Etude du jeu de données « Spambase »

Garance Chamalet

Présentation du jeu de données

Présentation du jeu de données

Informations générales

Le jeu de données « **Spambase** » présente des informations collectées sur **4601 mails**, ces derniers étant classés en deux catégories: **spam** ou **mail normal** (mail personnel, professionnel, etc).

Pour rappel, les **spams** (ou pourriel), est un « Envoi répété d'un message électronique, **souvent publicitaire**, à un grand nombre d'internautes **sans leur consentement** »*. Les systèmes de messagerie cherchent donc depuis des années à améliorer leurs filtres de spams, qui sont une **nuisance** pour les utilisateurs.

Ce jeu de données a été créé dans le but de réaliser un **filtre anti-spam personnalisé pour le donneur**, George Forman. En effet, il contient des variables **spécifiques aux mails que recevaient Forman** (la fréquence du mot « George » et celle du code « 650 » sont deux variables du jeu de données). Il n'y a pas assez de données dans ce jeu pour pouvoir créer un filtre général.

Une des méthodes utilisées pour réaliser ces filtres est de **créer une IA prédictive**, à qui l'on donne des mails issus de la vie réelle, classés en tant que mail ou spam.

* *Définition issue du dictionnaire LeRobert*

Présentation du jeu de données

Informations générales

- **Créateurs:** Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt
- **Donneur:** George Forman
- **Date de début de collecte des données:** inconnue
- **Année de soumission:** 1999
- **Source des données:** collecte de mails et spams venant de divers collègues et de l'administrateur du serveur de messagerie de l'entreprise
- **Lien d'accès:** https://archive.ics.uci.edu/ml/datasets/Spambase

Data Set Characteristics:	Multivariate	**Number of Instances:**	4601	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	57	**Date Donated**	1999-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	553561

Tableau issu de la page de présentation du jeu de données

Présentation du jeu de données

Variables

Nom: word_freq_<WORD>
Type de l'attribut: Réel, de 0 à 100
Nombre d'attributs: 48

Description: Pourcentage du mot « WORD » dans l'email

Formule de calcul: 100 * (nombre d'occurrences du mot WORD) / nombre total de mots

Nom: char_freq_<CHAR>
Type de l'attribut: Réel, de 0 à 100
Nombre d'attributs: 6

Description: Pourcentage du caractère « CHAR » dans l'email

Formule de calcul: 100 * (nombre d'occurrences du caractère CHAR) / nombre total de caractères

Nom: is_spam
Type de l'attribut: Booléen
Nombre d'attributs: 1

Description: Pourcentage du mot « WORD » dans l'email

Nom: capital_run_length_average
Type de l'attribut: Réel, de 1 jusqu'à l'infini
Nombre d'attributs: 1

Description: Longueur moyenne des séquences ininterrompues de majuscules

Nom: capital_run_length_longest
Type de l'attribut: Réel, de 1 jusqu'à l'infini
Nombre d'attributs: 1

Description: Longueur de la plus longue séquence ininterrompue de lettres majuscules

Nom: capital_run_length_total
Type de l'attribut: Réel, de 1 jusqu'à l'infini
Nombre d'attributs: 1

Description: Nombre total de lettres majuscules dans l'email

Limites du jeu de données

Premièrement, aucune information sur **la taille totale du mail** n'est accessible: on se retrouve avec des équations à solutions multiples si l'on essaie de la retrouver en prenant les différentes fréquences, de ce type:

$$\begin{cases} Y = 45*X \\ Y = 74*Z \end{cases}$$

Avec Y, le nombre total de mots dans le texte, et X et Z, le nombre d'occurrence de deux mots différents

Or la taille totale joue peut être un **rôle déterminant** pour repérer un spam ou non.

De plus, cela aurait permis de vérifier les lignes ayant des valeurs très élevées pour les fréquences en raison de la longueur d'un mail.

Par exemple, une des lignes classées comme spam à une fréquence de 9,09% pour le mot « email ». Si le mail **n'est constitué que de 10 mots** et qu'on l'inclue à la moyenne de fréquence du mot « email » pour des textes constitués d'une centaine de mots, ça peut faire un sacré écart.

Limites du jeu de données

- Il est dommage de ne pas avoir inclus **les textes des mails** utilisés pour récupérer les informations. Ces derniers peuvent aussi contenir des balises HTML (les couleurs utilisées peuvent être intéressantes à récupérer), ou encore des nombres et d'autres informations qui auraient pu être récupérées.

- Le choix des mots pour les fréquences est un **choix arbitraire** des créateurs, mais qui n'est pas forcément le meilleur. Par exemple, on ne voit aucune mention **de mots sexuels**, pouvant être utilisés dans les spams proposant de la pornographie (c'est à titre d'exemple: peut-être que les spams recensés ne contenaient absolument de pub pour des sites pornographiques). Il aurait peut-être été intéressant **d'inclure plus de mots pour les autres utilisateurs du jeu de données que G.Forman.**

- La **présence d'URL** dans le mail est aussi un critère qui aurait été intéressant. En effet, le jeu de données date de 1999 et le commerce en ligne avait déjà commencé (Amazon a lancé son site en 1995 pour référence). Bien que je ne puisse pas l'assurer, il est probable que les mails collectés contenaient aussi parfois des URL, **redirigeant vers les sites commerçants** et autres.

- La **présence d'une pièce-jointe** ou non est un critère qui aurait aussi pu aider à déterminer la catégorie du mail

- Il aurait aussi été intéressant de mentionner **l'adresse mail de l'expéditeur**, qui sont parfois très particulières pour les spams (en reprenant certains que j'ai reçu: « fmb6832@mabledfab.com » ou encore « info235@chance.candy-sales.com », qui contiennent **tous les deux des nombres**, contrairement à la majorité des adresses classiques)

Objectif du modèle d'IA à créer

Grâce à ce jeu de données, on peut tenter de modéliser une IA pouvant **prédire** si le texte donné en paramètre est **un spam ou non**.

Cependant, il ne faut pas oublier qu'un mail peut être considéré comme un spam par une personne A, **mais pas par une personne B**. C'est pourquoi il est difficile de réaliser un filtre de spam avec une haute précision sans connaître les habitudes de l'utilisateur.

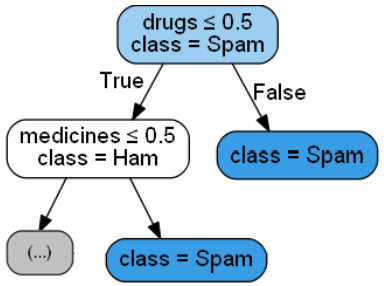
Le but est donc ici de développer un **modèle global de filtre de spam**, sans prendre en compte les centres d'intérêt d'un utilisateur (même si le modèle restera de toute façon **biaisé**, vu que l'on ne connaît pas **la méthode de classification des mails** utilisée par le jeu de données).

Objectif du modèle d'IA à créer

Pour classer un mail comme spam, on peut utiliser **trois approches***:

1. Utiliser un **arbre de règles**
2. Utiliser des **statistiques**
3. Utiliser **conjointement** les deux approches citées au-dessus

Ici, nous allons utiliser **l'approche statistique** en déterminant des coefficients pour chaque caractéristique utilisée.



Arbre de décision utilisé pour détecter les spams

ailments	buy	cheap	come	conference	drugs	follow	free	great	medicines	meeting	pills	price	today
-1.386294	-1.386294	-1.386294	-2.079442	-2.079442	-0.693147	-2.079442	-1.386294	-1.386294	-0.693147	-2.079442	-0.693147	-1.386294	-1.386294

Matrice de coefficient d'un modèle utilisé pour détecter les spams

* Basée sur le présent article

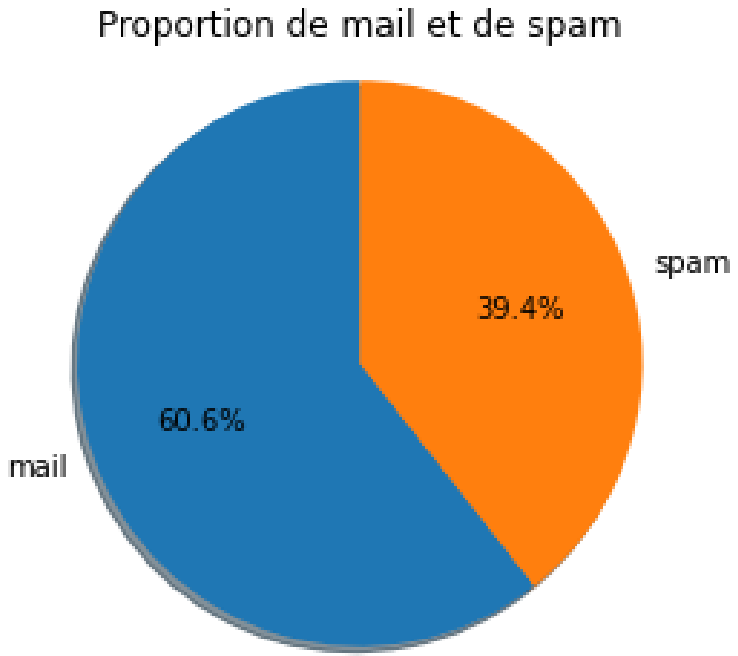
Exploration du jeu de données

Présentation générale

Nous avons un jeu de données avec plus de 60,6% textes classés en tant que mails, et 39,4% en tant que spams.

Sur 4601 valeurs, nous avons 2788 mails et 1813 spams.

```
is_spam
0    2788
1    1813
dtype: int64
```



Présentation générale

Renommage des attributs

Hormis la colonne "is_spam", nous avons **trois différents types** de nom de colonne:

- word_freq_[mot]
- char_freq_[caractère]
- capital_run_length_[type]

On va **raccourcir** ces noms, afin que cela soit plus lisible, ce qui nous donnera:

- wf_[mot]
- cf_[caractère]
- cap_rl_[type]

Présentation générale

Nettoyage des données

Dans le résumé du jeu de données, on peut noter qu'il manquerait apparemment des données.

Data Set Characteristics:	Multivariate	**Number of Instances:**	4601	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	57	**Date Donated**	1999-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	553561

Après avoir analysé le jeu de données, aucune valeur manquante n'a été trouvée*. Nous travaillerons donc avec l'entièreté du jeu par la suite.

** A voir plus en détails sur le notebook.*

Présentation générale

Statistiques générales (mail et spam confondus)

Toutes nos données étant numériques, nous pouvons utiliser la fonction « describe() » pour regarder de plus près quelques statistiques.

On peut voir ici que la majorité des critères ont les quartiles de 25% et 50% valant 0. N'oublions pas qu'ici, les mails et spams sont mélangés.

	count	mean	std	min	25%	50%	75%	max
wf_make	4601.0	0.104553	0.305358	0.0	0.000	0.000	0.000	4.540
wf_address	4601.0	0.213015	1.290575	0.0	0.000	0.000	0.000	14.280
wf_all	4601.0	0.280656	0.504143	0.0	0.000	0.000	0.420	5.100
wf_3d	4601.0	0.065425	1.395151	0.0	0.000	0.000	0.000	42.810
wf_our	4601.0	0.312223	0.672513	0.0	0.000	0.000	0.380	10.000

Présentation générale

Statistiques détaillées par groupe d'un attribut

Si l'on regarde l'attribut « wf_all » de plus près, on peut maintenant voir que le quartile 50% vaut 0.3 pour la classe spam, mais 0.0 pour la classe mail. Pour la suite de cette étude, nous allons étudier les deux groupes séparés et les comparer entre eux.

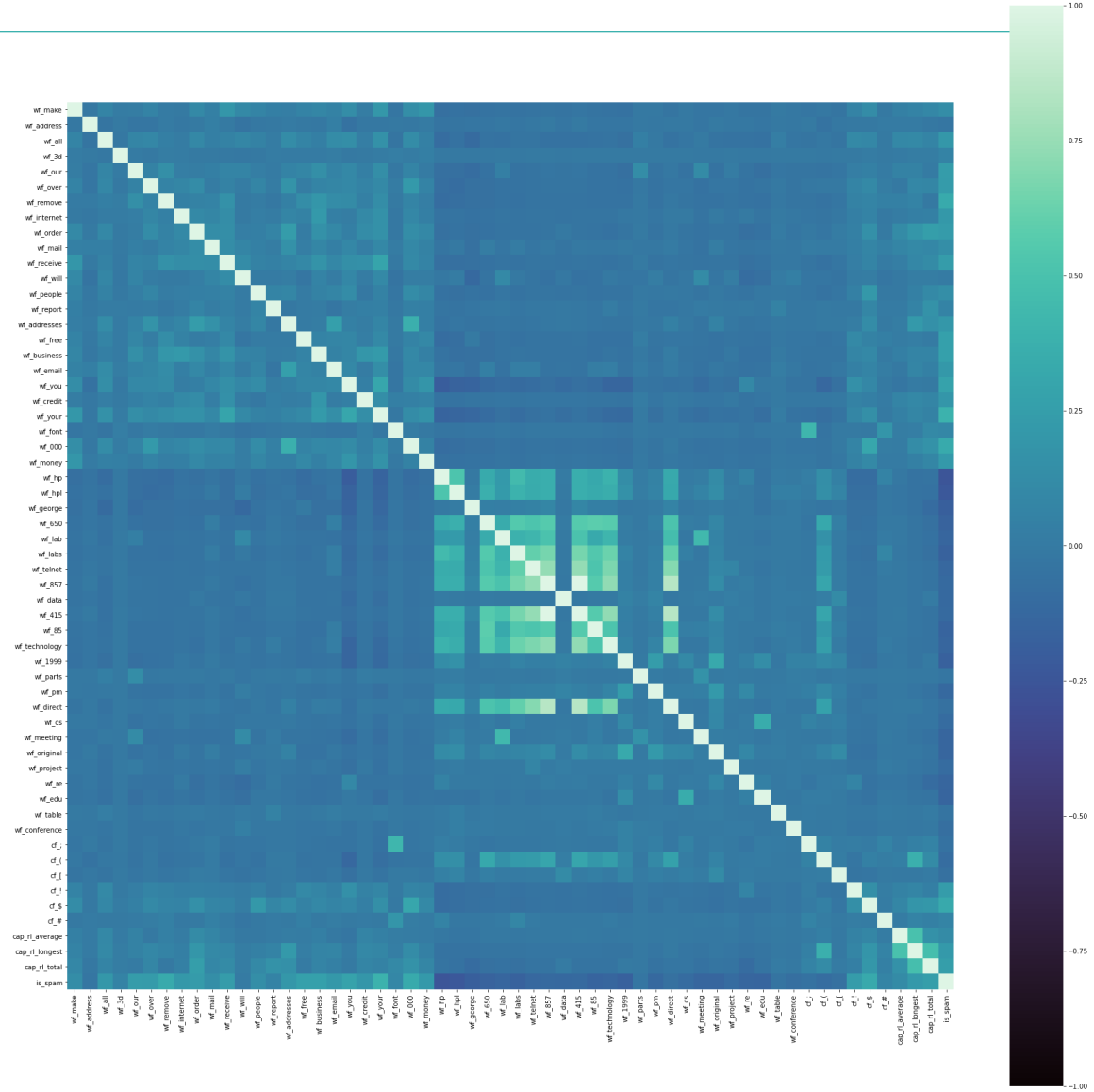
	wf_all							
	count	mean	std	min	25%	50%	75%	max
is_spam_label								
mail	2788.0	0.200581	0.502959	0.0	0.0	0.0	0.12	5.1
spam	1813.0	0.403795	0.480725	0.0	0.0	0.3	0.64	3.7

Présentation générale

Matrice de corrélation

Avant de nous lancer plus loin dans l'étude, ci-joint la matrice de corrélation des différents attributs entre eux.

Je vous recommande de la visualiser sur le notebook pour plus de lisibilité.



Regroupement par catégorie de variables

Comme nous l'avons vu précédemment, nous avons trois types de catégorie de variables:

- Fréquence des mots
- Fréquence des caractères
- Statistiques sur les majuscules

De plus, nous avons aussi la classe de spam et la classe de mail.
Nous allons donc créer chacun de ces groupes, comme ci-dessous.

```
# Récupération de la fréquence des mots
word_frequency_columns = dfData.columns[0:48].values.tolist()
```

```
# Récupération des statistiques sur les majuscules
cap_stats_columns = dfData.columns[54:57].values.tolist()
```

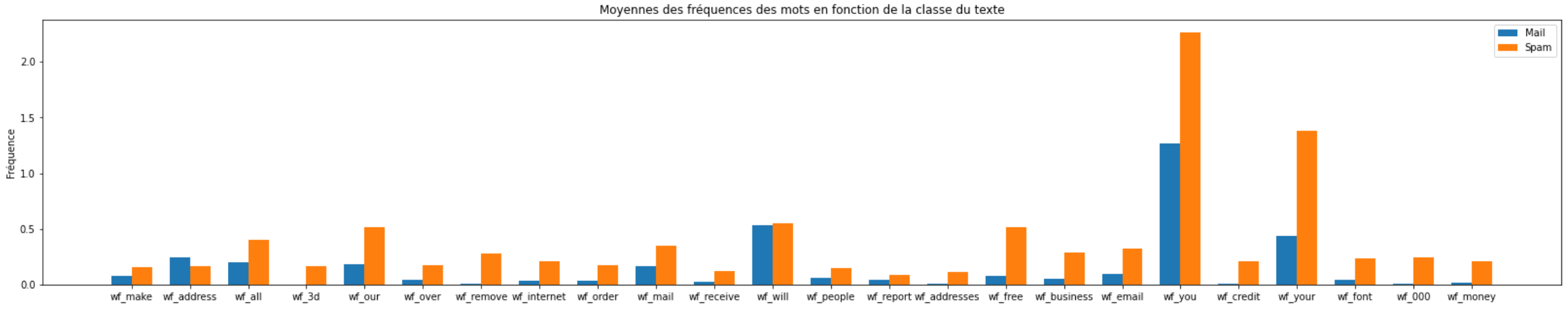
```
# Récupération de la fréquence des caractères
char_frequency_columns = dfData.columns[49:54].values.tolist()
```

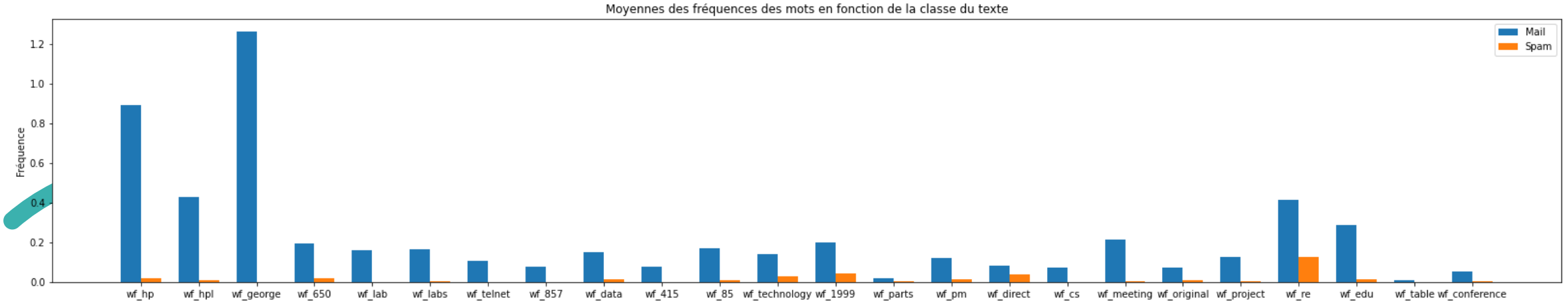
```
# Création des différents groupes étudiés
spams = dfData[dfData['is_spam'] == 1]
mails = dfData[dfData['is_spam'] == 0]
```

Etude de la fréquence des mots

Comparaison générale des moyennes

Après avoir récupérer la moyenne de chaque mot en fonction de sa classe, nous obtenons les deux graphiques suivants:

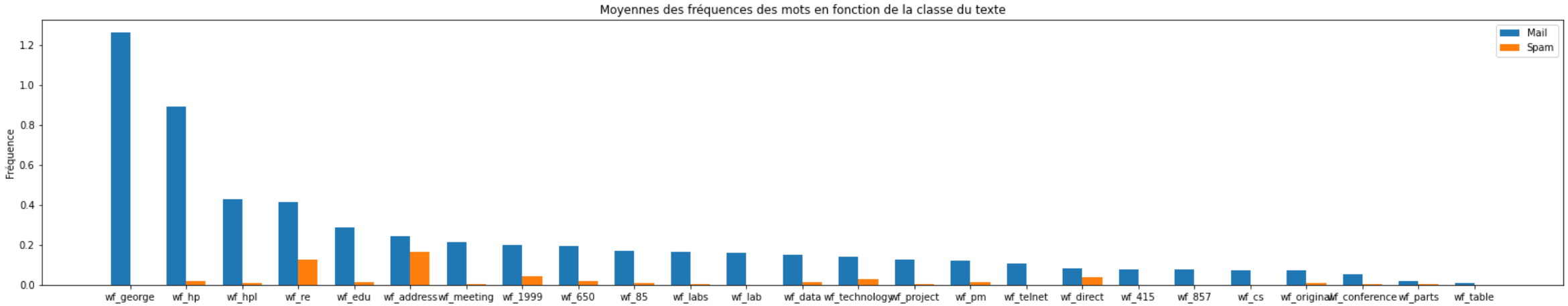


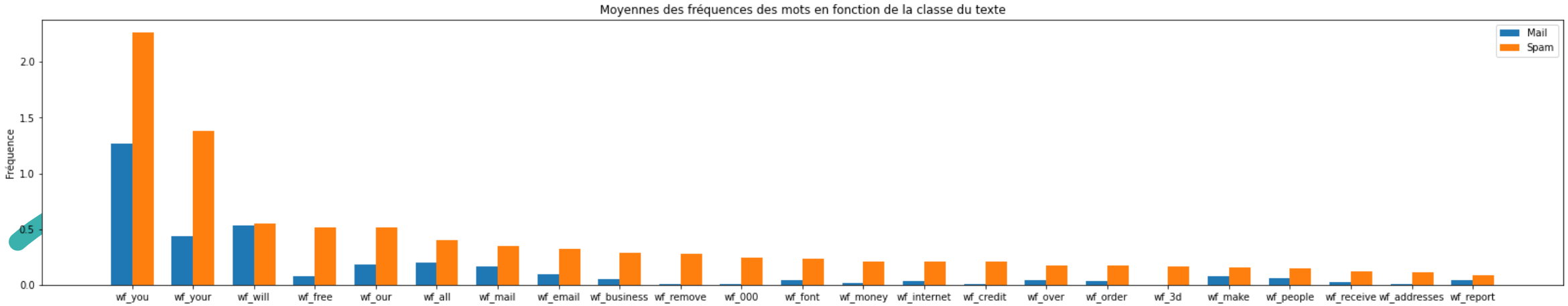


Etude de la fréquence des mots

Comparaison générale des moyennes

Pour plus de lisibilité, nous allons les afficher de manière ordonnée:



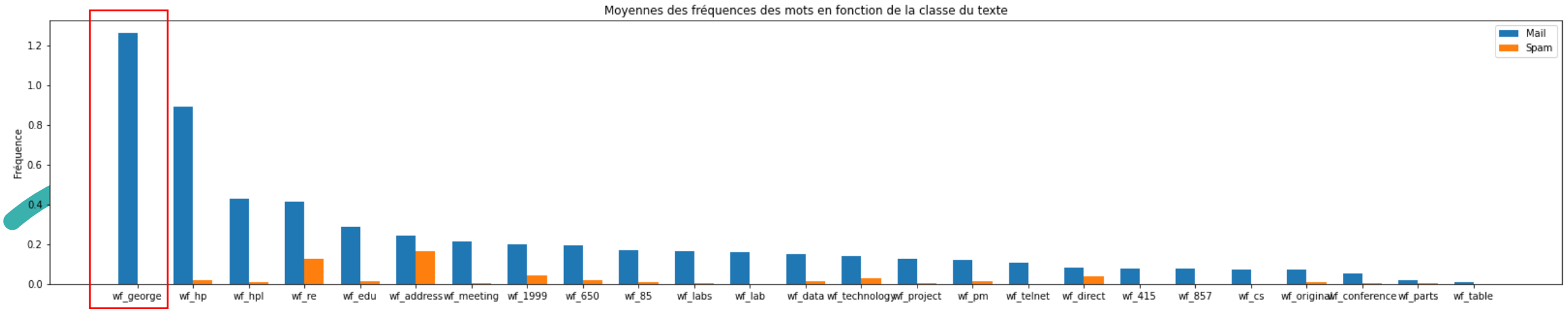


Etude de la fréquence des mots

Fréquences de mot supérieures dans les mails

On peut trouver le mot "**george**" en tête de liste. Cela n'a rien d'étonnant, vu que dans la description du jeu de données, on retrouve ces deux lignes:

« Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835. […] Our collection of non-spam e-mails came from filed work and **personal** e-mails, and hence the word '**george**' and the area code '650' are indicators of non-spam. »

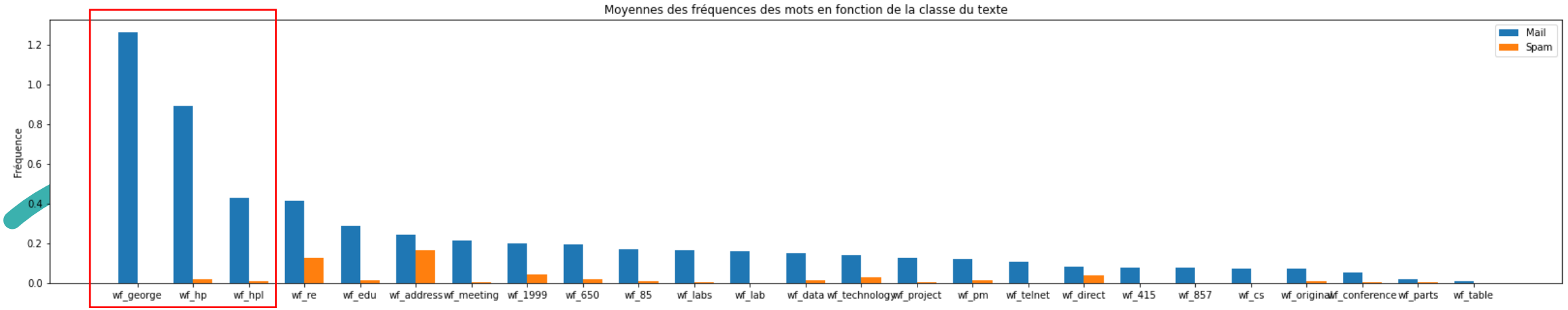


Etude de la fréquence des mots

Fréquences de mot supérieures dans les mails

Le **prénom** reste une information **personnelle**, que les spammeurs n'ont pas forcément la possibilité de récupérer à partir d'une simple adresse email. Aujourd'hui, c'est parfois plus facile quand des bases de données **hackées** sont accessibles de récupérer des adresses mails avec des informations personnelles liées à celle-ci. Mais ce jeu de données datant de **1999**, je me dis qu'il y en avait certainement moins qu'aujourd'hui, vu que l'on ne récupérait pas autant d'informations sur les personnes auparavant.

On voit aussi que les noms de domaine "**hp**" et "**hpl**" viennent tout juste après "**george**": cela reste des informations qui ne sont pas forcément utiles pour un spammeur, d'où leurs fréquences quasi-nulle dans les spams.

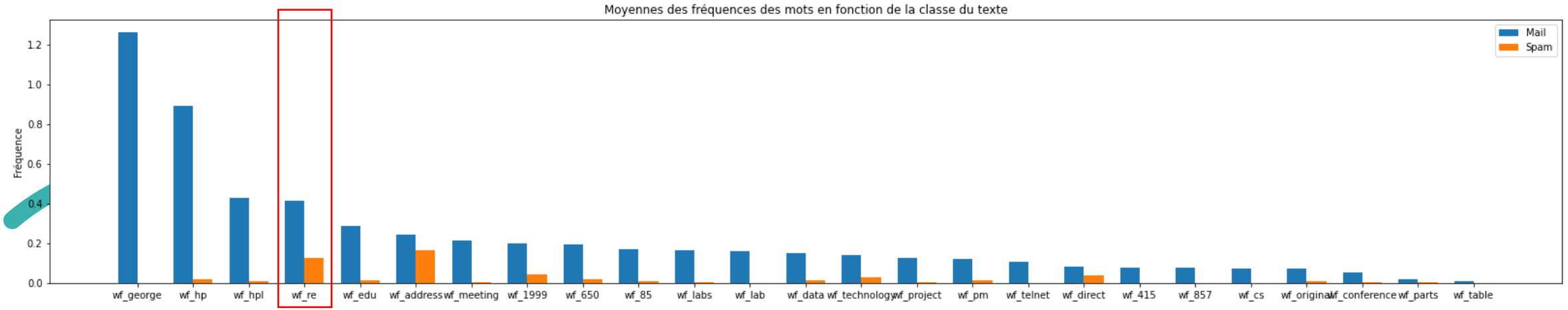


Etude de la fréquence des mots

Fréquences de mot supérieures dans les mails

Le mot "**re**" suit de très près la troisième plus grande fréquence, et il est intéressant de s'épancher un peu plus sur son cas. Je présume que celui-ci était déjà utilisé en 1999 dans les titres de message, tel que, par exemple "RE: info meeting", signifiant une **réponse au précédent mail** nommé "info meeting". Le jeu de données réunissant des mails personnels et professionnels, ce n'est pas étonnant de le voir apparaître.

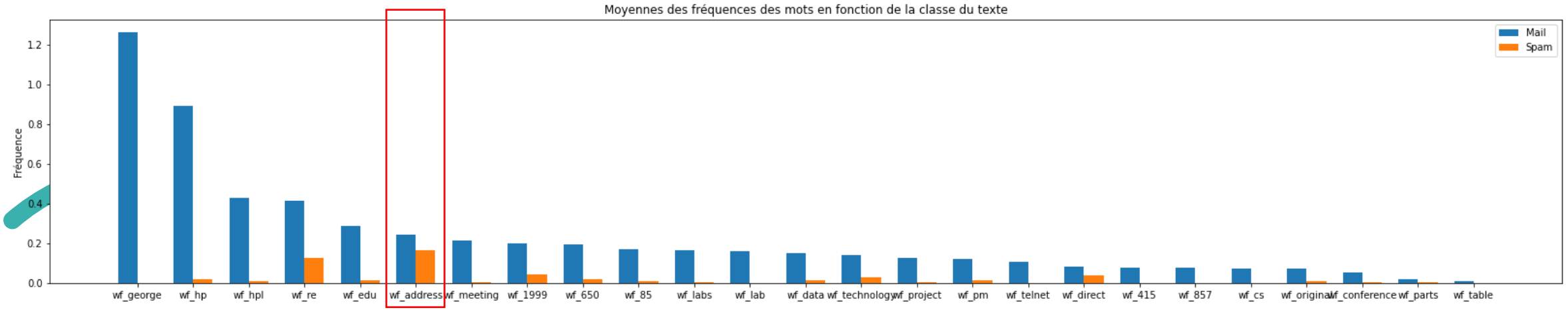
Par contre, on peut voir que ce dernier a une fréquence moyenne de **0.1%** par spam. Aujourd'hui, une méthode utilisée par les spammeurs est d'utiliser le terme "RE: ..." dans le titre pour inciter la cible à ouvrir celui-ci, comme si cette dernière avait **déjà participé à cette conversation**. On trouve parfois le mot "RE" dans les filtres de spam d'ailleurs. Il est donc probable qu'en 1999, certains spammeurs ont tenté d'utiliser cette technique eux aussi.



Etude de la fréquence des mots

Fréquences de mot supérieures dans les mails

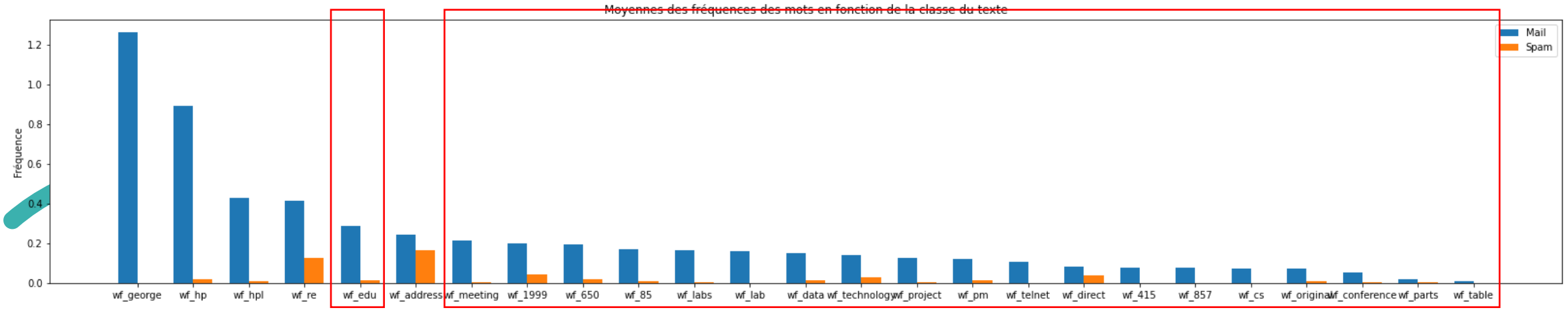
Ce qui est intéressant ici est que la fréquence du mot "**address**" reste plutôt proche entre le mail et le spam. C'est une information qui peut être demandé au travail ou par des proches, mais aussi par des spammeurs qui auraient besoin d'une adresse pour **envoyer de la pub papier** à la cible. Pour ce jeu de données, je doute qu'il soit donc effectif si l'on devait créer un filtre de spam.



Etude de la fréquence des mots

Fréquences de mot supérieures dans les mails

Le reste de mots sont constitués pour la plupart de chiffres et de termes professionnels (meeting, labs, conference...) et la fréquence dans les spams est **quasi-nulle**, ce qui ne semble pas étonnant.



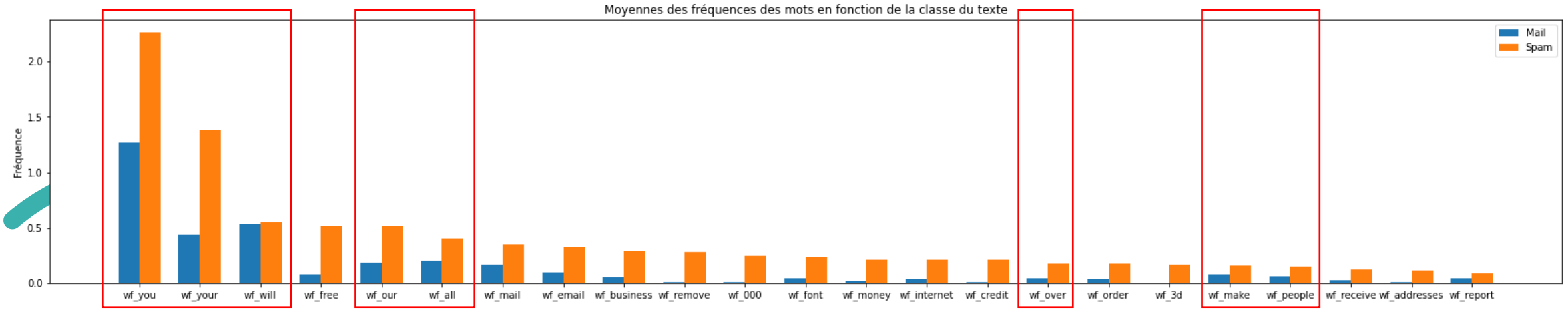
Etude de la fréquence des mots

Fréquences de mot supérieures dans les spams

Avant d'étudier plus en profondeur, il y a une certaine catégorie de mots sur ce deuxième graphique à prendre avec précautions: **les mots communs en anglais**. De nombreuses listes de vocabulaire sont existantes pour chaque langue, regroupant souvent les 100 à 500 mots les plus fréquents à l'écrit dans celle-ci. Vous pouvez d'ailleurs en retrouver plusieurs ici, basées sur différentes sources : https://en.wikipedia.org/wiki/Most_common_words_in_English

Si l'on s'en réfère à ces listes, on peut identifier certains mots appartenant à celles-ci dans notre graphique: **you, your, will, our, all, over, make et people**.

De ce fait, même si ces derniers apparaissent plus souvent dans les spams que dans les mails de ce jeu de données, il ne faut pas oublier qu'ils **restent fréquents en anglais écrit**. De plus, n'ayant pas le nombre de caractères total pour chaque mail, les fréquences sont aussi à prendre avec des pincettes.



Etude de la fréquence des mots

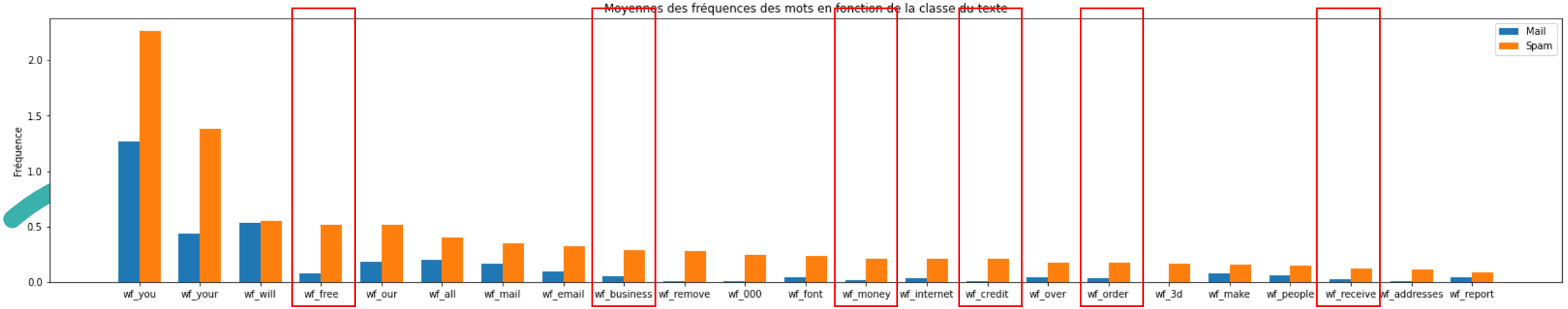
Fréquences de mot supérieures dans les spams

En mettant de côté les mots communs vus précédemment, on peut dégager un second groupe: **les mots commerciaux.**

On retrouve en effet les mots "**free**", "**business**", "**money**", "**credit**", "**order**" et "**receive**". On peut en dégager deux types de technique de spam:

- Parler d'un produit/commande gagné ou gratuit pour le correspondant (free, receive, order)
- Parler d'une affaire juteuse, qui pourrait rapporter de l'argent facilement (business, money, credit)

Il n'est donc pas étonnant de voir le mot **gratuit** (free) en première place des fréquences (si l'on écarte les mots communs devant lui), qui attire souvent le lecteur du message.



Comparaison en détails d’attributs

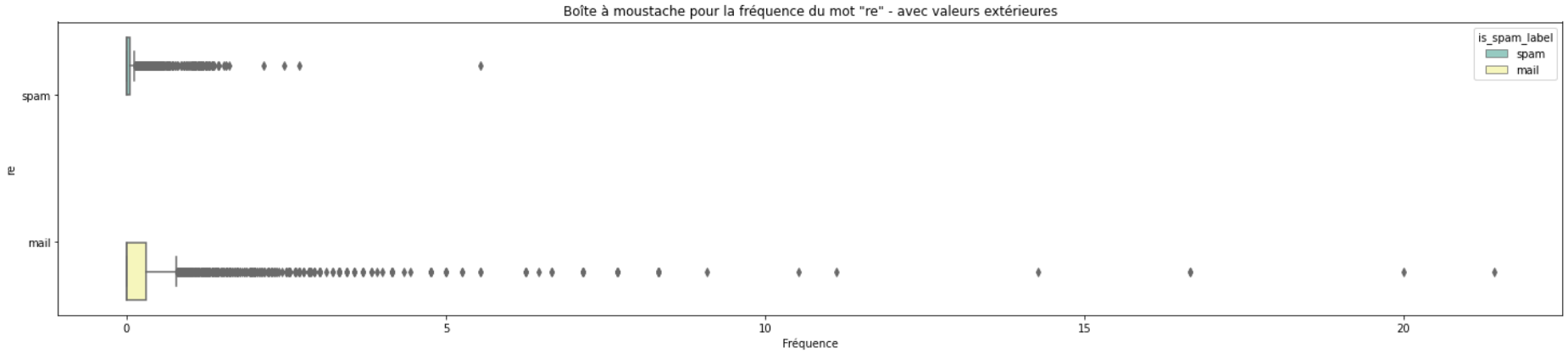
Nous allons devoir regarder de plus près certains mots, la moyenne ne suffisant pas pour se prononcer sur l’utilité d’un des attributs. Nous allons nous intéresser aux quartiles et médianes des mots suivants, pour déterminer s'ils seront intéressants pour détecter un mail d'un spam:

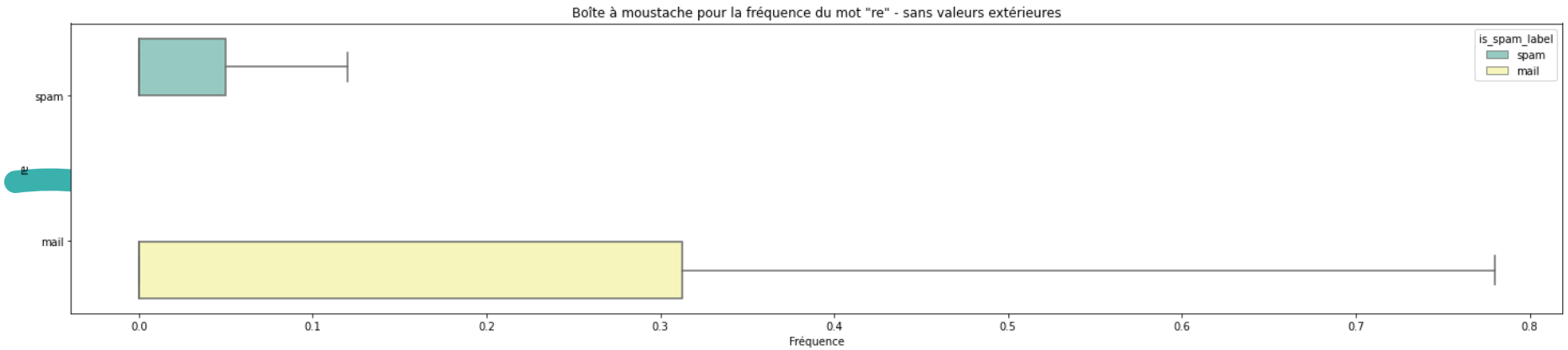
- re
- you
- your
- our
- all

Comparaison en détails d'attributs

Le mot « re »

Le mot "re" semble révélateur du mail. Plus de **25%** des fréquences de la classe mail sont supérieures à **0.3125**, tandis que celles de la classe spam sont supérieur à **0.0500**.



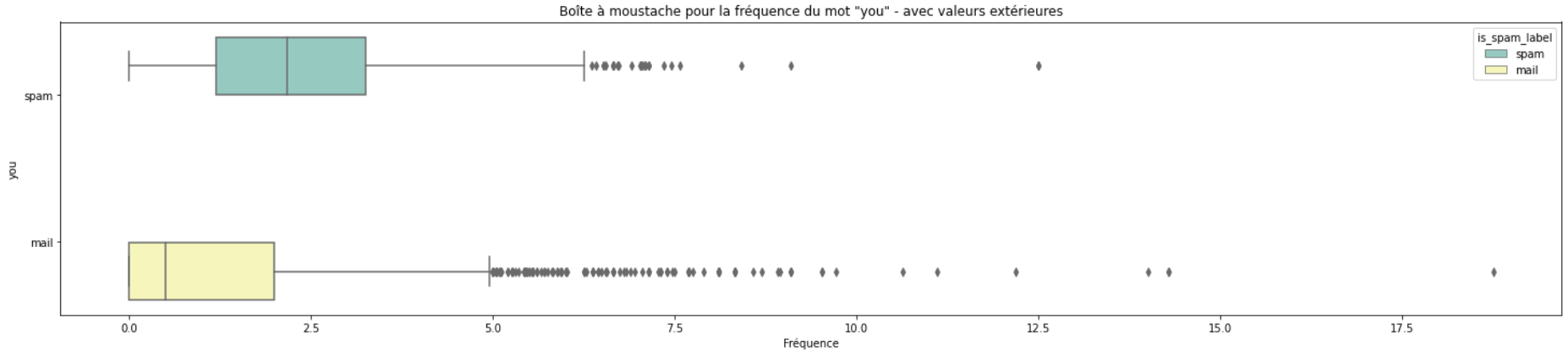


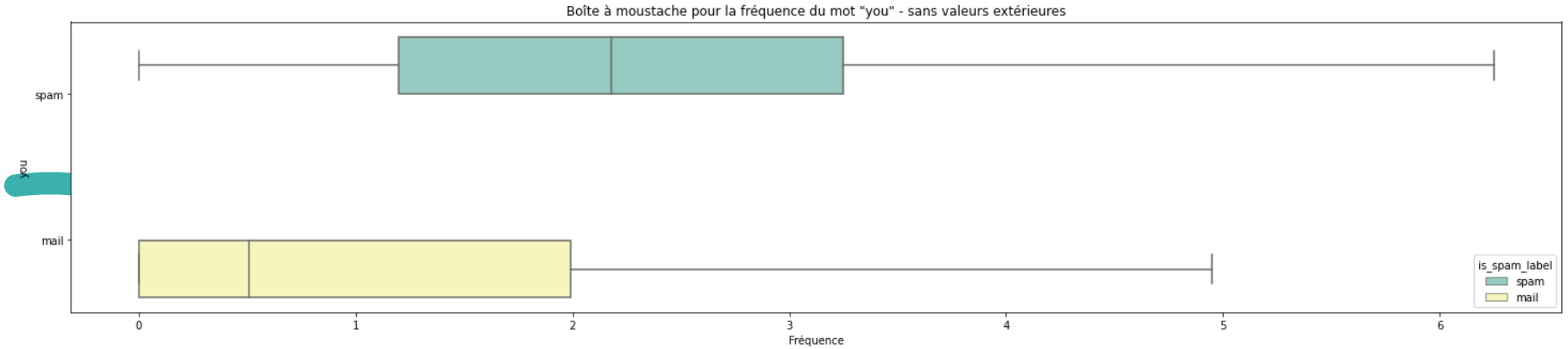
Comparaison en détails d'attributs

Le mot « you »

Le mot "**you**" semble révélateur de spam. Il est important de souligner que l'étude de la moyenne seule ici n'aurait pas suffit. En effet, pour une moyenne à 1.27 dans la classe mail, **50%** des valeurs sont tout de même inférieur à 0.51. Avec six valeurs extérieures supérieures à 10.0, notamment celle valant 18.75, la moyenne a été bien gonflée.

Ainsi, on peut se dire qu'on a **probablement de grandes chances d'avoir affaire à un spam** lorsque la fréquence est **supérieure à 2.18**.





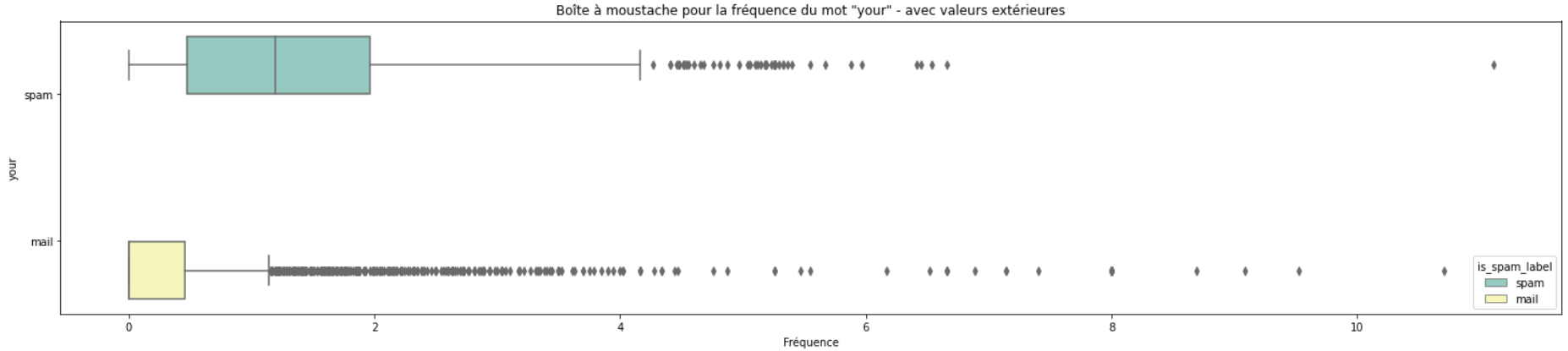
Comparaison en détails d'attributs

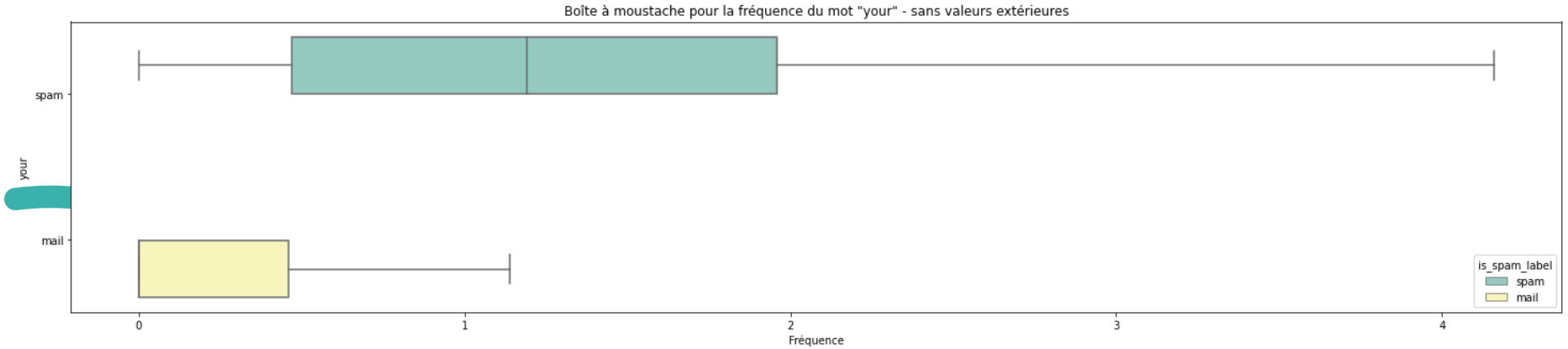
Le mot « your »

Ici aussi, nous avons un exemple flagrant d'une moyenne gonflée par quelques valeurs élevées.

Pour une moyenne de 0.43 de la classe mail, on se retrouve avec **plus de 75% des valeurs inférieures à 0.46**, alors que plus de 25% des valeurs de la classe spam sont **supérieur à 0.47**.

Le mot "**your**" est pour le moment, selon moi, un des meilleurs attributs utiles pour **la détection** entre les cinq mots étudiés.



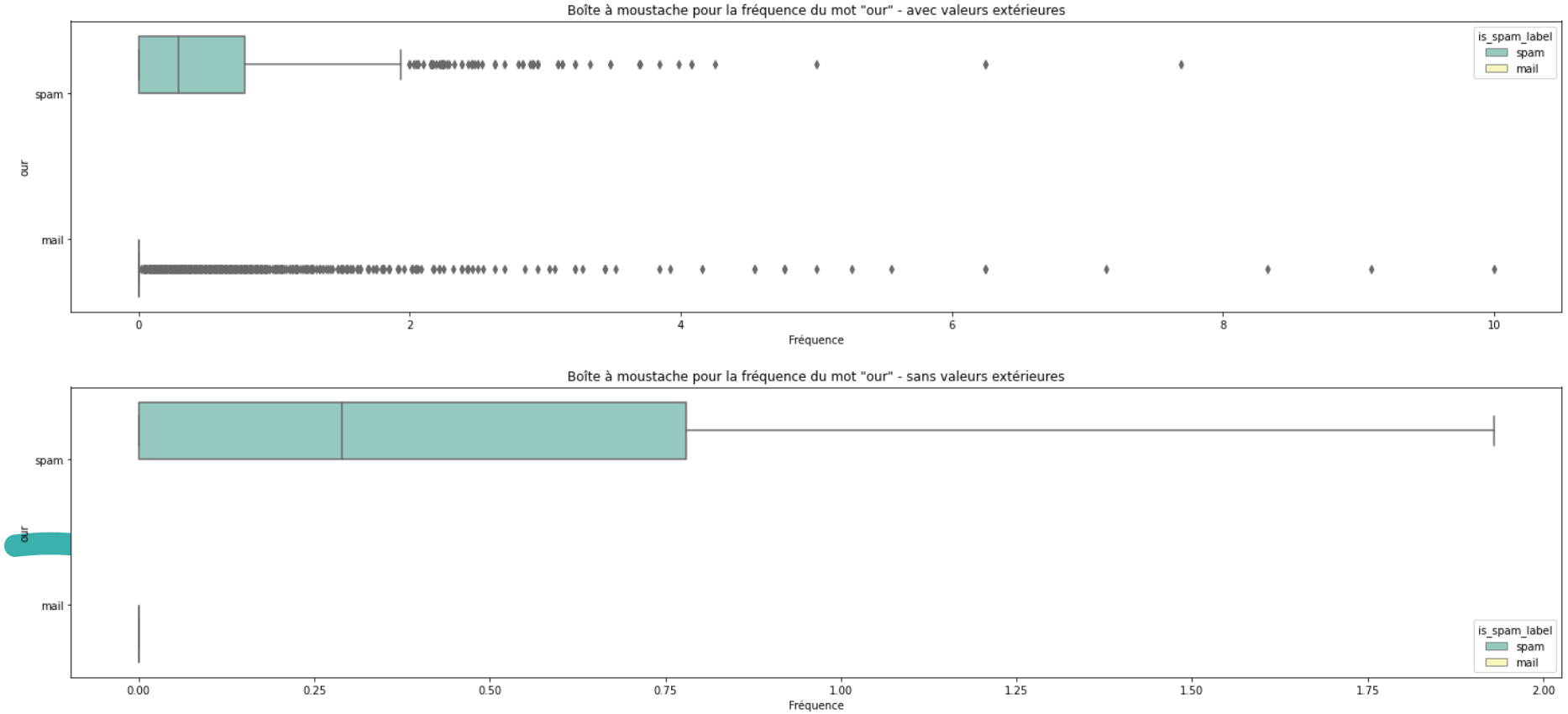


Comparaison en détails d'attributs

Le mot « our »

L'étude du mot "**our**" se révèle très intéressant. Tandis qu'on aurait pu croire avec le graphique des moyennes des fréquences que le mot "our" était quand même assez présent dans les mails, on voit ici que **plus de 75% des valeurs ont une fréquence égale à 0 pour les mails**. Tandis que pour les spams, plus de la moitié des valeurs sont supérieures à 0.29 et un tiers des valeurs supérieures à 0.78.

Avec le mot "your", "our" est définitivement intéressant.

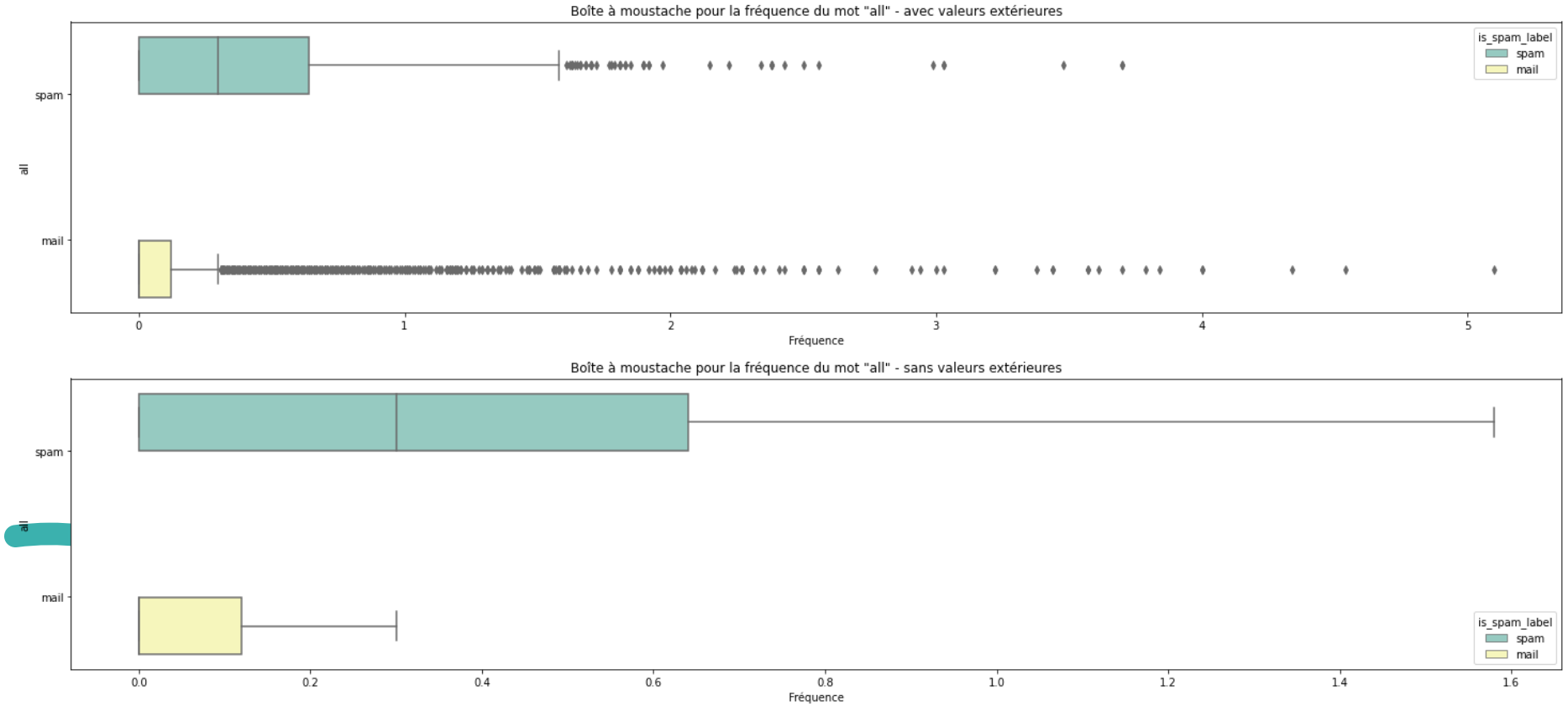


Comparaison en détails d'attributs

Le mot « all »

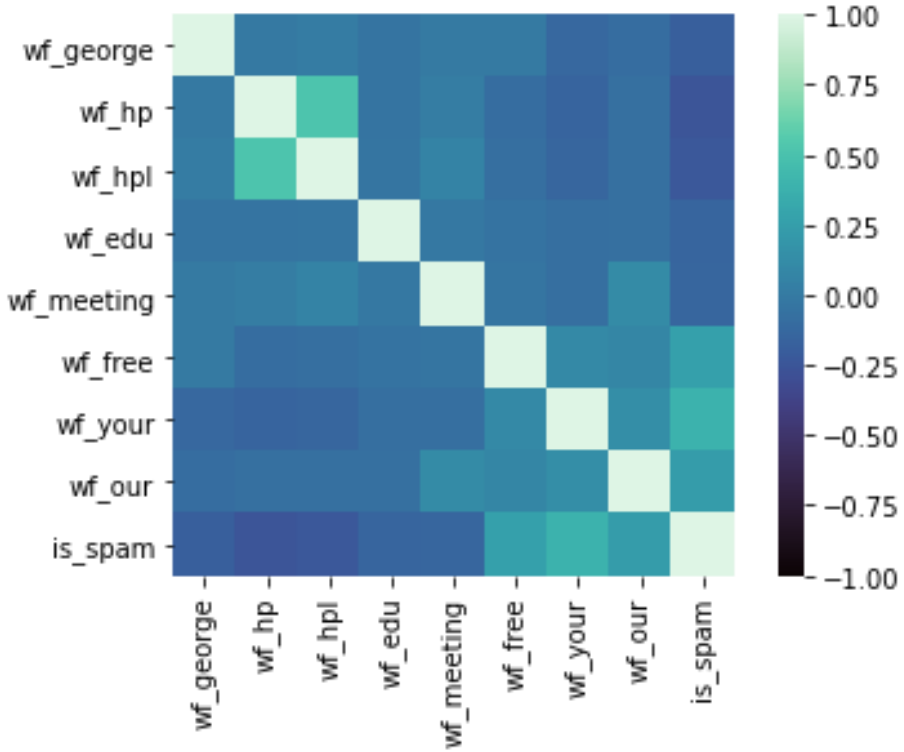
Encore une fois, la moyenne se révèle dangereuse pour la classe mail. Pour une moyenne de 0.2, seulement **75% des valeurs sont supérieures à 0.12**.

La moyenne des spams est beaucoup plus proche de sa médiane, respectivement 0.4 et 0.3. On peut donc en conclure que si la fréquence est supérieure à 0.64, il y a de fortes chances qu'on ait affaire à un spam.



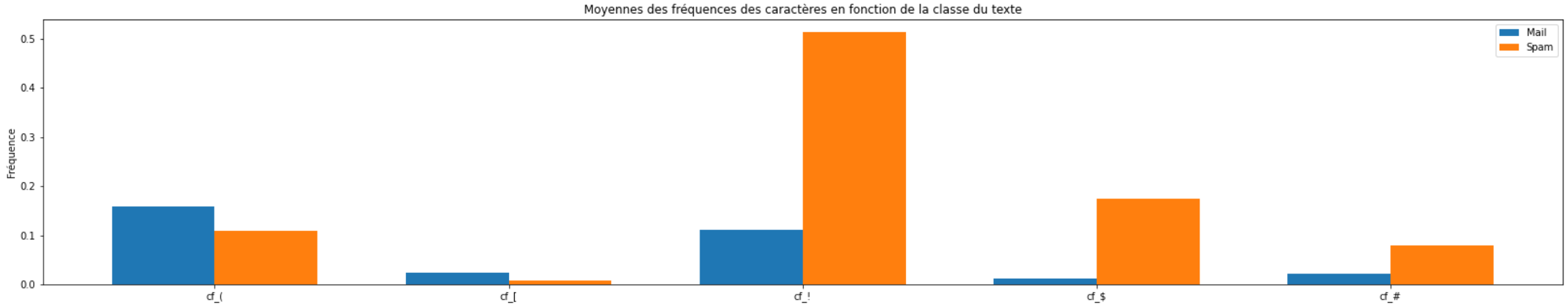
Matrice de corrélation pour les mots

En reprenant les mots semblants avoir le plus d'impact sur la classification, nous obtenons la matrice de corrélation suivante:



Etude de la fréquence des caractères

Vu que nous n'avons que cinq colonnes touchant aux caractères, nous allons pouvoir tous les regarder en même temps.



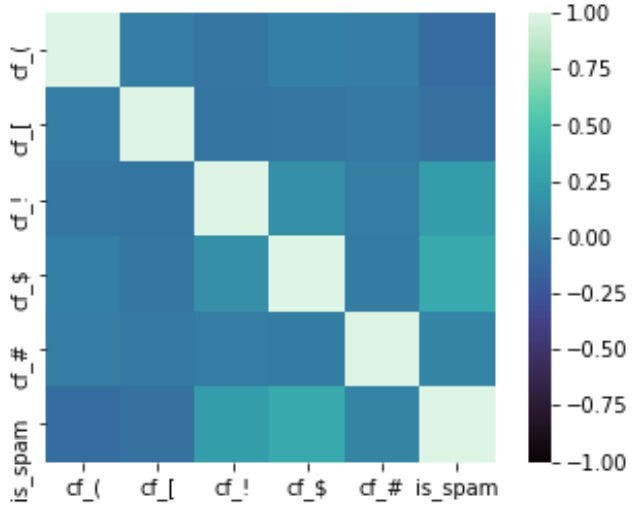
On peut déjà observer que l'utilisation de point d'exclamation "!" semble être très utile pour détecter un spam, ainsi que l'utilisation du signe dollar "$".

Etude de la fréquence des caractères

Le signe du dollar semble être légèrement plus utile pour détecter un spam que le point d'exclamation, mais ils restent proches.

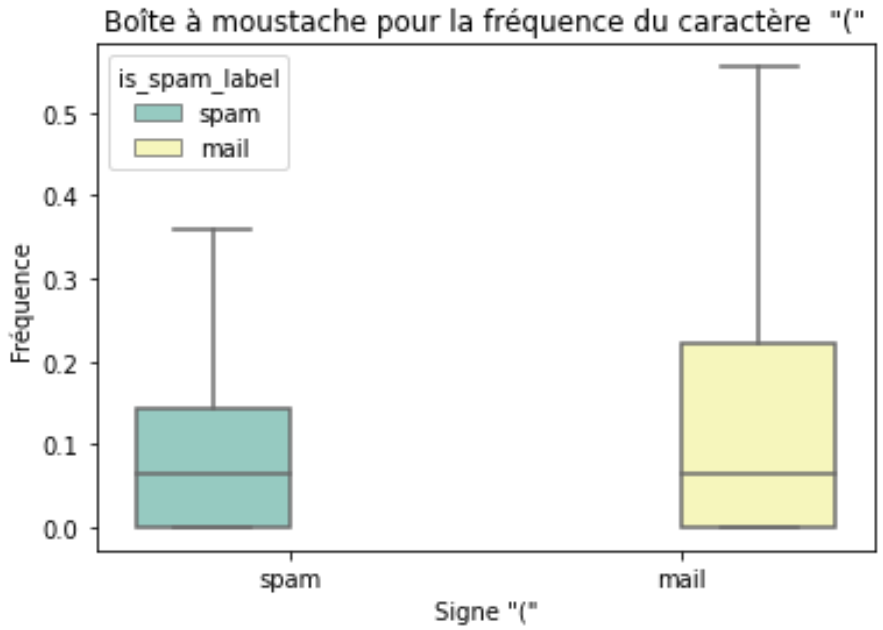
Pour détecter un mail, on peut voir que la parenthèse peut être utile, suivi par le crochet "[".

Le signe dièse quand à lui semble avoisiner autour des 0, nous le mettrons donc potentiellement de côté lors du modèle d'IA.



Etude de la fréquence des caractères

Pour ce qui concerne le signe de la parenthèse, on peut voir que la médiane est sensiblement la même entre les deux classes. Cependant, une valeur supérieure à 0.222 a un peu plus de chance de provenir d'un mail que d'un spam.



Etude des statistiques des majuscules

Pour les majuscules, nous avons trois attributs à étudier:

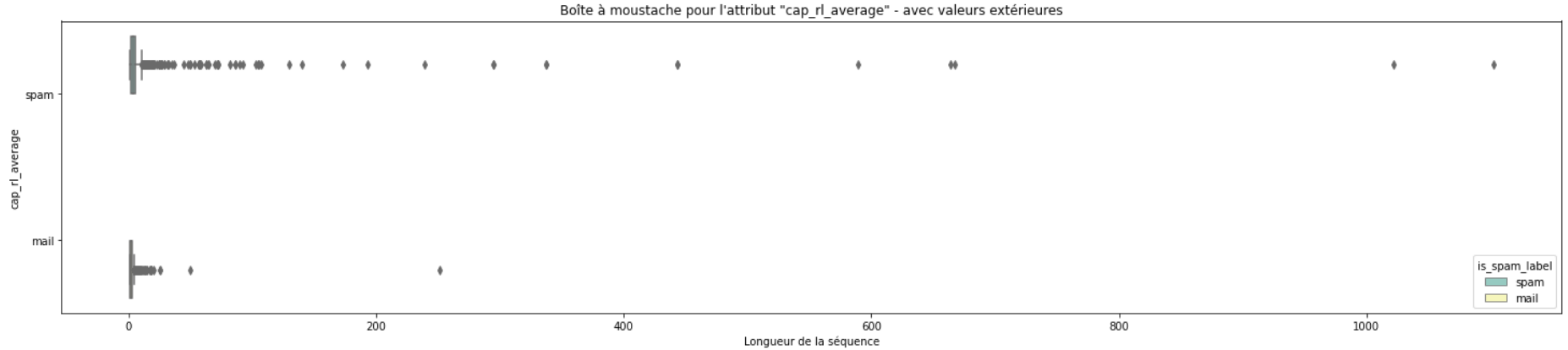
- **cap_rl_average**: longueur moyenne des séquences ininterrompues de majuscules
- **cap_rl_longest**: longueur de la plus longue séquence ininterrompue de lettres majuscules
- **cap_rl_total**: nombre total de majuscules

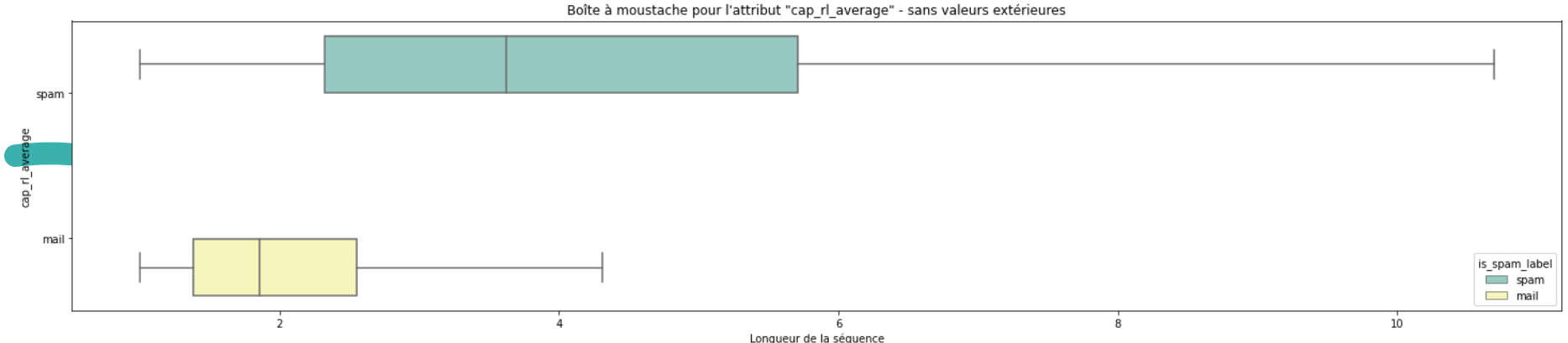
Etude des statistiques des majuscules

Attribut « cap_rl_average »

Attribut très intéressant, on peut en conclure qu'on a beaucoup plus de chance de faire affaire à un spam lorsque **la longueur moyenne est supérieure à 2.5**.

Cela n'est pas étonnant, dans le sens où l'emploi des majuscules apportent **un effet d'urgence/de danger/d'occasion à ne pas manquer**, que les spammeurs aiment employer.





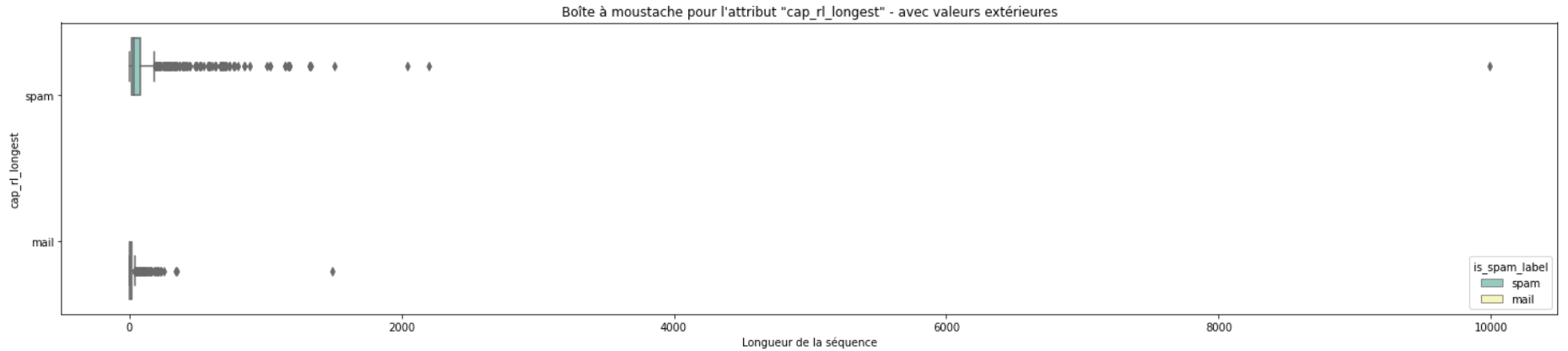
Etude des statistiques des majuscules

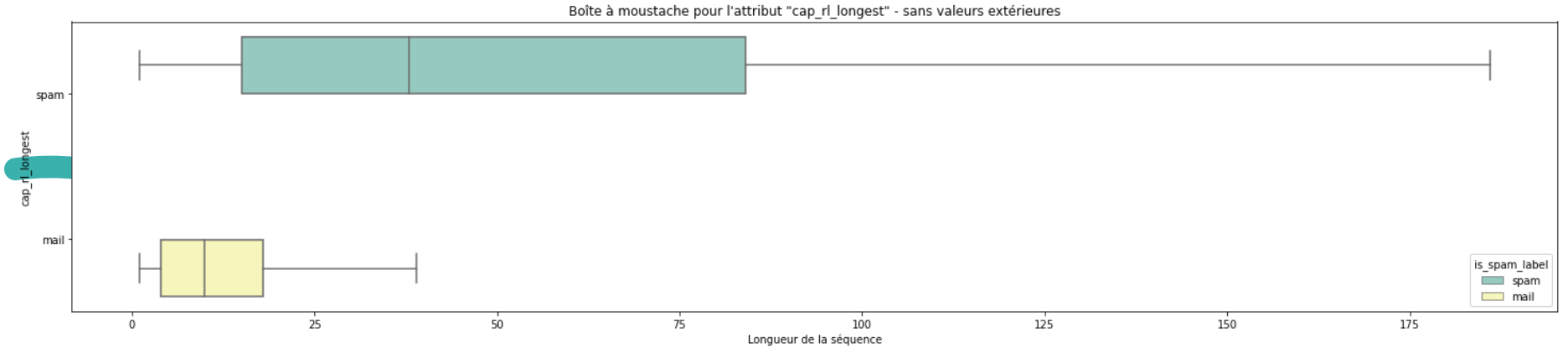
Attribut « cap_rl_longest »

Rien de surprenant par rapport à ce que l'on a appris avec l'étude de l'attribut précédent "cap_rl_average".

Cependant, je ne m'attendais pas au 3éme quartile de la classe mail d'être **aussi haut que 18**. De ce que j'ai pu trouver sur internet, la **taille moyenne d'un mot anglais** tourne aux alentours de **4.7 à 5 caractères**. Donc on aurait plus de 15% des mails qui emploieraient en moyenne 3 à 4 mots en majuscules d'affilée.

Cela n'est tout de même pas grand chose par rapport à la classe spam, qui a **plus de 50%** de ses textes contenant une séquence de majuscule ininterrompue de **38 caractères**, soit environ **7.6 mots d'affilée (38/5).**





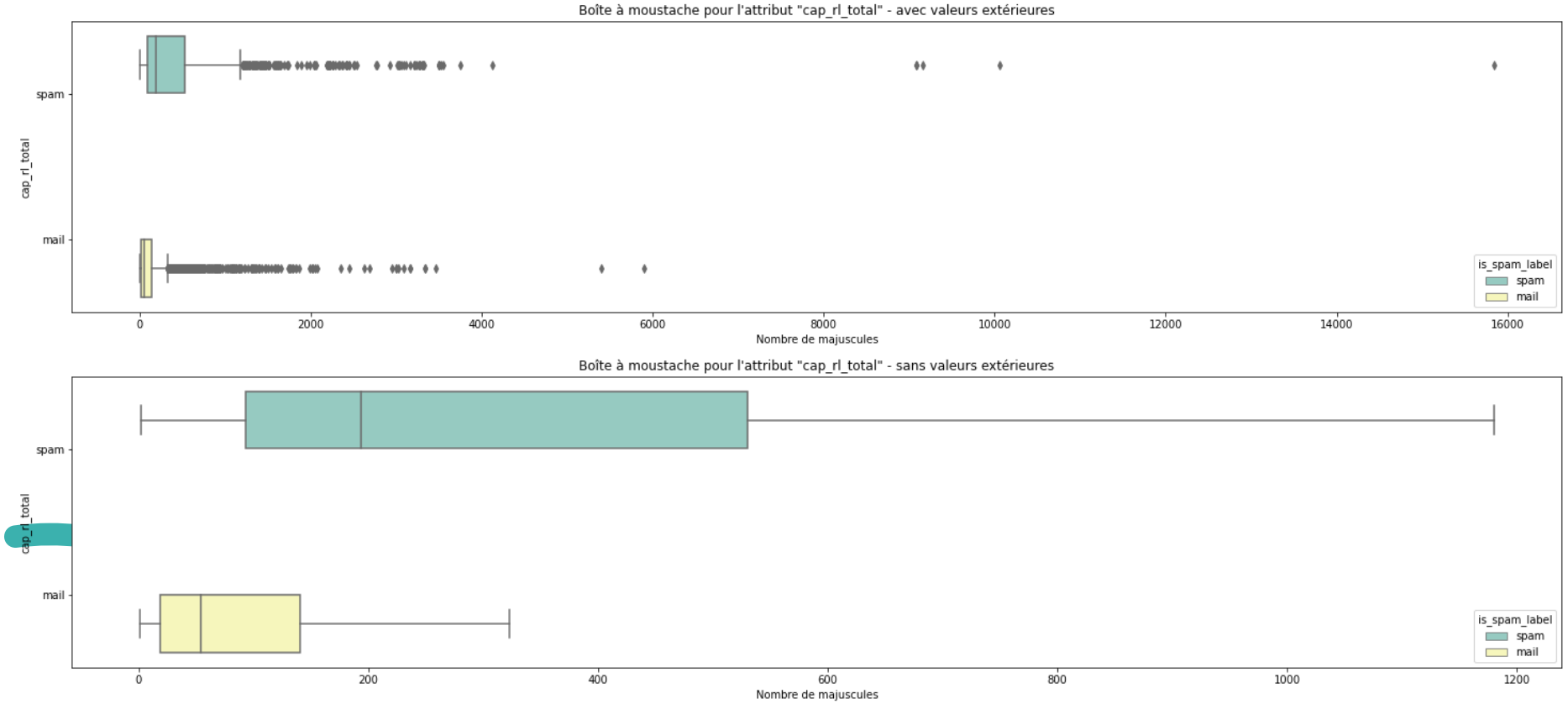
Etude des statistiques des majuscules

Attribut « cap_rl_total »

Il est vraiment dommage de ne pas avoir de moyen d'accéder à **la taille totale des textes** ici, et pourquoi pas le nombre de phrase au total dans le texte.

En effet, on peut voir que dans plus de 50% des cas pour les mails, on se retrouverait avec **54 majuscules au total**. Il ne faut pas oublier que dans un mail normal, **les phrases commencent par une majuscule**, ainsi que **les noms propres et certaines abréviations** (comme ASAP par exemple). Mais 54 me semble tout de même être un chiffre élevé. Et c'est encore plus impressionnant de se dire **que 15% de ces mails ont plus de 141 majuscules**.

Mis à part cela, comme les deux précédents attributs, on voit bien que **la classe spam se démarque** de la classe mail par son grand nombre.



Modélisation

Sélection des modèles à étudier

Après recherche, les meilleurs modèles utilisés pour les filtres anti-spam s'avèrent être de la famille de la classification naïve bayésienne (NB)*. Ainsi, nous nous contenterons de tester trois types de modèle NB:

- BernouilliNB
- GaussianNB
- MultinomialNB

* *Exemple d'article mentionnant la classification naïve bayésienne*

Rappel de ce que nous cherchons

Pour rappel, nous cherchons à **séparer les spams** des autres mails. Ainsi, nous serons plus regardant sur le **nombre de mail classé comme spam** que l'inverse.

En effet, si un spam ou deux sont identifiés comme mails, c'est désagréable pour l'utilisateur, mais cela s'arrête là.

Tandis que si l'on classe un mail en tant que spam, c'est beaucoup plus problématique pour l'utilisateur, vu que ce mail sera **directement placé dans la corbeille**, sans notifier l'utilisateur.

Préparation des données

On commence par organiser les données et les séparer dans les jeux de test et d'entraînement :

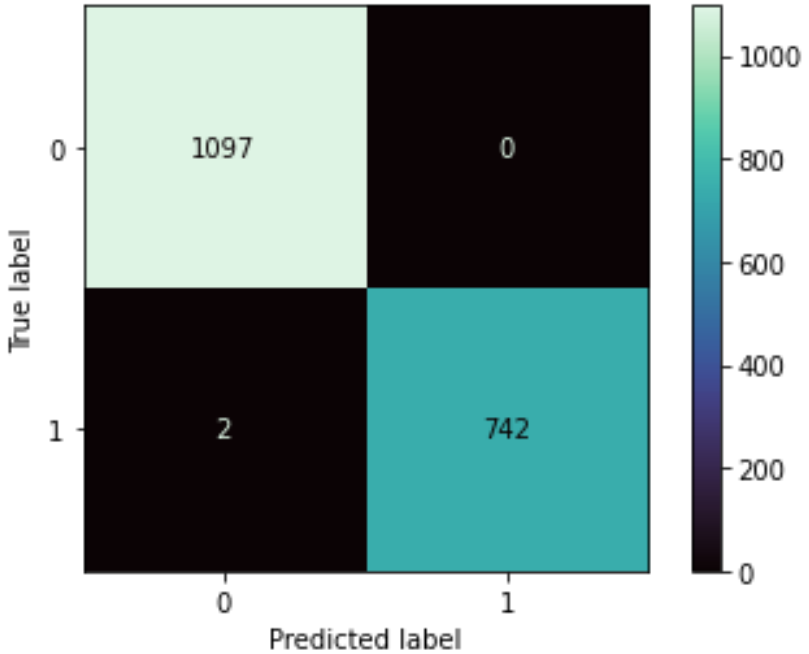
```
# Organisation des données
label_names = ['mail', 'spam']
labels = dfData['is_spam']
feature_names = columnsName
# On ne prend pas la dernière colonne de label
features = np.asarray(dfData[feature_names].values)
```

```
# Partage des données entre le dataset de train et de test
train, test, train_labels, test_labels = train_test_split(features,
                                                          labels,
                                                          test_size=0.4,
                                                          random_state=0)
```

Modèle GaussianNB

Le résultat est excellent: 99,89% de précision ! De plus, aucun mail n'a été classé comme spam, ce qui est excellent aussi.

```
Précision du modèle: 0.9989136338946225
Nombre de mails non identifiés correctement sur un total de 1841 mails : 2
              precision    recall  f1-score   support

        mail       1.00      1.00      1.00      1097
        spam       1.00      1.00      1.00       744

    accuracy                           1.00      1841
   macro avg       1.00      1.00      1.00      1841
weighted avg       1.00      1.00      1.00      1841
```

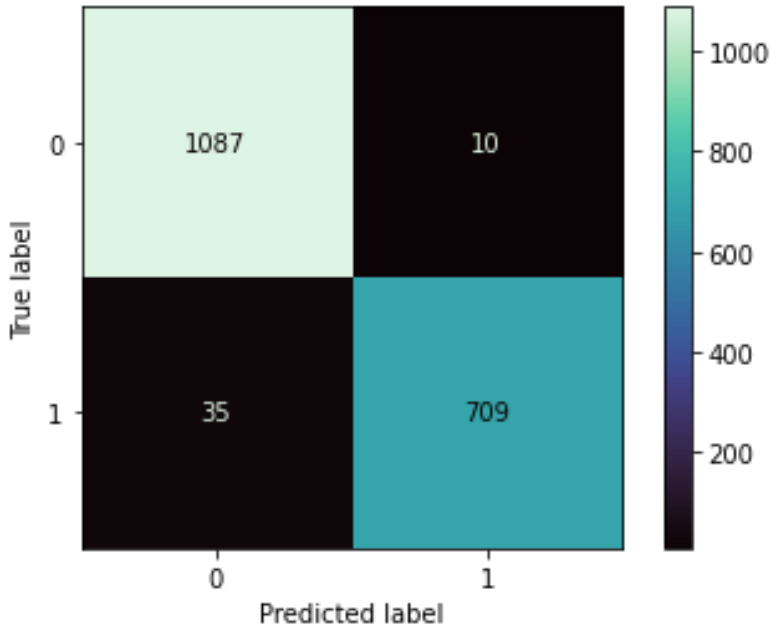


Modèle BernouilliNB

La précision est déjà moins bonne que pour le modèle Gaussien. De plus, en plus d'avoir 35 spams identifiés comme mails, on a surtout 10 mails identifiés comme spams !

Le modèle Gaussien reste en tête.

```
Précision du modèle: 0.9755567626290059
Nombre de mails non identifiés correctement sur un total de 1841 mails : 45
              precision    recall  f1-score   support

        mail       0.97      0.99      0.98      1097
        spam       0.99      0.95      0.97       744

    accuracy                           0.98      1841
   macro avg       0.98      0.97      0.97      1841
weighted avg       0.98      0.98      0.98      1841
```

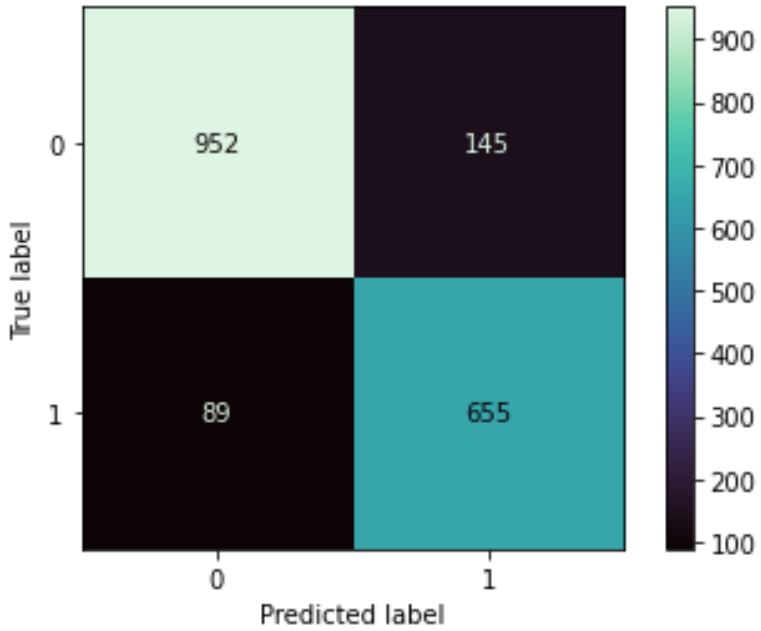


Modèle MultinomialNB

Par rapport aux deux modèles précédents, la précision chute de 10%.

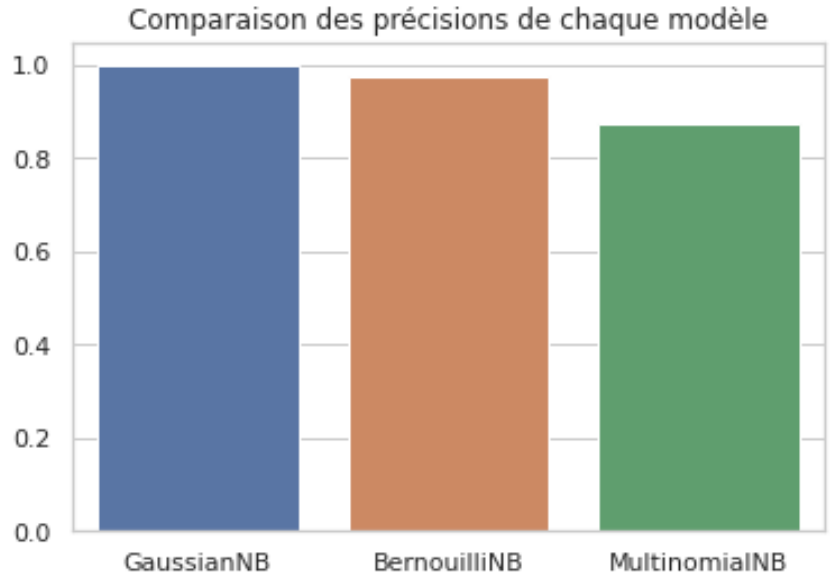
Et pour finir, on a beaucoup plus de mails identifiés comme spams que l'inverse. Ce modèle est à proscrire.

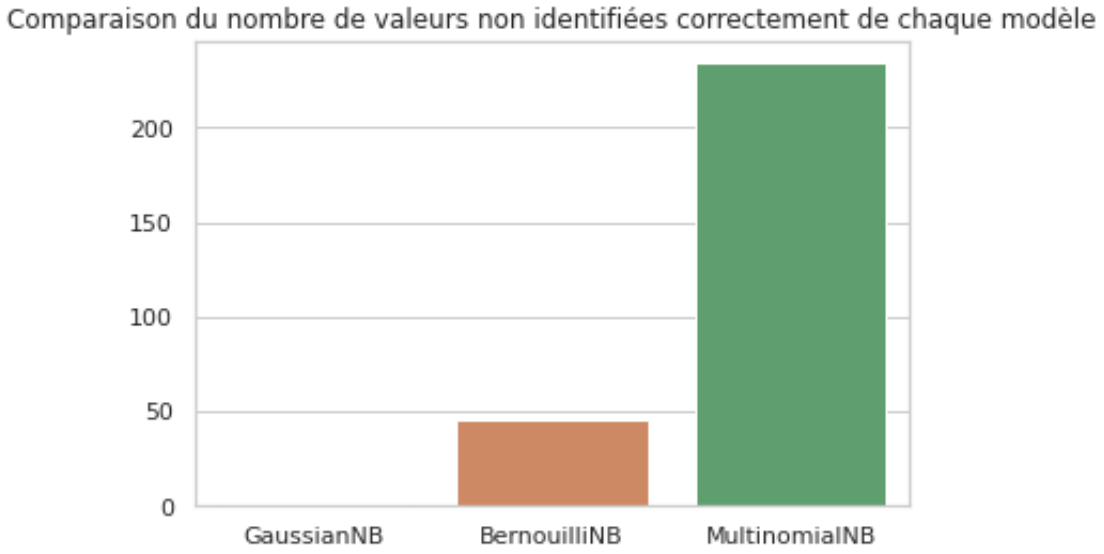
```
Précision du modèle: 0.872895165670831
Nombre de mails non identifiés correctement sur un total de 1841 mails : 234
              precision    recall  f1-score   support

        mail       0.91      0.87      0.89      1097
        spam       0.82      0.88      0.85       744

    accuracy                           0.87      1841
   macro avg       0.87      0.87      0.87      1841
weighted avg       0.88      0.87      0.87      1841
```



Comparaison finale

Il n'y a pas photo: le modèle Gaussien remporte haut la main la manche !





Conclusion

Conclusion

Il aurait été intéressant d'avoir les textes d'où proviennent les données du jeu, pour les comparer à des mails plus récents.

Cependant, certaines choses n'ont pas changé: l'utilisation de points d'exclamation et de majuscules par exemple, ou encore l'utilisation de certains mots.

Par exemple, vous pourrez trouver à ces deux liens des techniques pour éviter que le mail envoyé soit classé comme un spam par les filtres actuels:

https://www.yesware.com/blog/email-spam/

https://www.simplycast.com/blog/100-top-email-spam-trigger-words-and-phrases-to-avoid/#post

Mais pour reprendre la description du jeu de données, pour construire un filtre personnalisé, c'est au cas par cas. Tout le monde ne considère pas les mêmes choses comme spam par exemple.