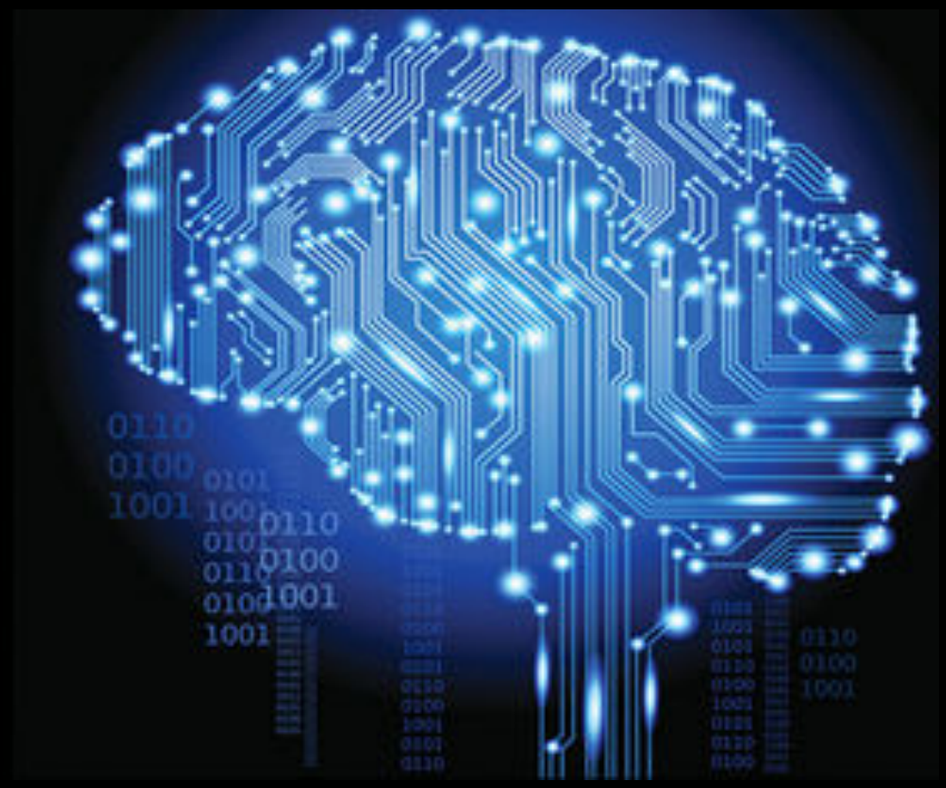
SPAM DETECTOR

HENRY ROMAIN
LANG JORDAN
SPIELDENNER JEREMY



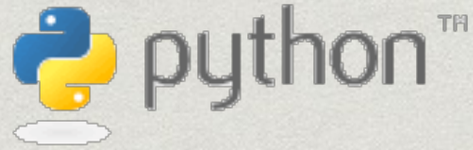
GITHUB.COM/SPIELDY/SPAMDETECT

Sommaire

Introduction

Etude du cahier des charges

- Python / Django
- K-Means
- Git / Github
- Architecture MVC
- JavaScript / Highcharts.js



Répartition des tâches

- Méthode agile : Extrem Programming
- Romain et Jordan : modèle et contrôleur
  - K-Means, normalisation, statistiques
- Jérémy : vue et contrôleur
  - Django, Highcharts, maquette

Normalisation et statistiques

- Normalisation des données
- Séparation des données en spam/non spam
- Statistiques sur les deux types de données
- Prospection des paramètres significatifs

2-Means

- Deux clusters (spam / non spam)
- Initialisation forcée des centroids
- Clusterisation des données
- Permettre l’extraction de données

Django et Highcharts.js

- Django
  - Web framework pour Python
  - Gestion des routes
  - Transfert de données
- Highcharts.js
  - Uniquement du JavaScript
  - Fonctionnalités plus complexe à notre portée

Tests unitaires

- Initialisation des clusters
- Assignation des clusters
- Mise à jour des clusters
- Normalisation des données
- Statistiques sur les données

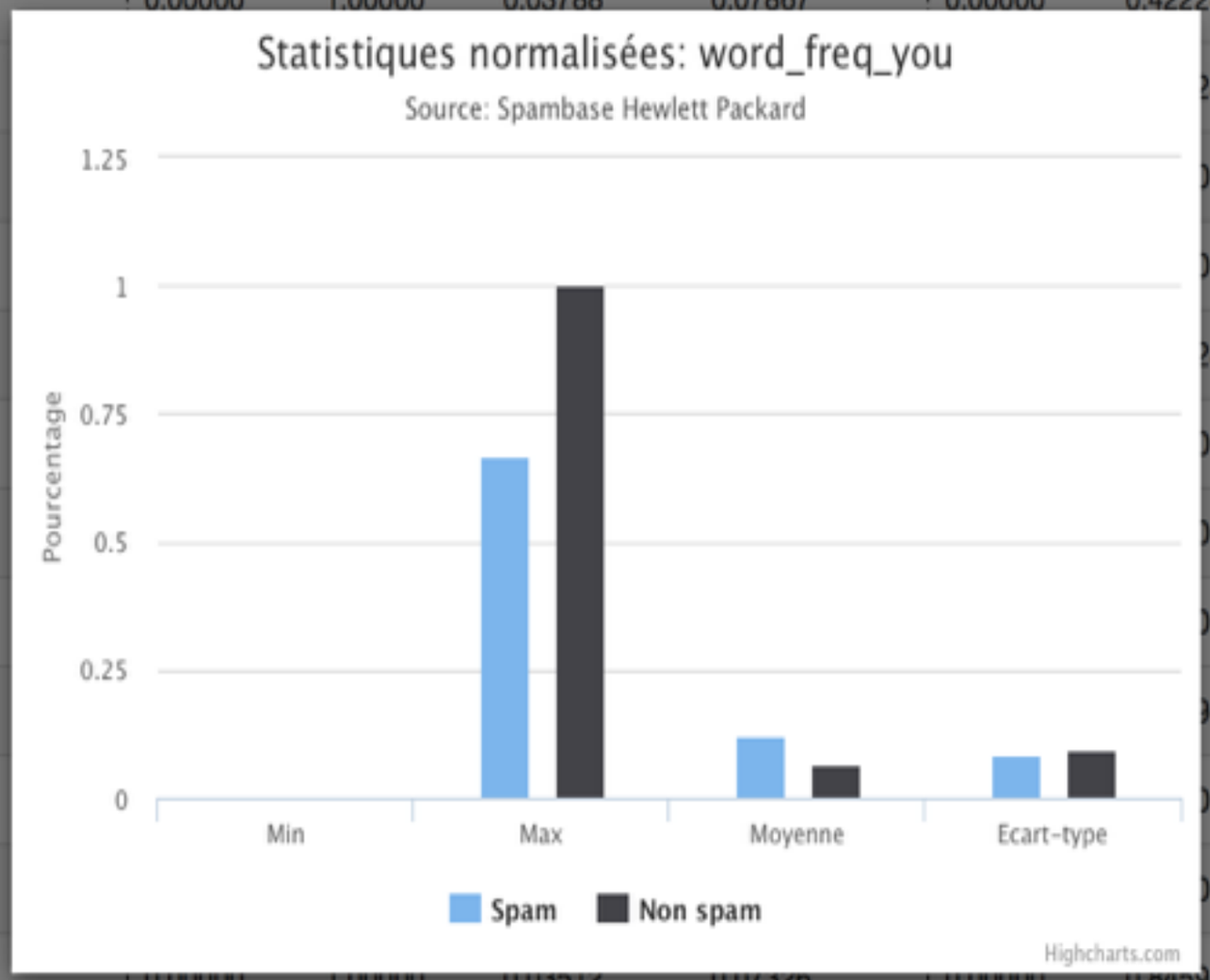
Conception graphique

Choix dataset | K-means

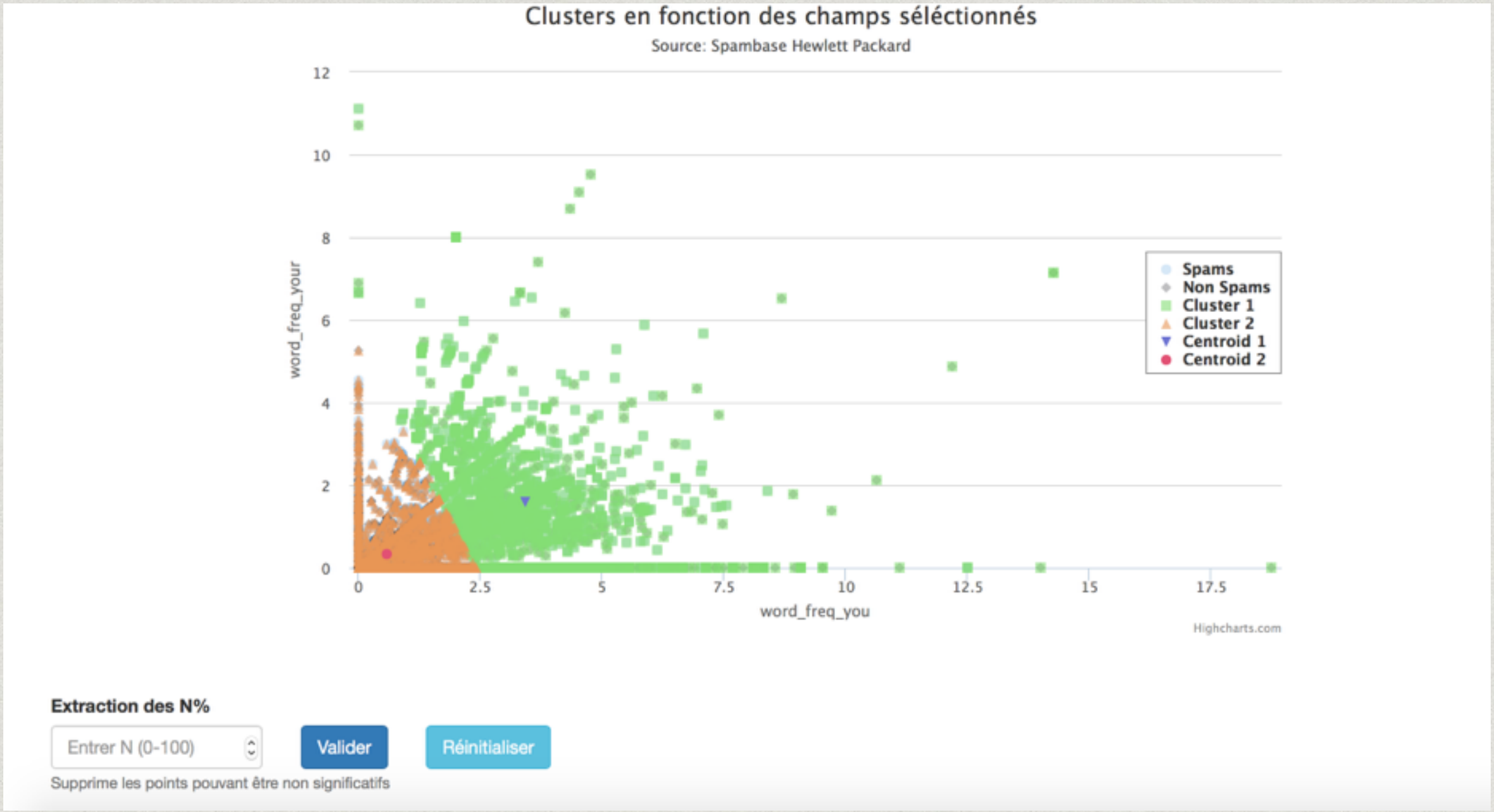
	Spam				Non Spam			
	Min	Max	Moyenne	Ecart-type	Min	Max	Moyenne	Ecart-type
0 - word_freq_make	0.00000	1.00000	0.03355	0.06841	0.00000	0.95595	0.01618	0.06559
1 - word_freq_address	0.00000	0.33333	0.01153	0.02443	0.00000	1.00000	0.01712	0.11435
2 - word_freq_all	0.00000	0.72549	0.07918	0.09423	0.00000	1.00000	0.03933	0.09860
3 - word_freq_3d	0.00000	1.00000	0.00385	0.05182	0.00000	0.02032	0.00002	0.00050
4 - word_freq_our	0.00000	0.76900	0.05140	0.07070	0.00000	1.00000	0.01810	0.06144
5 - word_freq_over	0.00000	0.43197	0.02974	0.05473	0.00000	1.00000	0.00758	0.03790
6 - word_freq_remove	0.00000	1.00000	0.03788	0.07867	0.00000	0.42228	0.00129	0.01519
7 - word_freq_internet	0.00000	1.00000	0.01873	0.04903	0.00000	0.52925	0.00346	0.02225
8 - word_freq_order	0.00000	0.63308	0.03233	0.06743	0.00000	1.00000	0.00723	0.03773
9 - word_freq_mail	0.00000	0.41529	0.01928	0.03472	0.00000	1.00000	0.00920	0.03537
10 - word_freq_receive	0.00000	1.00000	0.04538	0.09613	0.00000	0.76628	0.00832	0.05734
11 - word_freq_will	0.00000	0.64633	0.05687	0.06628	0.00000	1.00000	0.05546	0.10121
12 - word_freq_people	0.00000	1.00000	0.02586	0.06311	0.00000	1.00000	0.01111	0.04667

Conception graphique

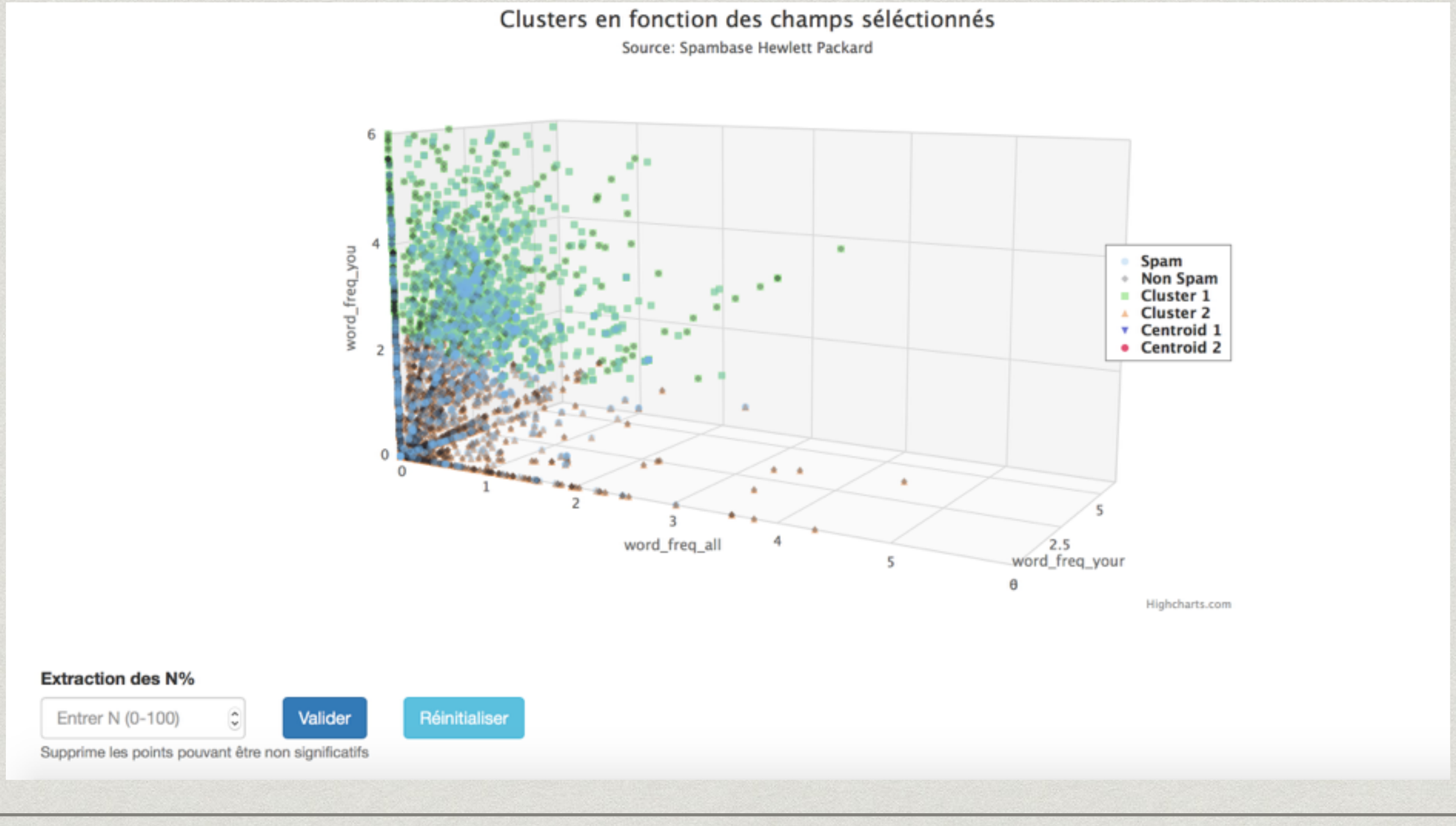
6 - word_freq_remove	0.00000	1.00000	0.03788	0.07867	0.00000	0.42228	0.00129	0.01519
7 - word_freq_internet						25	0.00346	0.02225
8 - word_freq_order						00	0.00723	0.03773
9 - word_freq_mail						00	0.00920	0.03537
10 - word_freq_receive						28	0.00832	0.05734
11 - word_freq_will						00	0.05546	0.10121
12 - word_freq_people						00	0.01111	0.04667
13 - word_freq_report						00	0.00424	0.03444
14 - word_freq_addresses						94	0.00189	0.02139
15 - word_freq_free						00	0.00368	0.03082
16 - word_freq_business						00	0.00677	0.03065
17 - word_freq_email	0.00000	1.00000	0.03512	0.07326	0.00000	0.84598	0.01070	0.04368
18 - word_freq_you	0.00000	0.66667	0.12078	0.08354	0.00000	1.00000	0.06775	0.09564
19 - word_freq_credit	0.00000	1.00000	0.01130	0.04335	0.00000	0.14851	0.00042	0.00532



Conception graphique

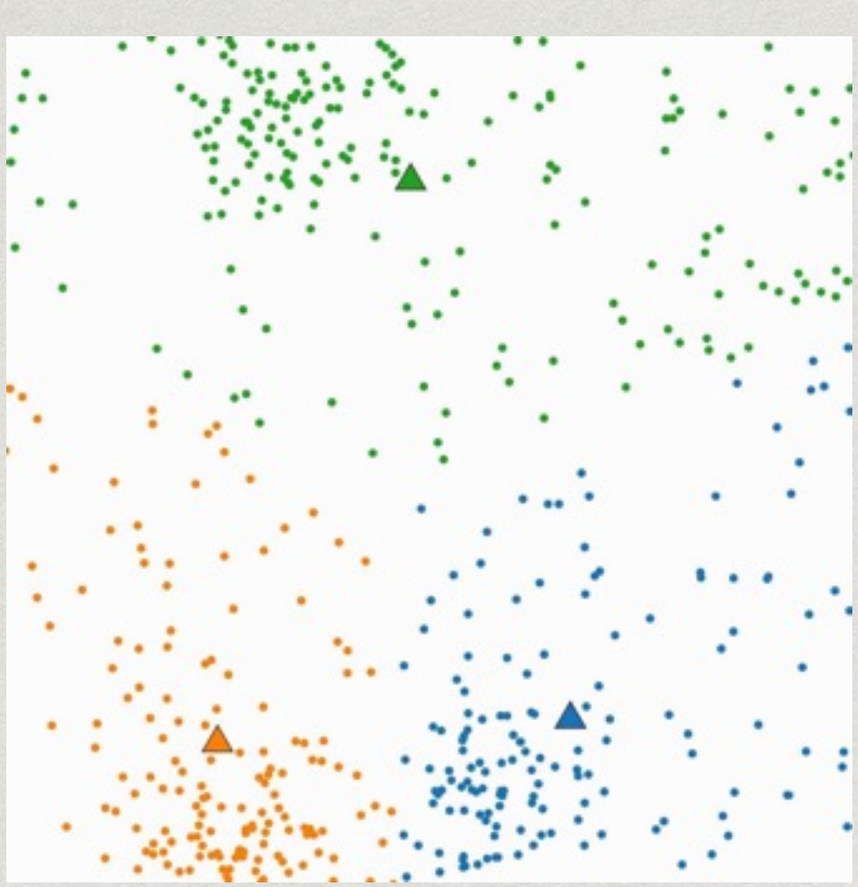


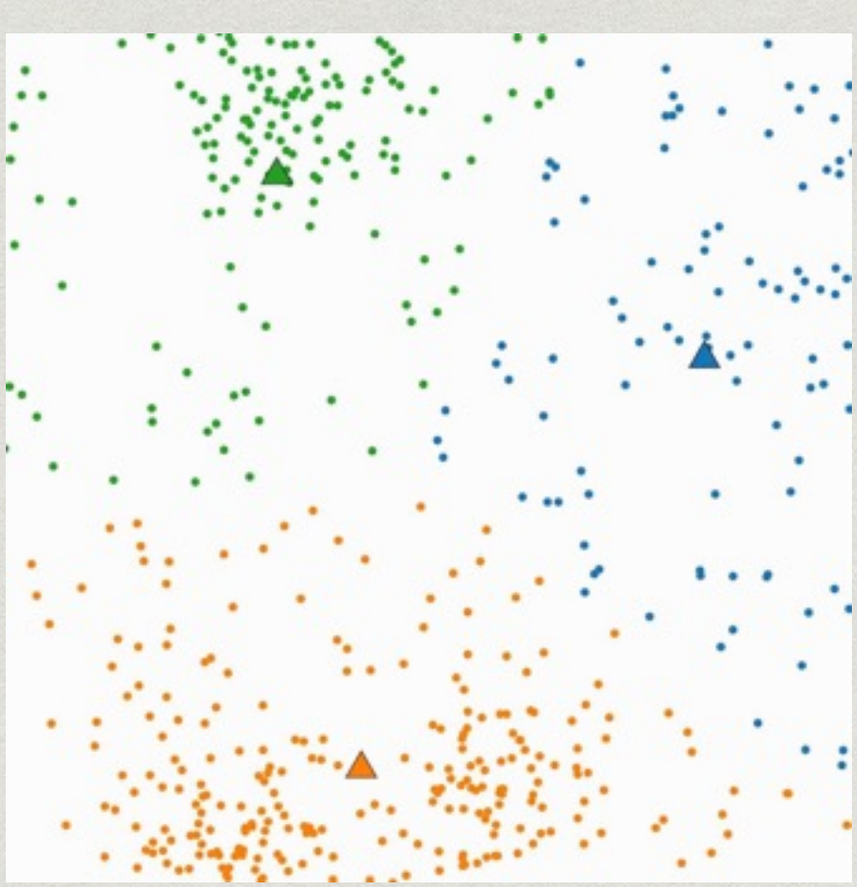
Conception graphique



Difficultés rencontrées

- K-Means et ses limites
- Trouver des champs significatifs





Conclusion

Cahier des charges

L'application développée

- est basée de préférence sur Python/Django
- affiche les anomalies en 2D ou 3D, au choix
- inclut 1 interface graphique, utilisant de préférence la bibliothèque. D3JS : http://d3js.org/. L'interface graphique pourra être composée de 2 écrans :
  - Optionnel : chaque log est représenté, et est accessible directement par l'écran de détection d'anomalie
- inclue 1 bibliothèque de classification (SVM) ou de clustering (K-Means) des données
- incluant des tests unitaires pyunit pour cette bibliothèque.

L'utilisation de langages et d'outil alternatifs est acceptée.

Le choix des champs de données du dataset à utiliser comme référence pour la détection d'anomalies fait partie du travail à réaliser par le groupe, ainsi que le choix des paramètres des algorithmes.