

Manitra RANAIVOHARISON
Olivier QUERET
Hugo LINCK

4AIBD



Neomail, le mail intelligent qui classifie vos mails

Table des matières

<i>Remerciement</i>	2
<i>Introduction</i>	3
<i>Projet Neomail</i>	4
<i>Introduction</i>	4
<i>Mise en place du projet</i>	5
<i>L'architecture du projet</i>	7
Environnement de travail	8
Docker.....	8
<i>Technologie utilisée</i>	9
Compte de service :	10
Identifiants API & Oauth	11
<i>La collecte de donnée</i>	12
Départ.....	12
Fin	12
<i>La transformation de donnée</i>	13
Départ	13
Fin	13
<i>La construction du modèle de machine learning de Labelisation</i>	14
Les algorithmes utilisés.....	14
Notion de similarité	15
Le choix du nombre de cluster.....	15
Algorithme K-means	18
Debut.....	18
Choisir aléatoirement K points (une ligne de la matrice de donnée). Ces points sont les centres des clusters (nommé centroïd).	18
Répéter	18
Jusqu'à convergence.....	18
Ou (stabilisation de l'inertie totale de la population)	18
Fin de l'algorithme.....	18
Remarques sur le K-Means	18
<i>Résultat</i>	19
<i>Personna</i>	19
Olivier.....	19
Résultats du labelling de ses mails.....	20
Victor	21
<i>Conclusion</i>	22
<i>Annexe</i>	23
<i>Définition :</i>	23

Remerciement

Nous adressons tout d'abord nos remerciements à notre enseignant M. WAJNBERG pour nous avoir donné la possibilité de travailler sur ce projet annuel, merci également pour ses conseils avisés à la fois technique et humaine.

On remercie M. Frédéric SANANES, directeur pédagogique de l'ingénierie de l'intelligence artificielle et Big-Data de nous avoir accepté et de nous avoir soutenu tout au long de notre formation.

De manière générale, nous remercions également l'ESGI et son corps enseignant pour l'enseignement qu'ils nous ont apportés, ce qui nous a permis d'être opérationnel pour le monde professionnel et élaborer ce projet annuel en autonomie.

Introduction

De nos jours, le monde de l'informatique et plus particulièrement le domaine du big-data et de l'intelligence artificielle évolue de manière rapide. De ce fait, les algorithmes de machine learning sont de plus en plus implémentés pour pouvoir créer de la valeur.

A cause de cette forte évolution également, les écoles d'ingénieur et/ou spécialisées en informatique poussent leurs étudiants à devenir des experts dans différents domaines de l'informatique. Étant en 4^{ème} année de Mastère en spécialisation Intelligence Artificielle et Big Data, nous sommes amenés pour la fin de notre année à réaliser un projet annuel.

Ce sujet étant libre, nous avons choisis de mettre en place une application permettant de trier les mails de compte Gmail qui s'intitule Neomail.

Ce rapport servira donc à vous présenter le travail que nous avons effectué et les expériences que nous avons acquis durant la réalisation de ce projet.

Dans un premier temps, nous procéderons à la présentation de notre projet, comment notre groupe s'est organisé pour mettre en place et avancer.

Puis dans un second temps, nous présenterons les différents outils et méthodes que nous avons utilisés et appliqués.

Ensuite, nous vous ferons part des difficultés que nous avons rencontrées et comment nous avons pu les résoudre.

Pour conclure, Nous ferons un bilan technique et humain, en mettant en valeur les compétences acquises et le lien avec les enseignements assimilés à l'École Supérieur de Génie Informatique.

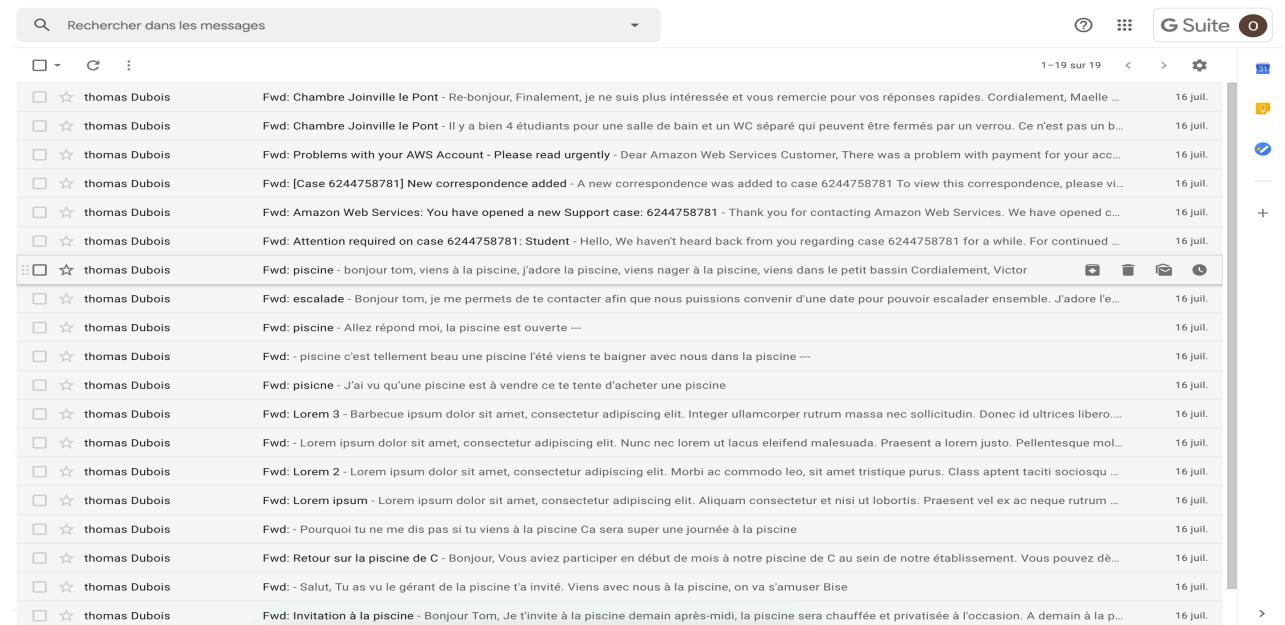
Projet Neomail

Introduction

Actuellement, plus d'un milliard de personnes utilisent un compte Gmail ; chacun des mails présents sur leur compte peuvent être classés via des labels que l'on peut créer. Tout cela se faisant actuellement manuellement.

Une personne reçoit chaque jour en moyenne 39 mails, ce qui donne plus de 14.000 mails par an. Classer les mails manuellement demande une rigueur journalière, sans quoi, le temps alloué à les classer sera dissuadant. De plus, il faut créer un label correspondant au mail ou bien lui affecter un label déjà existant (cela demande une recherche supplémentaire)

Exemple d'une boîte mail utilisateur non labélisé :



The screenshot shows a Gmail inbox with 19 messages. All messages are from 'thomas Dubois' and are in 'Fwd:' format. The subject lines are mostly generic, such as 'Fwd: Chambre Joinville le Pont - Re-bonjour', 'Fwd: Problems with your AWS Account - Please read urgently', and 'Fwd: Amazon Web Services: You have opened a new Support case'. The dates for all messages are '16 juil.' (July 16). The interface includes a search bar at the top, a toolbar with icons for reply, forward, etc., and a sidebar on the right with a 'G Suite' button and other account options.

Étant confronté également à cette problématique, notre groupe s'est proposé de créer le projet Neomail, une application qui crée et affecte automatiquement les mails de compte Gmail via des libellés clairs.

Cette application propose une interface pour que les utilisateurs puissent se connecter et lancer l'exécution de la labellisation.

Cette application a aussi un back-end en python (cf : annexe) qui fonctionne en micro-service (cf : annexe).

Nous proposerons aux développeurs de pouvoir faire des appels de nos micro-services (call API) à travers des méthodes HTTP.

Chaque micro-service travaille indépendamment les uns des autres.

Mise en place du projet

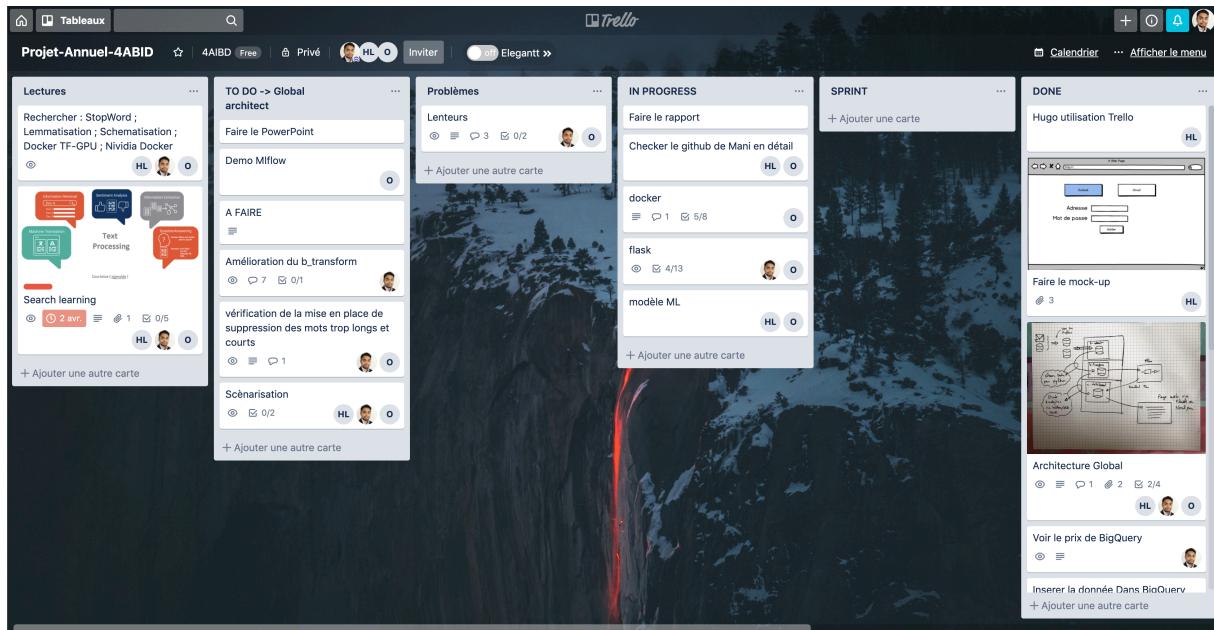
Pour mettre en place Neomail, nous nous sommes listé plusieurs tâches à effectuer sur Trello (voir cf : annexe) et nous avons mis en place des réunions hebdomadaires avec des objectifs à finir pour ces réunions.

Chaque personne du groupe s'est assignée lesdites tâches à faire et à présenter lors des réunions.

Les dates des réunions effectuées :

- 28/03/2019 : première ébauche de notre vision du projet ainsi que des services et outils que l'on utilisera. Différents POC et lectures sont prévues afin de se mettre à niveau sur ces services
- 11/04/2019 : premier test sur des services et retour d'expérience. Proposition de rencontrer une spécialiste en analyse sémantique
- 12/04/2019 : mise en place d'un datalab et d'un environnement de dev, le tout ne tournant pas sous docker. Le passage sous docker doit être mis en place afin que chacun puisse reproduire chez lui l'environnement global
- 14/04/2019 : Réunion avec Jainaïna Sabino, consultante en analyse sémantique, qui nous a donné des conseils et méthodologie sur le traitement de texte
- 18/04/2019 : Mise en place d'un cycle de passage sous docker des parties faites en local sur le système afin de tester le bon fonctionnement. Les nouvelles fonctionnalités seront directement développées sous docker
- Nous sommes ensuite passé sur des réunions hebdomadaires afin de croiser le développement et convenir de la suite du développement. Au cours de ces réunions, le choix d'abandonner certaines technologies au profit de nouvelles a été fait (Django vers Flask, Big Query vers google storage)

Les taches dans Trello :



En parallèle, avec les tâches que nous nous sommes assignées, nous avons effectué des recherches communes sur les appels d'interface de programmation d'application ou interface de programmation applicative (communément appelées API) pour bien comprendre leurs fonctionnements afin de les intégrer dans notre projet et les utiliser pour récupérer des mails depuis Gmail et Outlook.

Nous avons également cherché, en commun, comment fonctionne les appels d'API de Gmail et Outlook.

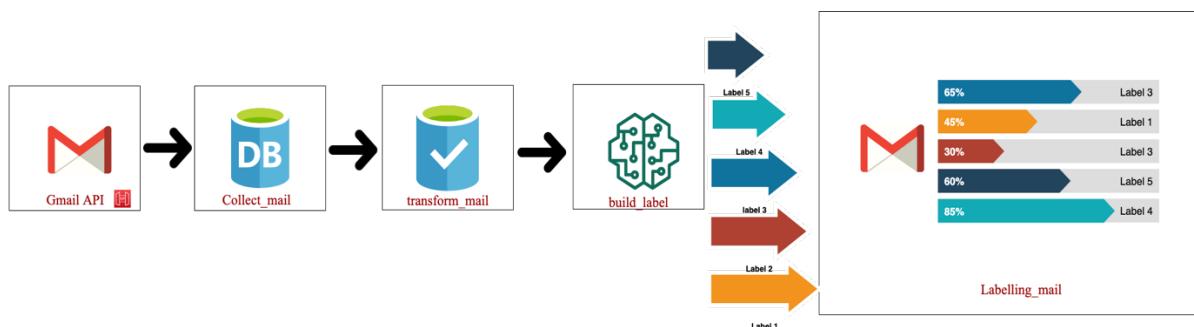
C'est suite à ces recherches et tests que nous sommes mis d'un commun accord à n'utiliser uniquement les comptes Gmail dans un premier temps afin de sortir une version bêta fonctionnelle qui pourra être adaptée.

L'architecture du projet

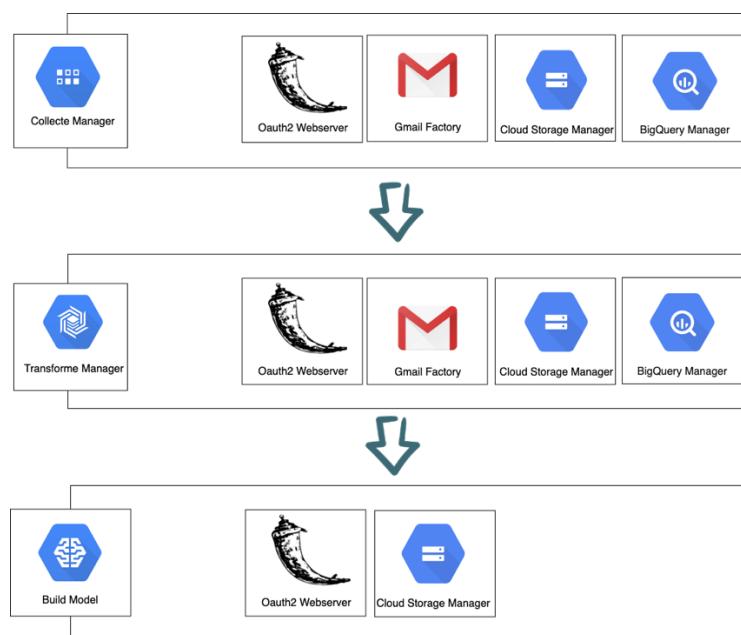
Nous avons donc choisi d'utiliser uniquement Gmail du fait de la complexité d'accès à l'API Outlook.

Nous avons mis en place un processus pour la récupération et le traitement des données qui se décompose en plusieurs phases :

- La phase d'Extraction (collect) consiste à collecter les données en provenance de Gmail à travers un call API
- La phase de transformation (Transform) consiste à reformater et à transformer les données afin de pouvoir éliminer les doublons et l'information superflue
- Enfin, la phase de clustering et de construction des labels grâce à notre modèle de machine learning
- La phase d'ajout des labels dans une boîte mail



Chaque partie est un micro-service dans notre architecture projet et représente une entité indépendante les uns des autres.



Environnement de travail

Datalab

Cet environnement sert à utiliser les données récupérées et traitées afin de tester et mettre en œuvre des modèles de machine learning.

Les tests ont été fait via des notebooks sur jupyter et les lancements via rundeck.

Production

L'environnement de production permet de lancer le site web via Flask ainsi que les différents traitements liés à la récupération des mails et à leurs traitements.

Docker

Docker est utilisé afin de mettre en place deux grandes parties de notre projet :

La première est le Datalab. Celui-ci est composé de trois conteneurs :

1. Un Anaconda avec jupyter dessus pour les tests de modèle
2. Un Rundeck afin d'automatiser le lancement des briques (collect ; transform)
3. Un python qui sert à Rundeck pour lancer les commandes python

Les trois répertoires permettant le stockage des csv sont montés sur les 3 conteneurs ainsi que le répertoire default regroupant les différentes fonctions et le répertoire resources ayant les différents fichiers d'accès aux services extérieurs (API GCP).

Le conteneur Rundeck a la permission de lancer le conteneur python et d'y exécuter des commandes.

Les ports 8888 pour Jupyter et 4440 sont ouverts afin d'y accéder de l'extérieur du réseau.

La seconde partie est la Production composée d'un conteneur python faisant tourner flask afin d'appeler les différentes fonctions du projet via ses routes.

Les trois répertoires permettant le stockage des csv sont montés sur les 3 conteneurs ainsi que le répertoire default regroupant les différentes fonctions et le répertoire resources ayant les différents fichiers d'accès aux services extérieurs (API GCP).

Le port 8080 est ouvert afin d'accéder au site de l'extérieur du réseau.

Technologie utilisée

Lors de nos tests sur l'API Gmail, nous avons constaté qu'il nous fallait ouvrir nos propres services API sur Google Cloud Platform afin que les utilisateurs externes puissent obtenir leurs Mail à travers nos APIs. De plus, BigQuery et Cloud storage sont des services proposés par Google qui sont directement dans Google Cloud Platform, ce qui a été un grand avantage pour notre projet par rapport au compte de service (cf : Annexe).

Dans le répertoire du projet vous trouvez un dossier ressources qui contient 5 sous dossiers :

- api_gcp_credential
- big_query_credential
- cloud_storage_credential
- gcp_credential
- gmail_credential

Chaque sous dossier contient des fichiers sous le nom '*service_account.json*' en extension json qui permet à chaque micro-service d'accéder au service de Google Cloud Platform.

Compte de service :

 Filtrer le tableau

<input type="checkbox"/>	E-mail	État
<input type="checkbox"/>	 neomail@appspot.gserviceaccount.com	
<input type="checkbox"/>	 141298797399-compute@developer.gserviceaccount.com	
<input type="checkbox"/>	 gcp-global-admin@neomail.iam.gserviceaccount.com	
<input type="checkbox"/>	 global-prod@neomail.iam.gserviceaccount.com	
<input type="checkbox"/>	 test-123@neomail.iam.gserviceaccount.com	

Un compte de service doit être créé pour chaque projet google cloud et doit être déposé dans le répertoire gcp_credential.

Identifiants API & Oauth

Clés de l'API

<input type="checkbox"/>	Nom	Date de création
<input type="checkbox"/>	⚠ Clé API 1	13 juin 2019

ID clients OAuth 2.0

<input type="checkbox"/>	Nom	Date de création
<input type="checkbox"/>	neomail_test	11 juin 2019
<input type="checkbox"/>	neomail	10 juin 2019

Clés de compte de service

<input type="checkbox"/>	ID
<input type="checkbox"/>	08c454e41c53fc9062c70f164643b1d01d395f65
<input type="checkbox"/>	6978f9b739b2aaaddbbb8e84d17aec1a276206

Ici, nous avons généré des identifiants pour se connecter à l'API mise en place sur Google Cloud Platform. Nous avons aussi configuré des ID pour le Oauth ; l'une pour nos tests en local et une autre pour la production.

Le fait de faire plusieurs ID permet de spécifier les IP et pages de redirection de l'API.

La collecte de donnée

La collecte de donnée est le premier micro-service que l'on utilise, ce micro-service fait appel à 3 fonctionnalités pour fonctionner :

- Gmail Factory
- Cloud Storage Manager
- BigQuery Manager

Fonctionnement de manière général :

Départ

Utilise Gmail Factory pour récupérer les identités de chaque mail (idMail : cf annexe) et stocker dans une liste.

Répéter

- o A partir de chaque idMail, un appel est lancé vers l'APIs Gmail afin de récupérer chaque donnée
- o Stocker chaque valeur récupérer dans une liste de dictionnaire.
Jusqu'à la fin de la liste
- o Écrire la liste finale dans un fichier en format csv avec comme format de nom 'nom_prenom@domaine_mail.com' et le stocker dans le dossier spécifique au collect
- o Envoyer le fichier créé dans Google Cloud Storage pour l'archiver
- o Créé une table dans BigQuery et remplir la table avec les données du fichier csv

Note : La boucle qui récupère chaque attribut d'un mail fonctionne en multithreading asynchrone avec des coroutines et des decorators qui sont orchestrés dans le *CollectManager*.

Fin

Une fois que la collecte de donnée s'est déroulée correctement et est terminée, le micro-service de transformation de donnée est lancé.

La transformation de donnée

La transformation de donnée est le second micro-service que l'on utilise, ce micro-service fait appel à 3 fonctionnalités pour fonctionner :

- Gmail Factory
- Cloud Storage Manager
- BigQuery Manager

Fonctionnement de manière général :

Départ

Répéter

- o A partir de chaque mail, on nettoie le corps du mail et on le stock dans une liste
- o Envoi du fichier généré dans Google Cloud Storage pour l'archiver
- o Création d'une table dans BigQuery et injection des données du csv dans la table

Fin

Le mico-service de construction du modèle est appelé.

La construction du modèle de machine learning de Labelisation

Après avoir transformé nos données via l'API de transformation, nous les utilisons dans la partie de création de labels. Pour mettre en place cette création de labels pour chaque mail, on utilise des algorithmes de machine learning.

Les algorithmes utilisés

Le clustering est l'un des algorithmes que nous utilisons pour créer des classifications dans nos mails. Cet algorithme est une méthode d'apprentissage non supervisée (unsupervised learning). Ainsi, on n'essaie pas d'apprendre une relation de corrélation entre un ensemble de features d'une observation et une valeur à prédire, comme c'est le cas pour l'apprentissage supervisé. L'apprentissage non supervisé va plutôt **trouver des patterns dans les données**.

Notamment, en regroupant les choses qui se ressemblent.

En apprentissage non supervisé, les données sont représentées comme ceci :

$$X = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & x_{(1,\dots)} & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,\dots)} & x_{(2,n)} \\ \dots & \dots & \dots & \dots \\ x_{(m,1)} & x_{(m,2)} & x_{(m,\dots)} & x_{(m,n)} \end{pmatrix}$$

Chaque ligne représente un individu (une observation, dans notre cas un mail). A l'issue de l'application du clustering, on retrouvera ces données regroupées par ressemblance.

Le clustering va regrouper en plusieurs familles (clusters) les individus/objets en fonction de leurs caractéristiques. Ainsi, les individus se trouvant dans un même cluster sont similaires et les données se trouvant dans un autre cluster ne le sont pas.

Il existe deux types de clustering :

- Le clustering hiérarchique
- Le clustering non-hiéarchical (partitionnement)

K-means

K-means est un algorithme non supervisé de clustering **non hiérarchique**. Il permet de regrouper en K clusters distincts les observations du dataset. Ainsi les données similaires se retrouveront dans un même cluster. Par ailleurs, une observation ne peut se retrouver que dans un cluster à la fois (exclusivité d'appartenance). Une même observation, ne pourra donc, appartenir à deux clusters différents.

Notion de similarité

Pour pouvoir regrouper un jeu de données en K cluster distincts, l'algorithme K-Means a besoin d'un moyen de comparer le degré de similarité entre les différentes observations. Ainsi, deux données qui se ressemblent, auront une distance de non similarité réduite, alors que deux objets différents auront une distance de séparation plus grande.

Les littératures mathématiques et statistiques regorgent de définitions de distance, la plus connue pour les cas de clustering est la distance Euclidienne :
C'est la distance géométrique qu'on apprend au collège. Soit une matrice X à n variables quantitatives. Dans l'espace vectoriel $\sum n$. La distance euclidienne entre deux observations x_1 et x_2 se calcule comme suit :

Le choix du nombre de cluster

Choisir un nombre de cluster K n'est pas forcément intuitif. Dans notre cas d'usage, nous ne pouvons pas établir d'avance les groupes de mail existant dans une boîte mail.

Spécialement, quand le jeu de données est grand et qu'on n'a pas d'à priori ou des hypothèses sur les données. Un nombre K grand peut conduire à un partitionnement trop fragmenté des données. Ce qui empêchera de découvrir des patterns intéressants dans les données.
Par contre, un nombre de clusters trop petit, conduira à avoir, potentiellement, des clusters trop généralistes contenant beaucoup de données. Dans ce cas, on n'aura pas de patterns "fins" à découvrir.

Pour un même jeu de données, il n'existe pas un unique clustering possible. La difficulté résidera donc à choisir un nombre de cluster K qui permettra de mettre en lumière des patterns intéressants entre les données. Malheureusement il n'existe pas de procédé automatisé pour trouver le bon nombre de clusters.

La méthode la plus usuelle pour choisir le nombre de clusters est de lancer K-Means avec différentes valeurs de K et de calculer la variance des différents clusters. La variance est la somme des distances entre chaque centroid d'un cluster et les différentes observations incluses dans le même cluster. Ainsi, on cherche à trouver un nombre de clusters K de telle sorte que les clusters retenus minimisent la distance entre leurs centres (centroids) et les observations dans le même cluster. On parle de minimisation de la distance intra-classe.

La variance des clusters se calcule comme suit :

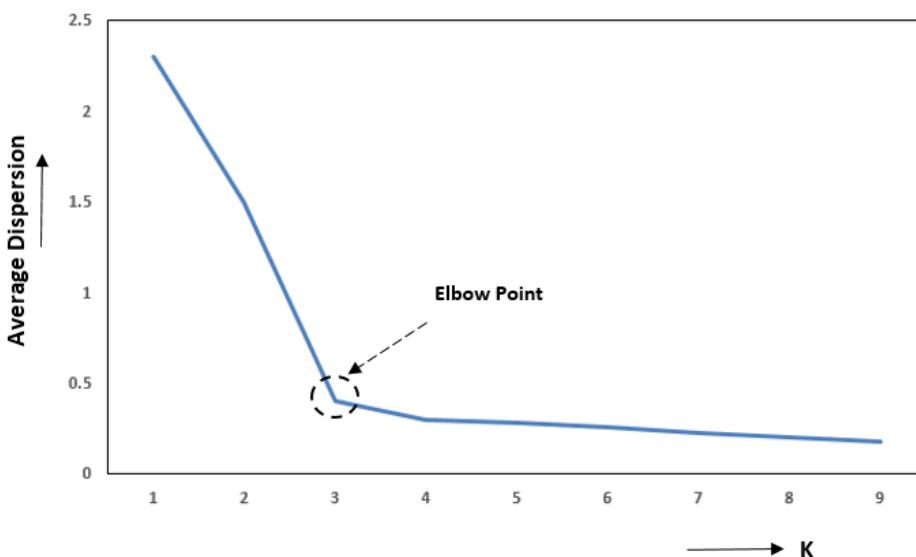
$$V = \sum_j \sum_{xi \rightarrow cj} D(cj, xi)^2$$

Avec :

- cj : Le centre du cluster (le centroïd)
- xi : La i ème observation dans le cluster ayant pour centroïd
- $D(cj, xi)^2$ La distance (euclidienne ou autre) entre le centre du cluster et le point

Généralement, en mettant dans un graphique les différents nombres de clusters K en fonction de la variance, on retrouve un graphique similaire à celui-ci :

Elbow Method for selection of optimal “K” clusters



On remarque sur ce graphique, la forme d'un bras où le point le plus haut représente l'épaule et le point où K vaut 9 représente l'autre extrémité : la main. Le nombre optimal de clusters est le point représentant le coude. Ici le coude peut être représenté par K valant 2 ou 3. C'est le nombre optimal de clusters.

Généralement, le point du coude est celui du nombre de clusters à partir duquel la variance ne se réduit plus significativement.

En effet, la "chute" de la courbe de variance (distortion) entre 1 et 3 clusters est significativement plus grande que celle entre 5 clusters et 9 clusters.

Le fait de chercher le point représentant le coude, a donné nom à cette méthode : La méthode Elbow (coude en anglais).

Fonctionnement de l'algorithme K-Means

k-means est un algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd. Le choix initial des centroïdes conditionne le résultat final.

Admettant un nuage d'un ensemble de points, K-Means change les points de chaque cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés, sous réserve de choisir la bonne valeur K du nombre de clusters.

Algorithme K-means

Entrée :

- K le nombre de cluster à former
- Le Training Set (matrice de données)

Debut

Choisir aléatoirement K points (une ligne de la matrice de donnée). Ces points sont les centres des clusters (nommé centroïd).

Répéter

- Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche par rapport au centre du groupe de cluster
- Recalculer le centre de chaque cluster et modifier le centroïd

Jusqu'à convergence

Ou (stabilisation de l'inertie totale de la population)

Fin de l'algorithme

Note 1: Lors de la définition de l'algorithme, le terme "point" est un point au sens "donnée/data" qui se trouve dans un espace vectoriel de dimension n . Avec n : le nombre de colonnes de la matrice de données.

Note 2 : La convergence de l'algorithme K-Means peut être l'une des conditions suivantes :

- Un nombre d'itérations fixé à l'avance, dans ce cas, K-means effectuera les itérations et s'arrêtera peu importe la forme de clusters composés.
- Stabilisation des centres de clusters (les centroids ne bougent plus lors des itérations).

L'affectation d'un point K à un cluster se fait en fonction de la distance de ce point par rapport aux différents K centroids. Par ailleurs, ce point K se fera affecté à un cluster K s'il est plus proche de son centroid (distance minimale). Finalement, la distance entre deux points dans le cas de K-Means se calcule par les méthodes évoquées dans le paragraphe "notion de similarité".

Remarques sur le K-Means

Optimums locaux

En analysant la façon de procéder de l'algorithme de K-means, on remarque que pour un même jeu de données, on peut avoir des partitionnements différents. En effet, L'initialisation des tous premiers K centroids est complètement aléatoire. Par conséquent l'algorithme trouvera des clusters différents en fonction de cette première initialisation aléatoire. De ce fait, la configuration des clusters trouvés par K-Means peut ne pas être la plus optimale. On parle d'optimum local.

Résultat

Personna

Olivier

**24 ans, adepte de natation et ingénieur louant son appartement de temps en temps.
Il utilise aussi les services d'Amazon Web Service.**

<input type="checkbox"/>	 thomas Dubois	Fwd: Chambre Joinville le Pont - Re-bonjour, Finalement, je ne suis plus intéressée et vous remercie pour vos réponses rapides. C...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Chambre Joinville le Pont - Il y a bien 4 étudiants pour une salle de bain et un WC séparé qui peuvent être fermés par un ver...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Problems with your AWS Account - Please read urgently - Dear Amazon Web Services Customer, There was a problem with ...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: [Case 6244758781] New correspondence added - A new correspondence was added to case 6244758781 To view this corre...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Amazon Web Services: You have opened a new Support case: 6244758781 - Thank you for contacting Amazon Web Service...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Attention required on case 6244758781: Student - Hello, We haven't heard back from you regarding case 6244758781 for a ...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: piscine - bonjour tom, viens à la piscine, j'adore la piscine, viens nager à la piscine, viens dans le petit bassin Cordialement, ...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: escalade - Bonjour tom, je me permets de te contacter afin que nous puissions convenir d'une date pour pouvoir escalader ...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: piscine - Allez répond moi, la piscine est ouverte ---	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: - piscine c'est tellement beau une piscine l'été viens te baigner avec nous dans la piscine ---	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: piscine - J'ai vu qu'une piscine est à vendre ce te tente d'acheter une piscine	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Lorem 3 - Barbecue ipsum dolor sit amet, consectetur adipiscing elit. Integer ullamcorper rutrum massa nec sollicitudin. Do...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc nec lorem ut lacus eleifend malesuada. Praesent a lorem ju...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Lorem 2 - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ac commodo leo, sit amet tristique purus. Class a...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Lorem ipsum - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam consectetur et nisi ut lobortis. Praesent vel ...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: - Pourquoi tu ne me dis pas si tu viens à la piscine Ca sera super une journée à la piscine	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Retour sur la piscine de C - Bonjour, Vous aviez participer en début de mois à notre piscine de C au sein de notre établissement...	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: - Salut, Tu as vu le gérant de la piscine t'a invité. Viens avec nous à la piscine, on va s'amuser Bise	16 juil.
<input type="checkbox"/>	 thomas Dubois	Fwd: Invitation à la piscine - Bonjour Tom, Je t'invite à la piscine demain après-midi, la piscine sera chauffée et privatisée à l'occa...	16 juil.

Résultats du labelling de ses mails

<input type="checkbox"/>		thomas Dubois	charge oliver Fwd: Chambre Joinville le Pont - Re-bonjour, Finalement, je ne suis plus intéressée et vous remercie pour vos rép...	16 juil.
<input type="checkbox"/>		thomas Dubois	charge oliver Fwd: Chambre Joinville le Pont - Il y a bien 4 étudiants pour une salle de bain et un WC séparé qui peuvent être fe...	16 juil.
<input type="checkbox"/>		thomas Dubois	amazon customer Fwd: Problems with your AWS Account - Please read urgently - Dear Amazon Web Services Customer, There ...	16 juil.
<input type="checkbox"/>		thomas Dubois	amazon using Fwd: [Case 6244758781] New correspondence added - A new correspondence was added to case 6244758781 T...	16 juil.
<input type="checkbox"/>		thomas Dubois	amazon using Fwd: Amazon Web Services: You have opened a new Support case: 6244758781 - Thank you for contacting Ama...	16 juil.
<input type="checkbox"/>		thomas Dubois	amazon using Fwd: Attention required on case 6244758781: Student - Hello, We haven't heard back from you regarding case 62...	16 juil.
<input type="checkbox"/>		thomas Dubois	piscine victor Fwd: piscine - bonjour tom, viens à la piscine, j'adore la piscine, viens nager à la piscine, viens dans le petit bassin...	16 juil.
<input type="checkbox"/>		thomas Dubois	adore victor Fwd: escalade - Bonjour tom, je me permets de te contacter afin que nous puissions convenir d'une date pour pouv...	16 juil.
<input type="checkbox"/>		thomas Dubois	adresse piscine Fwd: piscine - Allez répond moi, la piscine est ouverte ---	16 juil.
<input type="checkbox"/>		thomas Dubois	piscine victor Fwd: - piscine c'est tellement beau une piscine l'été viens te baigner avec nous dans la piscine ---	16 juil.
<input type="checkbox"/>		thomas Dubois	adresse piscine Fwd: piscine - J'ai vu qu'une piscine est à vendre ce te tente d'acheter une piscine	16 juil.
<input type="checkbox"/>		thomas Dubois	amet elit Fwd: Lorem 3 - Barbecue ipsum dolor sit amet, consectetur adipiscing elit. Integer ullamcorper rutrum massa nec soll...	16 juil.
<input type="checkbox"/>		thomas Dubois	amet elit Fwd: - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc nec lorem ut lacus eleifend malesuada. Praesen...	16 juil.
<input type="checkbox"/>		thomas Dubois	amet elit Fwd: Lorem 2 - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ac commodo leo, sit amet tristique pur...	16 juil.
<input type="checkbox"/>		thomas Dubois	amet elit Fwd: Lorem ipsum - Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam consectetur et nisi ut lobortis. ...	16 juil.
<input type="checkbox"/>		thomas Dubois	piscine victor Fwd: - Pourquoi tu ne me dis pas si tu viens à la piscine Ca sera super une journée à la piscine	16 juil.
<input type="checkbox"/>		thomas Dubois	adresse piscine Fwd: Retour sur la piscine de C - Bonjour, Vous aviez participer en début de mois à notre piscine de C au sein d...	16 juil.
<input type="checkbox"/>		thomas Dubois	piscine Fwd: - Salut, Tu as vu le gérant de la piscine t'a invité. Viens avec nous à la piscine, on va s'amuser Bise	16 juil.
<input type="checkbox"/>		thomas Dubois	adresse piscine Fwd: Invitation à la piscine - Bonjour Tom, Je t'invite à la piscine demain après-midi, la piscine sera chauffée et...	16 juil.

Victor

42 ans, adapte de découvertes culinaires via Uber Eats et fin connaisseur de musique classique sur Spotify

The screenshot shows a Gmail inbox with the following message list:

- Google 3: Alerter de sécurité - Quickstart a désormais accès à votre compte Googlecaroubier1@gmail.com (16 juil.)
- Uber: commande eats overflow been visible victor, cet été, on part ensemble ? (10 juil.)
- developer-accounts: Twitter developer account application [ref_00DA0KO8_5004A1ibmG9.ref] - Hello, Thanks for your interest in buil... (3 juil.)
- developer-accounts: Twitter developer account application [ref.00DA000000KO8.5004A0001ibmG9.ref] - Hello, We've received your a... (2 juil.)
- Twitter Developer A.: Verify your Twitter Developer Account - Email verification Hi caroubier! Thanks for applying for a Twitter Developer a... (2 juil.)
- Spotify: choisissez conditions premium spotify Dernière chance : 0,99 € pour 3 mois entiers de Spotify Premium. - Dépêchez-vous, il ne vous reste plus que qu... (29 juin)
- Spotify: choisissez conditions premium spotify Derniers jours : 0,99 € pour 3 mois de Spotify Premium - Envie de musique sans pub, hors connexion, et de pou... (26 juin)
- Uber Eats: commande eats overflow been visible Pizza Hut et Sprite vous régalent. - Frais de livraison offerts ! Summer vibes. Ça y est. On y est. ENFIN. Ça sent l'été ... (24 juin)
- Reçu Uber: [REDACTED] normal Votre commande Uber Eats de dimanche soir Total: 16,10 € dim., juin 23, 2019 Merci d'avoir passé commande, victor Voici votr... (23 juin)
- Spotify: choisissez conditions premium spotify 0,99 € pour 3 mois de Spotify Premium : c'est presque trop tard - Captez les bonnes ondes. Écoutez vos playli... (23 juin)
- Uber Eats: commande eats overflow been visible J'adooore les sushis ! - Des promotions tout le week-end sur les sushis ! Envie de makis ? Pas de sushi. Voilà une... (22 juin)
- YouTube: caroubier centre conditions gmail Modifications des Conditions d'Utilisation de YouTube - Afin de continuer à améliorer la transparence et la commu... (19 juin)
- Spotify: choisissez conditions premium spotify 0,99 € 3 mois de Spotify Premium : l'offre s'achève bientôt - Spotify Premium, ce sont des millions de chanson... (17 juin)
- Reçu Uber: [REDACTED] normal Votre commande Uber Eats de dimanche soir Total: 18,40 € dim., juin 16, 2019 Merci d'avoir passé commande, victor Voici votr... (16 juin)
- Spotify: choisissez conditions premium spotify 0,99 € pour 3 mois de Spotify Premium. Laissez-vous porter. - Parfois, un seul titre suffit. Choisissez-le avec S... (11 juin)
- Reçu Uber: [REDACTED] normal Votre course de vendredi soir en Uber - Total: 6,07 € ven., juin 07, 2019 victor, merci d'avoir utilisé Uber Nous espérons que vous... (8 juin)
- Uber Eats: commande eats overflow been visible 1 acheté, 1 offert chez KFC pour soutenir les Bleus. - La coupe à la maison ! Elles vont nous mettre des étoil... (7 juin)
- Spotify: choisissez conditions premium spotify 0,99 € pour 3 mois de Spotify Premium : zappez à volonté - Zappez de chanson en émotion avec Spotify Prem... (4 juin)

Conclusion

Ce projet nous a permis d'appréhender le métier de data engineer grâce au pipeline de micro service mis en place.

De plus, nos acquis en machine learning ont été consolidés sur la partie de création de modèle.

Le développement d'une solution complète a permis de mettre en évidence les problématiques liées au monde de l'entreprise avec la mise en production de projet.

Nous avons aussi appliqué la méthodologie Agile sur ce projet afin de mieux répartir les tâches et leurs importances propres. Les avancements ont été suivis de manière hebdomadaire, permettant ainsi une résolution rapide des problèmes.

Le projet ayant été pensé directement au départ comme modulable, il sera facile d'ajouter de nouveaux micro-services et ainsi continuer à faire vivre le projet à l'avenir.

Annexe

Définition :

Trello est un outil de gestion de projet en ligne, lancé en septembre 2011, et inspiré par la méthode Kanban de Toyota. Il est basé sur une organisation des projets en planches listant des cartes, chacune représentant des tâches.

Flask : Flask est un framework open-source de développement web en Python. Son but principal est d'être léger, afin de garder la souplesse de la programmation Python, associé à un système de templates. Il est distribué sous licence BSD.

Python : Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.

Docker : Docker est un logiciel libre permettant facilement de lancer des applications dans des conteneurs logiciels.

Micro-service : En informatique, les micro-services sont un style d'architecture logicielle à partir duquel un ensemble complexe d'applications est décomposé en plusieurs processus indépendants et faiblement couplés, souvent spécialisés dans une seule tâche. Les processus indépendants communiquent les uns avec les autres en utilisant des API indépendantes du langage de programmation. Des API REST sont souvent employées pour relier chaque micro-service aux autres. Un avantage avancé est que lors d'un besoin critique de mise à jour d'une ressource, seul le micro-service contenant cette ressource sera mis à jour, l'ensemble de l'application restant compatible avec la modification, contrairement à la totalité de l'application dans une architecture classique, par exemple une architecture trois tiers. Cependant, le coût de mise en place, en raison des compétences requises, est parfois plus élevé.

Google Cloud Platform : Google Cloud Platform est une plateforme de cloud computing fournie par Google, proposant un hébergement sur la même infrastructure que celle que Google utilise en interne pour des produits tels que son moteur de recherche.

idMail : un identifiant unique correspond à un mail.

