous ne gardions que les termes utiles à notre algorithmes; les stop words par exemple sont utilisé presque de la même fréquence dans les spam et les ham

(3)Feature Extraction: en réduisant notre dataset en se restituant à une base unitaire nétoyer

Pourquoi nous avons choisis le filtre baisien ?

Après avoir fait un benchmarking sur les filtres déjà existant sur internet soit sur des bibliothèques tel “*scikit learn*” et ceux qui ont été déjà développer lors des travaux de recherches implémentant des solutions anti spam dont la plupart deux choisissaient le filtre Baisien, nous avons décider de réaliser une étude et appliquer sur notre data enron [2] les algorithmes qui ont été les plus implémentés et comparer entre leurs performances suivant des critères bien précis BernoulliNB [3], MultinomialNB [6], Svm [4] et [5] et K-Means [7].

[2] <https://www.cs.cmu.edu/~enron/> [CALO Project]

[3] <https://www.kaggle.com/clydewang/a-naive-bayes-way-for-spam-classification> [Clyde Wang]
[Feb, 17, 2017]

[4] https://github.com/nshelly/NV_SVM_SpamFilter [Nicholas Shelly]

[5]http://www.bogotobogo.com/python/scikit-learn/scikit_learn_Support_Vector_Machines_SVM_spam_filtermachine_learning_.php [K Hong]

[6] <https://github.com/abhijeet3922/Mail-Spam-Filtering>

[7] [Spam Filtering using K mean Clustering with Local Feature Selection Classifier] [Anand Sharma & Vedant Rastogi]

Modèle Bayésien :

Il y’a deux façons différentes de représenter les caractéristiques dans un classificateur bayésien naïf.Ces deux méthodes vont donner des résultats différents en ce qui concerne la classification.Pour montrer la différence de ces deux modèles, nous allons tester dessus un message que nous pourrions recevoir comme ham “christmas tree farm pictures\n”

Selon McCallum et Nigam [8] qui ont réalisé des expériences sur 5 corpus différents, le modèle de Bernoulli multivarié donne de bons résultats avec des


vocabulaires de petites dimensions, mais le modèle multinomial donne de meilleurs résultats avec des vocabulaires de dimensions plus grandes.

[8][A Comparison of Event Models for Naive Bayes Text Classification] [Andrew MCCALLUM et Kamal NIGAM]

Modèle de Bernoulli multivarié [9]

Dans le modèle de Bernoulli multivarié, le document d est un vecteur binaire sur l'espace des mots. Soit un vocabulaire V , chaque attribut X_i est égal à 0 ou 1 selon que le token U issu de ce dictionnaire est présent ou non dans le courriel.

[9] [Une nouvelle approche pour la détection des spams se basant sur un tritement de données catégorielles] [Yassine Z. Parakh Ousman]

$$\vec{x} = \langle x_1, x_2, x_3, \dots, x_m \rangle = \langle 0, 1, 1, \dots, 0 \rangle$$


money rich ! unsubscribe

Le vecteur x dans un modèle de Bernoulli multivarié


Avec une telle représentation, nous faisons l'hypothèse de Bayes naïf que la probabilité d'apparition de chaque mot dans un document est indépendante de l'occurrence des autres mots. Ainsi, la probabilité d'un mot étant donné sa classe c_i est le produit des probabilités de chaque mot de ce document selon la classe donnée.

$$p(d|c_i) = \prod_{j=1}^m p(x_j|c_i)$$

On peut donc voir un document comme étant une suite de plusieurs expériences de Bernoulli, une pour chaque mot du vocabulaire V . On peut remarquer également que ce modèle ne considère pas la fréquence d'apparition des mots dans un document.

Modèle multinomial

Dans le modèle multinomial, au contraire, on considère la fréquence d'apparition des mots dans un document. Ainsi, chaque attribut x_i du vecteur x représente le nombre d'occurrences dans le message de chaque token t_i du vocabulaire V . Cette fois notre vecteur prend la forme ci-dessus:

$$\vec{x} = \langle x_1, x_2, x_3, \dots, x_m \rangle = \langle 0, 1, 3, \dots, 0 \rangle$$


The diagram shows four arrows pointing from words below to specific components of the vector $\langle 0, 1, 3, \dots, 0 \rangle$. The first arrow points from 'money' to the first '0'. The second arrow points from 'rich' to the '1'. The third arrow points from '!' to the '3'. The fourth arrow points from 'unsubscribe' to the final '0'.

Le vecteur x dans un modèle multinomial

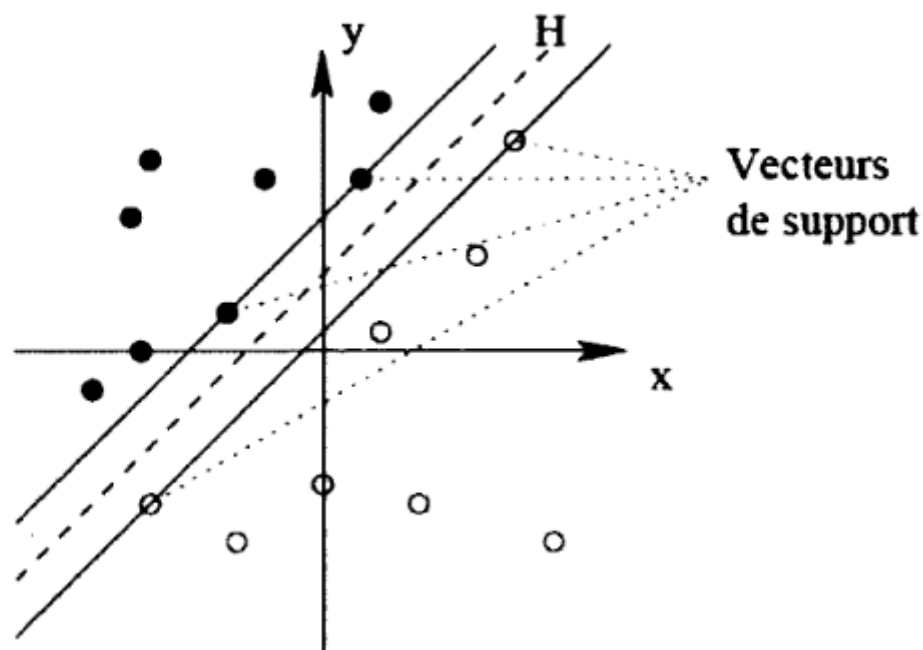
Dans ce modèle, un document est une séquence ordonnée de mots construits à partir du vocabulaire V . On suppose que la longueur du document ne dépend pas de sa classe et on fait l'hypothèse de Bayes naïf que la probabilité d'apparition d'un mot dans un document est indépendante de sa position dans le document et des autres mots du document. On peut donc calculer la probabilité d'un mot selon la classe, avec la formule suivante où $N_j = \sum x_j$ le nombre de mots du courriel:

$$p(d|c_i) = p(|d|)d! \prod_{j=1}^m \frac{p(x_j|c_i)^{N_j}}{N_j!}$$

Modèle SVM (Machines à Vecteurs de Support ou Séparateur à Vaste Marge):

Un séparateur à vaste marge (SVM) est une méthode de classification binaire par apprentissage supervisé, elle a été introduite par Vapnik en 1995. Cette technique se base sur l'existence d'un classificateur linéaire dans un espace approprié. Étant donné un problème de classification à deux classes, elle fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonctions dites noyaux (kernel) qui permettent une séparation optimale des données. Le but des SVM est de trouver un hyperplan (H) qui va séparer deux ensembles de points. De plus, les deux points les plus proches de cet hyperplan sont appelés vecteurs de support.

Il existe plusieurs hyperplans valides mais grâce aux SVM on obtient l'hyperplan optimal. Formellement, cela revient à chercher un hyperplan dont la distance minimale à l'ensemble d'apprentissage est maximale. On appelle cette distance marge entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise cette marge, d'où l'appellation séparateurs à vaste marge. En effet, le fait d'avoir une marge plus grande procure plus de sécurité lorsque l'on classe un nouvel élément. dans la figure ci-dessus la partie droite illustre le fait qu'avec un hyperplan optimal, un nouvel élément reste bien classé alors qu'il tombe dans la marge et au contraire dans la partie de gauche on constate qu'avec une plus petite marge, le nouvel élément est mal classé à travers l'augmentation du paramètre C .



Hyperplan et vecteurs de support

Modèle k-means où $k=2$:

Le clustering Premièrement est en mathématique, une méthode qui a comme objectif à partir un ensemble de données non étiqueté contrairement à la classification chercher une typologie ou partition des individus en cluster ou catégorie.

K-Means (MacQueen, 1967) est l'algorithme de clustering le plus populaire et le plus utilisé de l'apprentissage non supervisé. Cet algorithme a pour but de répartir en k clusters un ensemble de données de telle sorte à avoir dans chaque cluster les données qui ont des caractéristiques de similarité forte, et que les ensembles entre eux doivent être différents les uns des autres.[10]

[10] [mémoire fin d'étude - DSpace à Université abou Bekr Belkaid Tlemcen][M KOUDRI]

k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïde. Le choix initial des centroïdes conditionne le résultat final. Lorsque on obtient un nuage de point l'étape qui suit est de construire des catégories où le paramètre k détermine le nombre de catégories qu'on désire avoir. Notre algorithme change les points de chaque cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés, mais sous réserve de choisir la bonne valeur de k . Cet algorithme permet ainsi d'afficher les mots apparaissant le plus dans chacun des clusters.

Afin de pouvoir comparer la performance de ces algorithmes il existe des indices qui nous permettent ceci en répartissant notre data en train et en test suivant deux méthodes :

18

Validation empirique d'un apprentissage:

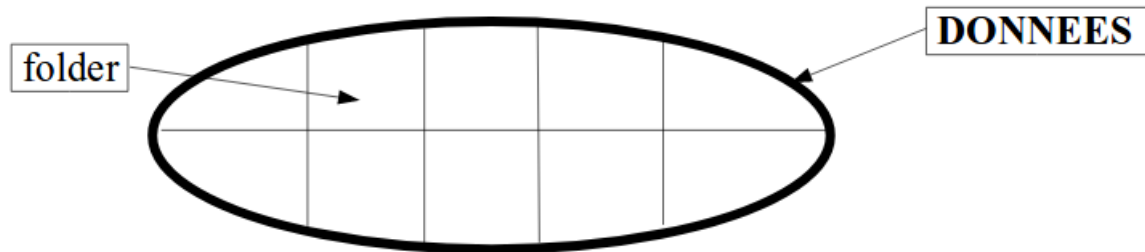
[11] Plusieurs méthodes permettent de valider (ou d'infirmer) la valeur d'un processus d'apprentissage. Une des approches consiste à n'utiliser qu'une part des données pour apprendre et à se servir des autres données pour tester le résultat.

Différentes mesures permettent alors de comparer des processus (erreur, F-score, etc.)



Validation croisée:

La cross-validation est une généralisation de la méthode précédente Elle consiste à diviser les données en k folders , à en enlever un pour l'apprentissage puis à l'utiliser pour la phase de test Le processus est ensuite réitéré. L'erreur moyenne tend alors vers l'erreur en généralisation.



Matrice de confusion

notre matrice de confusion est présentée comme suit :

| | Classé + | Classé - |
|-----------|----------|----------|
| Exemple + | V_p | F_p |
| Exemple - | F_n | V_n |

L'erreur

L'erreur (ou taux d'erreur) ne fait pas de distinction entre les erreurs : pas forcément une bonne mesure de qualité d'un apprentissage.

$$\text{Erreur : } accuracy = \frac{F_p + F_n}{F_p + F_n + V_p + V_n}$$

F_Score

Le F-score rend compte de la qualité d'une classification en fonction des classes, mais ne tient pas compte de l'éventuel déséquilibre entre les classes ,mais ne tient pas en compte les faux négatifs, calculé à travers la précision et le rappel

$$\text{F-score} = 2 \frac{\text{précision} * \text{rappel}}{\text{précision} + \text{rappel}}$$

Précision

$$\text{Précision} = \frac{V_p}{V_p + F_p}$$

Rappel

$$\text{Rappel} = \frac{V_p}{V_p + F_n}$$

[11] [Classification,Apprentissage,Décision] [Rémi Eyraud]

Comparaison entre les algorithmes de classification importé de la bibliothèque scikit

Total des emails classifiés : 5172

| | MultinomialNB | Bernoulli | SVM avec C=1 | SVM avec C=10 | SVM avec C=100 |
|--------|---------------|--------------------|----------------|--------------------------------|----------------|
| Vp | 3628 | 3627 | 3665 | 3665 | 3662 |
| Vn | 1428 | 1425 | 119 | 119 | 122 |
| Fp | 44 | 45 | 7 | 7 | 10 |
| Fn | 72 | 75 | 1381 | 1381 | 1378 |
| erreur | 0.02242846094 | 0.02320185615 | 0.2683681361 | 0,2683681361 | 0,2683681361 |
| Score | 0.96115983853 | 0.95969133692 3 | 0.145314516992 | 0.145314516992 145314516992 | 0.14916197263 |

Mise en oeuvre

Pour développer l'application, nous avons utilisé l'environnement Jupyter notebook, l'interface graphique a été réalisé avec la librairie

Les outils utilisés:

Python:

Python est un langage de programmation objet, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ;

C'est un langage où la syntaxe, clairement séparée des mécanismes de bas niveau, permet une initiation aisée aux concepts de base de la programmation[12]

[12] [https://fr.wikipedia.org/wiki/Python_\(langage\)](https://fr.wikipedia.org/wiki/Python_(langage))

les bibliothèque principale implémentées :

scikit-learn:

Scikit-learn est une bibliothèque libre Python dédiée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs³ notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria⁴ et Télécom ParisTech. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec des autres bibliothèques libre Python, notamment NumPy et SciPy.[13]

[13] <https://fr.wikipedia.org/wiki/Scikit-learn>

nlTK:

Natural Language Toolkit (NLTK) est une bibliothèque logicielle en Python permettant un traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'Université de Pennsylvanie. En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation (API).[14]

[14] https://fr.wikipedia.org/wiki/Natural_Language_Toolkit

Tkinter:

Tkinter est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk écrite pour Tcl.

[15]

[15] <https://fr.wikipedia.org/wiki/Tkinter>

Réalisation

Il est porté à la connaissance de l'utilisateur que pour gmail cette opération d'obtention de liste de contacts ce fait d'une manière très fluide suivant les étapes suivantes[16] :

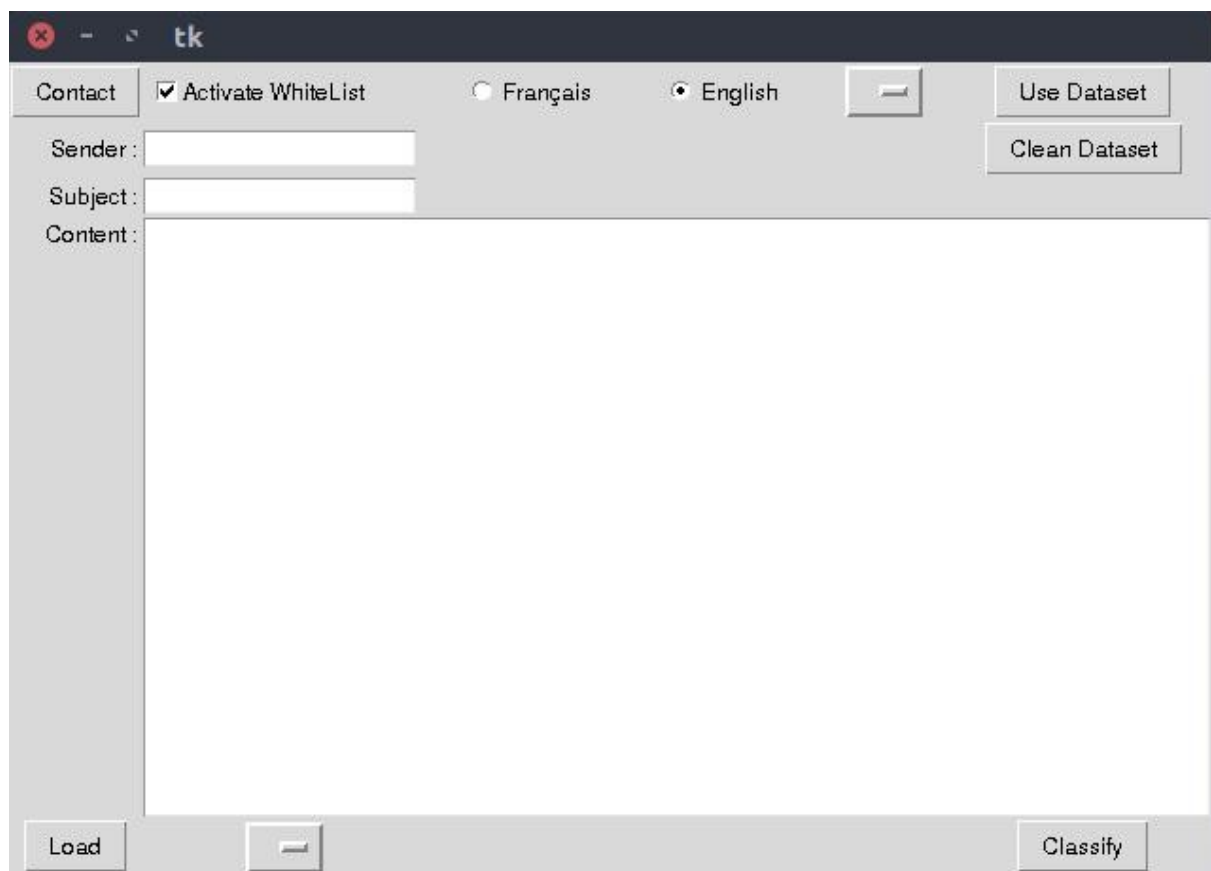
[16]<https://support.office.com/fr-fr/article/importer-des-contacts-gmail-dans-outlook-edbacfde-f48c-49da-a6a3-bcbb8f4f4819>

À partir de votre compte Gmail, choisissez Gmail >Contacts.

1. Sélectionnez Plus >Exporter.
2. Choisissez le groupe de contacts à exporter.
3. Choisissez le format exportation Format CSV Outlook (importation dans Outlook ou une autre application).
4. Choisissez Exporter.

Une fois le téléchargement a eu fin on prend notre fichier « contacts.csv » et on le place dans le répertoire de notre application

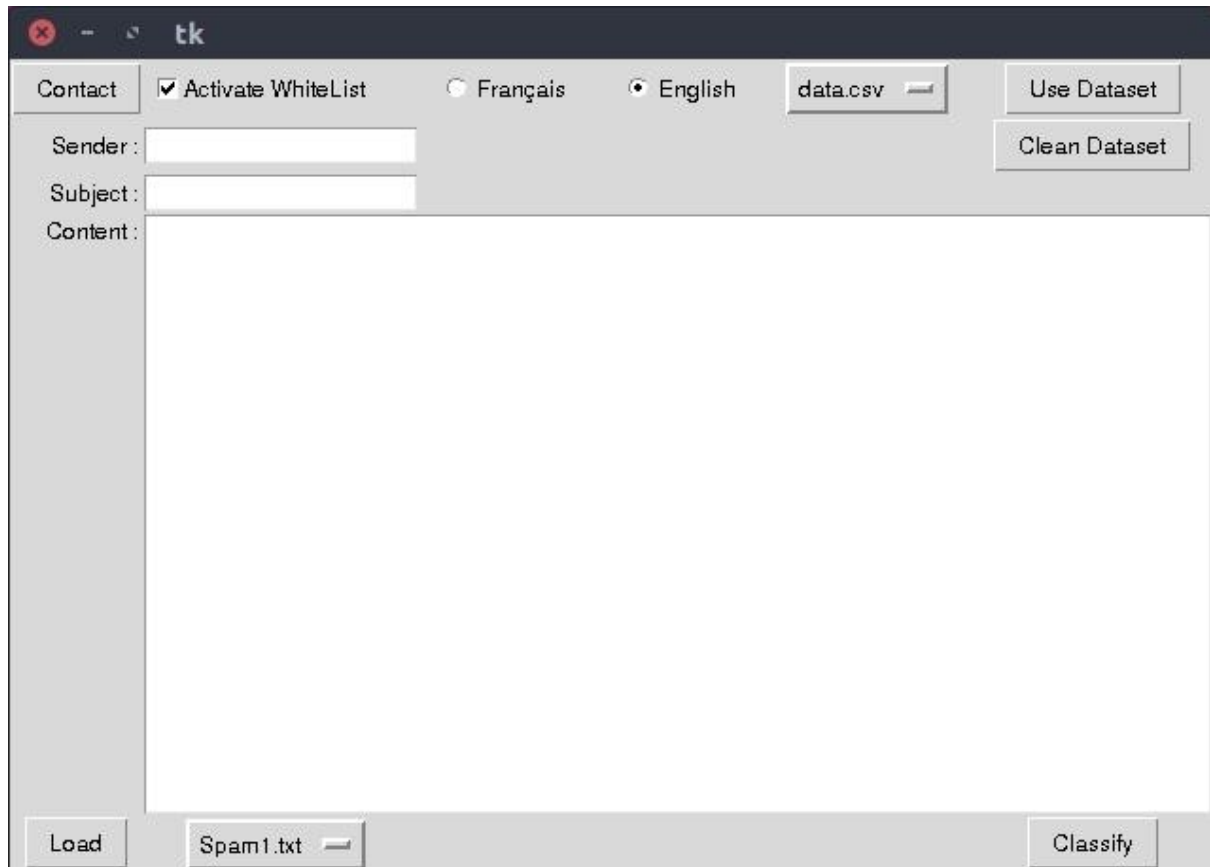
à chaque fois qu'il charge de nouveaux contacts ils sont téléchargé sur le répertoire Téléchargement il le prend et on le met dans notre répertoire cité avant et on le renome "contacts.csv" si il portait un autre nom par hasard.



La premiere fenetre de l'application nous affiche les champs necessaires pour la prediction,

le champ "Sender" qui nous renseigne sur l'expediteur de l'email, "subject" pour le sujet et "Content" pour le corps de l'email

le bouton "contacts" permet d'extraire une liste de contacts d'un fichier csv nommé "contacts.csv" (norme de google) le fichier en question contient plusieurs mais seul la liste d'adresses nous interesse, la fonction "importContacts" permet de garder la liste qui nous interesse en négligeant le reste



l'application dispose de deux boutons radio qui permettent de choisir la langue de traitement, ainsi le résultat de la prédiction dépend de la langue choisie, mais pour une prédiction optimale en français on doit insérer une base de données en français avec la liste déroulante suivante, et en cliquant sur "Use Dataset", ce qui permet de générer un modèle qui dépend de la langue et de la BD choisie.

tk

Contact ☒ Activate WhiteList ☐ Français ☒ English generated.csv Use Dataset

Sender: marie.dupuis@noos.fr Clean Dataset

Subject: JOKER KENO LOTERIE

Content: JOKER KENO LOTERIE
 \r
 Bonjour Madame/Monsieur, \r
 La JOKER KENO LOTERIE le plaisir de vous informer du tirage au sort du programm
 e de la loterie Anglaise qui s'est tenu Londres. \r
 Votre adresse lectronique attache au numro identifiant vous a permis degagner l
 a 1me place la somme forfaitaire hors taxe de 65.000. \r
 \r
 Veuillez trouver en annexe votre Notification de gain. \r
 Pour la rclamation de ce lot acquis par la JOKER KENO Loterie, nous vous prions
 de bien vouloir entrer en contact avec lhuissier accrdit afin quil puisse vous d
 ictez la procure suivre jusqu' la remise de votre gain. \r
 \r
 \r
 PROCEDURE DE REMISE DE GAIN \r
 \r
 Veuillez envoyer par mail les informations vous concernant sous 48 heures \r
 a Ladresse lectronique de Mtre Claude Mentenon \r
 charg de vous indiquer les Conditions gcales de remise de votre gain. \r
 \r
 Email : me-claudementenon@hotmail.fr \r
 CEL : (+229) 480 72 105 \r
 FAX : (+229) 208 75 405 \r
 \r

Load Spam1.txt Classify

le Bouton "Clean Dataset" permet de produire une bd copie de la bd choisi, mais qui subit en plus un traitement sur texte, avec la fonction "preprocess" permettant d'optimiser le taux de réussite, la BD nouvellement créer est ajouter a la liste de BD disponibles et peut etre utilisé immédiatement (Le processus de création prend entre 10-15min) le fichier csv correspondant est généré dans l'emplacement de l'application et nommé "generated.csv"

le Bouton load avec la liste de fichier text permettent de prendre les données(sender, subject, content) d'un email contenu dans un fichier text, et disponible dans l'emplacement de l'application et de les afficher dans notre fenetre pour une application immédiate.

tk

Contact

☐ Activate WhiteList
 ☐ Français
 ☒ English
 generated.csv
 Use Dataset

Sender:

alduc@numericable.fr

Result =

HAM

Clean Dataset

Subject:

ACTE DE DONATION

Content:

Bonsoir Monsieur/Madame,

Je n'ai pas un autre moyen de communiquer avec vous que de vous faire parvenir ce mail. Je suis dans le dsespoir et mon [UTF-8?]cœur saigne au moment ou je vous cris ce message qui j'espere retiendra votre attention. \r

Je vous contacte aujourd'hui car bien vrai que l'on ne se connaisse pas mais cel a n'empche ce geste de ma part. Je me nomme Monsieur ALBERI DUCROZ n le 03 Octob re 1953 en Normandie en France, mais pour une raison particulire, j'ai du tre un aventurier la recherche de je ne sais quoi. La raison qui me pousse vous est la suivante: Je voudrais passer par votre canal pour faire une ouvre de charit d ans votre dpartement. \r

C'est une donation en quelque sorte et elle s'lve la somme de sept cent cinquante milles euros [UTF-8?](750.000.00€). \r

Ma situation matrimoniale est telle que je n'ai ni femme et encore moins d'enfan ts qui je pourrais lguer cet hritage, et je souffre prsentement d'un cancer bro ncho-pulmonaire, je suis donc condamn une mort certaine. \r

C'est pour cela que, je voudrais de manire gracieuse et dans le souci d'aider le s enfants dmunis vous donner ce dit hritage pour raliser cette ouvre de charit. \r

Si vous tes d'accord faire ce que je vous ai demand faite le moi savoir. \r

Je vous prie d'accorder une oreille attentive ma proposition car je compte sur votre bonne volont et aussi le bon usage de ces fonds pour cette ouvce. \r

Fraternellement, \r

Mr ALBERI DUCROZ \r

Load

Spam3.txt

Classify

Le bouton "Classify" permet de faire le calcul avec la méthode bayesienne, sur un email tapé manuellement ou importé avec l'outil "loadEmail" et afficher en haut au milieu le résultat de la prédiction

ps: avant l'opération de classification l'utilisateur doit avoir au préalable déjà choisis une BD et importé un fichier text

ps2: un email n'est pas classifié si la case d'activation de la whitelist est coché et que l'expéditeur n'appartient pas a la liste de contacts7

tk

Contact

☒ Activate WhiteList
 ☐ Français
 ☒ English
 generated.csv
 Use Dataset

Sender: marie.dupuis@noos.fr
 WhiteList Filtered
 Clean Dataset

Subject: JOKER KENO LOTERIE\r
 Content: JOKER KENO LOTERIE\r
 \r
 Bonjour Madame/Monsieur, \r
 La JOKER KENO LOTERIE le plaisir de vous informer du tirage au sort du programme de la loterie Anglaise qui s'est tenu Londres. \r
 Votre adresse électronique attache au numéro identifiant vous a permis de gagner la 1^{ère} place la somme forfaitaire hors taxe de 65.000. \r
 \r
 Veuillez trouver en annexe votre Notification de gain. \r
 Pour la réclamation de ce lot acquis par la JOKER KENO Loterie, nous vous prions de bien vouloir entrer en contact avec l'huissier accrédité afin qu'il puisse vous indiquer la procédure à suivre jusqu'à la remise de votre gain. \r
 \r
 \r
 PROCEDURE DE REMISE DE GAIN \r
 \r
 Veuillez envoyer par mail les informations vous concernant sous 48 heures \r
 à l'adresse électronique de M^{re} Claude Mentenon \r
 chargée de vous indiquer les Conditions générales de remise de votre gain. \r
 \r
 Email : me-claudementenon@hotmail.fr \r
 CEL : (+229) 480 72 105 \r
 FAX : (+229) 208 75 405 \r
 \r

Load

Spam1.txt

Classify

En cochant la case "Activate whitelist" on ajoute un traitement préliminaire, qui compare l'adresse de l'expéditeur à notre liste de contacts et qui filtre les emails qui parviennent de personnes étrangères à notre liste de contact

Conclusion

Le projet « Réalisation d'un anti spam » était une concrétisation des acquis durant les cours du module

de l'Intelligence artificielle, et la mise en pratique avec Python de ces algorithmes pour un approfondissement de nos connaissances en ce langage et une première expérience avec la pratique des algorithmes des Réseau bayésiens dans ses deux phases Multinomiale et de bernouli vu en cours , ainsi que d'autres tel svm et k-mean dans le but de choisir celui le plus adéquat .

A travers notre, nous avons élaboré une application de détection de Spam , où nous avons mit deux passage le filtre par contacts dont l'utilisateur peut choisir ne pas utiliser et celui Baisien après avoir fait passer notre data et notre message à traiter par un processus basé sur *Natural language Processing* pour n'avoir que des données unitaire non redondante pas en terme de fréquence car le filtre implémenté Multinomiale est bien basé sur cette dernière , mais en éliminant des mots unitules , et on normalisant notre data

Puisque le principe est acquis, notre application peut être élargie pour couvrir toutes les langues en insérant d'autres dataset de même langue , où même de combiner des algorithmes tel celui bayésien et Svm pour améliorer de plus en plus les performances.pour utiliser ces mail par la suite de nouveau pour l'apprentissage de notre algorithme.