Projet L3
Ingénierie des langues
Fouille de données

GOEHRY Martial
16711476

18 septembre 2022

Table des matières

1 Introduction **2**

2 Récolte des données **2**
2.1 Recherche de dataset 2
2.2 Téléchargement des données 2

3 Pré-traitement **2**
3.1 Extraction des corps des mails 2
3.2 Nettoyage 2
3.3 Mise en base 2
3.4 Recherche de caractéristiques 2
3.5 Analyse préliminaire 2

A Développement visualisation distribution de Zipf **2**

B Tableau des choix technologiques **3**

C Bibliographie **3**

D Sitotec **3**

1 Introduction

Ce projet à pour but de permettre de détecter les mails comme étant 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

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

2 Récolte des données

2.1 Recherche de dataset

2.2 Téléchargement des données

3 Pré-traitement

3.1 Extraction des corps des mails

3.2 Nettoyage

Par regex

Par module

3.3 Mise en base

Stockage des données : ElasticSearch

Stockage des données statistiques du traitement : SQLite

3.4 Recherche de caractéristiques

Références

3.5 Analyse préliminaire

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^{eme}, 3

fois plus fréquent que le 3^{eme} etc.
La formulation finale de la 1^{ere} 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 permettant d'ajuster l'équation.

B 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	Oui	Corpus d'un million de texte en Anglais publié depuis 1961	Dataset utilisé pour 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.
SQLite	Oui	Base de données légère pour stocker uniquement les données statistiques des étapes	Rapide à mettre en place et déjà intégrée

C Bibliographie

D Sitotec