



UNIVERSITÉ LUMIÈRE LYON 2  
MASTER 2 SISE

STAGE DE FIN D'ÉTUDES CHEZ AGORA OPINION

---

## Analyse de données et Initiation d'une démarche de Machine Learning

---

**agoraopinion<sup>®</sup>**

The logo for Agora Opinion consists of the word 'agoraopinion' in a bold, lowercase, sans-serif font. The letters are primarily blue, with the 'o's being red and green respectively. A registered trademark symbol (®) is positioned above the final 'n'.

---

*Stagiaire :*  
Jean-Luc YONGWE

*Tuteurs*  
Mr. C-E. SERRE  
Mr. Julien AH-PINE

12 septembre 2017

## Table des matières

<b>1 Présentation de l'entreprise</b>	<b>5</b>
1.1 Une entreprise innovante . . . . .	5
1.2 Activité de l'entreprise . . . . .	5
<b>2 Introduction</b>	<b>8</b>
<b>3 Missions</b>	<b>10</b>
3.1 Analyse de données . . . . .	11
3.1.1 Graphical User Interface (Data Analysis) . . . . .	14
3.2 Démarche Machine Learning borne Notio . . . . .	20
3.2.1 Schématisation du projet . . . . .	21
3.2.2 Cahier des charges . . . . .	28
3.2.3 Fonctionnalité : classification sémantique . . . . .	38
3.2.4 Fonctionnalité : ratings . . . . .	40
3.2.5 Fonctionnalité : random . . . . .	41
3.2.6 Fonctionnalité : détections . . . . .	41
3.2.7 Fonctionnalité : news . . . . .	44
3.2.8 Clustering des Key Performance Indicators . . . . .	45
3.3 Use case : NeWays . . . . .	53
<b>4 Conclusion</b>	<b>61</b>
<b>5 Lexique</b>	<b>63</b>
<b>6 Annexe</b>	<b>64</b>
<b>7 Technologie</b>	<b>65</b>
<b>8 Bibliographie</b>	<b>66</b>

## NOTE DE CONFIDENTIALITÉ

*Le présent rapport est classé comme confidentiel. En conséquence, la divulgation de son contenu à une personne extérieure au corps professoral du Master de Statistiques Informatique pour la Science des donnEes de l'Université Lumière Lyon 2 ou à une personne extérieure à l'entreprise Agora Opinion est interdite*

## Remerciements

Le présent rapport est le fruit d'une belle collaboration avec les équipes de la société Agora Opinion pendant toute la durée du stage.

Mes remerciements vont en premier lieu à mon tuteur de stage Monsieur Claude-Emmanuel Serre directeur des opérations de chez Agora Opinion pour la confiance accordée en me choisissant pour les accompagner dans cet ambitieux projet présenté dans ce document , pour son implication dans ma bonne compréhension du but à atteindre , pour son ouverture d'esprit vis-à-vis des initiatives que j'ai pris et de m'avoir faciliter la tache dans la mesure de son possible et surtout par dessus tout son approche *business* de l'application de mes travaux.

A Monsieur Ricco Rakotomalala sans qui je ne serais pas ici et n'aurait pas eu l'opportunité d'acquérir les connaissances et informations utiles pour mener un tel projet.

(notre rencontre du 13 Septembre 2016 restera à jamais gravé dans ma mémoire)

Bruno Garcia pour son regard extérieur à la finition des travaux.  
Je tiens aussi à remercier mes collègues Corentin Thiercelin mon chef de projet pour avoir tenu compte et faciliter ma participation dans les projets menés en collaboration , Adonis Harouk pour avoir accédé à mes demandes de dernières minutes.. Marie Vinot , Jérémy Deleat et Simon Mazoua pour la facilitation de mon intégration dans l'équipe, pour avoir fait preuve de leur intérêt dans mes travaux m'amenant ainsi à chercher à les vulgariser.

A Monsieur Julien Ah-Pine mon tuteur pédagogique à l'université pour son apport dans ma compréhension du domaine de l'apprentissage machine et du traitement du signal.

A mes soeurs et mes parents et mes amis pour leur soutien indéfectible "*through thick and thin*"...

## Résumé

Ce document constitue le rapport du stage de fin d'année effectué au sein de l'entreprise *Agora Opinion* , une entreprise spécialisée dans la mesure de satisfaction client/collaborateur grâce à des bornes de sondage physiques ou connectés munies de boutons ayant l'apparence de smileys .

Ce stage a été mené dans le but de les accompagner dans leurs premières démarches de leur projet d'intégration de méthodes de *Machine Learning* dans leur(s) produit(s) futurs.

Les utilisateurs des bornes connectées sont amenés à voter sur des questions affichées sur un écran dont l'alternance se faisait de manière aléatoire.

Le but de l'intégration du Machine Learning sur ces bornes de sondages est d'améliorer d'une part la pertinence de la question qui est posée et d'autre part le moment où elle est posée aux utilisateurs en se basant sur des données de votes remontées en temps réel.

Dans ce document il y sera présenté les besoins des clients,les outils utilisés, les différentes méthodes d'apprentissages , les choix techniques, les problèmes rencontrés et leur résolution.

## Abstract

This document provide a report about the internship at Agora Opinion, a company that has specialized in the customer experience measurement through terminal either physical or connected with smileys look-alike buttons. The main goal of this internship was to lend a hand in their efforts to integrate Machine Learning approaches in their future products/services.

The end users of the connected terminals are usually called upon to vote on different questions displayed on a screen but their alternation was random-based.

The point of integrating Machine Learning is to improve on one side the relevancy of the question surveyed , and the moment where it is surveyed to the end users thanks to a real-time data analysis on the other. In this document it will be reported the client's needs , technologies used , machine learning implementation, technical choices , the issues and how they have been resolved.

# 1 Présentation de l'entreprise

## 1.1 Une entreprise innovante

Agora Opinion est une start-up né en 2015 opérant dans le domaine du *Customer Experience* et qui s'est spécialisé dans la mesure de satisfaction client en exploitant la technologie des objets connectés appelé IoT (*Internet of Things*) au travers différents types de réseaux de télécommunications et de bornes de sondage , l'IoT est actuellement une technologie qui monte en puissance en France.

## 1.2 Activité de l'entreprise

Les bornes de satisfaction commercialisées par Agora Opinion sont diverses et adaptables selon le besoin du client en termes de technologie, de taille et de forme. On trouve essentiellement les bornes connectées et les bornes dites USB.

### Les types de bornes



Un objet connecté est un objet électronique connecté sans fil et partageant des informations avec un ordinateur, une tablette électronique, un smartphone ou autre appareil.

## 1.2 Activité de l'entreprise

---

Les bornes de sondage connectés utilisent différentes technologies de télécommunication afin de remonter les données de votes vers une plateforme web, ces technologies sont LoRA P2P , LoRaWAN , Sigfox et GSM ; des protocoles essentiellement dédiées au domaine des objets connectées. Le chemin, la méthode d'envoi et de traitement des votes se diffèrent d'un protocole à l'autre.

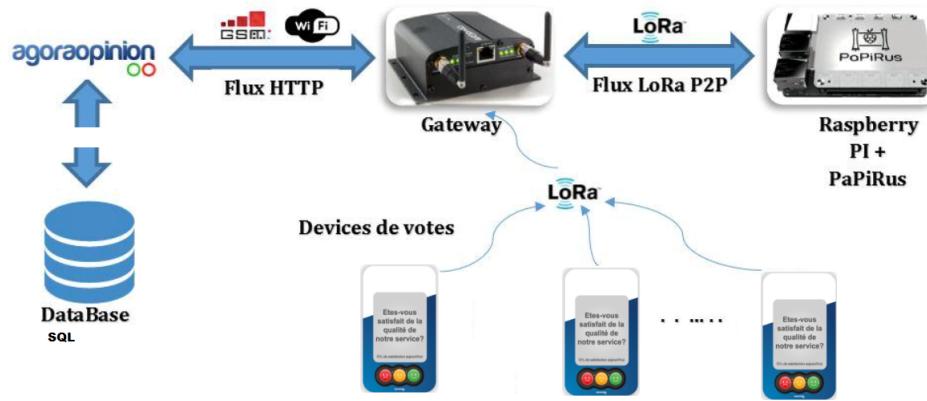
Un protocole de communication est l'établissement et la spécification de règles pour un type de communication particulier. Ces règles permettent aux machines de communiquer entre elles suivant une procédure bien définie.

Les réseaux IoT, Sigfox et Lora répondent à la demande des objets de faire passer des signaux de faibles puissances sur de longues distances et ce, à moindres coûts en terme d'énergie. Cela permet aux objets de garder une autonomie sur batterie importante allant jusqu'à 10 ans pour certains types d'objets.

Les bornes utilisant le protocole LoRa P2P, doivent obligatoirement passer par une gateway (passerelle) qui est capable de communiquer en plusieurs protocoles et qui sert à récolter les votes issus des bornes et de les envoyer sur internet.

(voir schéma ci-dessous)

## Réseaux type LoRa P2P



Les informations remontées depuis ce type de réseaux sont :

- le nombre de pushs sur un bouton *yes* (vert) , *neutre* (jaune) et *no* (rouge) ou *yes* et *no* (selon qu'il s'agisse d'une borne à trois ou deux boutons)
- le nombre total de votes
- l'identifiant de la borne
- un datetime (ex : 2017-05-09 10 :35 :29)

L'intervalle de temps sur lequel les informations (appelés trames) sont remontés peut être paramétré, par borne de sondage soit de 10 minutes ou 30 minutes. La batterie aura une plus grande durée de vie pour le paramétrage où l'intervalle est le plus grand, soit 30 minutes ici.

Quant aux bornes de sondage "*USB*", l'envoie et la collecte des données de votes ne se basent pas sur la communication radiofréquence mais sur un enregistrement direct sur une clé usb.

Les résultats des votes sont visualisables sur une plateforme web indépendante de celles des bornes connectés par Drag and Drop.

## 2 Introduction

A une époque où l'image d'une entreprise auprès du grand public est son atout majeur, disposer de moyens de la mesure de satisfaction de ses clients ou collaborateurs vis-à-vis du service fournie est capital.

C'est en faisant de cette activité son cœur de métier que Agora Opinion est devenu un des leaders français de la mesure de la satisfaction dans l'expérience client "Cx" au travers une gamme de bornes de sondage qui peuvent être connectés ou non , son avantage principale sur ses concurrents est que Agora Opinion conçoit et fabrique elle-même ses bornes.

Les clients majeurs de la société Agora Opinion commencent à montrer un réel besoin relatif au sondage de leurs clients ou leurs collaborateurs sur des questions dont l'alternance se ferait une période donnée de façon automatique (exemple une fois ou plusieurs fois par jour) et dont la pertinence du choix leur permettrait de mettre le doigt sur un aspect qui s'avère être sujet d'insatisfaction.

C'est face à un tel besoin que Agora Opinion a publié une offre de stage pour un poste de "Data Analyst/Data Scientist" sur la plateforme de stages de l'Université Lumière Lyon 2 , au travers laquelle j'ai eu l'opportunité d'y postuler.

L'offre de stage s'avérait intéressante parce qu'elle mêle deux disciplines pour lesquels j'ai eu l'occasion d'être formé à l'université à savoir le Text-Mining (extraction des connaissances selon un critère de nouveauté ou de similarité dans des textes) et le Machine Learning (l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches par des moyens algorithmiques plus classiques.)

L'objectif principal du sujet de stage est de concevoir un système autonome permettant d'alterner l'affichage de questions à sonder sur une période donnée ledit système se basera sur une analyse des données récupérés en temps réel (ie : analyse faite de manière automatique et mise à jour de façon régulière) pour affiner le choix de la question à sonder, cette dernière tache fut mon plus grand challenge.

En parallèle j'ai participé à la rédaction de rapports d'analyse pour les clients de la société Agora Opinion ce qui m'a amené à chercher à automatiser leur process d'analyse de données clients au travers un Graphical User Interface spécifiquement dédié en complément du dashboard de la société.

L'apport principale de ce stage a été non seulement l'exploitation et conversion des connaissances acquises en cours vers le sujet de stage , cela englobe un vaste enrichissement en termes de compétences techniques sur les outils employés , mais aussi des compétences humaines nécessaires en environnement d'entreprise.

---

### 3 Missions

La société Agora Opinion a mis sur le marché depuis l'année 2016 une gamme de borne de sondage connecté équipé d'un écran où apparaît une question sur laquelle le public visé est sondé, ladite question est modifiable à distance via une plateforme web spécialement développé à cet usage.

En Avril 2017 elle décide de lancer le projet "*Notio*" dans le but d'intégrer du Machine Learning dans le système de gestion à distance de cette nouvelle borne.

Le présent stage s'est donc déroulé aux cotés des équipes de développement web et de Recherche et Développement dont faisait partie l'auteur de ce document (en tant que stagiaire) qui était en charge du développement de la partie Machine Learning du projet.

L'auteur de ce document était sous la tutelle du directeur des opérations Monsieur Claude-Emmanuel Serre , mais le projet s'est déroulé sous la responsabilité de Monsieur Corentin Thiercelin.

La mission principale qui a été confié se déroule dans un contexte de "*Big Data et Machine Learning*".

A l'heure actuelle ce projet étant encore au stade de développement toute l'attention et les efforts sont portés sur le versant "Machine Learning" mais néanmoins il doit être planifié , mis en place et déployé en anticipant (en gardant en tête) le versant "Big Data" d'un tel projet; car en effet une technologie comme celle de Agora Opinion capable de récupérés des données en temps réels (toutes les 10 minutes) peut accumuler beaucoup de données (coté *Volumétrie* de la Big Data) en peu de temps ce qui fera beaucoup de données à traiter pour le système de Machine Learning.

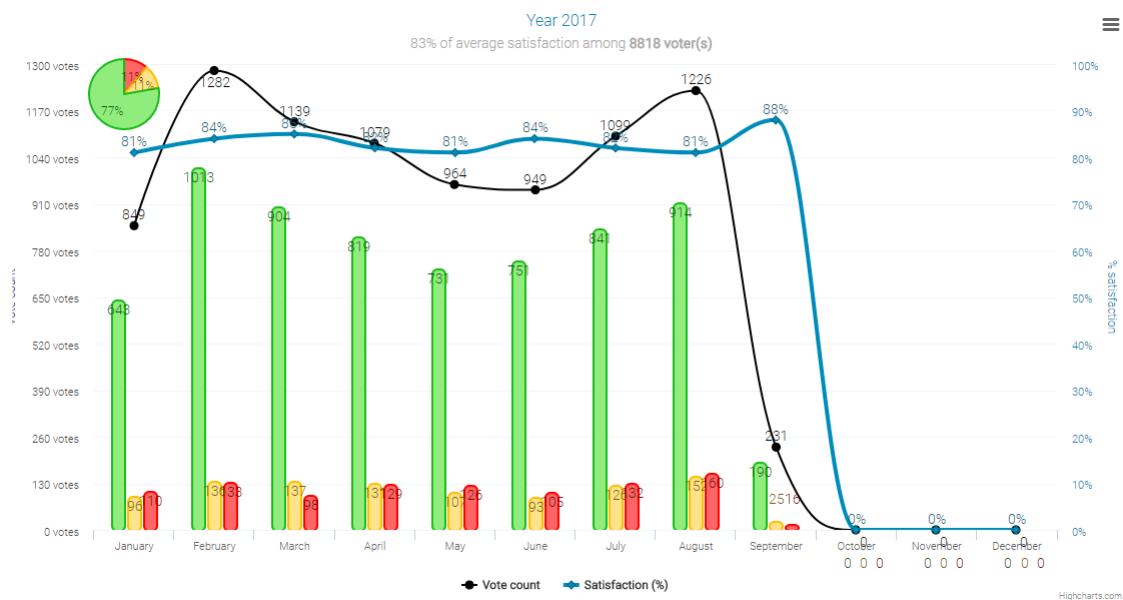
La mission secondaire qui a été confié a été d'aider l'équipe commerciale à mieux exploiter les données de votes horodatés à disposition dans l'optique de la rédaction de bilans (rapport de l'évolution de la satisfaction sur les questions sondées) qui sont remis aux clients.

En marge du dashboard (une plateforme web de data visualisation disponible via le site de Agora Opinion) il a été conçu un Graphical User Interface permettant de traiter la donnée de façon différente avec des options de visualisations de données non disponible sur le dashboard.

### 3.1 Analyse de données

#### 3.1 Analyse de données

De manière général ce sont les équipes commerciales de la société Agora Opinion en contacts avec les clients qui se chargent de la rédaction des rapports d'analyses. L'ensemble de la data visualisation de ces rapports est constitués de graphiques pris sur le dashboard de la société sur le compte du client concerné.



Le graphique ci-dessus est un exemple d'une data visualisation disponible sur le dashboard de la société ; à bien y regarder le client aura trois indicateurs réunis sur le même graphique :

- la courbe de satisfaction (en bleu)
- la courbe du nombre de votes (en noir)
- le diagramme en bâtons des pushs (vert : yes , jaune : neutre , rouge : no)

une manière de faire serait de séparer ces trois indicateurs.

Après la participation à deux rapports rédigés pour des clients de la société ; une automatisation de cette partie de leur process s'est profilé comme potentiellement intéressante , les rapports clients contenant essentiellement les mêmes types de graphiques en termes d'échelles temporelles

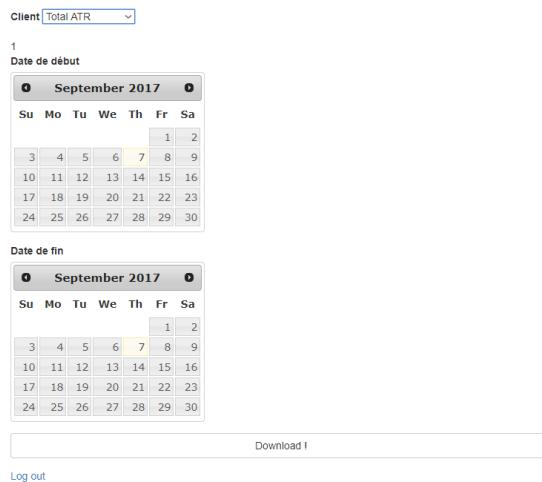
### 3.1 Analyse de données

---

Pour pouvoir aider les commerciaux dans leurs démarches de rédaction de rapports , il a été construit un Graphical User Interface (GUI) permettant de faire de la visualisation de données avec différentes options.

#### Origines des données

Un outil d'extraction de données (baptisé "EXTRACT") directement connecté à leur base de données de production a été conçus par les équipes de développement de Agora Opinion , son objectif double qui est d'un coté de permettre un accès aux données les plus couramment demandés par les équipes de la société (données portant sur les votes émanant de clients de l'entreprise sur les bornes qui sont déployés) , mais aussi de se garder du moindre incident qui serait causé à l'accès de leur base de productions par un non - expert en base de données.



Ci-dessus se trouve un aperçu de cet extracteur.

- L'onglet *client* sert à sélectionner le nom du client désiré
- le *premier calendrier* sert à indiquer à partir de quelle jour on veut commencer l'extraction
- le *second calendrier* sert à désigner jusqu'à quel jour l'on voudrait arrêter l'extraction

### 3.1 Analyse de données

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	client_name aire	section	lieu_exact	utc_datetime	yes	neutre	no	total vote	satisfaction	alert	check	batt	id_vote	id_device				
2	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143633	497				
3	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	1	0	2.75000		0	0		143634	497				
4	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	0	0	21.00000		0	0		143635	497				
5	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		3	1	0	4.087500		0	0		143667	497				
6	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		0	0	1	10.00000		0	0		143686	497				
7	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143717	497				
8	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	0	0	21.00000		0	0		143718	497				
9	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	1	0	3.083333		0	0		143719	497				
10	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		5	0	0	51.00000		0	0		143739	497				
11	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143791	497				
12	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	0	0	21.00000		0	0		143792	497				
13	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	1	0	2.75000		0	0		143793	497				
14	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	2	0	40.75000		0	0		143809	497				
15	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143836	497				
16	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143839	497				
17	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	0	0	21.00000		0	0		143840	497				
18	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	1	0	2.75000		0	0		143841	497				
19	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		4	1	0	50.90000		0	0		143861	497				
20	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143898	497				
21	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	0	11.00000		0	0		143899	497				
22	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	0	0	21.00000		0	0		143902	497				
23	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		2	2	1	50.60000		0	0		143934	497				
24	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		0	0	1	10.00000		0	0		143959	497				
25	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		0	1	0	10.50000		0	0		143960	497				
26	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		1	0	1	20.50000		0	0		143961	497				
27	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		0	0	1	10.00000		0	0		143976	497				
28	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		0	1	0	10.50000		0	0		143985	497				
29	GPCC Veniss Pharmacie	Espace empl Arrivée	██████████		4	0	0	41.00000		0	0		143989	497				

Ci-dessus se trouve un échantillon d'une extraction.

Les données auxquelles ont accès les utilisateurs (membres de la société Agora Opinion) de l'extracteur sont uniformisées , i.e l'entête des champs fournis ne sont pas dépend du client de Agora Opinion qui est consulté mais les données que ces dits champs contiennent ; ces derniers sont les suivants :

- *clientname* : champ désignant le nom du client
- *lieuexact* : champ désignant le lieu où est posé une et une seule borne
- *utcdatetime* : champ désignant le moment exact de la journée où est enregistré la trame de données
- *yes* : champ contenant le nombre de pushs sur le bouton vert
- *neutre* : champ contenant le nombre de pushs sur le bouton jaune
- *no* : champ contenant le nombre de pushs sur le bouton rouge
- *total vote* : champ contentant le nombre de votes émis entre la trame précédente et la trame suivante
- *check* : champ contenant le nombre badging (passage d'un badge des équipes de ménages) , option conçus dans le cas où le client utilise la borne de sondage pour sonder la propreté des toilettes de son bâtiment
- *alert* : champ contenant les alertes envoyés (par sms ou par mail) lorsque l'évolution de la satisfaction est critique sur le dashboard
- *batt* : champ contenant le niveau de batteries d'une borne de sondage
- *iddevice* : champ contenant l'identifiant unique d'une borne de sondage

### 3.1 Analyse de données

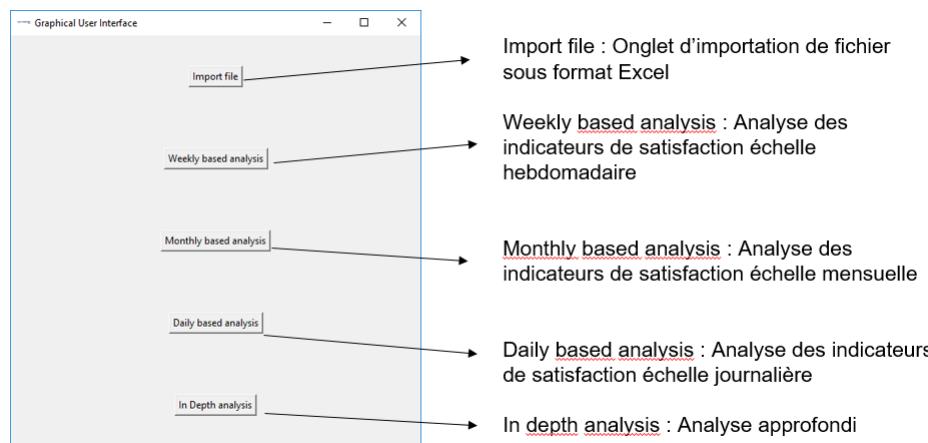
---

#### 3.1.1 Graphical User Interface (Data Analysis)

C'est en s'appuyant sur des données extraites avec l'EXTRACT que le GUI a été construit et validé.

#### Tutoriel d'utilisation du GUI

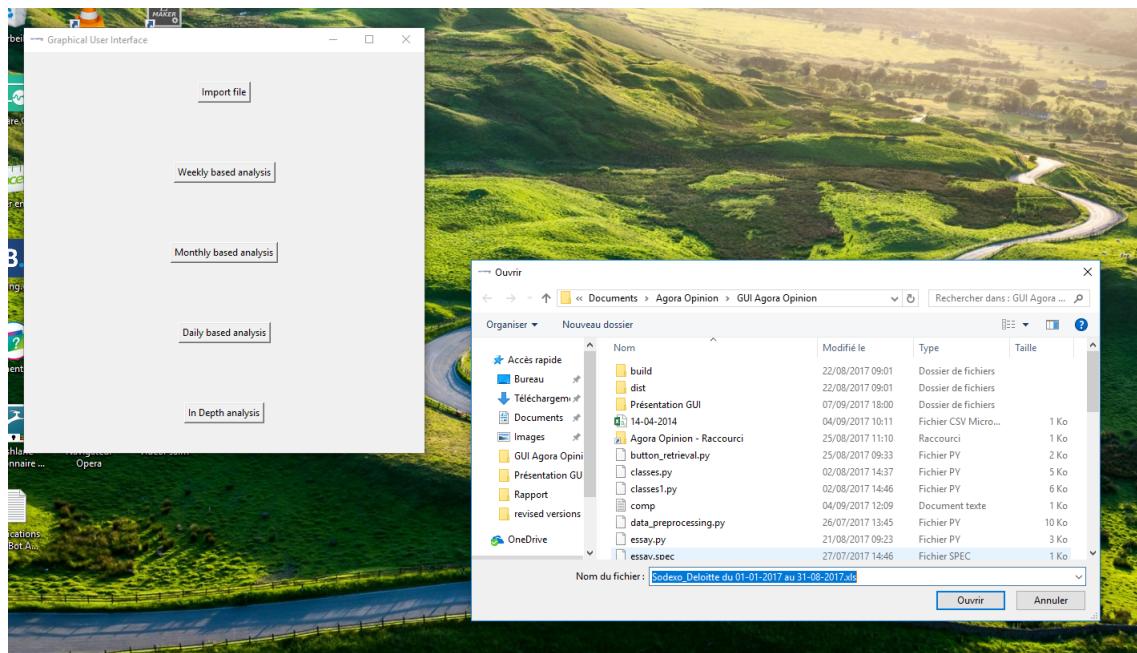
L'utilisateur de l'outil commence par extraire des données depuis l'extracteur (si l'utilisateur est un collaborateur de chez Agora Opinion) , ou en téléchargeant les données concernant son entreprise sur le site web de Agora Opinion



Il est impératif pour tout utilisateur de cet outil de toujours commencer par appuyer par l'onglet "*Import file*" pour importer le fichier de données , dans le cas contraire il se retrouverait à ne rien regarder car il n'y a pas de données incorporé par défaut.

### 3.1 Analyse de données

A l'appui sur l'onglet Import file l'arborescence de l' ordinateur s'ouvre et il suffira à l'utilisateur de y naviguer jusqu'à trouver le fichier de données qui l' intéresse.



Une fois cette étape passée , l'utilisateur a le choix entre 4 onglets qui correspondent chacun à une échelle de temps précise d'analyse.

Chacun de ses onglets d'échelles de temps est muni d'un menu déroulant permettant de voir l'évolution des différents indicateurs du client sur chacune des bornes se trouvant dans ses bureaux (entreprise cliente de Agora Opinion) voir image suivante

Pour la démonstration qui va suivre nous nous servirons des données extraites de la base de données de chez Agora Opinion concernant leur client Sodexo - Deloitte (Sodexo-Deloitte est la division de la société Sodexo qui se charge de la propriété des sanitaires chez la société Deloitte) , qui s'étendent sur une période allant de Janvier 2017 à Aout 2017

### 3.1 Analyse de données

#### Monthly based analysis



Il est visible que tous les indicateurs sont séparés. Il est possible d' avoir accès aux indicateurs pour chacune des bornes déployés chez Sodexo-Deloitte grâce à l'onglet supérieur de la page.

Une analyse rapide montre qu'au mois de Mars la satisfaction sur cet étage est la plus basse , ce qui s'explique par le traffic (nombre de votes total sur la période en abscisse concerné) (graphique en dessous) dans les sanitaires de cet étage en question car plus d'utilisateurs implique plus de risques d'être moins propres ..

On verra aussi que au mois d'Août il y a le moins de votes , ce qui est logique puisque les collaborateurs sont pour la plupart en congés à cette période de l'année.

### 3.1 Analyse de données

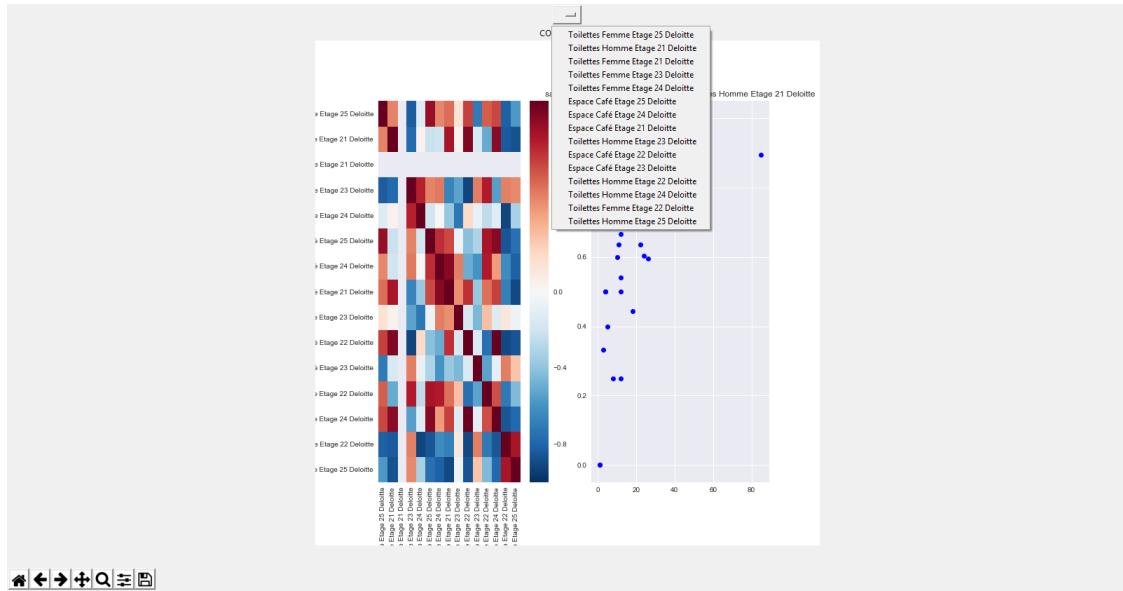
#### Weekly based analysis



Le graphique du traffic montre les collaborateurs de cet étage ne travaillent pas la journée du samedi et dimanche ainsi les satisfactions calculés sur ces jours précis sont issues de "trames perdues" (arrivant en retard dans la base de données)

La nouveauté pour cette échelle de temps est "l'alerting" , (l'alerting est un système d'envoi d'un message ou un mail par Agora Opinion aux équipes de ménages de Sodexo ou autre client pour les prévenir que la satisfaction sur la propreté du lieu est critique ou que il y'a eu plus de 5 pushes rouges à la suite) , les entiers au niveau de l'abscisse représente le numéro de la semaine dans l'année (exemple : 10 correspond à la seconde semaine du mois de Mars).

## In depth analysis

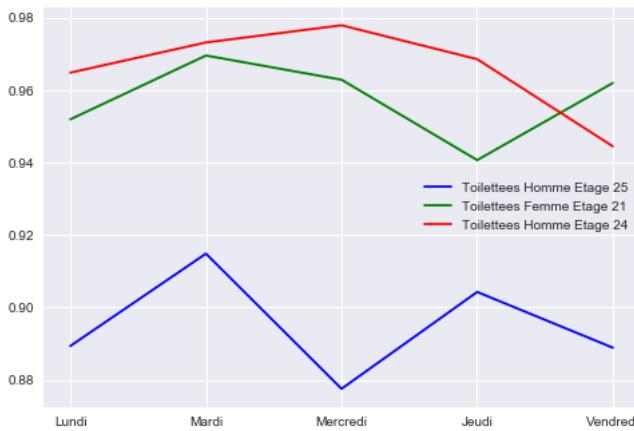


Le graphique de gauche est une représentation de la carte de chaleur de la corrélation qui lie l'évolution hebdomadaire (en termes jours) des différentes bornes.

Ce graphique est appelé une *heatmap*, cela est une (carte thermique) représentation graphique de données statistiques qui fait correspondre à l'intensité d'une grandeur variable (les taux de satisfactions) sur une gamme de tons ou un nuancier de couleurs sur une matrice à deux dimensions. Cette façon de faire permet de donner à des données un aspect visuel plus facile à saisir qu'un tableau de chiffres.

Au plus le rectangle sera rougis au plus la grandeur (taux de corrélation) qui lie les deux bornes sera intense et positive (évolution commune), dans l'autre sens au plus le rectangle sera bleuis au plus la grandeur qui lie les deux bornes sera intense mais négative (évolution inversé)

En zoomant sur la partie inférieure droite de la carte des chaleurs, elle indique que les bornes des étages "*Toilettes Femme Étage 21*" et "*Toilettes Homme Étage 25*" évoluent dans le même sens, tandis que pour la borne "*Toilettes Homme Étage 24*" l'évolution est contraire aux deux premières.



Ce graphique semble bien confirmer ce que nous a été entrevue avec la carte des chaleurs.

Le graphe de droite du widget "*In depth analysis*" cherche à montrer le lien possible entre la satisfaction et le trafic agrégés au niveau jour.

On aura trois cas possible :

- Le nuage de points a une forme ovale et est positionné dans le sens  $y = x$  (première bissectrice) alors cela signifie qu'il existe une relation linéaire entre les deux mesures.

$$satisfaction = coefficient * \text{nombre total de votes}$$

où coefficient est un nombre réel positif.

- Le nuage de points a une forme ovale et est positionné dans le sens  $y = -x$  (perpendiculaire de la première bissectrice) alors cela signifie qu'il existe une relation linéaire entre les deux mesures.

$$satisfaction = -coefficient * \text{nombre total de votes}$$

- Le nuage de points n'entre pas dans aucune des deux catégories précédemment défini et on ne peut rien en dire.

### 3.2 Démarche Machine Learning borne Notio

Devenir le leader incontesté de son domaine d'activité requiert une innovation régulière dans les services qui sont fournis et surtout faire preuve d'une valeur ajoutée rare sur le marché. C'est dans cette optique que les équipes dirigeantes de Agora Opinion ont décidés de faire les deux (innover et valoriser) pour révolutionner leurs services et se distinguer de leurs concurrents.

Ils ont innovés en se lançant dans l'exploitation de la technologie de l'IoT depuis 2015, ils ont à cœur de valoriser les données recueillis auprès de leurs clients en intégrant le Machine Learning dans le système de gestion à distance des bornes connectés.

Le produit alpha (sens premier à être lancé) de cette nouvelle vision est la borne Notio

Le nom "Notio" est le mot latin qui se traduit par notion i.e qui à rapport à la connaissance..

C'est en connaissance de cause qu'un tel nom a été choisis car c'est leur premier produit qui saura s'adapter à son client en fonction de ses besoins.

De façon général la société Agora Opinion peut regrouper ses clients par typologies grâce à leurs différents besoins.

Des typologies qui ont des attentes et des besoins différents les uns des autres, ont menés à la réflexion de concevoir différents types de méthodes.

Pour plus de détails voir la partie sur le cahier des charges.

### 3.2.1 Schématisation du projet

#### Philosophie du projet *Notio*

L'idée originelle du projet Notio est de traduire dans notre nouveau monde de "*L'Intelligence Artificielle*" né dans cette nouvelle ère de la "*Big Data*", la philosophie fondatrice de la société Agora Opinion qui est d'aider les entreprises-clientes à améliorer leur offre et services auprès de leur public grâce à une méthode de "*feedback*" peu couteuse en temps et en énergie.

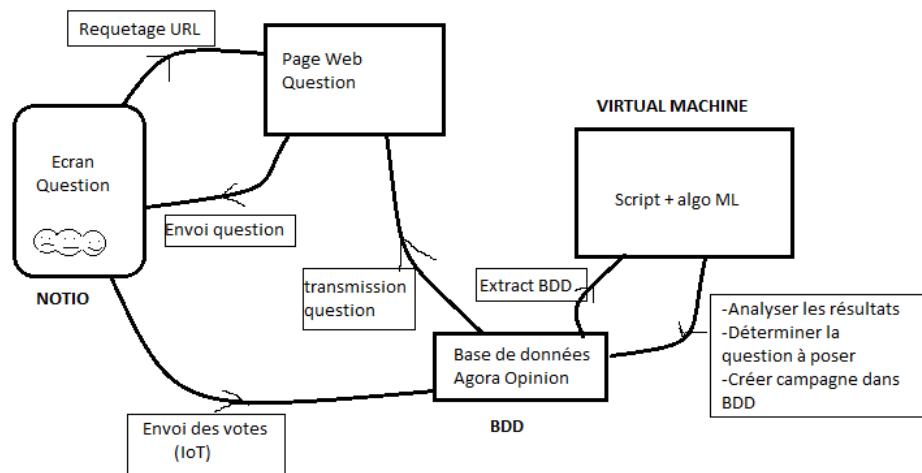
Une première manière de faire était le sondage classique par bornes physiques (question posée sur papier A4) , qui ont l'avantage d'être un moyen de sondage relativement court (3-5 secondes maximums) et qui ne nécessite pas de se connecter au site d'entreprise (ce que les personnes ne font pas forcément).

En introduisant une borne avec écran , il était nécessaire d'automatiser tout le process de sondage du public cible ; étant déjà en possession d'une technologie permettant de récupérer les données en temps réels l'étape suivante est d'acquérir une dimension d'analyse de données jamais exploiter jusqu'ici au travers les méthodes d'apprentissage machine (*Machine Learning*) permettant de faire parler la *Volumétrie* de leurs bases de données et la *Vélocité* de leur technologie.

Ainsi toute la réflexion du projet s'est tournée vers des questions tels que :

- *Comment automatiser les sondages de questions ?*
- *Comment déceler des tendances , en temps-réel , qui auraient été visibles après coup ?*
- *Comment réagir de manière plus rapide , à d'éventuelles défis ?*
- *Comment interroger le public sur la question la plus pertinente au meilleur moment ?*

Le schéma suivant est un résumé imagé du fonctionnement global du projet *Notio*



Lorsque le public cible votes sur la borne en appuyant sur le bouton de la couleur désirée (choix entre un bouton vert : yes , bouton jaune : neutre , bouton rouge : no).

Leurs votes sont transmis à la base de données de production de la société Agora Opinion au travers les réseaux de télécommunications de l'IoT ; en fonction des votes la méthode d'apprentissage procédera à une analyse adéquate et enverra dans la file d'attente des indicateurs pour une nouvelle campagne de sondage (plage temporelle [date de début et de fin] , identifiant de la question , identifiant de la borne [borne où afficher la question]).

Un site web (spécialement conçu) requêtera la base de données et enverra la question à la borne.

## Cadre

Les différentes méthodes dont le fonctionnement sera expliqué pour chacune d'elle dans la suite de ce document reposent sur différentes tables de données dont disposent la société Agora Opinion ,ces tables de données sont les suivantes :

1. *nao\_question\_bank* est une table contenant les identifiants des questions et leur format textuelle
2. *nao\_prog\_notio* est une table contenant les spécifications (paramétrisation personnalisée ) des clients
3. *nao\_kpi* est une table permettant de lier l'identifiant d'un client avec les KPIs qu'il désire exploiter (sonder)
4. *nao\_votes* est une table où sont enregistrés les votes émis sur les bornes en "temps réel"
5. *nao\_campaign* est une table permettant de préparer une campagne (la borne s'y réfère dans le but de savoir quelle question sondé)
6. *nao\_device* est une table contenant l'identifiant client accompagné de l'identifiant du device en exploitation dans ses bureaux ou locaux

On appelle "*KPI*" (Key Performance Indicator) un ensemble de questions portant sur un même aspect ou sujet d'un service qui est fourni par le client de la société Agora Opinion.

exemple :

KPI : espace de travail

questions :

- Les locaux contribuent à me rendre productif
- Que pensez-vous de votre/vos espace(s) de restauration ?

On appelle une "*campagne*" une durée de temps durant laquelle on interroge les utilisateurs de la borne Notio sur une question d'un sujet (appelé Key Performance Indicator) précis.

## Présentation du contenu des tables

### 1.nao\_question\_bank

NomChamp	Description	Format
idquestion	identifiant de la question	entier
question	question à poser	texte
idkpi	identifiant du sujet à sonder	entier
kpiname	sujet ou aspect à sonder	texte
threshold	satisfaction minimale de la question	real
rate	note de la pertinence de la question	real
undertreshold	temporisateur pour la mise à jour de la note	entier
status	profondeur de la question	text

exemple :

id_question	question	id_kpi	kpi_name	sector	rate	threshold	under_threshold
Filtre	Filtre	Filtre	ass	☒	Filtre	Filtre	Filtre
141	La réception f...	14	assurance	accueil	0.85	0.8	5

### 2.nao\_prog\_notio

NomChamp	Description	Format
idclient	identifiant client	entier
votetreshold	nombre de votes minimale pour significativité	entier
dureecampaign	durée de sondage sur une question	entier
profil	profil du client	entier
usagetype	type d'usage (courte ou longue durée)	entier

exemple :

id_client	vote_threshold	duree_campaign	profil	usage_type
Filtre	Filtre	Filtre	Filtre	Filtre
201	25	1	3	2

### 3.nao\_kpi

NomChamp	Description	Format
idclient	identifiant du client	entier
idkpi	identifiant du kpi	entier
kpiname	sujet ou aspect à sonder	texte

exemple :

<b>id_client</b>	<b>id_kpi</b>	<b>kpi_name</b>
Filtre	Filtre	Filtre
100	13	tangibilité

### 4.nao\_votes

NomChamp	Description	Format
iddevice	identifiant de la borne	entier
datetime	heure d'enregistrement d'une série de votes	datetime
yes	nombre de pushs bouton vert	entier
neutre	nombre de pushs jaune	entier
no	nombre de pushs rouge	entier

exemple :

<b>device</b>	<b>datetime</b>	<b>yes</b>	<b>neutre</b>	<b>no</b>
Filtre	Filtre	Filtre	Filtre	Filtre
497	2016-08-01 0...	1	0	0

### 5.nao\_campaign

NomChamp	Description	Format
idcampaign	remplir	entier
iddevice	identifiant de la borne en exploitation	entier
idclient	identifiant du client	entier
idquestion	identifiant de la question sonder	entier
idkpi	identifiant du sujet sonder	entier
frombegincampaign	date de début du sondage	datetime
toendcampaign	date de fin du sondage	datetime

exemple :

id_campaign	id_device	id_client	id_question	id_kpi	m_begin_campaign	o_end_campaign
Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre
2405	507	101	143	14	2016-08-01	2016-08-05

### 6.nao\_device

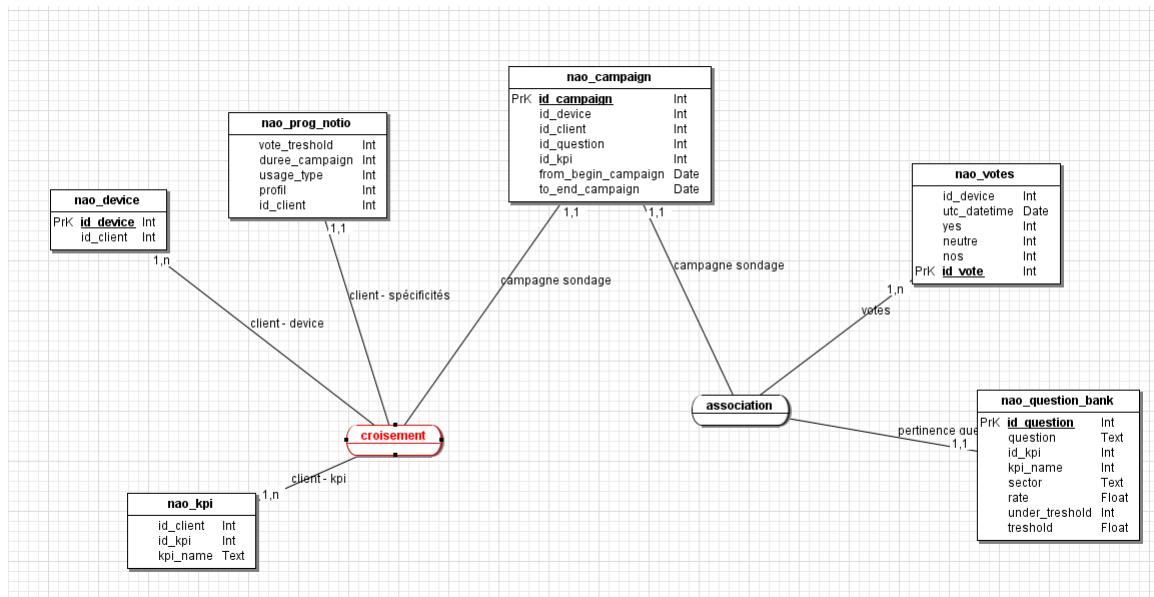
NomChamp	Description	Format
idclient	identifiant du client	entier
iddevice	identifiant du device en exploitation chez le client	entier

exemple :

id_client	id_device
Filtre	Filtre
100	494

## Interaction des tables

### Modèle Conceptuel de Données



Le système de gestion de campagne commencera par croiser les tables **nao\_prog\_notio**, **nao\_kpi**, **nao\_device** dans le but de trouver les spécificités de chaque client , à savoir quels sont les identifiants des devices en exploitation , les KPIs à sonder et les paramètres de système d’alternance des questions ; ensuite il regardera les données de votes de la campagne en cours dans la table **nao\_votes** et agira en fonction en extrayant la question adéquate dans la table **nao\_question\_bank**.

### 3.2.2 Cahier des charges

Une analyse rapide de la clientèle de Agora Opinion permet de faire émerger 3 profils (typologies) distincts :

#### Profil 1 : Retail

Par définition le retail est un terme anglais pour désigner l'activité de commerce de détail, à savoir une activité commerciale effectuée à destination du consommateur final qui consiste à vendre un bien dans l'état où il a été acheté.

#### Profil 2 : Facility Management

Les services généraux désignent l'ensemble des services nécessaires au fonctionnement normal d'une entreprise.

On peut citer par exemple :

- la gestion du courrier
- les achats de fournitures, la gestion de l'entretien des bâtiments, des espaces verts
- gestion des locaux techniques, des systèmes d'incendie, de sécurité, des droits d'accès, des énergies : électricité, froid, chauffage.

L'expression désigne généralement une activité interne à une entreprise.

#### Profil 3 : Ressources Humaines

Les services RH ont de multiples tâches au sein d'une entreprise (ex : gérer les problèmes, la paie ..) dont la principale est d'être un médiateur avisé entre le personnel et la direction générale, c'est aussi à ce service qu'est confié la gestion de la "vie en entreprise" des collaborateurs , à savoir :

- Faciliter l'accès à l'information au sein de l'entreprise
- Favoriser l'appropriation des nouvelles compétences nécessaires à l'entreprise. etc..

### Besoins clients par profils :

#### *Retail :*

Un tel profil aura principalement besoin de faire de l'identification de points d'insatisfaction pour pouvoir entreprendre des actions d'amélioration de l'expérience client , i.e :

- identifier si le service est rapide
- satisfaction de l'accueil (interaction initiale avec les employés)
- expérience magasin (environnement → propreté , orientation..)
- réponse au besoin client (correspondance requête client - réponse apporter)

#### *Facility Management :*

Un tel profil aura principalement besoin de sonder les utilisateurs de la borne (collaborateurs de leur entreprise-cliente) pour avoir une vision précise de la qualité des prestations fournies auprès de leurs clients , dans les divers domaines où il intervient chez lui (client du prestataire FM) , i.e :

- qualité de la propreté dans les lieux "publics" (bureaux , sanitaires , espaces de vies..)
- satisfaction de la maintenance des installations
- satisfaction de la maîtrise des couts (éco-responsabilité)

#### *Ressources Humaines :*

Le but principale dans ce cas précis est de sonder la "vie" en entreprise des collaborateurs dans le but de :

- connaître / évaluer le bonheur au travail des collaborateurs
- prévenir des risques psycho-sociaux liés au travail (burnout, dépression..)
- prévenir des grèves / arrêt de travail
- prévenir des accidents de travail

### Description du cas d'utilisation

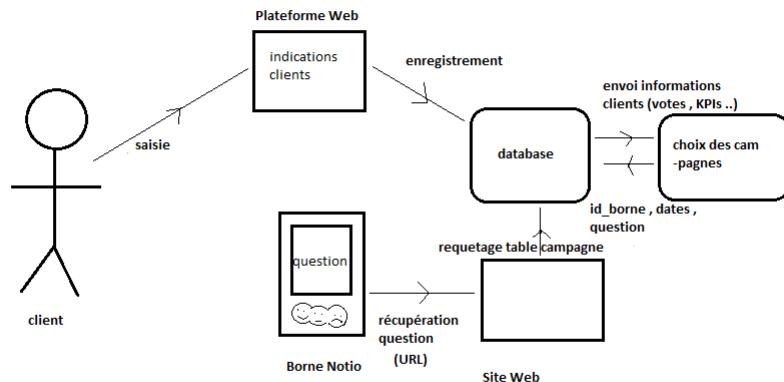
La gestion des campagnes de manière automatisée permet au client en possession de la borne Notio de ne pas avoir à se préoccuper de changer la question lui-même , le seul effort à fournir de sa part est de renseigner certaines indications le concernant.

Ces indications sont par exemple :

- la durée moyenne des campagnes
- le nombre de collaborateurs par lieux de déploiement de la borne
- les sujets de sondages

Le renseignement de ses indications fait partie du système de gestion de campagnes de sondages.

*Interaction client-système Agora Opinion*



## Objectifs et besoins d'affaires

Pour les clients d'Agora Opinion la gestion automatisé des campagnes de sondage fait partie des clés de l'amélioration de leurs services.

Plus il y aura de personnes qui feront appel à ce service plus il sera amélioré et deviendra pertinent.

Spécifiquement la solution de gestion de campagnes doit répondre aux objectifs :

1. *Precision* : poser la bonne question au bon moment de la journée
2. *Rapidité* : Mettre en évidence les points à améliorer parmi les services du client le plus vite possible
3. *Statistique* : Les données associées à une campagne doivent pouvoir être analysé dans le but d'aider à l'amélioration du choix de la campagne
4. *Interaction* : Les questions posées sont en liens avec des événements et données externe ou interne au bâtiment. Par exemple la température du bâtiment, ce qu'il y a eu au repas du midi, une news de l'entreprise...

## Solutions

La méthode de gestion automatisé des campagnes pourra se décliner en deux types d'utilisation , soit une utilisation dite de "courte durée" et une utilisation dite de "longue durée"

Une utilisation de courte durée sera une utilisation d'une durée de 1 à 2 mois (par exemple) , tandis qu'une utilisation de longue durée sera une utilisation d'une durée de 1 an au moins.

Différents profils de clients avec une offre de gestion qui se décline en deux formules peuvent éventuellement entraîner le développements de plusieurs systèmes de gestion.

Il se peut que d'autres découvertes / constatations dans le futur mène au développement de versions spécifiques ou tout simple "upgrader" les versions en exploitation .

Voici une description de leur fonctionnement tel que mise en adéquation avec les besoins clients dans un *premier temps* (premiers développements)

Version courte durée :

L'entreprise cliente aura à sa disposition la borne de sondage pendant 1 à 2 mois, elle aura donc besoin de rapidité et de précision (en terme de mettre le doigt sur l'aspect d'amélioration)..

1 à 2 mois de recueil de données ne sont pas estimés comme suffisants pour clairement déceler des tendances.

Une manière possible de faire pourrait être de croiser les données que le client fait émaner de ses utilisateurs avec ceux d'autres entreprises clientes ayant la même typologie / profil qui auraient déjà sondées les mêmes sujets.

Le choix d'une campagne peut dépendre de divers éléments que l'on peut essayer de résumer en répondant aux questions suivantes, qui peuvent être perçus comme des critères de discriminations :

Il faut savoir que les questions sondées ont toutes un "*status*" qui est un critère de distinction de la profondeur d'interrogation..

exemple : une question de status *main* sera de type générale :

*Êtes-vous en forme aujourd'hui ?*

Une question de status *secondaire* est une question un peu plus pertinente :  
*A quelle fréquence vous sentez-vous stressé au travail ?*

*Comment choisir une question ?*

En se basant sur un système de note qui permettra de classer les questions dont le *status* sera "*secondaire*" par ordre de priorités en fil d'attente de sondage

*Quand choisir une question ?*

Une nouvelle campagne est lancée à la fin de chaque campagne

*Comment choisir la question suivante ?*

La question suivante sera la question ayant la plus grande note , en dehors de celle qui est sondé au moment du changement

*Combien de temps afficher une question ?*

En théorie dépendamment du nombre de sujets (KPIs) requis par le client et le nombre de question qu'on lui préconise de poser , le temps d'une campagne sera le résultat de la division du temps total de déploiement de la borne (en terme de jours) divisé par le nombre de KPIs multiplier par le nombre de questions préconiser. ( $\frac{\text{nbjours}}{\text{nbkpis} \times \text{nbquestions}}$ )

Version longue durée :

Le client aura à sa disposition la borne pendant plus d'un an , il aura besoin de précision et de statistiques.

Un an ou plusieurs mois de données est suffisant pour commencer à déceler des tendances.

Une manière possible de faire est d'agir en deux temps , un premier temps dit "apprentissage" et un second temps dit "application".

*Période d'apprentissage*

Durant la période d'apprentissage la borne sera déployé sans contrainte aucune si ce n'est la prise en compte des spécificités clients (ceux qu'il aura saisie dans la plateforme de Agora Opinion conçu à cet effet)

Quant aux critères de discriminations ils seront les suivants :

*Comment choisir une question ?*

Choisir une question de manière aléatoire parmi les sujets de sondage préalablement choisis par l'entreprise cliente

*Quand choisir la question ?*

à la fin de chaque campagne en débuter une autre

*Combien de temps afficher une question ?*

Une semaine par défaut , si le client ne spécifie pas un temps que lui estime comme étant nécessaire de son côté.

Mais de manière réaliste cela va dépendre de la typologie client :

- Retail : formulation théorique valable (jamais les mêmes votant)
- FM : 2 jours au maximum par question
- RH : 1 ou 2 jours au-delà les gens ne regardent plus la borne

Dans le cas où le profil de l'entreprise cliente correspond au profil RH

*Comment choisir une question ?*

Choisir une question par classification sémantique

*Comment choisir la question suivante ?*

La question la plus proche en terme de sémantique

*Pourquoi une telle différence entre le profil RH et les autres profils ?*

Une telle différence est dû au fait que les responsables des ressources humaines qui seront en charge de déployer la borne doivent creuser les aspects divers et variés de la vie en entreprise des collaborateurs en faisant particulièrement attention à ne pas heurter les sensibilités.

*Période application*

Lancement des campagnes tout au long de la journée en adéquation avec les résultats de calcul de *détections*.

On appellera *détection* le passage de la satisfaction d'une question sondé en dessous d'un seuil fixe pendant au moins un temps définit par convention avant qu'elle ne repasse au dessus de ladite satisfaction seuil

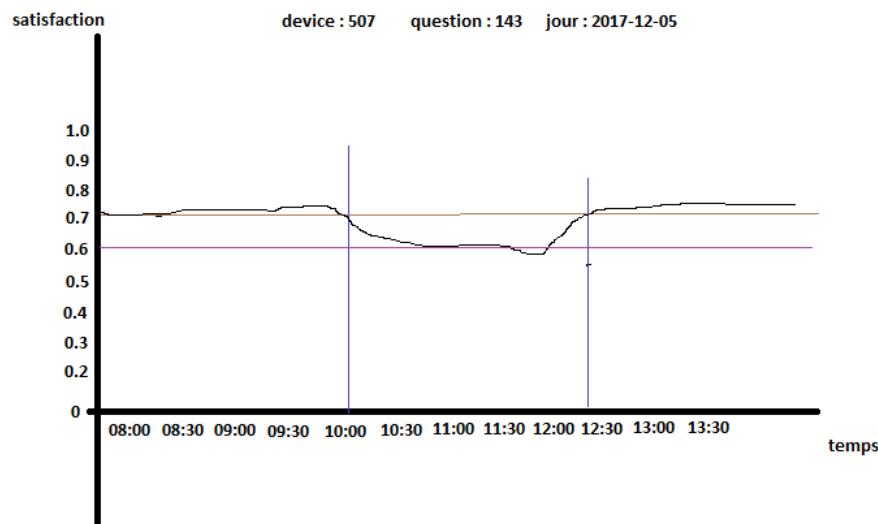
(sorties calcul détection)

ex :

- device : 507 , question : 143 , début : 09 :30 , fin : 10 :45 , jour : 2
- device : 507 , question : 143 , début : 15 :00 , fin : 17 :00 , jour : 2
- device : 507 , question : 132 , début : 13 : 00 , fin : 14 : 30 , jour : 4

*début* et *fin* définissent une plage horaire de la journée durant laquelle la satisfaction était en dessous en continu du seuil fixé (i.e entre début et fin la satisfaction n'est pas remonté au dessus du seuil)

*jour* sera l'entier correspondra au jour de la semaine où la détection est repérée



exemple graphique d'une détection.

Dans cet exemple la satisfaction seuil est à 70 pourcent et on remarquera que la satisfaction courante est restée en dessous de la satisfaction seuil de 10h00 à 12h30 ..

si cet état de fait se répète de nombreuse fois un jour de la semaine particulier ou juste sur ce créneau on pourra en conclure que cette plage horaire sera la plus indiquée pour la question qui était sondé

## Description de fonctions

Il s'agit d'une brève description de fonctions spécifiquement développés ou en cours de développement pour le fonctionnement des deux types d'utilisation du système de gestion de campagnes (courte et longue durée).

*période apprentissage profil RH , utilisation longue durée*

**nom fonction** : indexation sémantique (développée)

**entrées** : une liste de question (en ordre quelconque)

**sortie** : une liste de question d'un KPI (remis en ordre par sémantique)

**fonctionnement** : pour chacune des questions

1. étape 1 : extraction des mots vide
2. étape 2 : calcul de la distance entre chaque question deux par deux (en se basant sur la fréquence des mots)
3. étape 3 : reconstitution d'une liste de question en se basant sur la distance minimale

*période apprentissage tous profils hors RH , utilisation longue durée*

**nom fonction** : random choice (développée)

**entrées** : liste des identifiants questions

(questions de tous les KPIs choisis par le client)

**sorties** : une liste de questions mélangés aléatoirement

**fonctionnement** : sonder une question de manière aléatoire en veillant à ne pas la poser une deuxième fois avant que toutes les autres questions n'ait été sondé.

### 3.2 Démarche Machine Learning borne Notio

---

*période application tous profils , utilisation longue durée*

**nom fonction** : détections (développée)

**entrées** : identifiant de la question , identifiant de la borne , horodatage , nombre de votes yes ( pushes bouton vert) , nombre de votes neutre (pushs bouton jaune) , nombre de votes no (push bouton rouge)

**sortie** : numéro de device , identifiant de la question , horodatage 1 , horodatage 2 , jour de la semaine.

- horodatage 1 : moment de la journée où la satisfaction sur une question sondé passe en dessous de la satisfaction seuil (satisfaction seuil associé à la question sondé)
- horodatage 2 : moment de la journée où la satisfaction sur une question sondé passe au dessus de la satisfaction seuil (satisfaction seuil associé à la question)

(indicateurs à conserver dans une table pour pouvoir savoir à quel moment lancé une campagne sur une question précise , chez un client précis)

**nom fonction** : news (à développer)

**entrées** : nom client

**sorties** : Une news de la tendance du moment concernant le client.

Extraction des tweets , les plus récents sur l'entreprise-cliente +/ ou web scraping sur des sites de journaux en ligne

*application tous profil , utilisation courte durée*

**nom fonction** : rating (en cours de développement)

**entrées** : question , satisfaction seuil

**sorties** : note

La note sera fixé par palier en fonction de la différence entre le nombre de votes minimale et le nombre de votes cumulés par campagne d'une part et la différence de la satisfaction question sondé et satisfaction seuil associé à la question d'autre part.

### 3.2.3 Fonctionnalité : classification sémantique

La méthode de *classification sémantique* est une méthode obtenu en combinant différentes méthodes , elle ne s'applique que sur un Key Performance Indicators à la fois.

Considérons le Key Performance Indicator *organisation du travail* , les questions associées sont les suivantes :

1. Mon équipe est bien organisée
2. Mon équipe est efficace et performante
3. Avez-vous du temps pour vous en dehors de vos horaires de travail ?
4. Vous sentez-vous utile au sein de votre service ?
5. Je pense que la plupart des réunions sont utiles

### Extraction des caractères spéciaux et stopwords

Cette étape a pour but de débruiter le la liste de question (conjugaisons,accords ...)

1. 'équip' , 'bien' , 'organis'
2. 'équip' , 'efficac' , 'perform'
3. 'avez' , 'temp' , 'horair' , 'travail'
4. 'sent' , 'util' , 'sein' , 'servic'
5. 'pens' , 'réunion' , 'util'

### Word occurency

Cette étape a pour but de calculer l'occurrence d'apparition d'un terme dans une question.

1. 'équip' : 1, 'bien' : 1, 'organis' : 1
2. 'équip' : 1, 'efficac' : 1, 'perform' : 1
3. 'avez' : 1, 'temp' : 1, 'horair' : 1, 'travail' : 1
4. 'vous' : 1, 'sent' : 1, 'util' : 1, 'sein' : 1, 'servic' : 1
5. 'pens' : 1, 'réunion' : 1, 'util' : 1

### Cosine similaity

Cette mesure sert à calculer le degré de similarité entre deux vecteurs.

	1	2	3	4	5
1	1	0.5	0	0.258	0
2	0.5	1	0	0	0
3	0	0	1	0	0
4	0.258	0	0	1	0.258
5	0	0	0	0.258	1

Notre matrice de similarité étant symétrique il est suffisant de considérer la sous-diagonale de la matrice , soit la sous diagonale supérieur ou la sous-diagonale inférieur.

Les indices des valeurs non nulles de la matrice constituent les indices de notre nouvelle liste de questions.

#### Nouvelle liste de questions

1. Mon équipe est efficace et performante
2. Mon équipe est bien organisée
3. Vous sentez-vous utile au sein de votre service ?
4. Je pense que la plupart des réunions sont utiles
5. Avez-vous du temps pour vous en dehors de vos horaires de travail ?

Cette méthode se basant uniquement sur un calcul de distance , il n' a pas été trouvé d'indicateur pouvant prouver sa pertinence en fonction de différentes combinaison de liste de questions ; on ne peut que l'étonner par une expertise humaine de façon générale.

### 3.2.4 Fonctionnalité : ratings

(*Cette fonctionnalité permet de classer les questions à sonder par ordre de priorités*)

La note sera un nombre réel appartenant à l'intervalle  $[0, 1]$

Dans le système de gestion de campagne utilisant le rating l'une des conditions de mise à jour de la note associée à la question sont :

- comparaison entre les votes cumulés de la campagne et le nombre de votes minimal
- comparaison de la satisfaction courante avec la satisfaction seuil associée à chaque question.

Dans le cas où le nombre de votes est supérieur au nombre de votes minimal et la satisfaction courante est inférieure à la satisfaction seuil , en plus que la variable de temporisation undertreshold est supérieur à 5 alors la note de la question est remis à jour.

soit :

$$x = |\text{différence des votes}|$$

et

$$y = |\text{différence des taux de satisfaction}|$$

la fonction de rating  $f$  est tel que :

- la fonction  $f$  (fonction à deux variables) est croissante ,i.e :
- $x_1 = \text{différence de votes question 1 et } x_2 = \text{différence de votes question 2}$
- $y_1 = \text{différence de satisfaction question 1 et } , y_2 = \text{différence de satisfaction question 2}$
- $x_1 > x_2, y_1 = y_2 \Rightarrow f(x_1, y_1) > f(x_2, y_2)$
- la fonction  $f$  est continue et bornée sur l'intervalle  $[0, 1]$

NB : Cette fonctionnalité ne s'appliquera que sur des questions de "status" secondaire ,i.e question d'enquête.

### 3.2.5 Fonctionnalité : random

(*Cette fonctionnalité s'applique aux identifiants des questions de Key Performance Indicator en exploitation sur la borne*)

Le but de cette méthode est de pouvoir tirer un entier au hasard , et de simultanément veillé à ne pas le tirer une seconde fois avant d'avoir tiré au moins une fois tous les autres.

*(difficulté)*

Un simple tirage aléatoire d'une liste ordonnée a un fort risque de répétition au niveau du tirage , un tirage sans remise demande la spécification de plus d'instructions.

(exemple : 1,2,3,4 → 2 , 2 , 2 ,1)

*(solution)*

Battre la liste des identifiants comme des cartes , mélanger la liste une fois et lancer les campagnes en une fois

(solution la plus simple et moins couteuse)

(exemple : [1,2,3,4] → [4,2,1,3])

### 3.2.6 Fonctionnalité : détections

(*Cette fonctionnalité est le point névralgique de toute la structure Machine Learning du sondage intelligent , elle ne s'applique qu'aux informations remontés par les trames*)

**Comment étudier les résultats de votes en temps-réel ?**

L'indicateur principale que Agora Opinion étudie et avec lequel elle communique pour son client est la satisfaction dont la formule est la suivante :

$$\frac{nbyes + 0.5 * nbneutre}{nbyes + nbneutre + nbno}$$

où :

- *nbyes* est le nombre de votes verts (un entier)
- *nbneutre* est le nombre de votes neutre (un entier)
- *nbno* est le nombre de votes no (un entier)

Dans l'objectif de pouvoir suivre l'indicateur de satisfaction de manière courante (en temps-réel), il a été utilisé la technique du "**Windowing**", cette technique permet de mettre à jour un indicateur en le lissant sur l'échelle de temps observé.

Voici son fonctionnement dans notre cadre :

Considérons les trois vecteurs de votes extraits de la table de votes  
 $yes = [1, 2, 5, 1]$  ,  $neutre = [0, 0, 3, 1]$  ,  $no = [0, 0, 1, 0]$

Définition de la fenêtre de calcul sur 30 minutes (trames remontées toutes les 10 minutes, donc calcul mise à jour sur trois valeurs)

- $yes = [1, 2, 5]$
- $neutre = [0, 0, 3]$
- $no = [0, 0, 1]$
- $satisfaction = \frac{8+1,5}{12} = \frac{9,5}{12} = 0.79$

### **mise à jour de la data**

- $yes = [2, 5, 1]$
- $neutre = [0, 3, 1]$
- $no = [0, 1, 0]$
- $satisfaction = \frac{8+2}{13} = \frac{11}{13} = 0.77$

Comme le montre l'exemple précédent la mise à jour se fait en retirant la donnée la plus ancienne de la file (ou fenêtre temporelle) et en y ajoutant la plus récente.

**détection**

datetime	debut	fin	satisfaction	seuil
$t_1$			0.77	0.7
$t_2$			0.65	0.7
$t_3$	$t_2$		0.69	0.7
$t_4$	$t_2$		0.74	0.7
$t_5$	$t_2$	$t_4$	0.75	0.7

**Sortie :** iddevice ,  $t_2$  ,  $t_4$

Comme cette simulation le montre la méthode "*flagge*" (détecte) le premier moment où la satisfaction courante passe sous la barre de seuil et vice-versa pour la sortie de détection

### 3.2.7 Fonctionnalité : news

Pour cette fonctionnalité , il sera utilisé les modules d'extractions de tweets pré-existent sur le logiciel *Python* ; l'extraction sera suivi d'une analyse de sentiment ( processus qui permet de déterminer la tonalité émotionnelle qui se cache derrière une série de mots) avec des outils pré-existent sur ce même logiciel

On appelle *polarité* un indicateur d'orientation d'opinions , il peut être positif , négatif ou neutre.

Dans le cas présent l'objectif est d'informer sur les tendances de l'entreprise, il nous faudra filtrer l'ensemble des tweets extraits sur une polarité neutre après une analyse de sentiment.

Cette partie nécessitera l'intervention d'un humain pour sélectionner "la" news à afficher sur la borne.

Cette fonctionnalité devra être activée après avoir sonder le public cible sur la question :

*Êtes-vous au fait de l'actualité de l'entreprise ?*

Cette question sera sondé pendant environ 1h30 (par exemple) si son taux de satisfaction est inférieur à 75 pourcent alors il sera lancé la fonctionnalité de "*News*".

Une méthode de web scraping d'articles de journaux résumés a été trouvé dans la documentation web , elle est actuellement en cours de travail pour pouvoir avoir un rendu exploitable. (elle ne se montre efficace que pour des articles en langue anglaise)

### 3.2.8 Clustering des Key Performance Indicators

Dans la perspective d'être plus que un simple fournisseur de services (instruments de mesures de la satisfaction clients/collaborateurs) , une des cordes à ajouter à l'arc de la société Agora Opinion serait d'avoir une dimension de recommandations pour aider à un meilleur suivi du client.

Pour les aider dans ce sens la première pierre qui a été posé durant le stage a été de travailler sur une méthode de regroupement des sujets de sondages (appelés *Key Performance Indicators*) au travers une méthode de classification automatique.

Un avantage dans la structuration des données la société est que les questions portant sur les différents sujets sont semi-classifiées en ce sens que toutes les questions d'un même sujet sont réunies ensemble et séparées des questions des autres sujets de sondages.

L'échantillon (ensemble de KPIs) de travail n'étant pas de grand en terme de taille (une dizaine de sujets de sondage) , il a été fait le choix d'utiliser la méthode de la classification ascendante hiérarchique , dont la sortie graphique sera sous la forme d'un dendrogramme.

La *classification ascendante hiérarchique* (CAH) est une méthode de classification automatique utilisée en analyse des données qui à partir d'un ensemble de  $n$  individus les répartit en un certain nombre de classes (groupes). La méthode suppose que les individus sont dissimilaires (différents les uns des autres); dans le cas de points situés dans un espace *euclidien*\* on utilise la distance comme mesure de dissimilarité.

Un espace euclidien est un espace géométrique munit d'une distance euclidienne.

une distance euclidienne est une distance qui se calcule comme la racine carrée de la somme des carrés des coordonnées de deux points.

exemple soit deux points  $a, b$  dans un espace à deux dimensions , leurs coordonnées seront donc de types  $a = (x_1, y_1)$  et  $b = (x_2, y_2)$  la distance euclidienne entre  $a$  et  $b$  sera de la forme

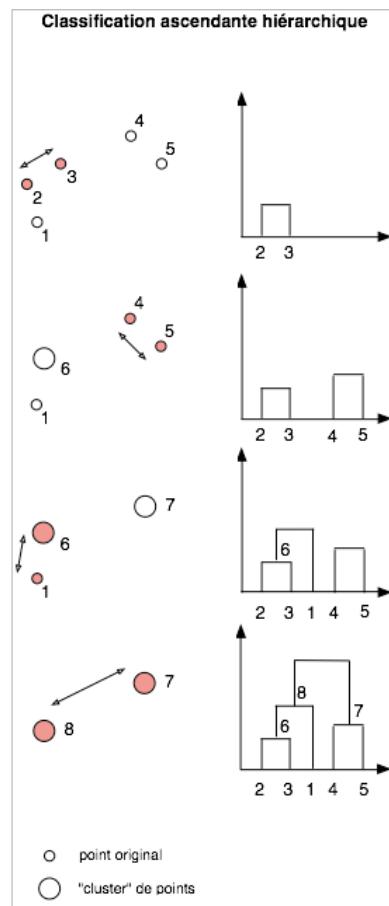
$$d(a, b) = \sqrt{(x_1^2 - x_2^2) + (y_1^2 - y_2^2)}$$

La classification ascendante hiérarchique est dite ascendante car elle part de l'hypothèse que tous les individus sont seuls dans une classe (i.e chaque individu est seul dans sa propre classe) , puis sont rassemblés en classes de plus en plus grandes (en termes de nombre d'individus)

Son principe de fonctionnement est le suivant :

1. sélectionner les deux individus dont la distance est minimale
2. calculer le nouveau point résultant de la fusion de ces deux individus
3. itérer avec les  $n - 1$  unités restantes

voir figure ci-dessous



Un *dendrogramme* est la représentation graphique d'une classification ascendante hiérarchique , il est présenté comme un arbre binaire dont les feuilles (individus en entrées) sont alignés sur l'axe des abscisses. (c'est le résultat auquel arrivera la figure précédente)

### Clustering Agora Opinion

Avant de procéder au clustering des Key Peformances Indicators disponible dans la table des questions chez Agora Opinion , il faut procéder à des étapes de pré traitement pour chaque Key Performance Indicators.  
(dans la suite , on expliquera chaque étape accompagné d'un exemple sur un Key Performance Indicator)

### Étapes de pré-traitement

Les KPI "tests" seront ceux portant sur *évolution et stratégie et vision* leurs listes de questions sont les suivantes :

— "stratégie et vision"

1. Pensez-vous être au courant de tout ce qui est important à savoir dans l'entreprise ?
2. Partagez-vous les valeurs de l'entreprise ?
3. Êtes-vous au fait de la stratégie commerciale de la société ?
4. Il y a une forte cohérence entre la stratégie de l'entreprise et mes tâches quotidiennes

— "évolution"

1. Avez-vous appris quelque chose de nouveau aujourd'hui
2. Avez-vous le sentiment de développer une nouvelle compétence en ce moment ?
3. Avez-vous évolué comme vous le souhaitez depuis votre arrivée ?
4. Avez-vous suivi une formation au cours de l'année 2016
5. Avez-vous le sentiment d'avoir besoin d'une formation ?

### Extraction de caractères spéciaux et stopwords

L'extraction des caractères et des stopwords (Mots vide) spéciaux servent à débruiter le dataset.

Un *mot vide* est un mot commun qu'il est inutile d'indexer.

liste de questions (stratégie et vision) :

1. 'pensez', 'etre', 'courant', 'important', 'savoir', 'entreprise'
2. 'partagez', 'valeurs', 'entreprise'
3. 'etes', 'stratégie', 'commerciale', 'société'
4. 'forte', 'cohérence', 'entre', 'stratégie', 'entreprise', 'taches', 'quotidiennes'

liste de questions (évolution) :

1. 'avez', 'apris', 'quelque', 'chose', 'aujourd', 'hui'
2. 'avez', 'évolué', 'souhaitez', 'arrivée'
3. 'avez', 'sentiment', 'développer', 'nouvelle', 'compétence', 'moment'
4. 'avez', 'suivi', 'formation', 'cours', 'année'
5. 'avez', 'sentiment', 'besoin', 'formation'

### Stemmization (Racinalisation)

La racinalisation est un procédé de transformation qui fait correspondre un mot à la partie restante une fois que l'on a supprimé son ou ses préfixes et suffixes.

exmple : 'policier' devient 'polici'

liste de questions (stratégie et vision) :

1. 'pens', 'etre', 'cour', 'import', 'savoir', 'entrepris'
2. 'partag', 'valeur', 'entrepris'
3. 'ete', 'strateg', 'commercial', 'societ'
4. 'fort', 'cohérent', 'entre', 'strateg', 'entrepris', 'tach', 'quotidien'

liste de questions (évolution) :

1. 'avez', 'appris', 'quelqu', 'chos', 'aujourd', 'hui'
2. 'avez', 'sent', 'développ', 'nouvel', 'compétent', 'moment'
3. 'avez', 'évolu', 'souhait', 'arriv'
4. 'avez', 'suiv', 'format', 'cour', 'anné'
5. 'avez', 'sent', 'besoin', 'format'

### TF-IDF (pondération)

Le TF-IDF (*term frequency - inverse document frequency*) est une méthode de pondération qui fournit une mesure statistique qui permet d'évaluer l'importance d'un terme contenu dans un document ( dans le cas présent une liste de questions) , le poids augmente proportionnellement au nombre d'occurrences.

C'est une méthode de scoring permettant de transformer des données brutes (données textuelles) en données utiles (vecteur de nombre réels)

Cette méthode appliquée aux 2 listes de questions , le résultat sera le suivant :

	évolution	stratégie et vision
anné	0.027726	0.000000
appris	0.027726	0.000000
arriv	0.027726	0.000000
aujourd	0.027726	0.000000
avez	0.138629	0.000000
besoin	0.027726	0.000000
chos	0.027726	0.000000
cohérent	0.000000	0.033007
commercial	0.000000	0.033007
compétent	0.027726	0.000000
cour	0.000000	0.000000
développ	0.027726	0.000000
entre	0.000000	0.033007
entrepris	0.000000	0.099021
ete	0.000000	0.033007
etre	0.000000	0.033007
format	0.055452	0.000000
fort	0.000000	0.033007
hui	0.027726	0.000000
il	0.000000	0.033007
import	0.000000	0.033007
moment	0.027726	0.000000
nouvel	0.027726	0.000000
partag	0.000000	0.033007
pens	0.000000	0.033007
quelqu	0.027726	0.000000
quotidien	0.000000	0.033007
savoir	0.000000	0.033007
sent	0.055452	0.000000
societ	0.000000	0.033007
souhait	0.027726	0.000000
strateg	0.000000	0.066014
suiv	0.027726	0.000000
tach	0.000000	0.033007
valeur	0.000000	0.033007
évolu	0.027726	0.000000

Il s'agit du calcul du tf-idf pour chacun des termes présent dans les deux listes si ils étaient pris un à un , ce qui a pour résultats de rendre deux vecteurs de nombres réels.

On voit que les termes non-présents dans une liste sont remplacés par des zéros

exemple : le terme 'anné' vaut 0 pour *stratégie et vision* et une valeur  $> 0$  pour *évolution*

### Cosine similarity

*La similarité cosine* est une mesure permettant de calculer la similarité de deux vecteurs à  $n$  dimensions en déterminant le cosinus de l'angle entre eux.

La valeur du cosinus de tout angle est toujours compris entre  $[-1, 1]$

- la valeur 0 indiquera des vecteurs indépendants
- la valeur 1 indiquera des vecteurs similaires

Toutes valeurs intermédiaires permettra d'évaluer un degré de similarité.

La similarité de nos deux KPIs après calcul est de 0 , ce qui nous laisse entrevoir qu'ils auront de fortes chances d'être dans des groupes différents.

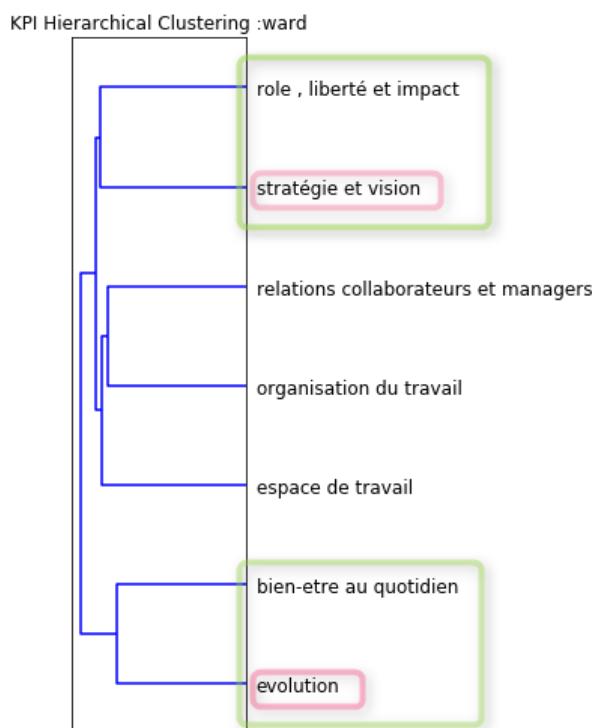
### Cophenetic correlation coefficient

Le coefficient de corrélation cophénétique mesure avec quel fidélité une classification ascendante hiérarchique représente la dissimilarité parmi les observations.

Au plus la valeur est proche de 1 au plus la classification respecte les distances originelles.

## Clustering de Key Performance Indicators

En déployant toutes les méthodes de pré-traitement précédentes , suivi d'une classification ascendante on obtient le graphe suivant :



La classification précédente obtient un coefficient de corrélation de Co-phenetic de 0.91 par la méthode de "ward" conserve donc assez bien les dissimilarités originelles. (c'est à la méthode de WARD qu'on doit l'aspect de fusion deux à deux , elle se base sur les KPIs dont la similarité cosine est proche 1)

La classification ascendante tel qu'elle est présentée elle se base uniquement sur la sémantique des questions , ce qui est une solution seulement utilisable pour les premiers clients de la borne *Notio*.

Une solution à long terme et plus pertinente serait de se baser sur une méthode de corrélations "*inter-Key Performance Indicators*" en se basant sur l'évolution de la satisfaction des dits KPIs

(exemple : si le client A , B et C ont pris en commun les KPIs *évolution*, *stratégie* et *vision* et *espace de travail*, et que après calcul *évolution* et *espace de travail* sont corrélés à plus de 70 pourcent ; si un client D arrive et demande les KPIs *évolution* et *bien-être quotidien* , il lui serait recommandé de sonder le KPI *espace de travail*).

### 3.3 Use case : NeWays



### 3.3 Use case : NeWays

---

#### introduction :

Le projet Notio est un projet qui a pour but de pouvoir mettre en place un *sondage intelligent*, i.e être en mesure de sonder la bonne question au bon moment.

#### Mise en situation

(*On se place dans une époque où le projet notio est déployé chez les premiers clients*)

La responsable des ressources humaines de la société *NeWays* contacte les équipes la société Agora Opinion pour passer commande d'une borne Notio et elle souhaite sonder les Key Performance Indicator suivants :

- “organisation du travail”
- “évolution”

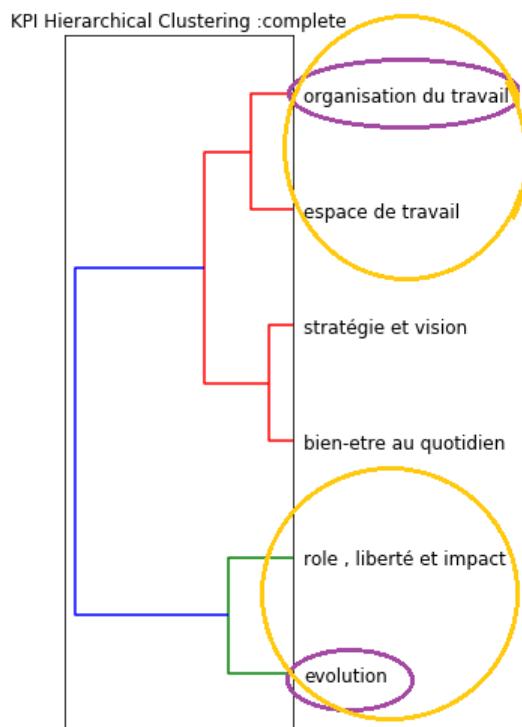
Dans un de leurs bureaux.

La société Agora Opinion a en sa possession des outils d'analyses de données qui lui permettent de pouvoir faire de la recommandation / conseil auprès des clients lors du choix des sujets de sondages.

Un de ses outils est une méthode de regroupement de KPI qui leur permet de faire de l' association de sujets de sondages ; dans une démarche d'accompagnement la société va fournir les outils nécessaires i.e une borne Notio et lancer les campagnes demandés , mais elle va aussi suggérer / recommander à la responsable des ressources humaines de sonder d'autres Key Performance Indicators.

Dans notre exemple :

Voici ce que la société Agora Opinion aura comme sortie d'un de leurs outils (voir graphique page suivante)



*N.B : en mauve sont entourés les sujets de sondages demandés par la responsable RH et en jaune sont entourés les KPIs recommandés par Agora Opinion.*

Au vu de la sortie graphique de cette méthode il en ressort que si la responsable des ressources humaines désire sonder ses collaborateurs sur leur "*organisation au travail*" et leur "*évolution*" dans la société *NeWays* , une recommandation faite par la société Agora Opinion serait qu'elle sonde ses collaborateurs en plus sur leur "*espace de travail*" et leur "*rôle , liberté et impact*".

Dans le but de pouvoir arriver à leur objectif ,i.e un sondage intelligent dans la société *NeWays* , il est nécessaire que la borne soit déployée pendant un certain temps avant de mettre en place un tel mécanisme (sonder la bonne question au bon moment)

### Utilisation borne notio longue durée :

Une borne notio à utilisation longue durée sera connecté à un système de type *apprentissage / application*, c'est-à-dire que le moment de passage des questions sur l'écran de la borne sera déterminée par une analyse des votes émis durant la période d'apprentissage et une meilleure gestion des campagnes (la bonne question au bon moment) de sondages sera déployé à la fin de cette période (période apprentissage).

La responsable RH devra renseigner quelques indications sur le dashboard de la société Agora Opinion :

*nombre de collaborateurs :*

il s'agira du nombre de collaborateurs qui seront amenés à être en interaction avec la borne.

*durée de la campagne :*

il s'agira de la durée moyenne par campagnes , à savoir le temps de sondage d'une question particulière.

*typologie :*

Il s'agit ici d'un entier représentant le profil du client (parmi les trois principaux profils présents dans la clientèle de Agora Opinion) :

- Retail → 1
- FM → 2
- RH → 3

*type d'usage :*

il s'agit du type d'utilisation désirée par le client :

- courte durée → 1
- longue durée → 2

*N.B : Nous nous appuierons sur les données d'une autre société qui a un profil RH , dont les votes sont disponibles chez Agora Opinion, pour la suite de notre étude. [l'objectif principal est de montrer le bon fonctionnement des outils d'analyses de Agora Opinion]*

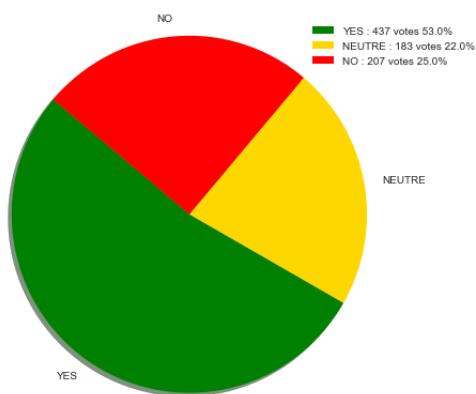
### 3.3 Use case : NeWays

---

La période d'études choisis par notre responsable des ressources humaines s'étend de Décembre 2016 à Février 2017 (une période de 51 jours de sondages au total) , période durant laquelle elle a sondée ses collaborateurs sur les Key Performance Indicators qui lui ont été suggérés par Agora Opinion.  
(23 questions au total , sondés deux fois chacune au cours de la période d'études)

Durant sa période d'apprentissage le système de gestion de campagne a lancé les campagnes en se basant sur une analyse sémantique des questions appartenant aux sujets choisis par le client NeWays.

Voici un diagramme résumant les votes sur la période apprentissage.



Parmi les sujets sondés se trouve le sujet portant sur “espace de travail” Concentrons-nous sur l'une de ses questions.

question : “*Je pense que la plupart des réunions sont utiles*”

Grâce à classification sémantique la question a été sondée sur les journées du 2016-12-12 /2016-12-13 et 2017-01-27 /2017-01-28

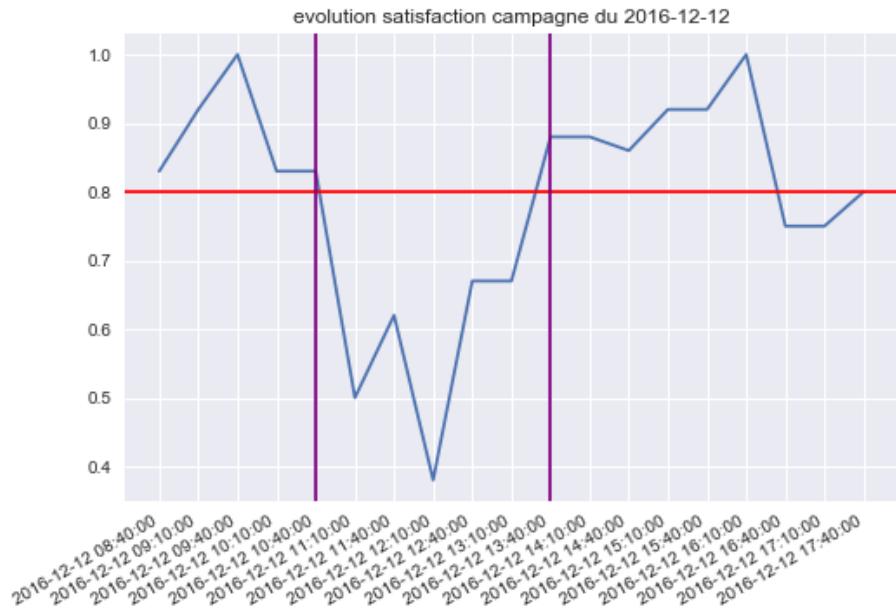
Après passage dans le système d'analyse il en ressort que la satisfaction à cette question est passée sous la satisfaction seuil associée à la question (satisfaction seuil de 80 pourcent) aux heures suivantes :  
(voir page suivante)

### Apercu de la méthode des détections :

**campagne du 2016-12-12**

device	heure	heure-debut	heure-fin	satisfaction	seuil
533	10 :09 :36			0.83	0.8
533	10 :39 :40			0.83	0.8
533	10 :49 :41	10 :39 :40		0.5	0.8
533	10 :59 :42	10 :39 :40		0.62	0.8
533	12 :39 :53	10 :39 :40		0.38	0.8
533	12 :49 :54	10 :39 :40		0.67	0.8
533	12 :59 :56	10 :39 :40		0.67	0.8
533	13 :30 :00	10 :39 :40	13 :30 :00	0.88	0.8

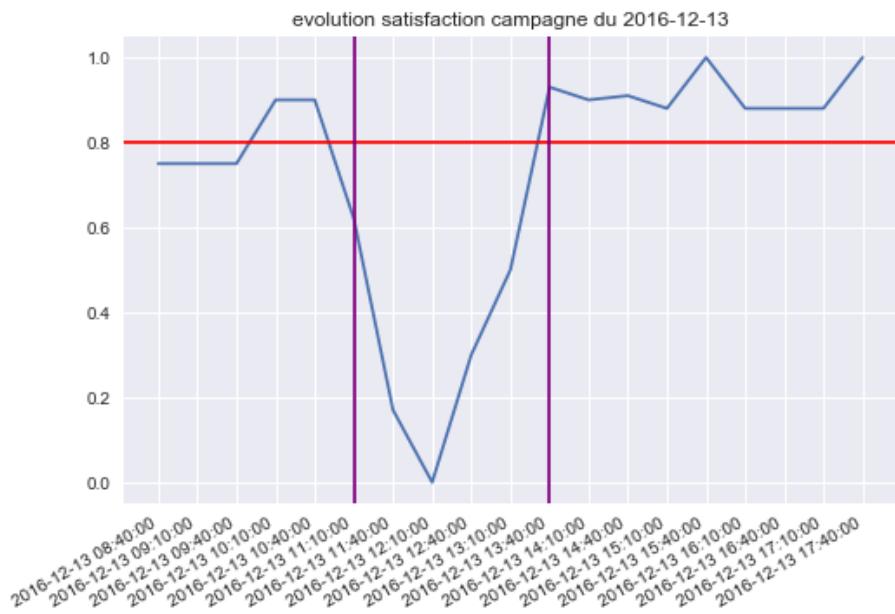
visualisation graphique de l'évolution de la satisfaction :



### campagne du 2016-12-13

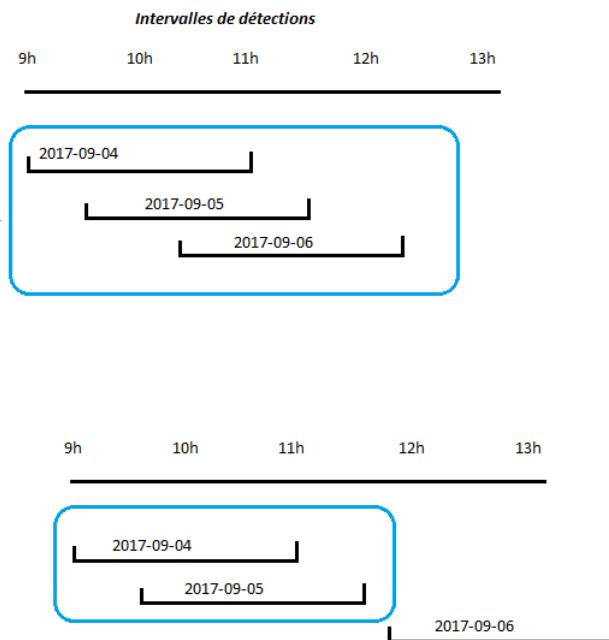
device	heure	heuredebut	heurefin	satisfaction	seuil
533	10 :42 :28			0.9	0.8
533	10 :52 :29			0.9	0.8
533	11 :02 :31	10 :52 :29		0.62	0.8
533	11 :42 :36	10 :52 :29		0.17	0.8
533	12 :32 :42	10 :52 :29		0.0	0.8
533	13 :12 :47	10 :52 :29		0.3	0.8
533	13 :22 :48	10 :52 :29		0.5	0.8
533	13 :32 :50	10 :52 :29	13 :32 :50	0.93	0.8

visualisation graphique de l'évolution de la satisfaction :



On peut commencer à supposer que le moment le plus propice pour sonder les collaborateurs de la société NeWays sur cette question en particulier est de commencer à les sonder à partir de 10-11h jusqu'à l'heure de la reprise après le déjeuner vers 14h.

L'idée Pour désigner les intervalles qui seront utilisés , est la suivante :



Dans le second cas on s'arrêtera au groupe d'intervales majoritaires (en termes de nombre de détections) , on désignera comme début du moment de sondage le début de la première détection du groupe , et la fin du sondage la fin de la dernière détection.

device	id_question	begin_hour	end_hour
Filtre	Filtre	Filtre	Filtre
533	56	10:00	13:00

Voici un aperçu de l'analyse des détections dans la société NeWays pour la question identifiée comme 56 (chez Agora Opinion) qui est :  
*"Je pense que la plupart des réunions sont utiles"*

---

## 4 Conclusion

### Perspectives pour l'entreprise

A la fin de ce stage la société Agora Opinion , une fois les méthode conçues sous python déployées , sera en mesure de pouvoir gérer les campagnes de sondages de manière automatisé et pertinente pour chacun de leur client qui en fera la demande.

En plus d'en tirer une dimension de traitement de données en temps réels, ils en tirent une dimension de "*consulting*" avec des outils leur permettant de visualiser d'une part et d'estimer la force de liaison entre les *Key Performance Indicator* de manière ponctuelle d'autre part.

L'entreprise Agora Opinion se trouvant à une période charnière , ce fut un privilège pour moi d'être compter parmi leur équipe l'espace d'une demi-année, et être au premier rang pour l'un de leur plus ambitieux projets depuis la création de l'entreprise.

### Bilan étudiant

J'ai suivi une spécialisation en *Data Science* parce que je désirais devenir *Data Scientist* principalement parce que passionné par les méthodes et applications dans le monde réel (les fruits du travail), aujourd'hui après ce stage j'ai découvert la responsabilité qui incombaît à une personne dans ma position dans une entreprise , les dilemmes auxquels on doit faire face quotidiennement, les compromis entre rigueur mathématique et nécessités opérationnels..

Les méthodes développés (constituent la V1 du projet) s'appuieront principalement sur les données émises par le client (trames) , il pourrait être pertinent de croiser les données émanant des clients , la sémantique des questions qu'ils sondent pour lui créer de nouvelles questions spécifiques.

J'ai acquis des compétences techniques principalement en informatique , en base de données et en méthodologie de projet en entreprise  
Il sera laissé sur place un document explicatif des méthodes (en détails) pour les rendre accessibles en vue d'itération de ma partie du projet.

## 5 Lexique

*Customer experience :*

L'expérience client désigne l'ensemble des émotions et sentiments ressentis par un client avant, pendant et après l'achat d'un produit ou service. C'est le résultat de l'ensemble des interactions qu'un client peut avoir avec la marque ou l'entreprise.

*Campagne :*

On appelle campagne une durée de temps durant laquelle on sonde une question d'un sujet (Key Performance Indicator)

*Key Performance Indicator :*

On appelle Key Performance Indicator un ensemble de questions portant sur un aspect précis d'un service fourni

*détection :*

On appelle détection la l'intervalle de temps durant laquelle la satisfaction observée devient inférieure à la satisfaction seuil

*Sondage :*

On appelle un sondage une enquête ponctuelle réalisée auprès d'un échantillon représentatif de la population étudiée

*trames :*

Une trame est une ligne d'information à décoder qui contient , les indicateurs de votes , l'horodatage ,identifiant de la borne

## 6 Annexe

### Formule similarité cosinus

Soit  $\theta \in [0, 2\pi]$ ,  $A$  et  $B$  deux vecteurs tel que  $\widehat{AB} = \theta$

$$\cos\theta = \frac{A \cdot B}{|A||B|}$$

### Formule TF-IDF

$$tf(t, d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

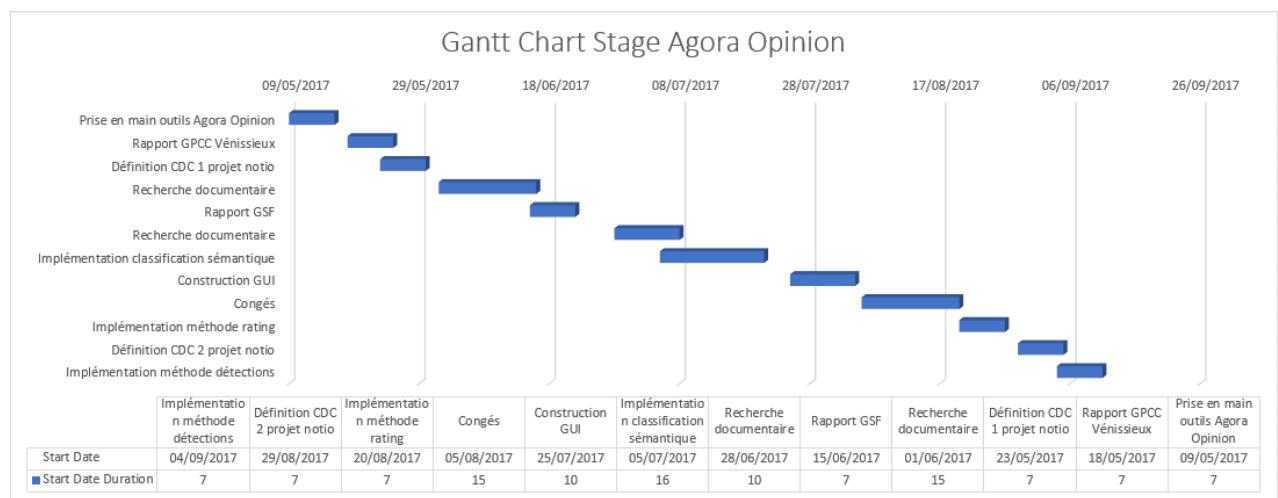
$$idf(t, D) = \ln \left( \frac{|D|}{|\{d \in D : t \in d\}|} \right)$$

$$tfidf(t, D) = tf(t, d) * idf(t, D)$$

où  $D$  = ensemble des documents,

$f_d(t)$  = fréquence du terme t dans le document d

### Gantt Chart du stage



---

## 7 Technologie

Les deux projets du stage ont été menés sur le langage de programmation Python qui est placé sous une licence libre et fonctionne sur la plupart des plates-formes informatiques de *Windows* à *Unix* avec notamment *GNU/Linux* en passant par *macOS*, ou encore *Android*, *iOS*, et aussi avec Java ou encore *.NET*.

### liens de téléchargements

lien : <https://www.anaconda.com/download/>

Le lien précédent est le lien de téléchargement vers le site de *Anaconda Continuum* distributeur officiel de *Spyder* un IDE (Interface Development Environment) du langage *Python* utilisée pour le développement de la majorité des projets en *Machine Learning* tel que celui dont a fait l'objet ce stage.

lien : <http://sqlitebrowser.org>

Le lien précédent est un lien de téléchargement de *DB Browser for SQLite*, il s'agit d'une application permettant de simuler le fonctionnement d'une base de données et de reproduire les interactions des différentes tables. C'est cette application qui a été utilisé pour simuler le fonctionnement des systèmes de gestion de campagne.

lien : "how to deploy a python script on azure server"

Le lien précédent mène des différents liens aidant à déployer des scripts du langage de programmation python sur un server Azure (Machine Virtuel présente chez Agora Opinion et sur lequel sera déployé le système de gestion de campagne)

## 8 Bibliographie

@Article, author = Danushka Tarupathi Bollegala,  
title = A Machine Learning Approach to Sentence Ordering For Multi-  
Document Summarization and its Evaluation, year = 2015

@Proceedings,  
title = Using TF-IDF to convert unstructured text to useful features,  
year = 2016

@Proceedings <http://www.jybaudot.fr/Analdonnees/cah.html>,  
title = La Classification Ascendante Hiérarchique

@Proceedings <http://brandonrose.org/clustering> ,  
title = Document Clustering Python

@Proceedings <https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/> ,  
title = SciPy Hierarchical Clustering and Dendrogram Tutorial

@Proceedings <https://stackoverflow.com/questions/14720324/compute-the-similarity-between-two-lists>,  
title = Compute the similarity between two lists

@Proceedings <http://www.fonctionnel.net>,  
title = L'analyse fonctionnelle expliquée