

# Projet L3

## Ingénierie des langues

## Fouille de données

GOEHRYS Martial  
16711476

18 avril 2023

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Phase 1 : Récupération des données</b>	<b>5</b>
2.1	Récolte des données . . . . .	5
2.2	Pré-traitement . . . . .	6
2.2.1	Importation . . . . .	6
2.2.2	Extraction des corps des mails . . . . .	6
2.2.3	Nettoyage . . . . .	7
2.2.4	Mise en base . . . . .	12
2.3	Données de la phase 1 . . . . .	17
<b>3</b>	<b>Phase 2</b>	<b>22</b>
3.1	Traitement . . . . .	22
3.1.1	Recherche de caractéristiques . . . . .	22
3.1.2	Analyse préliminaire . . . . .	22
<b>4</b>	<b>Phase 3</b>	<b>23</b>
<b>A</b>	<b>Développement visualisation distribution de Zipf</b>	<b>24</b>
<b>B</b>	<b>Déploiement des bases de données</b>	<b>30</b>
B.1	ElasticSearch . . . . .	30
B.1.1	Conteneurisation . . . . .	30
B.1.2	Initialisation de l'index . . . . .	34
B.2	PostgreSQL . . . . .	35
B.2.1	Conteneurisation . . . . .	35
B.2.2	Initialisation de la base de données . . . . .	36
<b>C</b>	<b>Tableau des choix technologiques</b>	<b>39</b>
<b>D</b>	<b>Modèles</b>	<b>40</b>
D.1	Naïves Bayes . . . . .	40
<b>E</b>	<b>Bibliographie</b>	<b>45</b>

<b>F</b>	<b>Sitotec</b>	<b>45</b>
F.1	Corpus . . . . .	45
F.2	Modules . . . . .	45
F.3	Modèles . . . . .	45
<b>G</b>	<b>Codes sources</b>	<b>46</b>
G.1	Analyse statistiques de la phase 1 . . . . .	46

# 1 Introduction

Ce projet a pour but de développer un modèle permettant de catégoriser des emails en spam ou ham. La définition d'un spam dans le dictionnaire *Larousse* est :

"Courrier électronique non sollicité envoyé en grand nombre à des boîtes aux lettres électroniques ou à des forums, dans un but publicitaire ou commercial."

Il est possible d'ajouter à cette catégorie tous les mails indésirables comme les tentatives d'hameçonnage permettant de soutirer des informations personnelles à une cible.

L'objectif est de travailler uniquement sur les données textuelles issues du corps du mail. Nous avons donc en point de départ les éléments suivants :

- langue : anglais
- corpus : monolingue écrit
- type : e-mail

**Déroulé** Le développement de ce projet s'articule autour de 3 phases majeurs

- Phase 1 : Récupération des données
- Phase 2 : Analyse des caractéristiques
- Phase 3 : Construction du modèle

**Phase 1** La phase 1 concerne la récolte des informations et les traitements minimums nécessaires pour la mise en base. Les objectifs de traitement de cette phase sont :

- Extraire les corps des mails et éliminer les méta-données superflues
- Eliminer les mails non anglais
- Eliminer les mails en doublons
- Eliminer les parties de textes non pertinente (liens, réponses, certaines ponctuations)

Cette phase se termine avec la mise en base des documents dans un index Elasticsearch et le stockage du nombre de liens dans une base PSQL.

**Phase 2** La phase 2 vise à générer les données statistiques et numériques à partir des corps de mails. Lors de cette phase une analyse statistique manuelle des données est réalisée

**Phase 3** La phase 3 regroupe tous les opérations d'exploitation des données et vise à développer et à créer un modèle de classement des mails et d'en évaluer les performances.

Afin de conserver une certaine cohérence dans le déroulé entre les phases et au vu du temps que j'ai pris pour réaliser ce projet l'ensemble des étapes est automatisé en langage Python. Seule la récolte initiale des mails a été réalisée à la main.

Le schéma ci-dessous donne une vue synthétique des étapes du projet

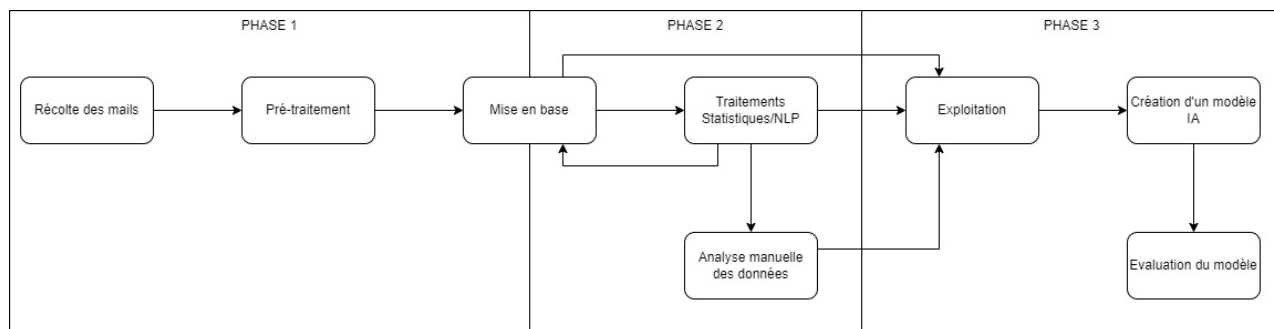


FIGURE 1 – Schéma des grandes étapes

## 2 Phase 1 : Récupération des données

La phase 1 est une pipeline qui vas permettre d'extraire un maximum d'information d'un email en essayant de pas dénaturer le fond ni la forme. Il pourra ensuite être stocké avec sa catégorie d'appartenance. Durant cette phase nous allons également initialiser les bases de données en créant les index (ES) et les tables (PSQL et SQLITE).

Ci-dessous le schéma général de cette phase.

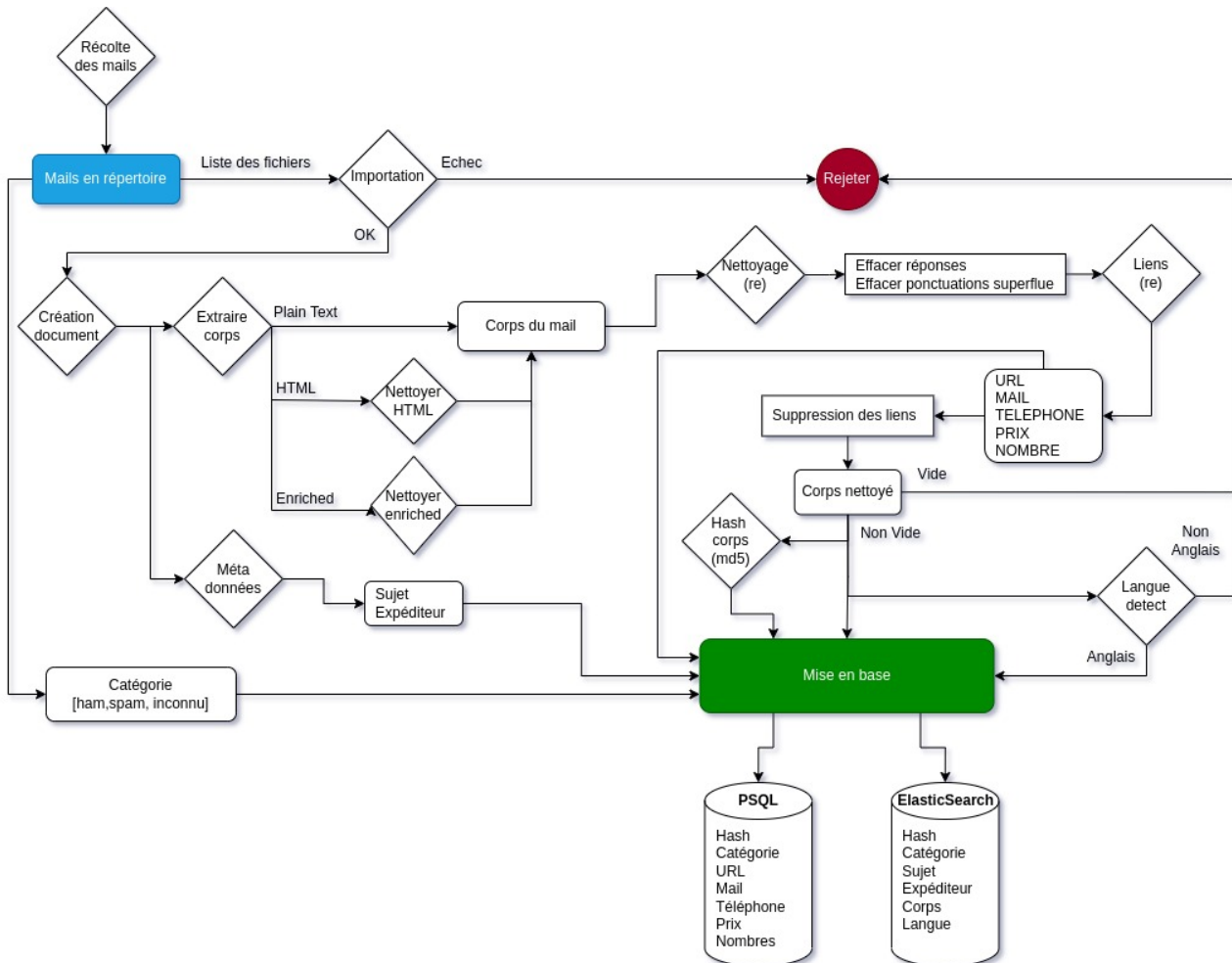


FIGURE 2 – Schéma des étapes de la phase 1

### 2.1 Récolte des données

**Recherche de dataset** J'avais dans l'idée de faire la recherche de mails en français. Cependant, je n'ai pas trouvé de dataset dans cette langue. Je me suis donc retourné vers les datasets de mails en anglais.

J'ai pu alors récupérer deux dataset :

- Enron company mails (voir : F.1)
- Dataset SpamAssassin (voir : F.1)

Les mails de SpamAssassin ont l'avantage d'être pré-triés, contrairement aux mails de la compagnie Enron. Ainsi le développement du moteur se fera uniquement avec les mails du SpamAssassin afin de pouvoir vérifier les résultats de l'analyse.

**Téléchargement des données** Le téléchargement du dataset Enron est possible à partir du moment où l'on possède un compte sur la plateforme Kaggle. Le dataset SpamAssassin est

ouvert, il suffit de télécharger les archives de chaque catégorie.

La récolte des données a été réalisée à la main sans automatisation. Les mails sont alors stockés dans plusieurs répertoires *HAM* et *SPAM* selon leur catégorie.

Format :

- *Enron* - 1 fichier CSV avec tous les mails
- *SpamAssassin* - 1 fichier texte par mail

## 2.2 Pré-traitement

Les étapes de pré-traitement regroupent toutes les étapes et actions réalisées avant la mise en base. L'objectif de ces étapes est d'extraire le message en retirant les métadonnées du mail. Il va être possible d'effectuer certain traitement de nettoyage et de récupération d'informations sommaires.

Les manipulations de messages dans Python se font principalement à l'aide du module *email* natif.

### 2.2.1 Importation

La fonction *email.message\_from\_binary\_file* permet de transformer un fichier mail en objet python manipulable :

Fonction d'importation des fichiers

---

```
1 def import_from_file(chemin):
2     try:
3         with open(chemin, 'rb') as data:
4             msg = message_from_binary_file(data, policy=policy.default)
5             return msg
6
7     except FileNotFoundError:
8         print("Fichier : '{}' non trouve".format(chemin), file=sys.stderr)
9         return None
```

---

### 2.2.2 Extraction des corps des mails

Une fois le fichier importé au format *EmailMessage*, il est possible d'en extraire le corps.

Le corps du mail peut être composé de plusieurs parties qui ne sont pas forcément du texte. Les parties non textuelles ne sont pas conservée.

Extraction du corps du mail

---

```
1 def extract_body(msg):
2     refused_charset = ['unknown-8bit', 'default', 'default_charset',
3                       'gb2312_charset', 'chinesebig5', 'big5']
4     body = ""
5
6     if msg.is_multipart():
7         for part in msg.walk():
8             if not part.is_multipart():
9                 body += extract_body(part)
10    return body
11
```

---

```

12     if msg.get_content_maintype() != 'text':
13         return ""
14
15     if msg.get_content_charset() in refused_charset:
16         return ""
17
18     if msg.get_content_subtype() == 'plain':
19         payload = msg.get_payload(decode=True)
20         body += payload.decode(errors='ignore')
21
22     if msg.get_content_subtype() == 'html':
23         payload = msg.get_payload(decode=True)
24         body += nettoyage.clear_html(payload.decode(errors='ignore'))
25
26     if msg.get_content_subtype() == 'enriched':
27         payload = msg.get_payload(decode=True)
28         body += nettoyage.clear_enriched(payload.decode(errors='ignore'))
29
30     return body

```

---

### 2.2.3 Nettoyage

Le nettoyage du texte utilise principalement les expressions régulières pour retirer un maximum d'éléments indésirables dans le texte. J'utilise 2 modules externes afin de traiter le code HTML et faire la détection des mails qui ne sont pas écrit en anglais.

**Par regex** J'utilise le module python *re* pour générer les suivantes :

**Suppression des réponses** Lorsque l'on réponds à un mail, le texte du message précédent est conservé dans le corps du mail. Afin de permettre la distinction avec les mails précédent le caractère '>' est ajouter en début de ligne. Je retire toutes les lignes correspondant à des réponses afin de limiter les doublons.

#### Nettoyage des réponses

```

1 def clear_reply(texte):
2     pattern = re.compile('^>.*$', flags=re.MULTILINE)
3     return re.sub(pattern, '', texte)

```

---

**Suppression des ponctuations** Afin de ne pas surcharger la base de données et pour se concentrer sur le texte, une grande partie des caractères de ponctuation seront retirés. L'idée est de se concentrer sur les ponctuations les plus présentes (.,?!)

#### Nettoyage des ponctuations

```

1 def clear_punctuation(texte):
2     pattern_ponct = re.compile('[*#\_\-=;<>\\[\]\\"'\~)(|/$+}{@%&\\\\]',
3     flags=re.MULTILINE)
4     return re.sub(pattern_ponct, '', texte)

```

---

**Suppression des balises pour les enriched text** Certaines parties du corps de mail sont de type *enriched text*. Les balises ne sont pas pertinente dans notre analyse et sont donc retirées.

## Nettoyage des balises enriched text

```
1 def clear_enriched(texte):
2     pattern = re.compile('<.*>')
3     return re.sub(pattern, '', texte)
```

**Suppression des liens** Certaines informations présentent dans le texte ne peuvent pas être utilisées dans l'analyse textuelle. Cependant il peut être intéressant de conserver une trace de leur présence. Nous allons donc modifier les liens url, mail et les numéros de téléphone qui seront comptabilisé avant d'être retiré du texte.

## Nettoyage des liens

```

1 def change_lien(texte , liens):
2     pattern_mail = re.compile(' [a-zA-Z0-9_+~]+@[a-zA-Z0-9_-]+\.[a-zA-Z0-9_-]+\.' )
3
4     pattern_url1 = re.compile(' (http|ftp|https)?://([w|-]+(?:[.:\\w|-]+@)?)?/?'
5                                , flags=re.MULTILINE)
6     pattern_url2 = re.compile(' (\\w+\\.)+\\w+', flags=re.MULTILINE)
7     pattern_tel1 = re.compile(' \\(\\d{3}\\)\\d+-\\d+' ) # (359)1234-1000
8     pattern_tel2 = re.compile(' \\d+( [.-]?\\d)+' ) # +34 936 00 23 23
9
10    temp, liens['MAIL'] = re.subn(pattern_mail , ' ', texte)
11
12    temp, liens['URL'] = re.subn(pattern_url1 , ' ', temp)
13    temp, nb = re.subn(pattern_url2 , ' ', temp)
14    liens['URL'] += nb
15
16    temp, liens['TEL'] = re.subn(pattern_tel1 , ' ', temp)
17    temp, nb = re.subn(pattern_tel2 , ' ', temp)
18    liens['TEL'] += nb
19
20    return temp

```

**Suppression des nombres** Comme pour les liens, les nombres sont comptabilisés et retirés. Je fais la distinction entre les nombres seuls et les nombres accompagnés de sigle monétaires.

```
monnaie = '€$£'
```

## Nettoyage des nombres

```

1 def change_nombres(texte, liens):
2     pattern_prix1 = re.compile(f'[{monnaie}]( )?\d+([\.,]\d+)? ', flags=
re.MULTILINE)
3     pattern_prix2 = re.compile(f' \d+([\.,]\d+)?( )?[monnaie]', flags=
re.MULTILINE)
4     pattern_nb = re.compile('\d+')

```



```

5
6     temp, liens['PRIX'] = re.subn(pattern_prix1, ' ', texte)
7     temp, nb = re.subn(pattern_prix2, ' ', temp)
8     liens['PRIX'] += nb
9
10    temp, liens['NOMBRE'] = re.subn(pattern_nb, ' ', temp)
11
12    return temp

```

---

**Par module** J'ai utilisé deux modules externes plus performant que ce que j'aurais pu faire avec simplement des expressions régulières.

**Suppression du code HTML** Certaines parties du corps du mail sont de type HTML. J'utilise le module *BeautifulSoup* pour parser le code et récupérer le texte affiché.

#### Nettoyage des nombres

---

```

1 from bs4 import BeautifulSoup
2
3 def clear_html(texte):
4     brut = BeautifulSoup(texte, "lxml").text
5     return brut

```

---

**Sélection des mails en anglais** Lors de mes tests, je me suis rendu compte que certains mails n'étaient pas en anglais. J'ai donc trouvé le module *langdetect* qui permet de détecter le langage utilisé dans un texte en utilisant un modèle Naïve Bayes avec une précision de 99% (voir F.2).

Je conserve dans les données à mettre en base le langage détecté dans l'idée de pouvoir traité plusieurs langues (en idée d'évolution).

La détection de la langue se fait dans la fonction s'occupant de créer le document pour la mise en base ElasticSearch.

#### Création d'un document

---

```

1 import langdetect
2
3 def create_document(mail, categorie):
4     corp = mail_load.extract_body(mail)
5     corp, liens = nettoyage.clear_texte_init(corp)
6     sujet, expéditeur = mail_load.extract_meta(mail)
7
8     if not corp:
9         return None
10
11    try:
12        lang = langdetect.detect(corp)
13    except langdetect.lang_detect_exception.LangDetectException:
14        return None
15
16    if lang != 'en':
17        return None
18

```

```

19     if categorie.lower() not in ['spam', 'ham']:
20         categorie = 'inconnu'
21
22     doc = {
23         'hash': hashlib.md5(corp.encode()).hexdigest(),
24         'categorie': categorie.lower(),
25         'sujet': sujet,
26         'expediteur': expediteur,
27         'message': corp,
28         'langue': lang,
29         'liens': liens
30     }
31     return doc

```

---

**Exemple de traitement** Les sections suivantes présentent des exemples de traitement de la phase 1.

#### Traitement initial

---

```

1  message = '''
2  Message dedicated to be a sample to show how the process is clearing the
   text.
3
4  Begin reply :
5  > He once said
6  >>> that it would be great
7  End of reply.
8
9  Substitutions :
10 spamassassin-talk@example.sourceforge.net
11 https://www.inphonic.com/r.asp?r=sourceforge1&refcode1=vs3390
12 hello.foo.bar
13 between $ 25 and 25,21 $
14
15 A number is : 2588,8 588
16 Phone type a : (359)1234-1000
17 Phone type b : +34 936 00 23 23
18 Punctuation : ——## ..
19 ~~~~~~
20 '''
21 text, liens = clear_texte_init(message)
22 print(liens)
23 print(text)
24

```

---

Résultat traitement initial :

```
{'URL': 2, 'MAIL': 1, 'TEL': 2, 'NOMBRE': 3, 'PRIX': 2}
```

Message dedicated to be a sample to show how the process is clearing the text.

Begin reply

End of reply.

Substitutions

between and

A number is ,  
Phone type a  
Phone type b  
Ponctuation ..

---

#### Traitement HTML

---

```
1 message_html = '''
2 <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
3 <html>
4 <head>
5   <title>Foobar</title>
6 </head>
7 <body>
8   I actually thought of this kind of active chat at AOL
9   bringing up ads based on what was being discussed and
10  other features
11   <pre wrap="">On 10/2/02 12:00 PM, "Mr. FoRK"
12   <a class="moz-txt-link-/rfc2396E" href="mailto:fork_
13   list@hotmail.com">&lt;fork_list@hotmail.com&gt;</a>
14   wrote: Hello There, General Kenobi !?
15 <br>
16 </body>
17 </html>
18 '''
19 print(clear_html(message_html))
20
```

---

Résultat traitement HTML :

Foobar

I actually thought of this kind of active chat at AOL  
bringing up ads based on what was being discussed and  
other features  
On 10/2/02 12:00 PM, "Mr. FoRK"  
<fork\_list@hotmail.com>  
wrote: Hello There, General Kenobi !?

---

#### Traitement enriched text

---

```

1 message_enriched = '''
2 <smaller>I'd like to swap with someone also using Simple DNS to take
3 advantage of the trusted zone file transfer option.</smaller>
4 '''
5 print(clear_enriched(message_enriched))
6

```

---

Résultat traitement enriched text :

I'd like to swap with someone also using Simple DNS to take  
 advantage of the trusted zone file transfer option.

## 2.2.4 Mise en base

Cette section détaille les éléments relatifs à la mise en base des informations récoltées. Dans ce projet, j'utilise 2 moteurs de bases de données pour stocker les extractions des mails.

1. un index Elasticsearch pour faire le stockage des données textuelles
2. une base PostgreSQL pour le stockage des données numériques

J'utilise des conteneurs *docker* pour héberger les services de bases de données. L'utilisation des conteneurs me permet de partager plus facilement mes configurations et limite les erreurs d'installations.

Pour chaque mail récolté le programme de la phase 1 va générer un *document* avec les informations suivantes :

- hash - signature md5 du texte nettoyé
- catégorie - Ham, Spam ou Inconnu
- sujet - correspond à l'objet du mail
- expéditeur - adresse mail
- corps - corps du mail nettoyé
- langue - la langue détectée du mail (en)
- liens - données non textuelles extraites du corps :
  - URL - liens URL
  - Mail - adresses mail
  - Téléphone - numéros de téléphone
  - Prix - nombres avec un symbole de devise
  - Nombres

Chaque document va générer une entrée dans la base Elasticsearch et une entrée dans la base PostgreSQL.

### Création d'un document

---

```

1 def create_document(mail, categorie):
2     corp = mail_load.extract_body(mail)
3     corp, liens = nettoyage.clear_texte_init(corp)
4     sujet, expediteur = mail_load.extract_meta(mail)
5
6     if not corp:
7         return None
8
9     try:
10        lang = langdetect.detect(corp)
11    except langdetect.lang_detect_exception.LangDetectException:
12        return None

```

```

13
14     if lang != 'en':
15         return None
16
17     if categorie.lower() not in ['spam', 'ham']:
18         categorie = 'inconnu'
19
20     doc = {
21         'hash': hashlib.md5(corp.encode()).hexdigest(),
22         'categorie': categorie.lower(),
23         'sujet': sujet,
24         'expediteur': expediteur,
25         'message': corp,
26         'langue': lang,
27         'liens': liens
28     }
29     return doc
30

```

Ci-dessous le schéma des bases de données avec les relations entre elles.

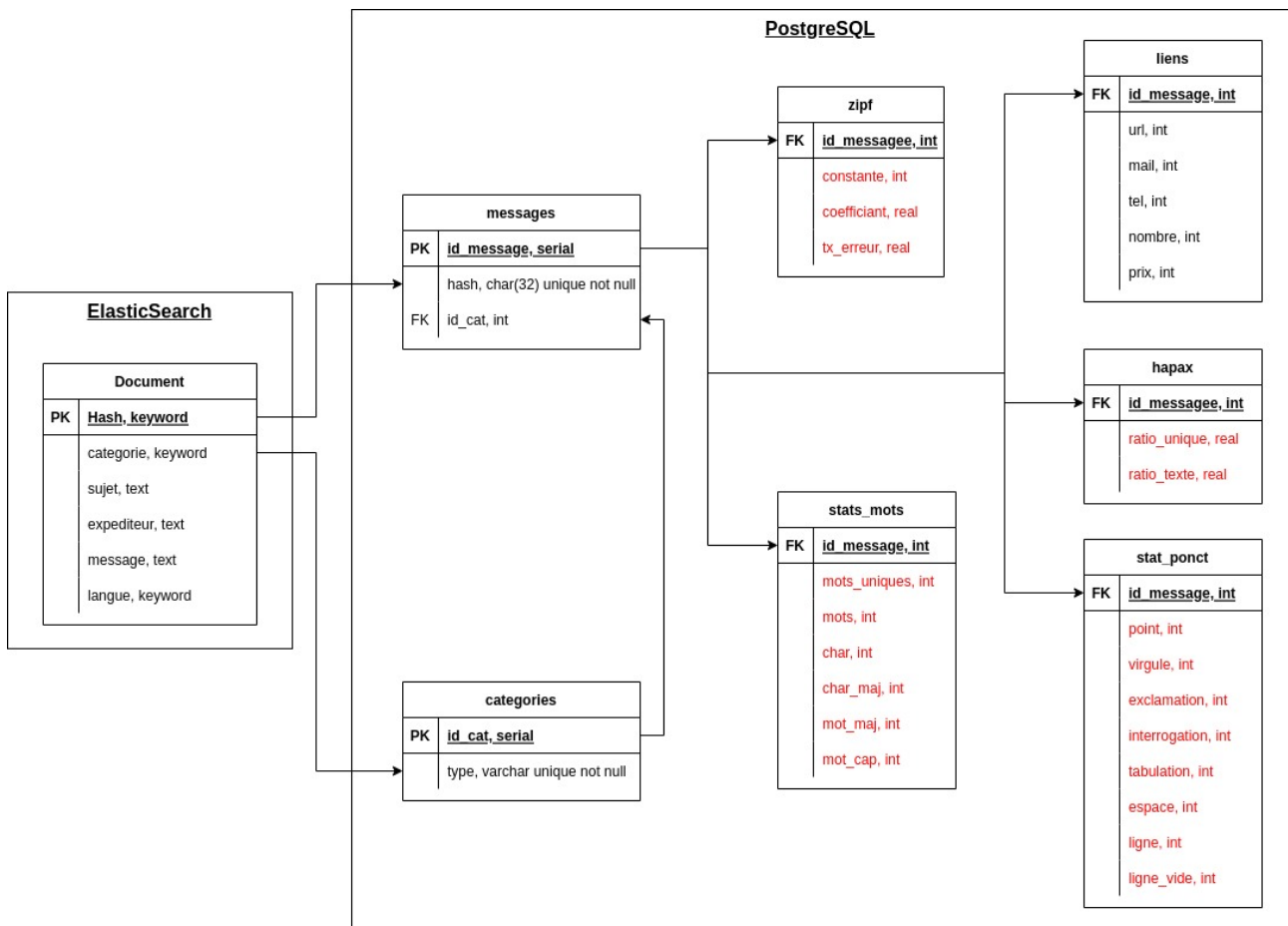


FIGURE 3 – Schéma des bases de données de l'application

Le *hash* calculé lors de la phase de traitement est l'identifiant unique du mail dans toutes les bases. La catégorie du mail est également présente dans les deux bases. Les champs en rouge sont des caractéristiques qui ne sont pas calculées lors de la phase 1.

**Stockage des données : Elasticsearch** Elasticsearch est un moteur de base de données NoSQL. Il intègre un moteur d'indexation des documents est assez performant pour stocker des données textuelles de taille aléatoire. Il est cependant assez compliqué de modifier le schéma d'un index une fois qu'il a été créé. Pour ces raisons, j'utilise cette technologie pour ne stocker les données textuelles après nettoyage car elles n'ont plus vocation à être modifiées.

Les corps de mail mise en base seront récupérés ultérieurement pour réaliser les opérations d'analyse. Les résultats seront stockés dans la base PostgreSQL plus souple.

Ci-dessous les fonctions principales pour l'ajout d'un document dans la base ES.

#### Fonctions basique pour la liaison Elasticsearch

---

```
1 def es_connect(server, creds, crt):
2     """ Connexion au serveur Elasticsearch """
3     client = Elasticsearch(server, api_key=creds, ca_certs=crt)
4
5     try:
6         client.search()
7         client.indices.get(index="*")
8         return client
9
10    except (exceptions.ConnectionError, AuthenticationException,
11            AuthorizationException) as err:
12        print("ES:conn - Informations client Elasticsearch :\n\t", err)
13        client.close()
14        return None
15
16 def es_index_doc(es_cli, index, doc):
17     """ Index un document dans la base ES """
18     id_doc = doc['hash']
19
20     if es_document_exists(es_cli, index, id_doc):
21         return 1
22
23     es_cli.index(index=index, document=doc)
24     es_cli.indices.refresh(index=index)
25     return
26
27
28 def es_document_exists(es_cli, index, hash):
29     """ Regarder dans l'index si le hash du document est deja present """
30     try:
31         resp = es_cli.search(index=index, query={"match": {"hash": hash}})
32     except elasticsearch.NotFoundError as err:
33         print("Error : hash", err, file=sys.stderr)
34         return None
35
36     return True if resp['hits']['total']['value'] == 1 else False
37
```

---

Le déploiement et les configurations de la base Elasticsearch sont disponibles dans l'annexe B.1.

**Stockage des premières informations statistiques : PostgreSQL** Pour le stockage des données statiques et d'analyse, j'ai décidé de m'orienter vers le système de gestion de base de données PostgreSQL. Une base de données relationnelle est plus flexible qu'un index Elastic-Search pour l'ajout de nouvelles caractéristiques.

Durant la phase 1, j'utilise que 3 tables :

- messages - liste des mails avec le hash permettant de faire le lien avec l'index ES
- categorie - liste des catégories de mails
- liens - chaque ligne contient toutes les nombres de liens retiré lors du nettoyage

Cette base a pour but de stocker les données formatées pour l'analyse, l'exploitation et l'entrainement du modèle.

#### Fonctions basiques pour la liaison PostgreSQL

---

```
1 def connect_db(database, user, passwd, host, port):
2     """ Connexion a la base de donnees Postgres """
3     try:
4         client_psql = psycopg2.connect(database=database, user=user,
5         password=passwd, host=host, port=port)
6     except psycopg2.Error as e:
7         print("Erreur de connexion : \n{}".format(e), file=sys.stderr)
8         return None
9
10    client_psql.autocommit = True
11    return client_psql
12
13 def insert_data(client_psql, table, data):
14     """
15     Insere les donnees d'un dictionnaire dans une table de la base de
16     donnees PSQL
17     Les cles du dictionnaire doivent correspondre aux colonnes de la table
18     """
19     cols = ','.join([str(c) for c in data.keys()])
20     vals = ','.join([str(v) if (type(v) != str) else f"'{v}'" for v in
21     data.values()])
22     query = f"INSERT INTO {table}({cols}) VALUES ({vals})"
23
24     exec_query(client_psql, query)
25
26 def get_data(client_psql, table, champs, clause=None):
27     """ Recupere les donnees de la base. """
28     query = f"SELECT {'','.join(champs)} FROM {table}"
29     if clause:
30         query += f" WHERE {clause}"
31
32     result = exec_query(client_psql, query)
33     return [dict(zip(champs, ligne)) for ligne in result]
34
35 def insert_document_init(client_psql, data, id_cat):
36     """ Insere un nouveau document dans la base PSQL. """
```

```

37     insert_data(client_psql, 'messages', {'hash': data['hash'], 'id_cat':
id_cat})
38     id_message = get_data(client_psql, 'messages', ['id_message'], f"hash
LIKE '{data['hash']}'")[0]['id_message']
39
40     liens = data['liens']
41     liens.update({'id_message': id_message})
42     insert_data(client_psql, 'liens', liens)
43
44 def exec_query(client_psql, query):
45     """ Execute une query dans la base PSQL """
46     cursor = client_psql.cursor()
47
48     try:
49         cursor.execute(query)
50         if query.upper().find("SELECT", 0, 6) >= 0:
51             return cursor.fetchall()
52         return []
53     except psycopg2.Error as e:
54         print("Erreur d'execution de la requete : {}".format(e), file=sys.
stderr)
55         print("requete : {}".format(query), file=sys.stderr)
56         return []
57

```

Le déploiement et les configurations de la base PostgreSQL sont disponibles dans l'annexe B.2.

**Stockage des données statistiques du traitement : SQLite** Les données présentes dans cette base permettent de suivre l'évolution du traitement lors des différentes étapes de nettoyage. A chaque grandes étapes de la phase 1 (Importation, Nettoyage, Mise en base), je calcule pour les HAM, SPAM et (HAM+SPAM) les éléments suivants :

- mails - nombre de mails
- mots - nombre de mots
- mots\_uniques - nombre de mots uniques

Ces données me permettent d'estimer la quantité de données nettoyées durant cette phase.

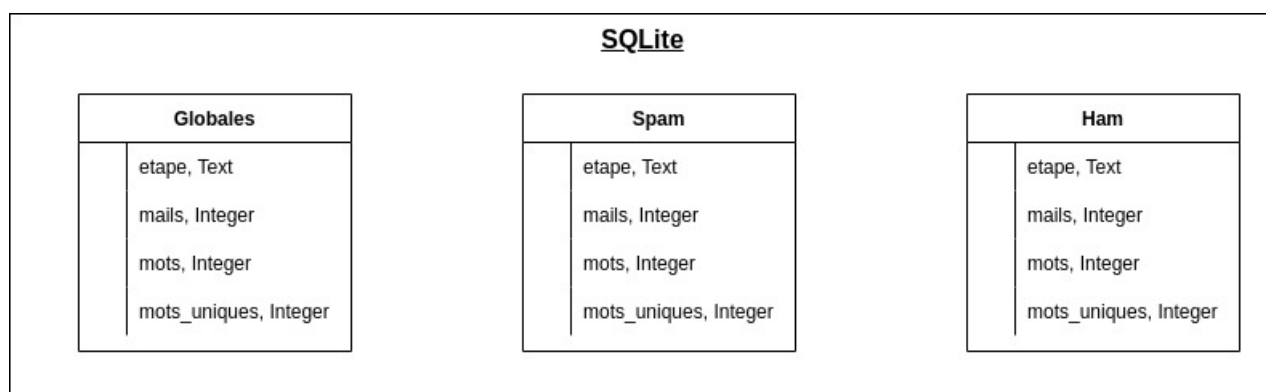


FIGURE 4 – Schéma de la base de données pour lors du traitement



## 2.3 Données de la phase 1

Sortie du programme : main\_collecte.py

```
=== Vérification des conteneurs Docker ===
* Conteneur 'docker-es01-1'... OK
* Conteneur 'docker-kibana-1'... OK
* Conteneur 'docker_pgadmin_1'... OK
* Conteneur 'docker_pgdb_1'... OK
=== Phase 1 : collecte & mise en base ===
== Création de la base SQLITE
SQLITE table globales : CREATED
SQLITE table spam : CREATED
SQLITE table ham : CREATED
== Recolte ==
-- Création de la liste des fichiers... OK
--- Process de statistiques après la récolte
-- Stats - étape : Récolte spam... OK
-- Stats - étape : Récolte ham... OK
--- Sauvegarde des stats de l'étape: recolte... OK
Données stats de l'étape: recolte:
HAM, mails: 4153 mots: 1863674 mots uniques: 178812
SPAM, mails: 1898 mots: 918256 mots uniques: 129601
GLOBALES, mails: 6051 mots: 2781930 mots uniques: 287637
== Création document ==
-- Importation - Création spam... OK
-- Importation - Création ham... OK
--- Process de statistiques après la création de document
-- Stats - étape : création documents spam... OK
-- Stats - étape : création documents ham... OK
--- Sauvegarde des stats de l'étape: creation document... OK
Données stats de l'étape: creation document:
HAM, mails: 4124 mots: 844151 mots uniques: 73725
SPAM, mails: 1784 mots: 584012 mots uniques: 40099
GLOBALES, mails: 5908 mots: 1428163 mots uniques: 97235
== Mise en base des documents ==
-- Création de l'index ElasticSearch... OK
-- Création de la base PostgreSQL
User 'data' déjà existant
-- Création des tables PostgreSQL... OK
-- Mise en base ES & PSQL des spam... OK (256 doublons)
-- Mise en base ES & PSQL des ham... OK (78 doublons)
-- Récupération des spam... OK
-- Récupération des ham... OK
--- Process de statistiques après la mise en base
-- Stats - étape : Mise en base spam... OK
-- Stats - étape : Mise en base ham... OK
--- Sauvegarde des stats de l'étape: mise en base... OK
Données stats de l'étape: mise en base:
HAM, mails: 4046 mots: 839368 mots uniques: 73725
SPAM, mails: 1528 mots: 529121 mots uniques: 40099
```

GLOBALES, mails: 5574 mots: 1368489 mots uniques: 97235  
 == FIN ==

**Évolution lors du traitement** J’effectue une petite analyse sur le volume de données conservés au cours des différentes étapes de la phase 1. Je vérifie pour chaque catégorie (ham, spam) ainsi qu’en globalité les éléments suivants :

- nombre de documents conservés
- nombre de mots (n-gram) dans le corpus
- nombre de mots (n-gram) uniques dans le corpus

Texte provisoire :

Il est alors possible de constater les comportements suivants :

- La diminution des documents spam est plus importante que celle des ham
- Le nombre de mots uniques ne diminue plus après la création de document
- La diminution du nombre de mots est plus importante dans les ham que dans les spam.
- le nombre de mot uniques est plus important dans les ham que dans les spam

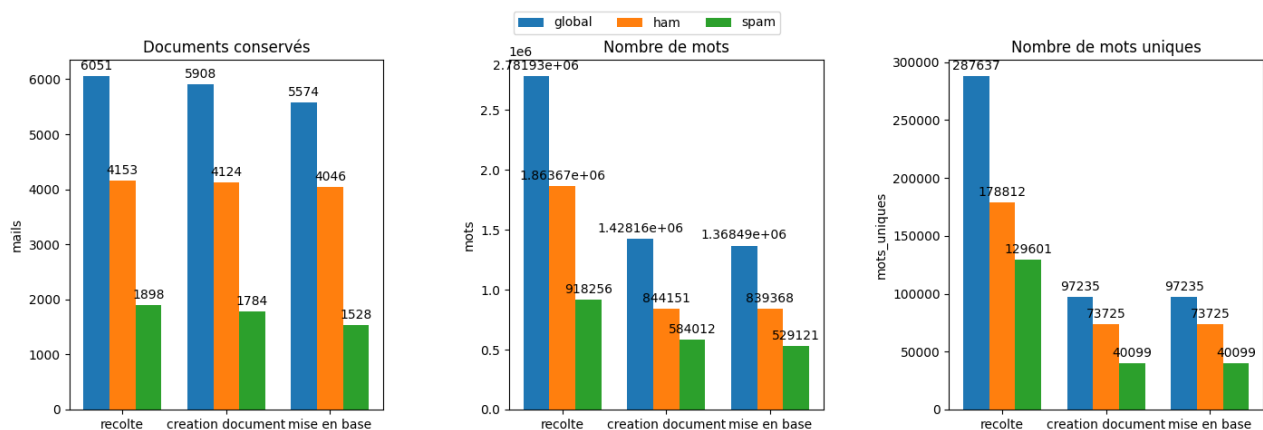


FIGURE 5 – Evolution des principaux marqueurs lors du traitement

**Conclusion de la récolte** A l’issue de la phase de récolte, on remarque qu’il y a un nombre plus important de document en double dans la catégorie *spam*. Il y a également une réduction importante du nombre de mots ainsi que du nombre de mots uniques dans les *ham*. Cette réduction peut s’expliquer par le nettoyage du corps des mails :

- Retrait des réponses
- Retrait de certaines ponctuations
- Suppression des balises HTML et enriched text
- Retrait des liens, et nombres

Enfin on peut voir que les *ham* utilise plus de mots uniques que les *spam*. Il est donc possible que le vocabulaire des *spam* soit plus restreint.

**Premières données statistiques** Les documents conservés à l’issue de cette phase de récupération se constitue à 72,6% de Ham et à 27.4% de Spam.

## Répartition des ham/spam

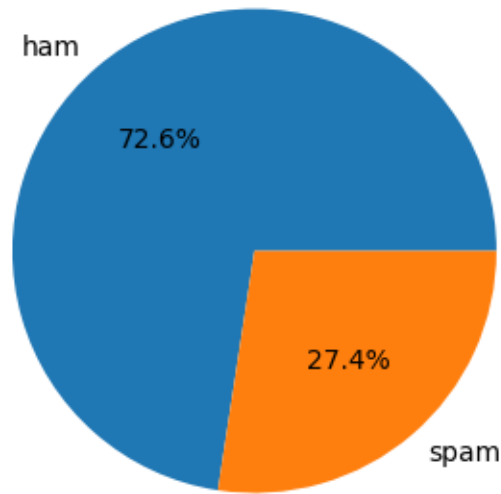


FIGURE 6 – Répartition des ham/spam dans le dataset

A partir des informations et des pré-traitement réalisés lors de cette première phase, il est possible de présenter quelques données statistiques. Notamment sur la présence et l'utilisation de liens (url, mail, téléphonique) et des données numériques (nombres et prix).

TABLE 1 – Statistiques sur les liens

	url			mail			tel		
	Globale	Ham	Spam	Globale	Ham	Spam	Globale	Ham	Spam
moyenne	5.47	5.54	5.26	1.16	1.21	1.01	0.16	0.20	0.64
écart-type	49.74	57.77	13.71	2.50	1.72	3.88	1.06	0.63	1.73
minimum	0			0			0		
25%	1			0			0		
médiane	2	2	3	1	1	0	0		
75%	4	4	5	2	2	1	0		
maximum	3557	3557	295	69	34	69	67	25	67

A partir des données des liens présentées dans le tableau ci dessus il est possible d'émettre les hypothèses suivantes.

On trouve en moyenne environ 5 url dans les deux types de mails. Cependant la dispersion est notablement plus faible pour les spams. On voit également que 75% des spam contiennent 5 liens ou moins.

La présence d'adresse mail est assez marginale environ 1 adresse mail par corps de texte. L'écart-type est plus faible dans les ham. On voit également que 75% des ham contiennent 2 adresse mails ou moins contre 1 pour les spam.

Enfin la présence de numéro de téléphone est quasiment nulle pour les deux catégories.

TABLE 2 – Statistiques sur les données numériques

	nombre			prix		
	Globale	Ham	Spam	Globale	Ham	Spam
moyenne	15.14	9.38	30.38	0.28	0.08	0.81
écart-type	219.16	49.35	410.53	1.54	0.66	2.66
minimum	0			0		
25%	1	1	2	0		
médiane	5	5	7	0		
75%	9	8	14	0		
maximum	15801	2794	15801	29	17	29

Les données numériques du dataset montrent les hypothèses suivantes.

Les nombres simples semblent être les données les plus présentes dans le corps du mail, environ 15 en moyenne par mail, cependant la dispersion est élevée. Cela étant 75% des spam ont 14 nombres ou moins contre 8 ou moins pour les ham.

Enfin la présence de prix est assez marginale par rapport aux nombres. Il y a cependant en moyenne 10 fois plus de prix dans les spam que les ham.

Distribution des mails en fonction du nombre de liens

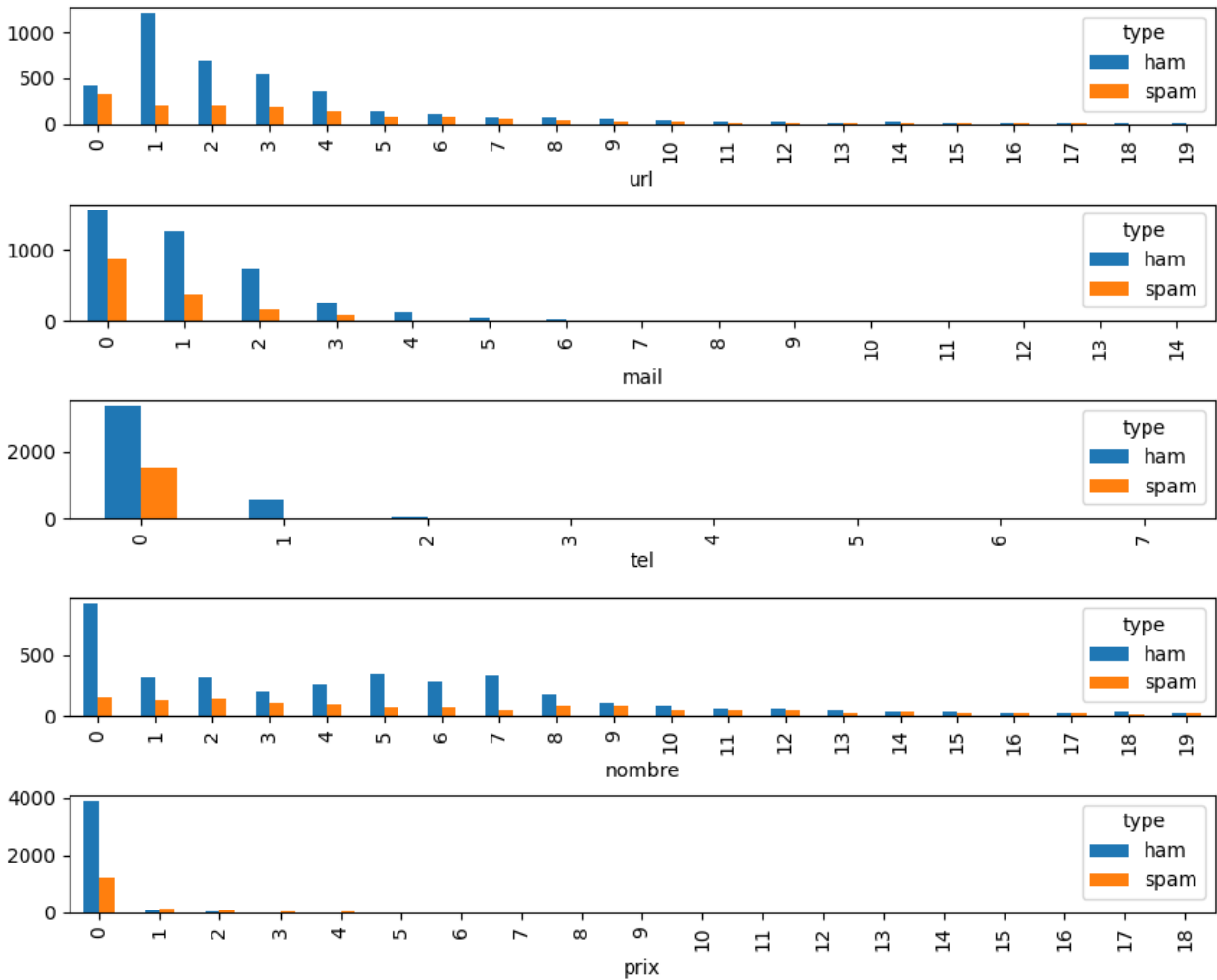


FIGURE 7 – Distribution du nombre de liens selon la catégorie du mail

Au vu des analyses précédents, on peut dire qu'il est statistiquement plus probable de trouver dans les spam plus de liens url, de nombre et de prix que dans les ham. La présence d'adresse mail est plus probable dans les ham. Il est enfin possible d'exclure la présence de numéro de téléphone qui est trop marginale.

Les codes sources de cette partie sont disponibles dans le fichier `./analyse/analyse_p1.py` ou dans la section G.1

## 3 Phase 2

### 3.1 Traitement

#### 3.1.1 Recherche de caractéristiques

#### Références

#### 3.1.2 Analyse préliminaire

4    Phase 3

# A Développement visualisation distribution de Zipf

**Présentation** La loi de distribution de Zipf est une loi empirique (basée sur l'observation) qui veut que le mot le plus fréquent est, à peu de chose près, 2 fois plus fréquent que le 2<sup>ème</sup>, 3 fois plus fréquent que le 3<sup>ème</sup> etc.

La formulation finale de la 1<sup>ère</sup> loi de Zipf est la suivante :

$$|mot| = constante \times rang(mot)^{k \approx 1}$$

avec  $|mot|$  la fréquence d'apparition d'un mot, *constante* une valeur propre à chaque texte,  $rang(mot)$  la place du mot dans le tri décroissant par fréquence d'apparition et  $k$  un coefficient proche de 1.

**Développement** Afin de pouvoir utiliser les résultats de cette distribution dans mon analyse, j'ai développé un ensemble de fonctions sur un corpus "*reconnu*". Mon choix s'est porté sur le corpus *Brown* (voir F.1) présent dans la librairie *nltk*. Ce corpus contient environ 500 documents contenant 1 millions de mot en anglais.

Le processus d'analyse se fait sur 2 versions de ce corpus.

— la première version contient tous les mots sans modifications

— la seconde version contient tous les mots sans les *stopwords*

Les *stopwords* sont des mots qui n'ont pas ou peu de signification dans un texte. Ces mots sont retirés dans la 2<sup>e</sup> version pour voir l'effet d'une réduction sur la distribution de Zipf.

Les paragraphes ci-dessous détaillent les étapes du développement :

**Étape 1 - Ordonner les mots** La première étape est de compter les occurrences de tous les mots des 2 corpus et de les ranger en fonction de leur nombre d'occurrence.

## Triage des mots

```
1 def frequence_mot(bag, freq=None):
2     """
3     Calcule la frequence de chaque mot dans un sac de mot
4     :param bag: <list> – liste de tous les mots d'un texte
5     :param freq: <dict> – dictionnaire avec {<str> mot: <int> frequence}
6     :return: <dict> – dictionnaire avec la frequence par mot {mot:
7     frequence}
8     """
9     if freq is None:
10         freq = {}
11     for mot in bag:
12         freq[mot] = freq.get(mot, 0) + 1
13     return freq
14
15 def classement_zipf(dico):
16     """
17     Trie un dictionnaire de mots : occurrence et leur assigne un rang en
18     fonction du nombre d'occurrence
19     :param dico: <dict> dictionnaire de mot: occurrences
20     :return: <list> {"rang": <int>, "mot": <str>, "frequence": <int>}
```



```

19     """
20     ranked = []
21     for rang, couple in enumerate(sorted(dico.items(), key=lambda item:
22     item[1], reverse=True), start=1):
23         ranked.append({"rang": rang,
24                         "mot": couple[0],
25                         "frequence": couple[1]})
26     return ranked

```

On obtient les représentations suivantes :

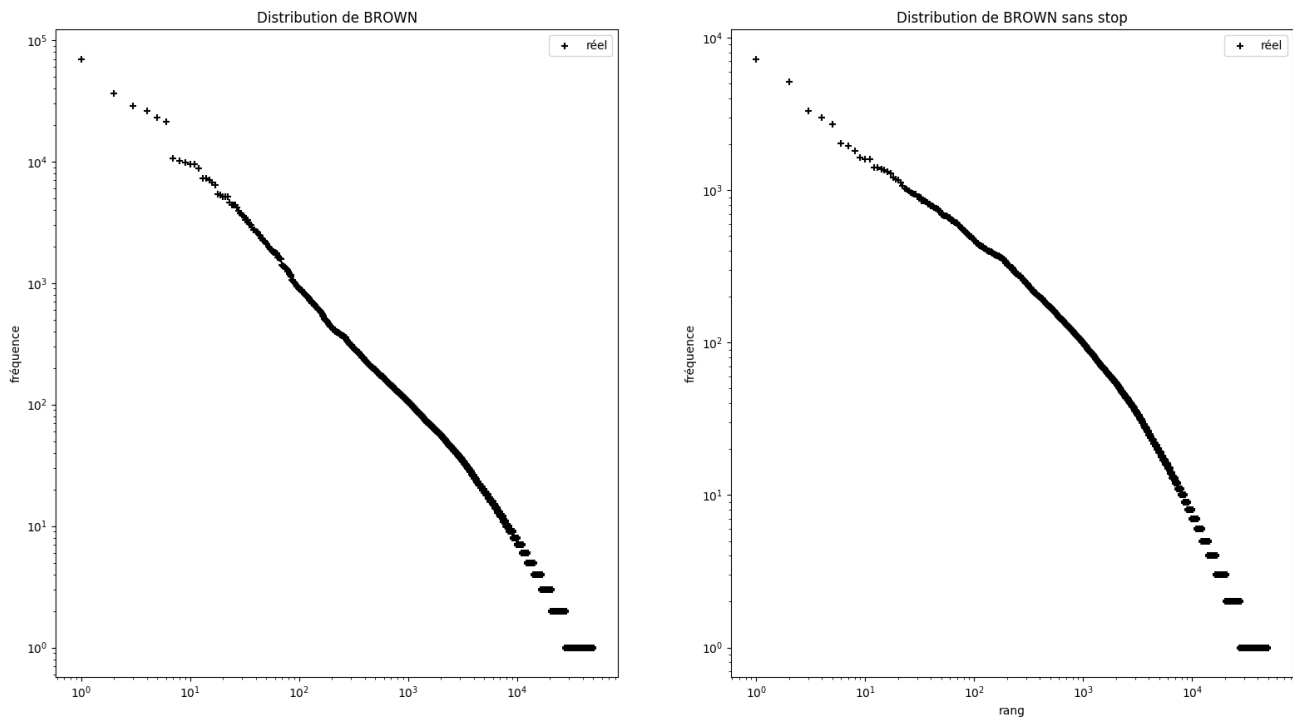


FIGURE 8 – Distribution de Zipf pour les deux corpus

- Nombre de mots dans brown : mots : 49398 occurrences : 1012528
- Nombre de mots dans brown stop : mots : 49383 occurrences : 578837

La distribution de la version complète du corpus semble à première vue plus fidèle à la représentation classique de la distribution de Zipf.

**Etape 2 - calcul de la constante** Le premier paramètre que je détermine est la *constante*. Pour ce faire j'effectue le calcul suivant pour tous les mots :

$$constante = |mot| \times rang(mot)$$

On obtient une liste de toutes les constantes théoriques pour chaque mot selon son rang. De cette liste, nous allons extraire la moyenne et la médiane.

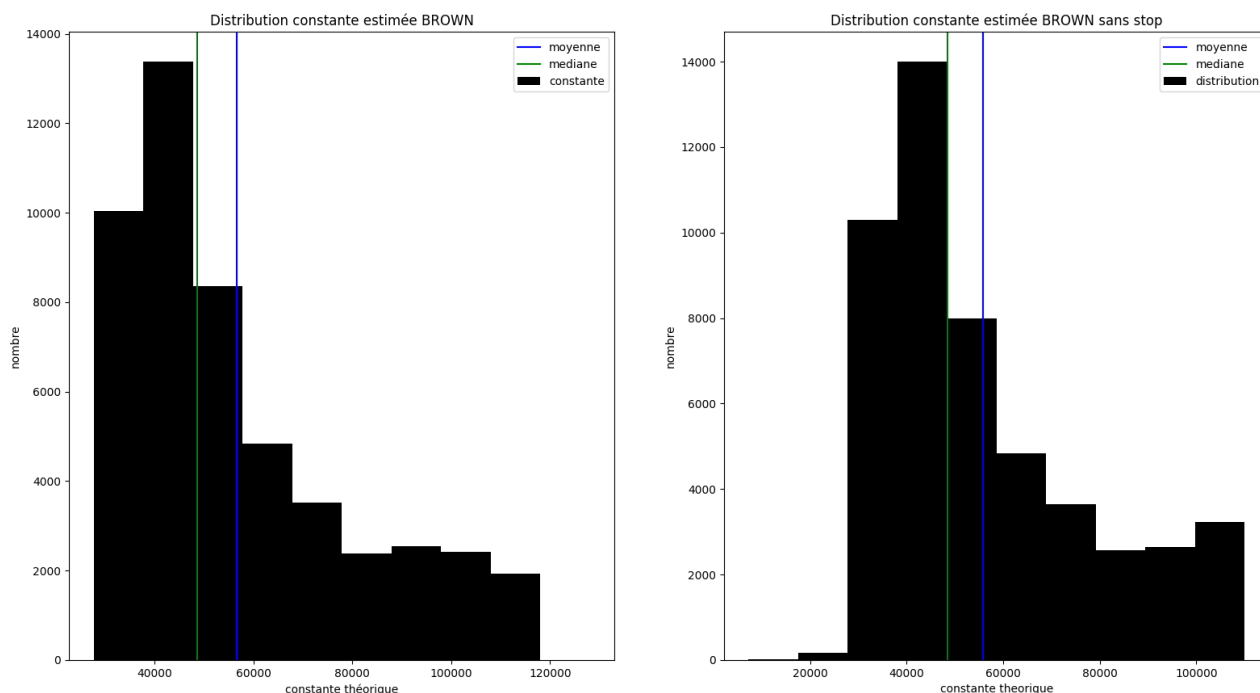


FIGURE 9 – Distribution des constantes théoriques pour les deux corpus

On voit qu'il y a une majorité de mots donnent une constante brute comprise entre 20.000 et 60.000. Dans les deux corpus La différence entre les moyennes et médianes des deux corpus n'est pas flagrante :

- Brown moyenne : 56525.81, médiane : 48601.50
- Brown (- stopwords) moyenne : 55809.97, médiane : 48494.00

**Etape 3 - recherche du coefficient** Le coefficient  $k$  permet d'ajuster le résultat, et pourra éventuellement donner une indication de complexité. La recherche de  $k$  se fera sur les deux corpus avec utilisant les moyennes et médianes.

Pour ce faire nous allons :

1. Faire la liste de tous les coefficients possibles dans l'intervalle  $[0.86, 1.3]$  avec un pas de  $0.01$ <sup>1</sup>.
2. Calculer toutes la fréquences théoriques de tous les rangs avec tous les coefficients possibles avec en constante la moyenne et la médiane de chaque corpus.
3. Calculer la moyenne des coûts absolus entre les fréquences théoriques par coefficient avec la fréquence réelle observée pour chaque corpus.

Le couple coefficient/constante avec le coup minimal sera retenu pour l'utilisation dans la phase de *feature engineering*.

#### Fonctions utilisées dans la recherche du coefficient

```

1 def zipf_freq_theorique(constante, rang, coef):
2     """
3     Calcul la frequence theorique d'un mot selon son rang, la constante du
    texte et un coeficiant d'ajustement

```

1. les bornes et le pas sont totalement arbitraire afin d'obtenir un graphique présentable

```

4      :param constante: <int> constante determinee par la distribution de
      Zipf
5      :param rang: <int> rang du mot selon sa frequence
6      :param coef: <float> variable d'ajustement
7      :return: <float> frequence theorique zipfienne
8      """
9      return constante / (rang ** coef)
10
11 def cout(l1, l2, methode):
12     """
13     Calcul le cout de l'ecart entre les elements de l1 et le l2, place par
14     place
15     :param l1: <list> liste d'entier
16     :param l2: <liste> liste d'entier
17     :param methode: <str> methode de calcul du cout
18     :return: <float> cout selon methode
19     """
20     if len(l1) != len(l2):
21         print("Erreur, fonction cout: l1 & l2 de taille differente", file=
22 sys.stderr)
23         return None
24
25     if len(l1) == 0:
26         print("Erreur, fonction cout: liste vide", file=sys.stderr)
27
28     if methode.lower() not in ['absolue', 'carre', 'racine']:
29         print("Erreur, fonction cout - methode '{}' inconnue".format(
30 methode), file=sys.stderr)
31         return None
32
33     if methode.lower() == 'absolue':
34         return np.mean([abs(x-y) for x, y in zip(l1, l2)])
35
36     if methode.lower() == 'carre':
37         return np.mean([(x-y)**2 for x, y in zip(l1, l2)])
38
39     if methode.lower() == 'racine':
40         return np.sqrt(np.mean([(x-y)**2 for x, y in zip(l1, l2)]))
41
42     return None

```

---

#### Calcul des fréquences par coefficient

---

```

1     ls_coef = list(np.arange(0.86, 1.3, 0.01))
2     zbmo_th = {coef: [stats.zipf_freq_theorique(zb_const_moyen, r, coef)
3 for r in zb_rang] for coef in ls_coef}
4     zbme_th = {coef: [stats.zipf_freq_theorique(zb_const_median, r, coef)
5 for r in zb_rang] for coef in ls_coef}
6     zbmoth_cmoy = [stats.cout(zb_freq, zbmo_th[coef], 'absolue') for coef
7 in ls_coef]
8     zbmeth_cmoy = [stats.cout(zb_freq, zbme_th[coef], 'absolue') for coef
9 in ls_coef]
10
11     zbsmo_th = {coef: [stats.zipf_freq_theorique(zbs_const_moyen, r, coef)

```

```

    for r in zbs_rang] for coef in ls_coef}
8   zbsme_th = {coef: [stats.zipf_freq_theorique(zbs_const_median, r, coef
) for r in zbs_rang] for coef in ls_coef}
9   zbsmoth_cmoy = [stats.cout(zbs_freq, zbsmo_th[coef], 'absolue') for
coef in ls_coef]
10  zbsmeth_cmoy = [stats.cout(zbs_freq, zbsme_th[coef], 'absolue') for
coef in ls_coef]

```

La recherche du coefficient nous retourne les éléments suivants :

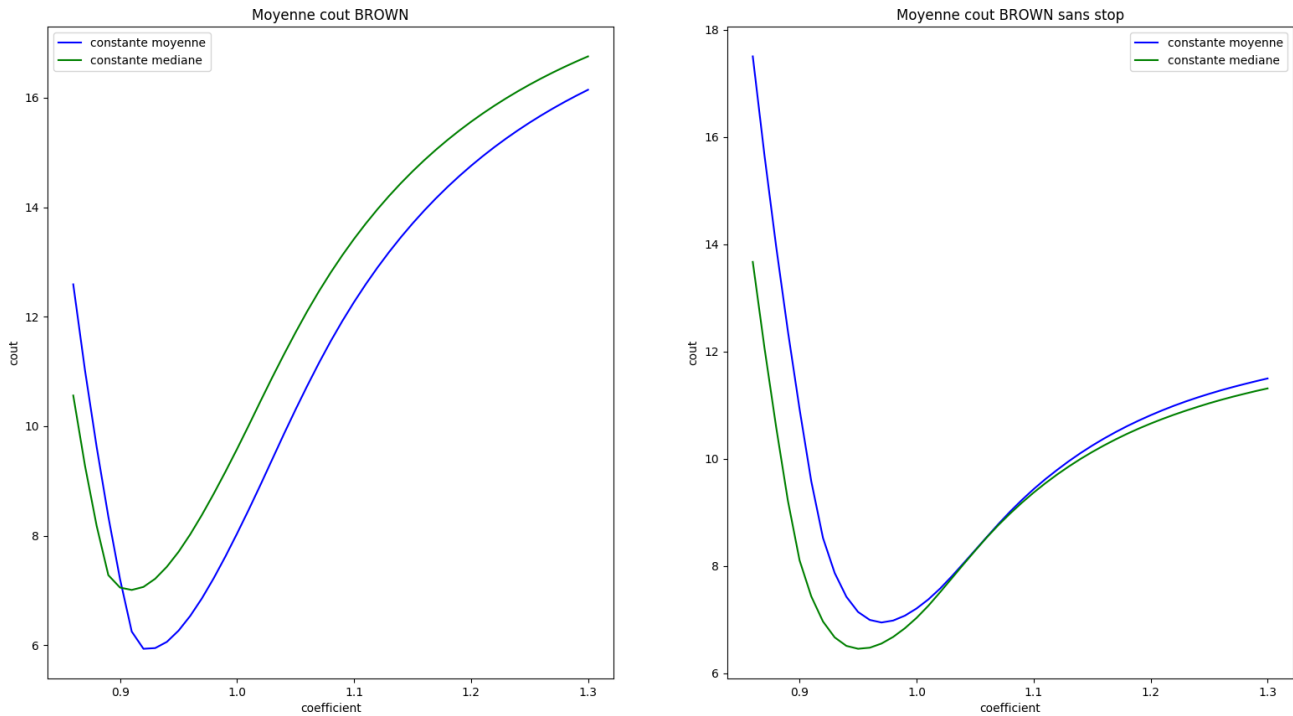


FIGURE 10 – Coût absolu moyen par coefficient

- Coût min brown moyenne : 5.93, median : 7.01
- Coût min brown (- stopwords) moyenne : 6.95, median : 6.46
- Coefficient min brown moyenne : 0.92, median : 0.91
- Coefficient min brown (- stopwords) moyenne : 0.97, median : 0.95

**Résultats** Le tableaux ci dessous rappelle les données récupérées au long de la recherche :

	BROWN avec stopwords	BROWN sans stopwords
nombre de mots uniques	49398	49383
nombre de mots total	1012528	578837
Constante moyenne	56525.81	55809.97
Constante médiane	48601.50	48494.00
Coefficient avec moyenne	0.92	0.97
Cout du coefficient moyenne	5.93	6.95
Coefficient avec médiane	0.91	0.95
Cout du coefficient médiane	7.01	6.46

D'après les données il est possible de dire que l'on obtient de meilleurs résultats si on conserve tous les mots du corpus. Dans ce cas l'utilisation de la moyenne des constantes génère un taux d'erreur plus faible que la médiane.

Ci-dessous la représentation des fréquences théoriques avec le coefficient optimal pour chaque corpus et chaque méthode. On voit que la courbe de la constante moyenne sur le corpus brute est celle qui suit le mieux les données réelles.

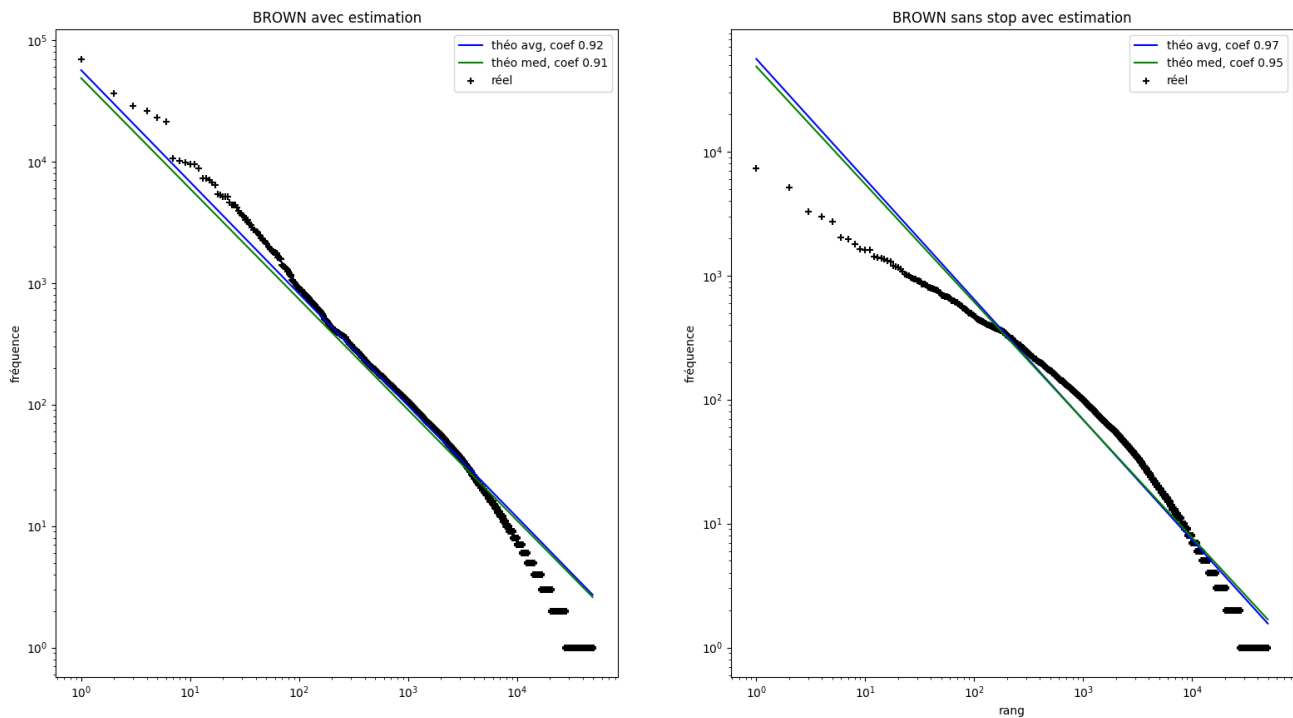


FIGURE 11 – Distribution de Zipf avec les estimations

En conclusion, j'utiliserais la moyenne des constantes sur un document complet afin de déterminer le coefficient dans ma recherche de spam.

**Notes :** L'ensemble des codes sources pour cette partie est disponible dans les fichiers :

- `./analyse/rech_zipf.py`
- `./traitement/stats.py`

## B Déploiement des bases de données

Cette annexe détaille la mise en place de l'infrastructure de base de données pour le projet. J'utilise 2 environnements de base de données en version conteneur (docker) :

- ElasticSearch
  - 1 Node Elastic, pour le service de base de données
  - 1 Service Kibana, pour la visualisation des données
  - 1 Service de certificat, pour sécuriser les échanges
- PostgreSQL
  - 1 service PostgreSQL, pour la base de données
  - 1 service PgAdmin, pour la visualisation de la base

Chaque environnement est indépendant, et possède une ouverture sur le PC hôte.

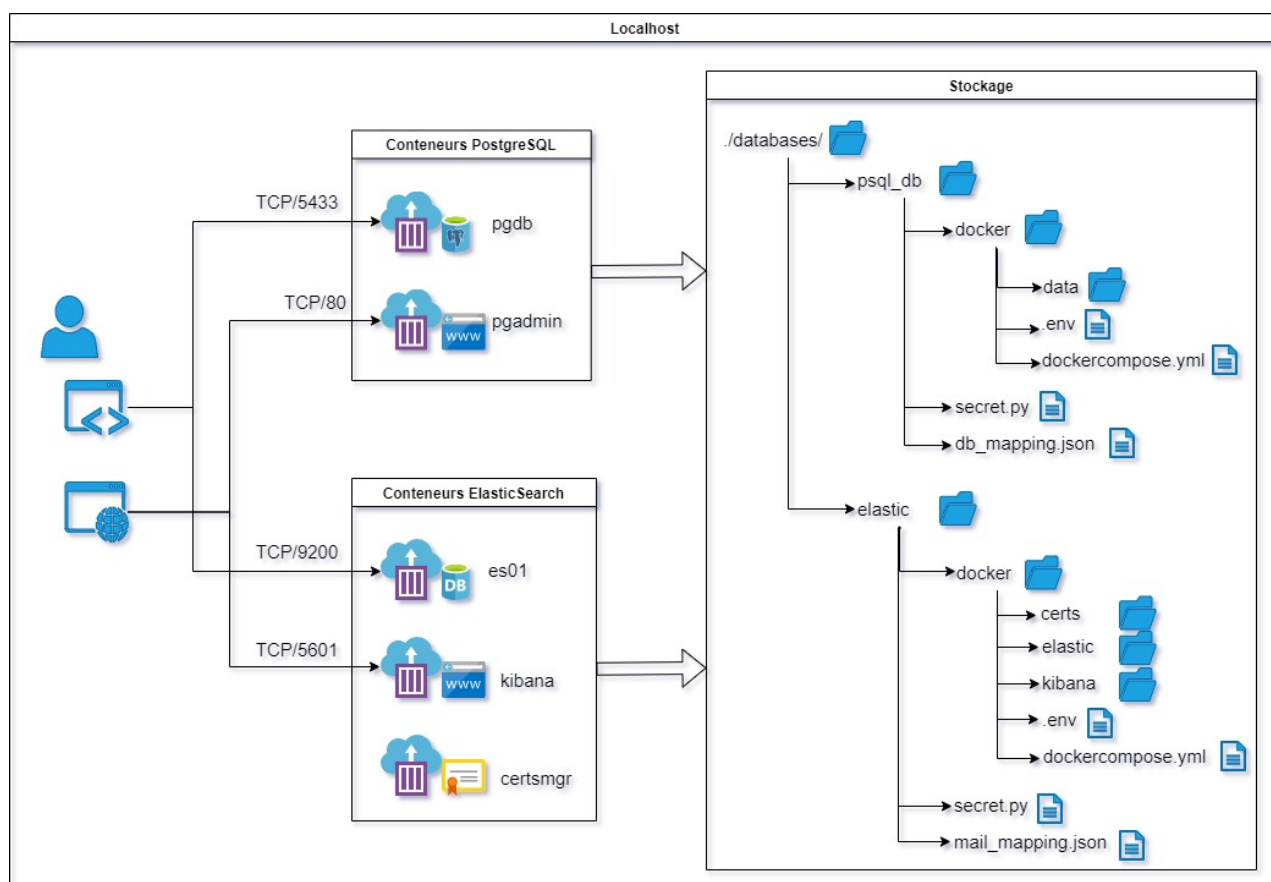


FIGURE 12 – Schéma de l'architecture Docker

### B.1 ElasticSearch

Cet environnement se lance avec la commande suivante :

```
docker-compose -f ./databases/elastic/docker/docker-compose.yml up -d
```

#### B.1.1 Conteneurisation

DockerCompose

```
1 version: "3.2"
2
3 services:
```

```

4      certsmgr:
5          image: elasticsearch:${VERSION}
6          volumes:
7              - ./certs:/usr/share/elasticsearch/config/certs
8          user: "0"
9          command: >
10             bash -c '
11                 if [ x${ELASTIC_PASSWORD} == x ]; then
12                     echo "Set the ELASTIC_PASSWORD
environment variable in the .env file";
13                     exit 1;
14                 elif [ x${KIBANA_PASSWORD} == x ]; then
15                     echo "Set the KIBANA_PASSWORD
environment variable in the .env file";
16                     exit 1;
17                 fi;
18                 if [ ! -f certs/ca.zip ]; then
19                     echo "Creating CA";
20                     bin/elasticsearch-certutil ca --
silent --pem --out config/certs/ca.zip;
21                     unzip config/certs/ca.zip -d
config/certs;
22                 fi;
23                 if [ ! -f certs/certs.zip ]; then
24                     echo "Creating certs";
25                     echo -ne \
26                         "instances:\n"\
27                         "  - name: es01\n"\
28                         "    dns:\n"\
29                         "      - es01\n"\
30                         "      - localhost\n"\
31                         "    ip:\n"\
32                         "      - 127.0.0.1\n"\
33                         > config/certs/instances.
yaml;
34                     bin/elasticsearch-certutil cert --
silent --pem --out config/certs/certs.zip --in config/certs/instances.
yaml --ca-cert config/certs/ca/ca.crt --ca-key config/certs/ca/ca.key;
35                     unzip config/certs/certs.zip -d
config/certs;
36                 fi;
37                 # echo "Setting file permissions"
38                 echo "chown -R root:root config/certs";
39                 echo "find . -type d -exec chmod 750 {\{\}
\;";
40                 echo "find . -type f -exec chmod 640 {\{\}
\;";
41                 echo "Waiting for Elasticsearch
availability";
42                 until curl -s --cacert config/certs/ca/ca.
crt https://es01:9200 | grep -q "missing authentication credentials";
do sleep 30; done;
43                 echo "Setting kibana_system password";

```

```

44         until curl -s -X POST --cacert config/
certs/ca/ca.crt -u elastic:${ELASTIC_PASSWORD} -H "Content-Type:
application/json" https://es01:9200/_security/user/kibana_system/
_password -d "{\"password\":\"${KIBANA_PASSWORD}\"} | grep -q \"^{}\";
do sleep 10; done;
45         echo "All done!";
46     ,
47     healthcheck:
48         test: ["CMD-SHELL", "[ -f config/certs/es01/es01.
crt ]"]
49         interval: 1s
50         timeout: 5s
51         retries: 120
52
53     es01:
54         depends_on:
55             certsmgr:
56                 condition: service_healthy
57         image: elasticsearch:${VERSION}
58         volumes:
59             - ./certs:/usr/share/elasticsearch/config/certs
60             - ./elastic/data:/usr/share/elasticsearch/data
61         ports:
62             - ${ES_PORT}:9200
63         environment:
64             - discovery.type=single-node
65             - ELASTIC_PASSWORD=${ELASTIC_PASSWORD}
66             - bootstrap.memory_lock=true
67             - xpack.security.enabled=true
68             - xpack.security.http.ssl.enabled=true
69             - xpack.security.http.ssl.key=certs/es01/es01.key
70             - xpack.security.http.ssl.certificate=certs/es01/
71 es01.crt
72             - xpack.security.http.ssl.certificate_authorities=
certs/ca/ca.crt
73             - xpack.security.http.ssl.verification_mode=
certificate
74             - xpack.security.transport.ssl.enabled=true
75             - xpack.security.transport.ssl.key=certs/es01/es01
.key
76             - xpack.security.transport.ssl.certificate=certs/
es01/es01.crt
77             - xpack.security.transport.ssl.
certificate_authorities=certs/ca/ca.crt
78             - xpack.security.transport.ssl.verification_mode=
certificate
79             - xpack.license.self_generated.type=${LICENSE}
80         mem_limit: ${MEM_LIMIT}
81         ulimits:
82             memlock:
83                 soft: -1
84                 hard: -1

```



```

85         healthcheck:
86             test:
87                 [
88                     "CMD-SHELL",
89                     "curl -s --cacert config/certs/ca/
ca.crt https://localhost:9200 | grep -q 'missing authentication
credentials'",
90                 ]
91             interval: 10s
92             timeout: 10s
93             retries: 120
94
95     kibana:
96         depends_on:
97             es01:
98                 condition: service_healthy
99         image: kibana:${VERSION}
100        volumes:
101        - ./certs:/usr/share/kibana/config/certs
102        - ./kibana/data:/usr/share/kibana/data
103        ports:
104        - ${KIBANA_PORT}:5601
105        environment:
106            - SERVERNAME=kibana
107            - ELASTICSEARCH_HOSTS=https://es01:9200
108            - ELASTICSEARCH_USERNAME=kibana_system
109            - ELASTICSEARCH_PASSWORD=${KIBANA_PASSWORD}
110            - ELASTICSEARCH_SSL_CERTIFICATEAUTHORITIES=config/
certs/ca/ca.crt
111        mem_limit: ${MEM_LIMIT}
112        healthcheck:
113            test:
114                [
115                    "CMD-SHELL",
116                    "curl -s -I http://localhost:5601
| grep -q 'HTTP/1.1 302 Found'",
117                ]
118            interval: 10s
119            timeout: 10s
120            retries: 120
121

```

---

#### Fichier d'environnement

---

```

1  # Password for elastic user
2  ELASTIC_PASSWORD=elastic123
3  # Password for kibana_system user
4  KIBANA_PASSWORD=kibana123
5  # Version elastic product
6  VERSION=8.1.2
7  # Licence to use
8  LICENSE=basic
9  # Port for elastic HTTP API
10 ES_PORT=9200

```

```

11 #ES_PORT=127.0.0.1:9200
12 # Port for kibana access
13 KIBANA_PORT=5601
14 # Memory available (bytes)
15 MEM_LIMIT=322122547
16

```

---

## B.1.2 Initialisation de l'index

### Exemples de secrets

```

1 serveur = "https://localhost:9200"
2 apiid = "XXXXXXXXXXXXXXXXXXXXXXXXXXXX"
3 apikey = "XXXXXXXXXXXXXXXXXXXXXXXXXXXX"
4 ca_cert = "databases/elastic/docker/certs/ca/ca.crt"
5

```

---

Il est possible de générer une clé API via l'interface Kibana > Stack Management > Sécurité > API keys > Create API Key > JSON.

### Mapping

```

1 {
2   "properties": {
3     "hash": {"type": "keyword"},
4     "categorie": {"type": "keyword"},
5     "sujet": {"type": "text"},
6     "expediteur": {"type": "text"},
7     "message": {"type": "text"},
8     "langue": {"type": "keyword"}
9   }
10 }
11

```

---

### Fonctions utiles

```

1 def es_connect(server, creds, crt):
2     """ Connexion au serveur Elasticsearch """
3     client = Elasticsearch(server, api_key=creds, ca_certs=crt)
4
5     try:
6         client.search()
7         client.indices.get(index="*")
8         return client
9
10    except (exceptions.ConnectionError, AuthenticationException,
11           AuthorizationException) as err:
12        print("ES:conn - Informations client Elasticsearch :\n\t", err)
13        client.close()
14        return None
15
16 def es_create_indice(es_cli, index, mapping):
17     """ Créer un indice s'il n'existe pas déjà """
18     indices = es_cli.indices.get(index='*')

```

```

19     if indices and index in indices:
20         print("Warning: Indice {} deja present".format(index), end=' ')
21         return
22
23     try:
24         res = es_cli.indices.create(index=index, mappings=mapping)
25     except elasticsearch.ApiError as err:
26         print(err)
27         return
28
29     if not res['acknowledged']:
30         print("Error : Echec de la creation de l'indice {}".format(index))
31

```

---

## B.2 PostgreSQL

Cet environnement se lance avec la commande suivante :

```
docker-compose -f ./databases/psql_db/docker/docker-compose.yml up -d
```

L'application se charge automatiquement de créer le compte utilisé par l'application pour se connecter à la base.

### B.2.1 Conteneurisation

#### DockerCompose

---

```

1  version: "3.3"
2  services:
3      pgdb:
4          image: postgres:${PS_VERSION}
5          restart: always
6          environment:
7              POSTGRES_PASSWORD: ${PS_PASSWORD}
8
9          volumes:
10             - ./data:/var/lib/postgresql/data
11
12         ports:
13             - ${PS_PORT}:5432
14
15         pgadmin:
16             image: dpage/pgadmin4:${PS_VERSION}
17             environment:
18                 PGADMIN_DEFAULT_EMAIL: ${PGA_MAIL}
19                 PGADMIN_DEFAULT_PASSWORD: ${PGA_PASSWORD}
20
21             ports:
22                 - ${PGA_PORT}:80
23
24             depends_on:
25                 - pgdb

```

---

#### Fichier d'environnement

---

```
1 PS_VERSION=latest
```

```

2
3 PS_PASSWORD=postgres
4 PS_PORT=5433
5
6 PGA_MAIL=data@data.org
7 PGA_PASSWORD=data
8 PGA_PORT=80
9

```

---

## B.2.2 Initialisation de la base de données

### Exemples de secrets.py

---

```

1 owner = "XXX"
2 owner_pw = "XXX"
3 admin = "XXXX"
4 admin_pw = "XXXX"
5 host = "localhost"
6 port = "5432"
7

```

---

### Mapping

---

```

1 {
2   "mail_features": {
3     "categories": {
4       "id_cat": ["SERIAL", "PRIMARY KEY"],
5       "type": ["VARCHAR", "UNIQUE", "NOT NULL"]
6     },
7     "messages": {
8       "id_message": ["SERIAL", "PRIMARY KEY"],
9       "hash": ["CHAR(32)", "UNIQUE", "NOT NULL"],
10      "id_cat": ["INT", "NOT NULL"],
11      "fk": ["fk_message", "id_cat", "categories(id_cat)", "SET NULL"]
12    },
13    "liens": {
14      "id_message": ["INT"],
15      "url": ["INT"],
16      "mail": ["INT"],
17      "tel": ["INT"],
18      "nombre": ["INT"],
19      "prix": ["INT"],
20      "fk": ["fk_liens", "id_message", "messages(id_message)", "CASCADE"]
21    },
22    "stats_mots": {
23      "id_message": ["INT"],
24      "mots_uniques": ["INT"],
25      "mots": ["INT"],
26      "char": ["INT"],
27      "char_maj": ["INT"],
28      "mot_maj": ["INT"],
29      "mot_cap": ["INT"],
30      "fk": ["fk_stats_mot", "id_message", "messages(id_message)", "
    CASCADE"]

```

```

31     },
32     "stat_ponct": {
33         "id_message": ["INT"],
34         "point": ["INT"],
35         "virgule": ["INT"],
36         "exclamation": ["INT"],
37         "interrogation": ["INT"],
38         "espace": ["INT"],
39         "tabulation": ["INT"],
40         "ligne": ["INT"],
41         "ligne_vide": ["INT"],
42         "fk": ["fk_stats_ponct", "id_message", "messages(id_message)", "
CASCADE"]
43     },
44     "zipf": {
45         "id_message": ["INT"],
46         "constante": ["INT"],
47         "coefficient": ["REAL"],
48         "tx_erreur": ["REAL"],
49         "fk": ["fk_zipf", "id_message", "messages(id_message)", "CASCADE"]
50     },
51     "hapax": {
52         "id_message": ["INT"],
53         "ratio_unique": ["REAL"],
54         "ratio_texte": ["REAL"],
55         "fk": ["fk_hapax", "id_message", "messages(id_message)", "CASCADE"]
56     }
57 }
58 }
59

```

---

### Fonctions utiles

---

```

1 def create_db(nom, owner, user, passwd, host, port):
2     """ Créer une nouvelle base de données """
3     client_psql = psycopg2.connect(user=user, password=passwd, host=host,
4     port=port)
5     client_psql.autocommit = True
6     cursor = client_psql.cursor()
7
8     cursor.execute("DROP DATABASE IF EXISTS {};".format(nom))
9     cursor.execute("CREATE DATABASE {};".format(nom))
10    cursor.execute("ALTER DATABASE {} OWNER TO {};".format(nom, owner))
11    client_psql.close()
12
13 def connect_db(database, user, passwd, host, port):
14     """ Connexion a la base de données Postgres. Penser a fermer la
15     connexion """
16     try:
17         client_psql = psycopg2.connect(database=database, user=user,
18         password=passwd, host=host, port=port)
19     except psycopg2.Error as e:
20         print("Erreur de connexion : \n{}".format(e), file=sys.stderr)

```

```

19         return None
20
21     client_psql.autocommit = True
22     return client_psql
23
24
25 def create_table(client_psql, nom, champs):
26     """ Créer une nouvelle table dans la base de données """
27     fk = ""
28     fields = []
29
30     curseur = client_psql.cursor()
31     curseur.execute(f"DROP TABLE IF EXISTS {nom}")
32
33     if 'fk' in champs.keys():
34         ls = champs.pop('fk')
35         fk = f"CONSTRAINT {ls.pop(0)} FOREIGN KEY({ls.pop(0)}) REFERENCES
36 {ls.pop(0)}"
37         if ls:
38             fk += f" ON DELETE {ls.pop(0)}"
39
40     for key, value in champs.items():
41         fields.append(f"{key} {' '.join(value)}")
42
43     query = f"CREATE TABLE {nom} ({', '.join(fields)})"
44     if fk:
45         query = query[:-1] + f", {fk}"
46     curseur.execute(query)
47
48 def create_index(client_psql, nom, table, colonne):
49     """ Index sur une colonne """
50     query = "CREATE UNIQUE INDEX {} ON {}({})".format(nom, table, colonne)
51     exec_query(client_psql, query)
52

```

---

## C Tableau des choix technologiques

Élément	Retenu	Raisons	Observations
Datasets			
Mail de la compagnie Enron	Non	Mails non classés	Non retenu pour la phase de développement car pas de moyen fiable de contrôler la sortie automatiquement
Mail du projet SpamAssassin	Oui	Mails déjà pré-triés	Mails principalement en Anglais déjà pré-trié en catégorie Spam et Ham
Brown dataset (nltk)	Oui	Corpus d'un million de mots en Anglais publié en 1961	Dataset utilisé pour le développement de la visualisation de la distribution de Zipf
Stopwords (nltk)	Oui	Corpus de mots commun non significatif dans un texte	Utilisation dans le développement de la visualisation de la distribution de Zipf
Langage et Modules			
Python	Oui	Langage polyvalent pour le traitement des données	
Module email	Oui	Module natif pour le traitement des mails	Grande flexibilité pour la lecture des mails
Bases de données			
ElasticSearch	Oui	Technologie utilisée dans mon entreprise. Présence d'une interface de visualisation des données Kibana.	Application dockerisée.
PostgreSQL	Oui	Moteur de base de données relationnelle plus facilement scalable que ElasticSearch pour l'ajout de nouvelle catégorie de données. Il n'est pas nécessaire de ré-indexer toute la base pour ajouter des champs	Application dockerisée
SQLite	Oui	Base de données légère pour stocker uniquement les données statistiques des étapes de la phase 1	Rapide à mettre en place et déjà intégrée

## D Modèles

### D.1 Naïves Bayes

Ce type de modèle est utilisé par le module *langdetect* qui me sert pour la détection des langues dans mon programme.

**Introduction** Les modèles Naïves Bayes se basent sur le théorème de probabilité de Bayes. Il permet de déterminer la probabilité conditionnelle d'apparition d'un événement A sachant qu'un événement B s'est produit. Le terme naïf fait référence au fait que l'on présuppose que les événements A et B ne sont pas corrélés.

Ces techniques sont utilisées pour des modèles de classification en apprentissage supervisé.

La formule mathématique de ce théorème est la suivante :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

On recherche ici  $P(A|B)$ , c'est à dire la probabilité d'apparition d'un événement A sachant que l'évènement B s'est produit.

Pour ce faire nous avons besoin des données suivantes :

- $P(B|A)$  est la probabilité que l'évènement B s'est produit sachant que l'évènement A s'est produit
- $P(A)$  est la probabilité d'apparition de l'évènement A
- $P(B)$  est la probabilité d'apparition de l'évènement B

**Exemples d'utilisation** Les exemples ci dessous vont permettre d'illustrer l'utilisation de cette technique. d'abord manuellement sur un petit jeu de données puis à l'aide d'un code pré-existant sur un autre jeu de données plus important.

**Manuel** Dans cet exemple nous allons déterminer la probabilité qu'a un joueur d'aller sur le terrain selon les conditions météorologiques. Cette probabilité sera calculée en fonction des données récupérées lors des match précédent.<sup>2</sup>

On recherchera ainsi la probabilité de présence sur le terrain d'un joueur selon la météo  $P(A|B)$ . Pour ce faire nous auront besoin de :

- $P(A)$  Probabilité de jouer quelque soit le temps
- $P(B)$  Probabilité de l'évènement météorologique
- $P(B|A)$  Probabilité de l'évènement sachant que le joueur a été sur le terrain

TABLE 3 – Données de présence sur le terrain

météo	soleil	soleil	couvert	pluie	pluie	pluie	couvert
présent	non	non	oui	oui	oui	non	oui
météo	soleil	soleil	pluie	soleil	couvert	couvert	pluie
présent	non	oui	oui	oui	oui	oui	non

---

2. Les données présentées sont inventées



TABLE 4 – Synthèse et probabilité simple  $P(A)$  et  $P(B)$ 

météo	oui	non	$P(B)$
couvert	4	0	$4/14$
soleil	2	3	$5/14$
pluie	3	2	$5/14$
$P(A)$	$9/14$	$5/14$	

On peut déterminer les probabilités de chaque météo en fonction de la présence du joueur sur le terrain  $P(B|A)$ . Pour ce faire on divise le nombre d'évènements de présence du joueur lors d'un évènement météo par le nombre total d'évènements de présence du joueur

TABLE 5 – Probabilité météo selon présence du joueur

météo	$P(B oui)$	$P(B non)$
couvert	$4/9$	$0/5$
soleil	$2/9$	$3/5$
pluie	$3/9$	$2/5$

On va maintenant calculer la probabilité qu'à un joueur d'être sur le terrain si la météo est couverte.

On commence par la probabilité du oui :

$$\begin{aligned}
 P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
 P(A|B) &= \frac{\frac{4}{9} \cdot \frac{9}{14}}{\frac{4}{14}} \\
 P(A|B) &= \frac{\frac{4}{14}}{\frac{4}{14}} \\
 P(A|B) &= \frac{4}{14} \cdot \frac{14}{4} \\
 P(A|B) &= 1
 \end{aligned}$$

On enchaîne sur la probabilité de ne pas jouer si le temps est couvert

$$\begin{aligned}
 P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
 P(A|B) &= \frac{\frac{0}{5} \cdot \frac{5}{14}}{\frac{4}{14}} \\
 P(A|B) &= 0 \cdot \frac{14}{4} \\
 P(A|B) &= 0
 \end{aligned}$$

On peut dire que si le temps est couvert le joueur très probablement sur le terrain On peut également déterminer la probabilité de jouer pour chaque évènement météo

TABLE 6 – Probabilité présence du joueur selon la météo

météo	oui	non	plus probable
couvert	1	0	oui
soleil	$2/5$	$3/5$	non
pluie	$3/5$	$2/5$	oui

*Cas polynomial* : Il est possible de déterminer la probabilité d'un évènement par rapport à plus autres. Dans ce cas, il faudra multiplier entre elles les probabilités de ces évènements selon l'apparition de l'évènement voulu.

Calcul pour un évènement (A) selon 2 autres évènements (B et C)

$$P(A|BC) = \frac{P(B|A)P(C|A)P(A)}{P(B)P(C)}$$

**En code** Dans cet exemple nous allons utiliser un code existant dans la librairie python `scikit-learn`[1]. Ce moteur Naïves Bayes va nous permettre cette fois-ci de catégoriser des variétés d'iris selon la longueur et la largeur des pétales et des sépales. Les données proviennent cette fois-ci d'un dataset également disponible dans `scikit-learn`.

Nous allons utilisé le modèle *GaussianNB* de `scikit-learn` qui est adapté lorsque les données utilisées suivent une distribution normale. Ce qui semble être le cas pour les longueurs et largeur des sépale.

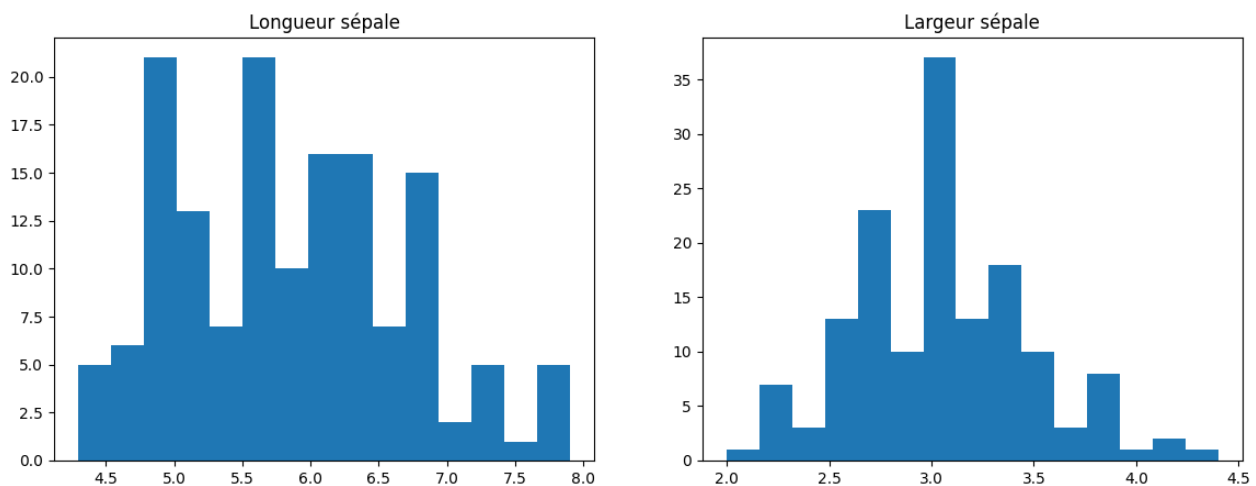


FIGURE 13 – Distribution des longueurs et largeurs des sépales

#### Progamme complet

---

```

1 from sklearn.datasets import load_iris
2 from sklearn.model_selection import train_test_split
3 from sklearn.naive_bayes import GaussianNB
4 from sklearn.metrics import accuracy_score, confusion_matrix,
   ConfusionMatrixDisplay, f1_score, \
5     recall_score
6
7 import matplotlib.pyplot as plt
8
9 X, y = load_iris(return_X_y=True)
10
11 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
12     random_state=0)
13 model = GaussianNB()
14 model.fit(X_train, y_train)
15
16 y_pred = model.predict(X_test)
```

```

16 precision = accuracy_score(y_pred, y_test)
17 recall = recall_score(y_test, y_pred, average="weighted")
18 f1 = f1_score(y_pred, y_test, average="weighted")
19
20 print("Precision:", precision)
21 print("Rappel:", recall)
22 print("Score F1:", f1)
23
24 plt.figure('Donnees du modele', figsize=(14, 5))
25 plt.subplot(1, 3, 1, title='Donnees du train set')
26 plt.scatter(X_train[:, 0], X_train[:, 1], c=y_train)
27 plt.xlabel('Sepale long.')
28 plt.ylabel('Sepale larg.')
29 plt.subplot(1, 3, 2, title='Donnees du test set')
30 plt.scatter(X_test[:, 0], X_test[:, 1], c=y_test)
31 plt.xlabel('Sepale long.')
32 plt.subplot(1, 3, 3, title='Donnees test apres evaluation')
33 plt.scatter(X_test[:, 0], X_test[:, 1], c=y_pred)
34 plt.xlabel('Sepale long.')
35 plt.show()
36
37 cm = confusion_matrix(y_test, y_pred, labels=[0, 1, 2])
38 disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=[0, 1,
39                               2])
40 disp.ax_.set_title('Matrice de confusion')
41 disp.plot()
42 plt.show()
43
44 plt.figure('Distribution des donnees Iris', figsize=(14, 5))
45 plt.subplot(1, 2, 1, title='Longueur sepale')
46 plt.hist(X[:, 0], bins=15)
47 plt.subplot(1, 2, 2, title='Largeur sepale')
48 plt.hist(X[:, 1], bins=15)
49 plt.show()

```

---

Les données du dataset ont été séparés en 2 jeux, un pour l'entraînement du modèle et un pour le test. On obtient alors la représentation suivantes après entraînement et test du modèle

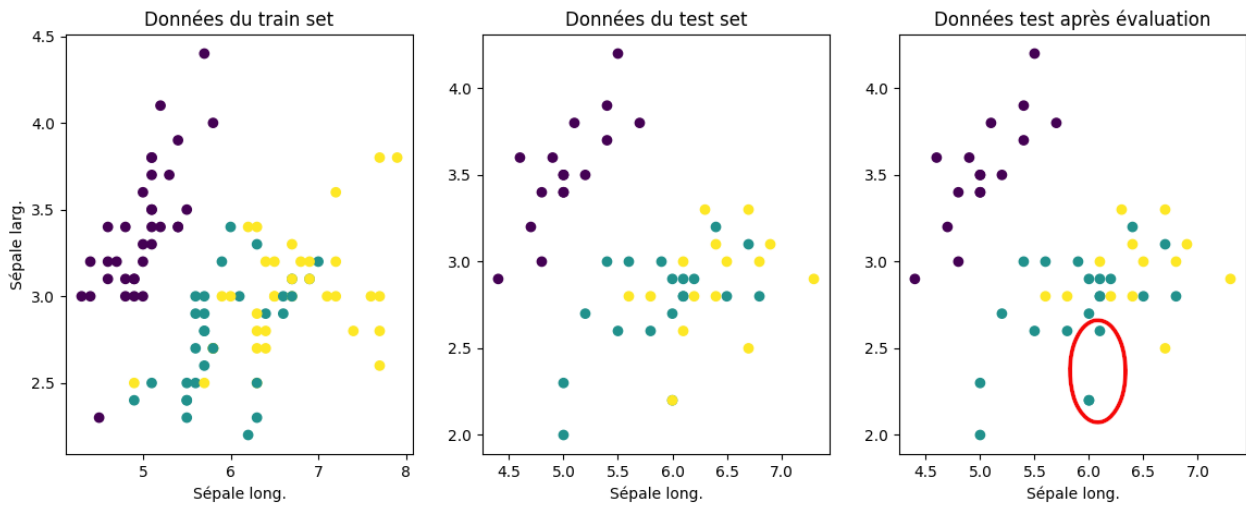


FIGURE 14 – Représentation des données

Dans les données de test nous avons 2 catégorisation qui n'ont pas été réalisée correctement. On obtient les scores suivants :

- Précision : 0.96<sup>3</sup>
- Rappel : 0.96<sup>4</sup>
- Score F1 : 0.9604285714285714<sup>5</sup>

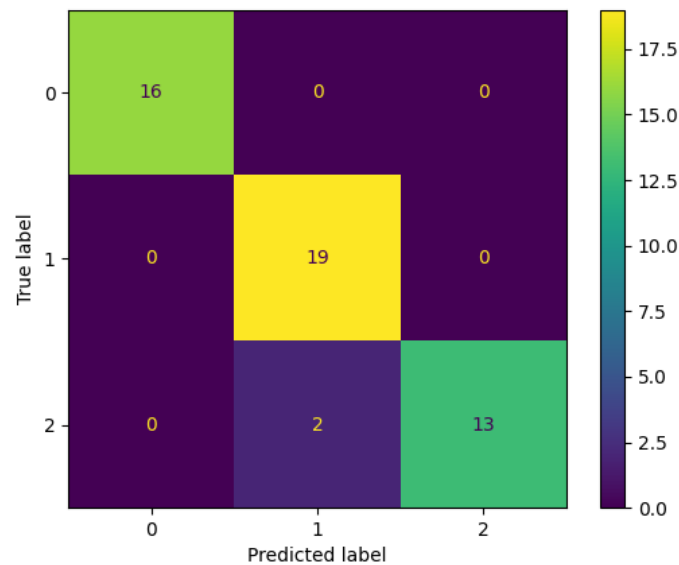


FIGURE 15 – Matrice de confusion

A l'aide de ce modèle nous devrions avoir une 96% de chance de déterminer la bonne variété d'iris en se basant sur la longueur et la largeur des sépales.

**Avantages et inconvénients** Le modèle Naïve Bayes est un modèle simple et rapide qui ne nécessite pas de grande capacités de calcul. De ce fait il permet de traiter une grande quantité

---

3. La précision est la proportion des éléments correctement identifiés sur l'ensemble des éléments prédit  
 4. Le rappel est la proportion des éléments correctement identifiés sur l'ensemble des éléments de la catégorie  
 5. Le Score F1 est la moyenne harmonique calculée de la manière suivante  $2 * (precision * rappel) / (precision + rappel)$

de données.

Cependant, les données qui lui sont fournies ne doivent pas être corrélées ce qui est rarement le cas dans les problèmes du monde réel. Ce type de modèle est limité à des problèmes de classification supervisée. Si on se fie à l'équation (1) la probabilité d'apparition de l'évènement  $B : P(B)$  ne peut pas être nulle.

## E Bibliographie

### Références

- [1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

## F Sitotec

### F.1 Corpus

- Enron company mails, fichier CSV contenant l'ensemble des mails d'une entreprise ayant fermée ses portes (33.834.245 mails) [en ligne], <https://www.kaggle.com/wcukierski/enron-email-dataset> (consulté le 27/01/2022)
- Mails project SpamAssassin, projet opensource de détection de spam (6065 fichiers email déjà trier en ham et spam) [en ligne], <https://spamassassin.apache.org/old/publiccorpus/> (consulté le 27/01/2022)
- Brown corpus, ensemble de texte en anglais publié en 1961 qui contient plus d'un million de mots <https://www.nltk.org/book/ch02.html> (consulté le 20/08/2022)

### F.2 Modules

#### Module langdetect

- Page Github du projet *langdetect* capable de différencier 49 langages avec une précision de 99%, [en ligne] <https://github.com/Mimino666/langdetect> (consulté le 04/12/2022)
- Language Detection Library, présentation du module (anglais) [en ligne] <https://www.slideshare.net/shuyo/language-detection-library-for-java> (consulté le 04/12/2022)

### F.3 Modèles

**Naïves Bayes** Le modèle Naïves Bayes est employé dans le module langdetect (F.2)

- Les algorithmes de Naïves Bayes, Explication sommaire du principe de ces type d'algorithme, [en ligne] <https://brightcape.co/les-algorithmes-de-naives-bayes/> (consulté le 26/03/2023)
- Naive Bayes Classification Tutorial using Scikit-learn, exemple d'utilisation de ce type de modèle avec python (anglais) [en ligne] <https://www.datacamp.com/tutorial/naive-bayes-scik> (consulté le 26/03/2023)

- Scikit learn Naive Bayes, description des types d’algorithmes disponibles dans le module Scikitlearn en python (anglais) [en ligne] [https://scikit-learn.org/stable/modules/naive\\_bayes.html](https://scikit-learn.org/stable/modules/naive_bayes.html) (consulté le 26/03/2023)

## G Codes sources

### G.1 Analyse statistiques de la phase 1

Code pour l’affichage des statistiques de la phase 1

---

```
1 import matplotlib.pyplot as plt
2 import pandas as pd
3
4 from databases.psql_cmd import connect_db, exec_query
5 from databases.psql_db import secrets as psql_secrets
6
7
8 def get_p1_data():
9     psql_cli = connect_db('mail_features_prod', psql_secrets.owner,
10                          psql_secrets.owner_pw,
11                          psql_secrets.host, psql_secrets.port)
12     column = ['id_message', 'type', 'url', 'mail', 'tel', 'nombre', 'prix'
13 ]
14     query = """select m.id_message, c.type, l.url, l.mail, l.tel, l.nombre
15 , l.prix from messages
16 as m
17 join categories as c on m.id_cat = c.id_cat
18 join liens as l on m.id_message = l.id_message"""
19
20     df = pd.DataFrame(exec_query(psql_cli, query), columns=column)
21     psql_cli.close()
22
23     return df.set_index('id_message')
24
25 def set_bar_graph(data, feat, subplot, pos):
26     df = data[data[feat] < 20].groupby(['type', data[feat]]).size()
27     df.unstack(0).plot(kind='bar', ax=subplot[pos])
28
29 if __name__ == '__main__':
30     df_all = get_p1_data()
31     df_spam = df_all[df_all['type'] == 'spam']
32     df_ham = df_all[df_all['type'] == 'ham']
33
34     d_pie = df_all.groupby(['type']).size()
35     fig, ax = plt.subplots()
36     fig.suptitle('Repartition des ham/spam')
37     ax.pie(d_pie, labels=d_pie.index, autopct='%1.1f%%')
38     plt.show()
39
40     print("Statistiques Liens")
```

```
41 print("Globales: \n", df_all.describe())
42 print("Ham: \n", df_ham.describe())
43 print("Spam: \n", df_spam.describe())
44
45 fig, ax = plt.subplots(nrows=5, ncols=1)
46 fig.suptitle("Distribution des mails en fonction du nombre de liens")
47 fig.tight_layout(pad=0.5)
48 position = 0
49 for feat in ['url', 'mail', 'tel', 'nombre', 'prix']:
50     set_bar_graph(df_all, feat, ax, position)
51     position += 1
52 plt.show()
53
```

---