



**UNIVERSITÉ
DE GENÈVE**

FACULTÉ DES SCIENCES



**UNIVERSITÉ
DE GENÈVE**

CENTRE UNIVERSITAIRE
D'INFORMATIQUE

Crowdsourcing and AI

Professeurs : Jose Marquez, François Grey

Rapport final

Poopidoo

Auteurs : Joao Filipe Costa da Quinta, Roze Defossez, Léa Heiniger

Printemps 2022

Table des matières

Introduction	3
Problématique	4
Brainstorming	4
Formulation du problème	4
Dangers	4
Solutions	6
État de l'art	6
Notre solution	7
Résultats attendus	8
Implémentation	10
Application	10
AI	10
Résultats	11
Étapes suivantes	25
Liens	26
Auteurs	27

Introduction

Ce projet se situe dans l'intersection de deux branches, le crowdsourcing et l'intelligence artificielle. L'objectif est de résoudre un problème, n'importe lequel, à l'aide de ces deux techniques.

Le crowdsourcing est l'externalisation d'une tâche, cette tâche peut être sous plusieurs formes, et la contribution extérieure peut être récompensé ou pas. La limite est l'imagination de l'entreprise qui met en place cette tâche.

Par exemple, "Genes in Space" est un jeu vidéo qui permet à ses joueurs d'aider des scientifiques qui font des recherches importantes sur le cancer. Comme récompense, un simple tableau de score est proposé. Sachant que les joueurs auront envie d'être parmi les meilleurs, ils vont continuer de jouer.

Nous voyons ici une forme de crowdsourcing qui est très imaginative, et gratuite!

L'intelligence artificielle (IA) doit être appliquée dans notre projet pour automatiser des tâches répétitives ou optimiser un problème donné.

Un exemple est la simple classification d'images, c'est une tâche qui est trivial pour un humain mais pas très engageante. Du coup si on a une IA capable de faire cette classification, on sera contents de lui laisser cette tâche.

Un exemple où on mélange les deux branches, c'est les tests CAPTCHA. Ce sont des tests que nous devons résoudre pour accéder à certains sites web. Ces tests ont pour but de vérifier que l'utilisateur est un humain.

Pour faire cette vérification, le site demande à l'utilisateur de résoudre une tâche qui est très intuitive pour un humain, par exemple d'identifier où est le passage piéton dans une image.

Si l'utilisateur réussi à répondre, on lui donne accès au site.

L'utilisateur résout une simple tâche de classification. Cette classification pourra à son tour aider des développeurs à d'entraîner des intelligences artificielles, qui seront capables d'identifier des passages piétons, des voitures, des panneaux de trafic dans toute nouvelle image.

Ce sont des capacités essentielles à toute voiture capable de conduire de façon autonome!

Problématique

Brainstorming

Le problème à résoudre devait être local, car ce sont nous, les créateurs, les premiers qui devront utiliser cette application et ainsi créer les premières données.

Lors de notre brainstorming nous avons pensé à des problèmes de classification d'image, car ce sont ce type de problèmes que nous avons le plus expérimenté en cours.

Assez vite nous sommes arrivés dans l'idée de recenser des objets trouvés par terre, comme par exemple des mégots. La contrainte c'est que nous avons rapidement trouvé plusieurs initiatives contre les mégots par terre.

En essayant de rester dans le même thème, nous avons pensé aux déjections canines non ramassées.

Formulation du problème

Notre but est donc de réduire, ou même complètement éliminer, les déjections canines des rues de Genève.

C'est un problème qui impacte fortement la ville de Genève, ainsi que ses citoyens. Dans la suite on explorera à quel point nous sommes réellement impactés par ce problème sans le savoir.

Malheureusement ce n'est pas un problème que l'on peut résoudre seul, c'est seulement en travaillant en tant qu'un groupe que nous pourrons le faire!

Dangers

On peut penser que la seule conséquence des déjections canines au sol est la "pollution" visuelle.

Malheureusement ce n'est pas le cas, il suffit de penser au cas de Paris, où plus de 600 personnes, annuellement, se sont fracturés un membre suite à un chute due à des glissements sur les déjections au sol. C'est un nombre non négligeable!

De plus, il est très facile de se faire infecter par des parasites dangereux pour les humains, bien sûr que pour que ça arrive, il faut être en contact avec ces déjections. En tant qu'adultes on a une bonne hygiène, du coup on pourra esquiver ces dangers. Mais imaginez un enfant qui joue dans le bac à sable du parc, un bac à sable où une déjection n'a pas été ramassé. Il est très possible que l'enfant mette ses mains à la bouche.

Finalement, on peut penser que la pluie pourra nettoyer le sol, malheureusement cela est tout simplement faux, quand il pleut, les déjections sont mélangées avec l'eau, ce qui veut

dire que toute flaque d'eau est remplie de ces parasites dangereux, qu'on peut ainsi les ramener chez nous.

Il y a tout plein d'autres scénarios où une infection est possible. On voit bien que c'est un problème qu'on ne peut pas ignorer!

Solutions

Il est souvent plus efficace et moins chère de prévenir un problème plutôt que de punir ceux qui en sont coupables. Par contre, il est bien plus compliqué de mettre en place toutes les plateformes nécessaires à cette prévention.

Dans cette section nous allons explorer les solutions proposées par plusieurs villes européennes face à ce problème. Le thème commun à toutes ses solutions, c'est qu'on explore plutôt la punition!

État de l'art

En Espagne on explore beaucoup l'idée de mettre des amendes élevées aux propriétaires des animaux. À Madrid on propose une amende de 1000 euros, alors qu'à Tarragona on va au point de faire des analyses ADN sur les déjections animales, les résultats des analyses seront enregistrés dans une base de données, une fois que le coupable est trouvé, il devra payer une amende, et rembourser les coûts des analyses ADN.

Dans les deux situations pour punir le coupable il faut que la police soit au bon endroit au bon moment.

À Édimbourg on explore un autre type de solution, quand un citoyen voit des déjections au sol, il peut faire un marquage à côté, en craie. Alors qu'à Grand-Bretagne on fait le marquage avec de la peinture aérosol. Ces marquages permettent de savoir où il faut nettoyer plus souvent, ainsi qu'où il faut faire plus de contrôles.

Un inconvénient c'est que les citoyens doivent être prêts à faire ces marquages, et avoir le matériel nécessaire pour le faire. De plus, la craie part facilement avec de l'eau ou la pluie, alors que la peinture aérosol ne partira pas facilement.

Dans quelques villes françaises, on explore la création d'une application qui permettrait à tout citoyen de recenser les déjections canines dans la rue avec leur propre position GPS. Avec ces données on cherche à faire des contrôles policiers plus focalisés dans les zones qui sont le plus affectés.

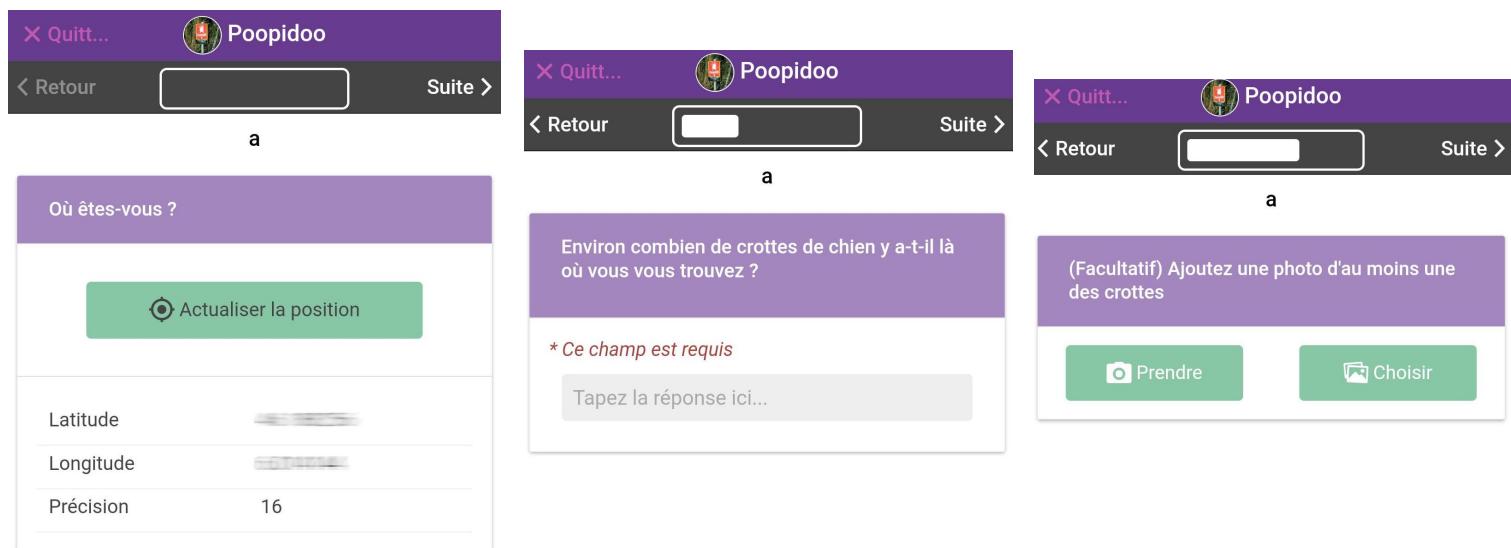
Cette application n'est pas encore arrivée aux grandes villes comme Paris, car elle dépend de l'accord du gouvernement de la ville.

À Wimblington, une petite ville anglaise, on a adopté une carte qu'on remplit collectivement, on cherche à recenser les endroits où des déjections canines ont été repérées. Cette solution a eu beaucoup de succès, en moins d'une semaine le nombre de déjections au sol a diminué drastiquement, sans mettre des amendes.

Malheureusement, pas tout individu pouvait mettre à jour la carte, on communiquait les localisations à travers des messages privés, puis les personnes qui avaient les droits mettaient à jour la carte. Cette solution n'est pas du tout possible dans une grande ville.

Notre solution

Notre solution se compose d'application mobile créée sur Epicollect5 dont des captures d'écran sont montrées ici:



Lorsque l'utilisateur de notre application se balade en ville et repère une déjection, il lui suffit d'appuyer sur le bouton principal. La localisation gps de la déjection est alors envoyée dans une base de données. Grâce à l'ensemble de ces localisations gps ainsi qu'un plan de la ville on peut alors facilement utiliser des algorithmes d'IA pour obtenir des informations intéressantes et pertinentes dans le cadre de la lutte contre l'incivilité.

Et en effet c'est ce que notre solution fait. En premier lieu, elle permet à la voirie de visualiser les zones concentrant le plus de déjections grâce à l'utilisation d'algorithmes de clustering. Puis grâce à des algorithmes d'optimisation de contraintes il permet d'indiquer l'emplacement optimal ou la voirie doit déployer des dispositifs de type sac à crotte.

Notre solution possède plusieurs avantages non négligeables comparée au système existant. En premier lieu, contrairement à Wimblington notre système de marquage ne prend que quelques secondes à être activé. En effet, il suffit à l'utilisateur de se connecter et d'appuyer sur un bouton pour indiquer la présence d'une crotte. En second lieu contrairement à Édimbourg notre marquage ne demande qu'un smartphone et très peu d'effort à l'utilisateur. De plus ce marquage étant numérique il ne disparaît pas avec les intempéries.

Résultats attendus

Bien évidemment pour obtenir des résultats intéressants nous ne pouvons pas nous limiter à nos simples données, c'est-à-dire les données rentrées par les trois créateurs du projet. Nous devons donc nous mettre en contact avec la voirie de Genève, et c'est ce qui a été fait.

Projet de cour Unige



Rose Defossez

To 'vvp@ville-ge.ch'

Cc ""Lea Marie Rose Heiniger""; ""Joao Filipe Costa Da Quinta""



jeu. 16.06

Bonjour,

Je vous contacte dans le cadre du cours Crowdsourcing and AI de l'unige qui est donné par Jose Luis Fernandez Marquez et François Grey. En effet pour notre projet final nous avons pour but de développer une application qui permettrait aux citoyens de noter les emplacements présentant des déjections canines. Puis avec les données récoltées l'on souhaiterait utiliser de l'intelligence artificielle pour indiquer les endroits nécessitant des nouveaux distributeurs de sac à déjections. On aimerait donc avoir votre avis sur notre projet, sur son utilité et si cela intéresserait la voirie

En vous remerciant d'avance, mes meilleures salutations.

Malheureusement au vu des délais, la réponse de la voirie arrivera au moment du rendu de ce rapport. Nous ferons donc dans la suite de cette section les hypothèses suivantes.

- la ville de Genève communique sur notre application
- la ville de Genève collabore avec nous sur le déploiement du nouveau dispositif

Dans ce cadre là on obtiendrait rapidement une carte des déjections mise à jour de manière continue. Avec ces données la ville de Genève pourra en premier lieu choisir où déployer les brigades de la voirie pour le nettoyage grâce à l'IA qui indiquait les lieux avec la plus grande concentration de déjection. En second lieu grâce à l'IA la ville de Genève pourra choisir ou placer de nouveau dispositif. En troisième lieu, la voirie pourra observer l'impact de différentes solutions. Pour finir ce genre de dispositif permet de sensibiliser la population.

Il nous paraît aussi important de parler du développement futur de notre solution. En effet, même si ici nous parlons principalement d'une collaboration avec la ville de Genève, d'autres choses sont possibles. La simplicité d'utilisation de notre application ajoutée au fait qu'il suffit de stocker de positions GPS nous permet en effet de proposer notre solution à d'autres cantons ou grandes villes de manière extrêmement simple. De plus, la simplicité de notre dataset collecté le rend utilisable de plein de manière différente. Pour finir cette solution et les algorithmes développés peuvent être utilisés pour tout autre sorte d'incivilité

(par exemple les mégots). Notre solution est donc souple, scalable et pertinente pour plein d'autres usages.

Implémentation

Application

Pour la collecte des données, nous avons décidé d'utiliser Epicollect5 pour héberger notre projet et créer l'application qui permet aux gens d'ajouter des données. Nous avons choisi Epicollect5 plutôt que CS logger car nous avons trouvé qu'il était plus facile d'obtenir rapidement une application fonctionnelle (qui peut être améliorée plus tard) sur cette plateforme et de pouvoir commencer à collecter des données.

Voici le lien vers notre projet sur Epicollect5 : <https://five.epicollect.net/project/poopidoo>

Lorsque quelqu'un veut ajouter des données, nous lui demandons d'indiquer sa localisation et ensuite de donner une estimation du nombre de crottes à cet endroit. Les gens peuvent également ajouter une photo des crottes s'ils le souhaitent, mais c'est optionnel. Nous avons fait le choix de ne pas demander de photos à chaque fois pour vérifier la véracité des données car on a supposé que peu de monde aurait envie de prendre des photos de crottes de chiens et comme c'est quelque chose de facilement identifiable il n'y a pas impérativement besoin de contrôler les données.

Une fois les données récoltées dans l'application, on les récupère dans un fichier json.

AI

Nôtre projet est assez compliqué car pour que notre IA fonctionne vraiment bien et que les résultats proposés soient utilisables il faut tenir compte de beaucoup de paramètres différents. Comme nous avions un temps assez limité pour faire ce projet nous avons choisi d'implémenter une version simplifiée du problème, afin de pouvoir déjà avoir quelques résultats à présenter, et qui pourrait être améliorée par la suite pour tenir compte de tous les paramètres.

Nous avons donc ramené notre problème à un simple problème de Clustering, notre modèle répartis les crottes de chiens dans différents clusters en fonction de la similitude de leur localisation puis on calcule le centroïde de chacuns des clusters afin de trouver le meilleur emplacement pour mettre un distributeur de sachets.

Nous nous sommes concentrés sur trois méthodes différentes de clustering :

- Affinity propagation : Ce modèle est intéressant pour notre projet parce qu'il choisit le nombre de clusters en fonction des données. Cela nous permet donc d'avoir une bonne estimation du nombre de distributeurs, panneaux de sensibilisation ou autre qu'il faudrait idéalement mettre en place.
- Mini-batch K-mean : Contrairement au modèle précédent celui-ci fonctionne avec un nombre prédéfini de clusters qu'on lui donne en paramètres et va les définir au mieux pour que dans chaque cluster la variance soit égale. Ce modèle pourrait nous être utile dans le cas où, par exemple pour des questions de budget, on aurait un nombre

limité de distributeurs à notre disposition et nous permettrait de trouver la meilleure façon de les placer.

- Agglomerative clustering : Ce modèle est intéressant car on peut soit l'utiliser comme le précédent en précisant le nombre de clusters qu'on veut, soit, comme ce modèle fonctionne en agglomérant petit à petit différents clusters, en précisant la distance à partir de laquelle les clusters ne sont plus assemblés. Cette deuxième possibilité nous permet par exemple de fixer une distance minimum entre deux distributeurs qu'on place.

Nous avons codé notre projet en python et nous nous sommes aidés des fonctions de la librairie *Sklearn*. Puis nous nous sommes servis de la librairie *Plotly* pour pouvoir afficher directement nos résultats sur une carte interactive.

Comme nous n'avons que très peu de vrai données nous avons également généré deux autres sets de données aléatoires (avec les fonctions de la librairie *Random*), un parfaitement aléatoire dans une zone donnée et un autre avec certaines parties de la zone qui contiennent plus de données que d'autres. Ces sets de données ne sont pas représentatifs de ce qu'on aurait avec de vraies données et ne tiennent pas compte de là où sont les bâtiments, les routes, etc. mais cela nous permet quand même de comparer un peu les différents modèles sur un plus grand nombre de données.

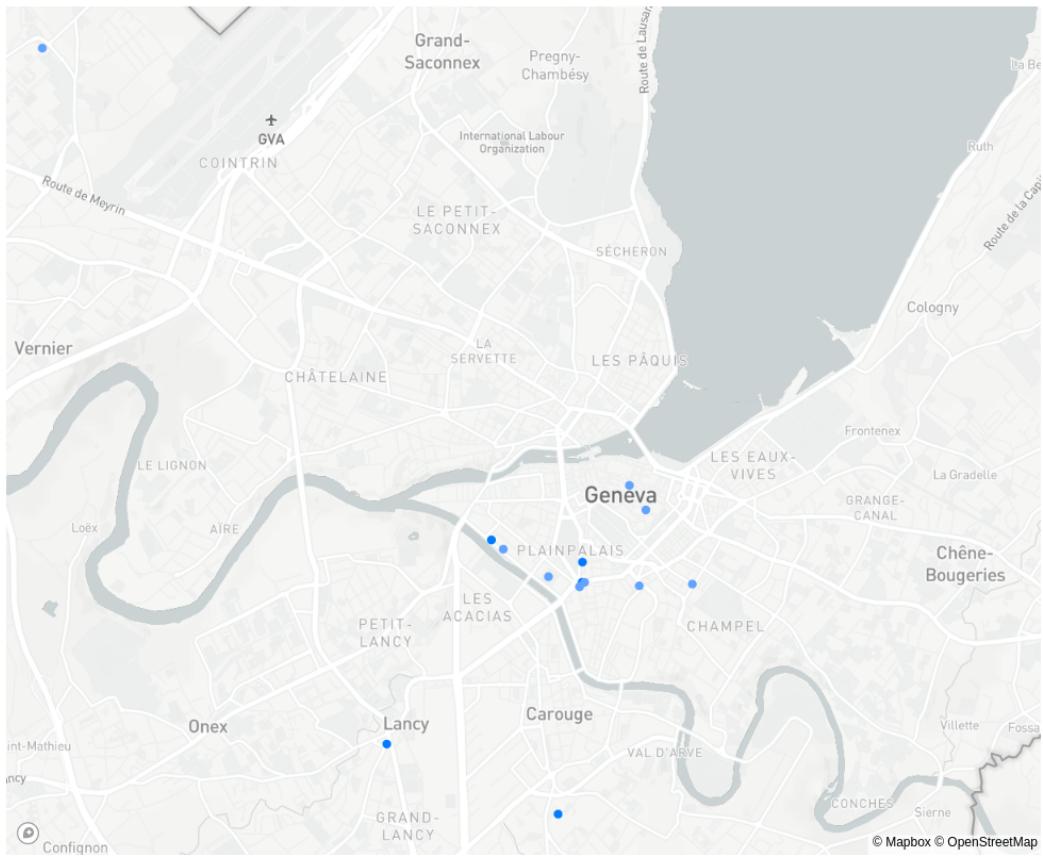
Résultats

Chacun des trois sets de données a été affiché sur une carte ainsi que les résultats obtenus avec chacun des modèles. Les points sont plus ou moins foncés en fonction du nombre de crottes indiqué par l'utilisateur, plus un point est bleu foncé plus il y a de crottes. Pour les sets de données créés aléatoirement le nombre de crottes est un nombre aléatoire entre 1 et 5. Les placements suggérés par les modèles pour les distributeurs de sachets sont indiqués en rouge sur les cartes.

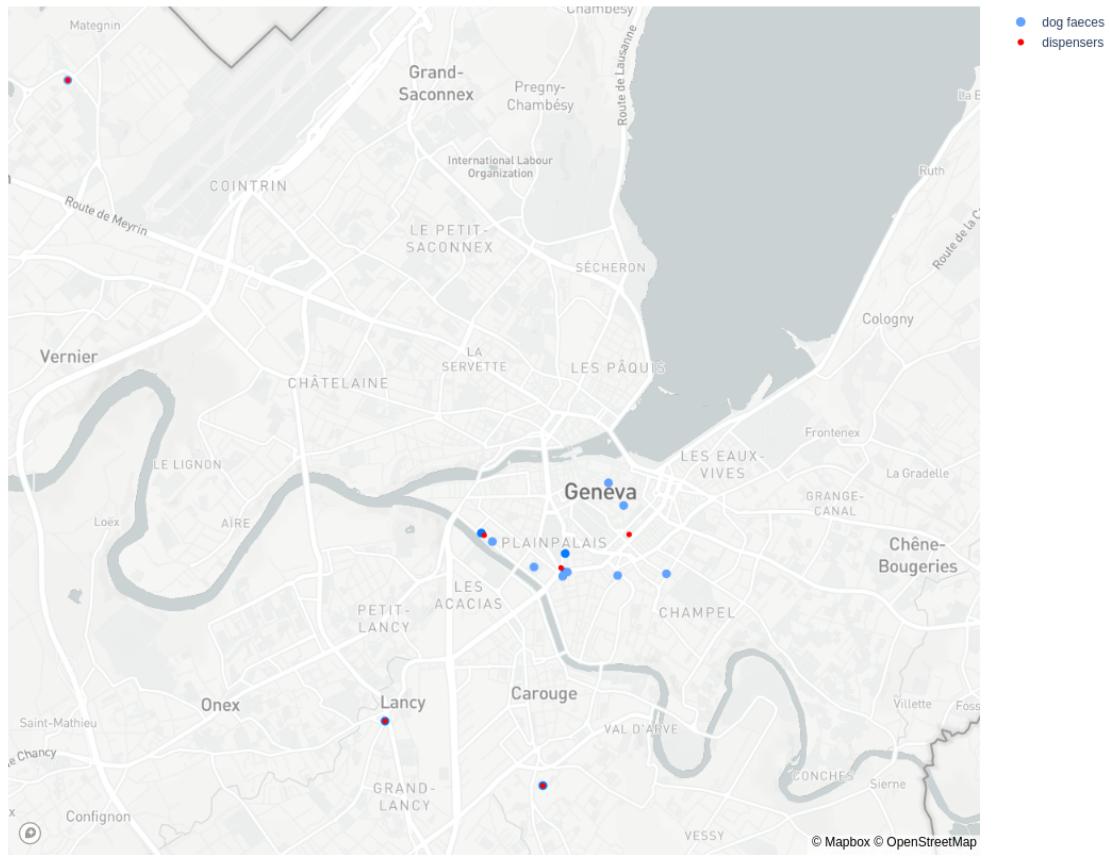
On a testé deux fois le modèle d'agglomerative clustering, une fois avec un nombre de clusters prédéfinis pour pouvoir comparer les résultats à ceux obtenus avec le modèle Mini-batch K-mean (dans les deux cas on a laissé la valeur par défaut qui est 8 clusters) et une fois avec une distance minimum en paramètres.

Voilà ce qu'on obtient avec les vraies données récoltées dans notre application :

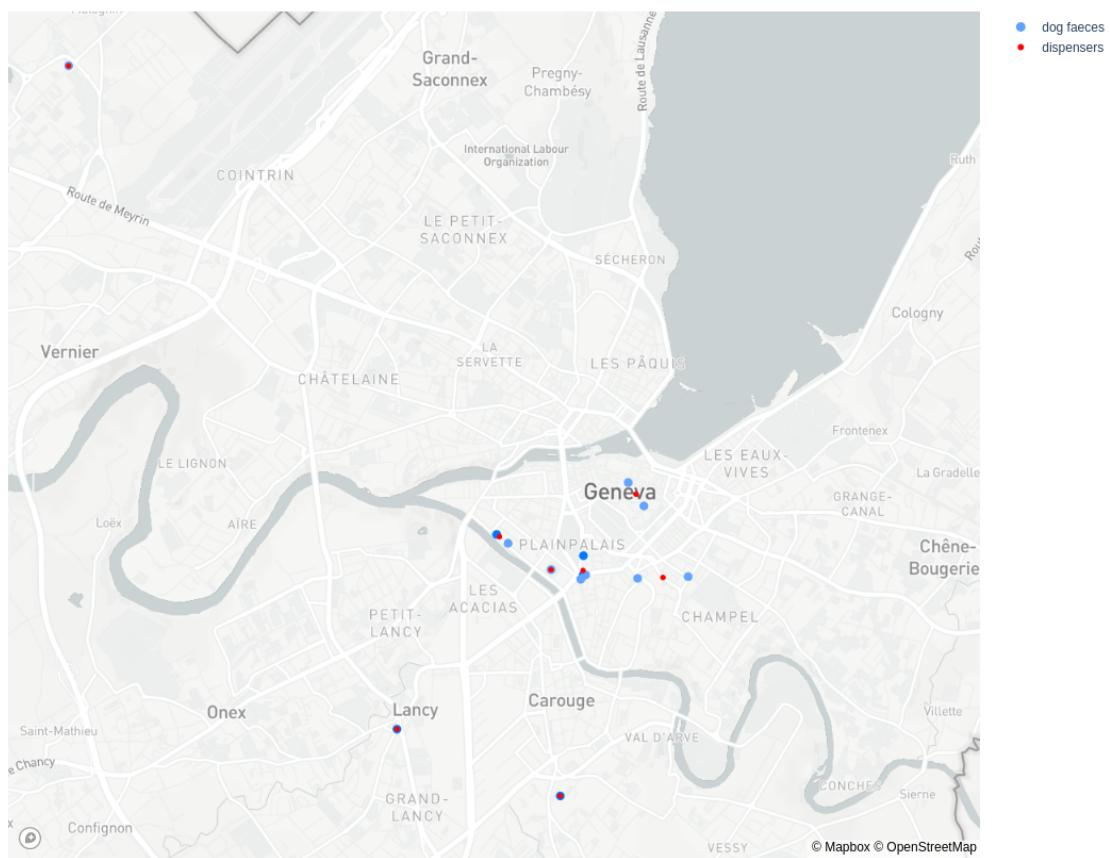
Données récoltées dans l'application



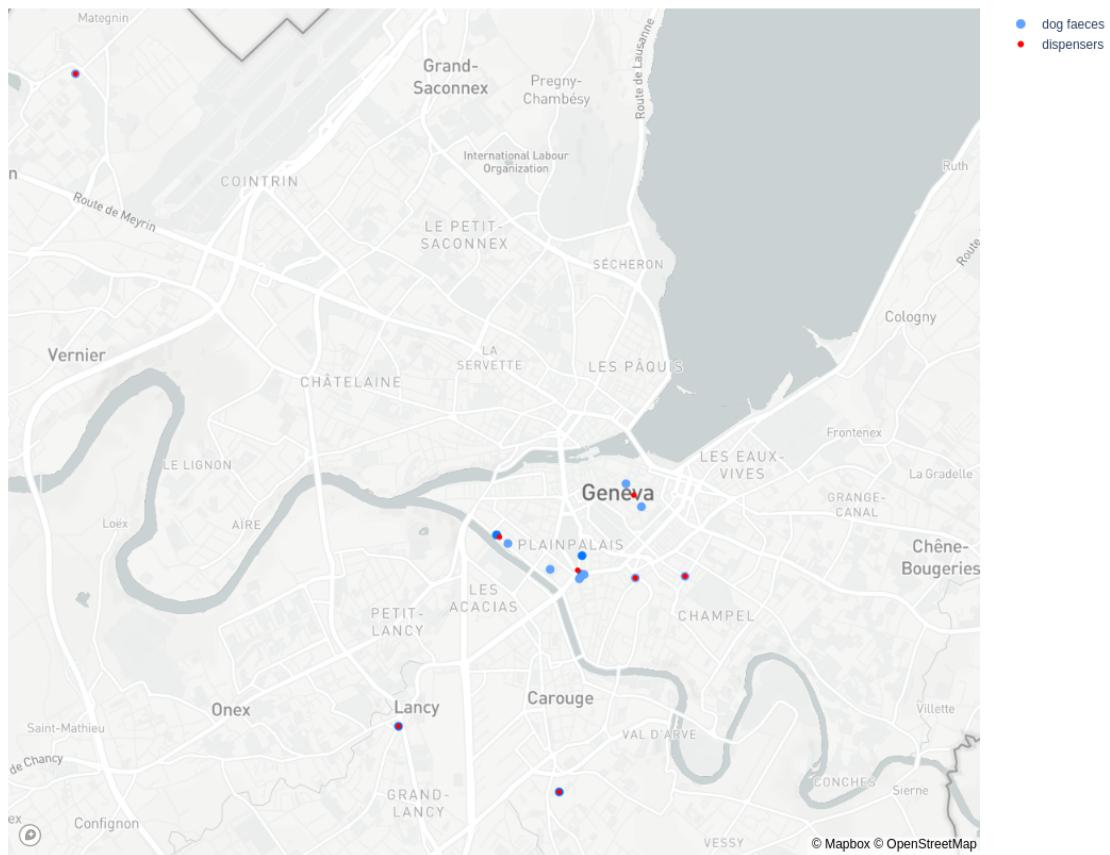
Affinity propagation



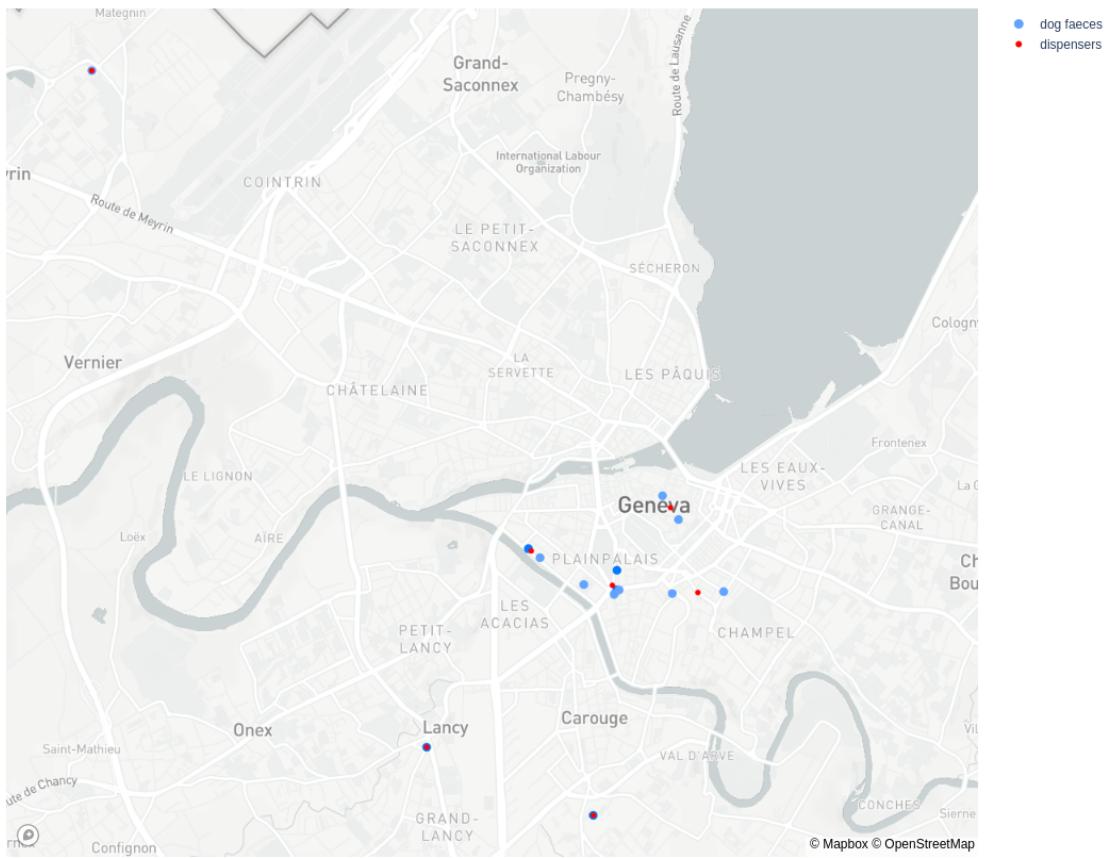
Mini-batch K-mean



Agglomerative Clustering



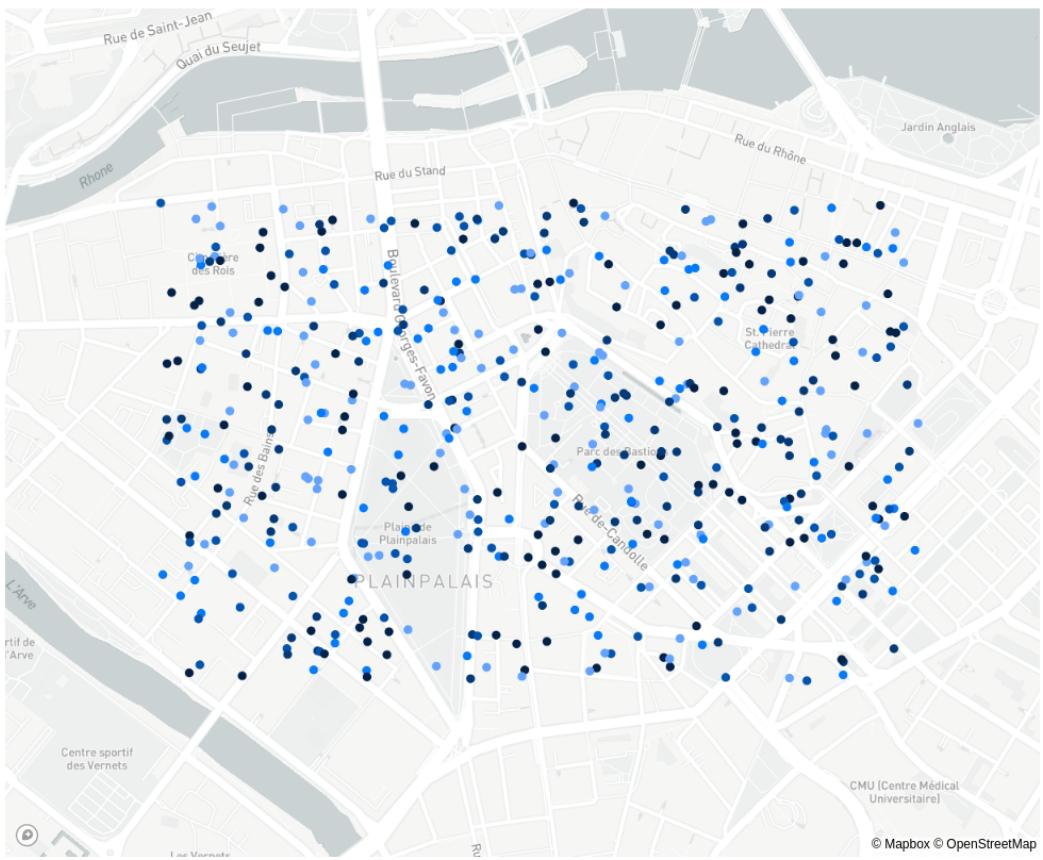
Agglomerative clustering avec une distance en paramètre



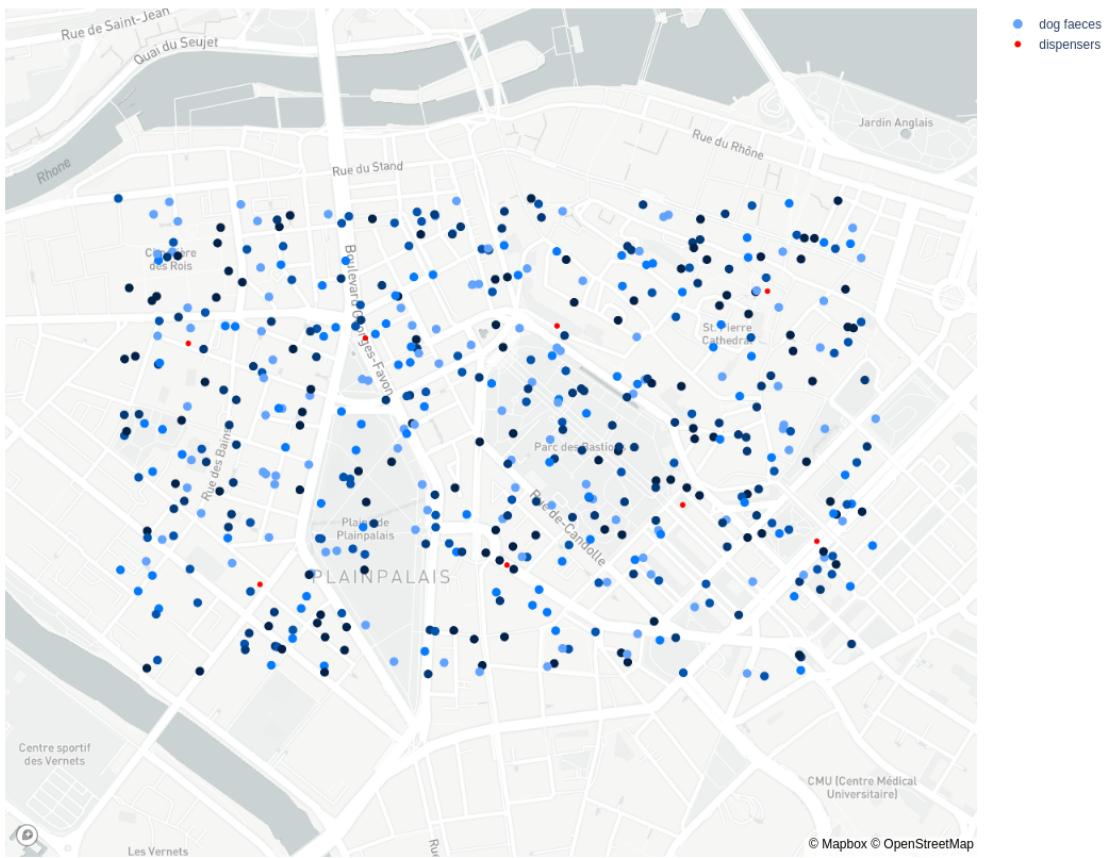
Pour les deux sets de fausses données, la méthode d'affinity propagation n'as pas réussi à converger car les données sont aléatoires, on a donc des résultats uniquement pour l'agglomerative clustering et pour le Mini-batch K-mean. Mais cette méthode devrait bien fonctionner pour un set de vrai données même plus grand, c'est pour celà que nous l'avons laissée.

Voilà les résultats obtenus pour le premier set aléatoire :

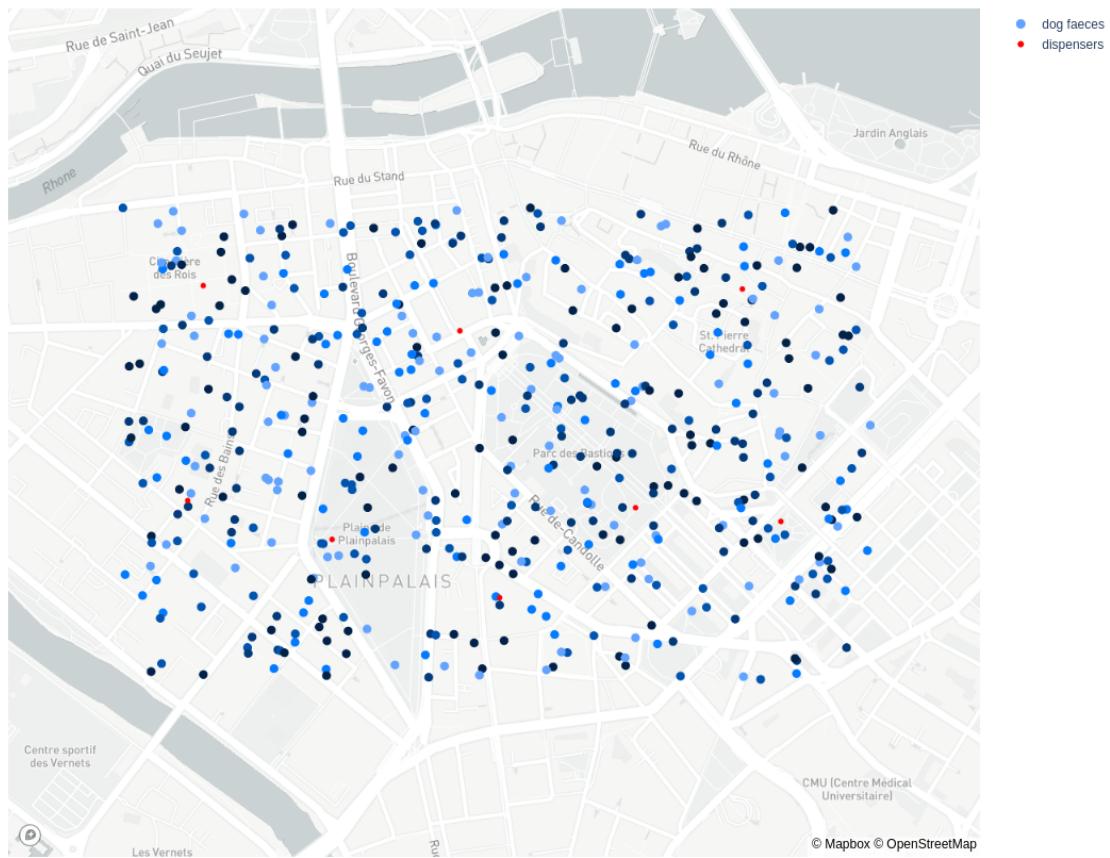
Premier set de données aléatoire



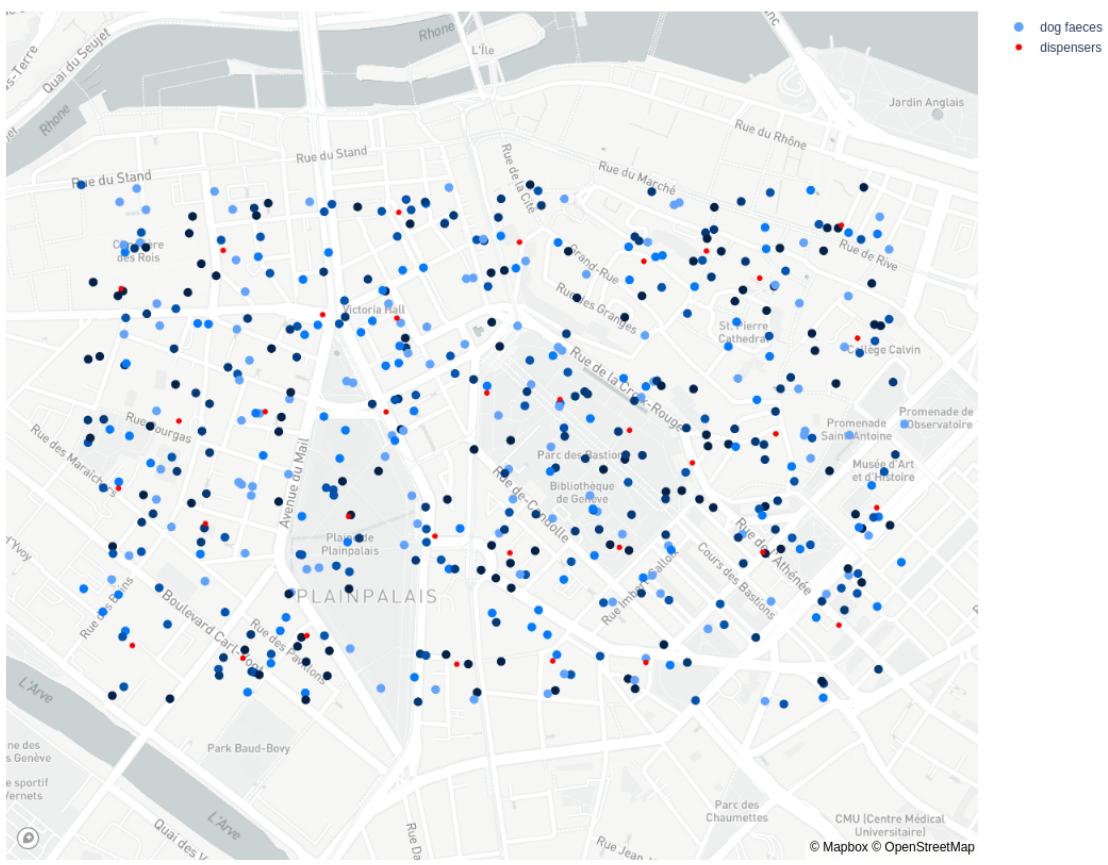
Mini-batch K-mean



Agglomerative clustering

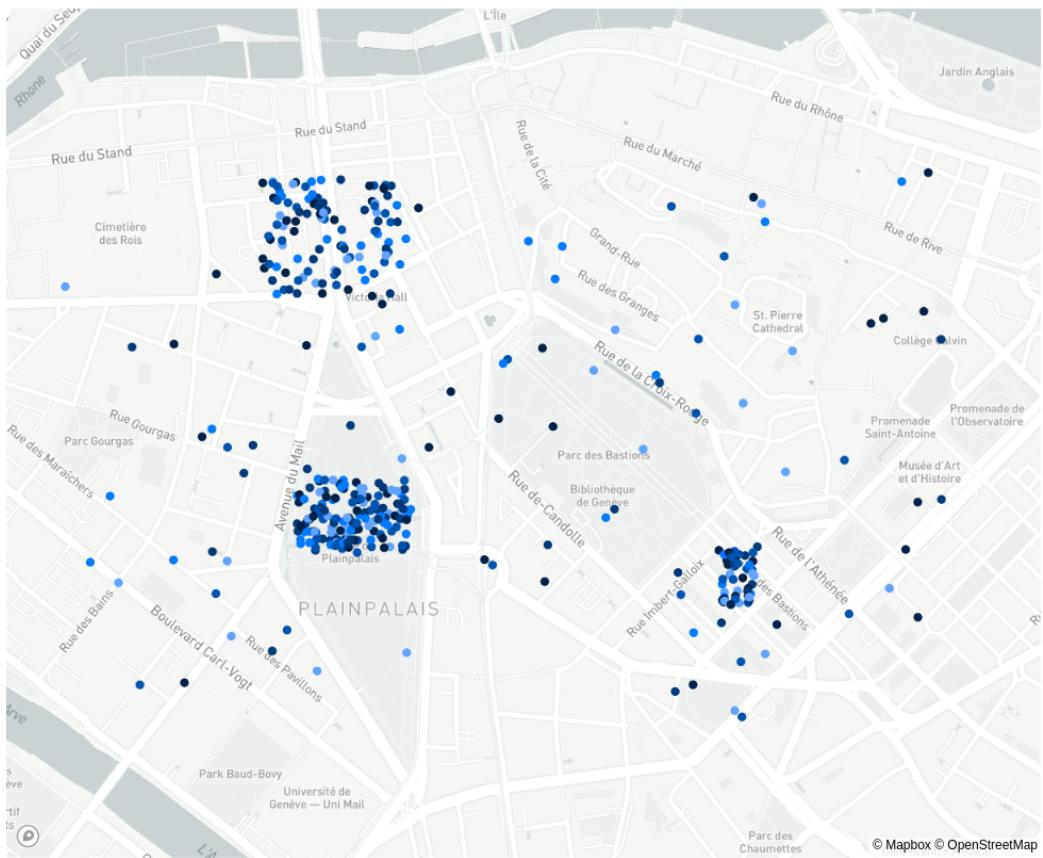


Agglomerative clustering avec une distance en paramètre

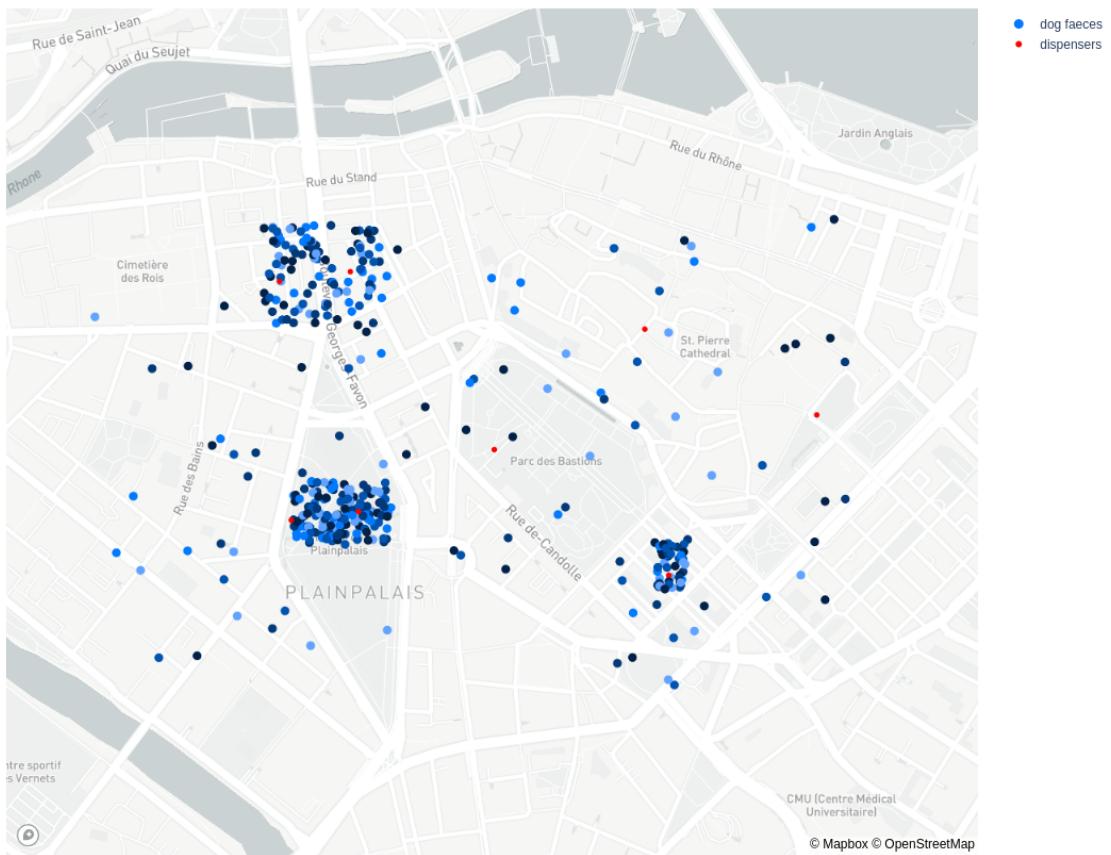


Et pour le deuxième set aléatoire :

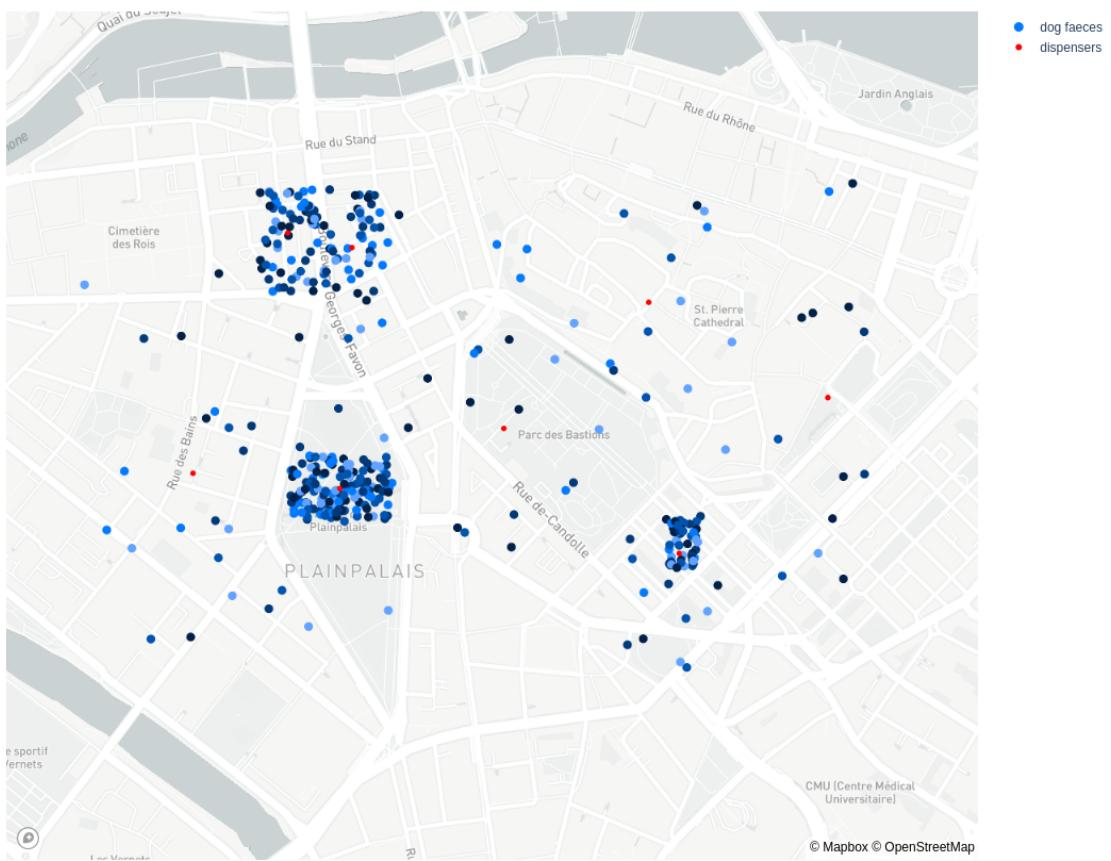
Deuxième set de données aléatoire



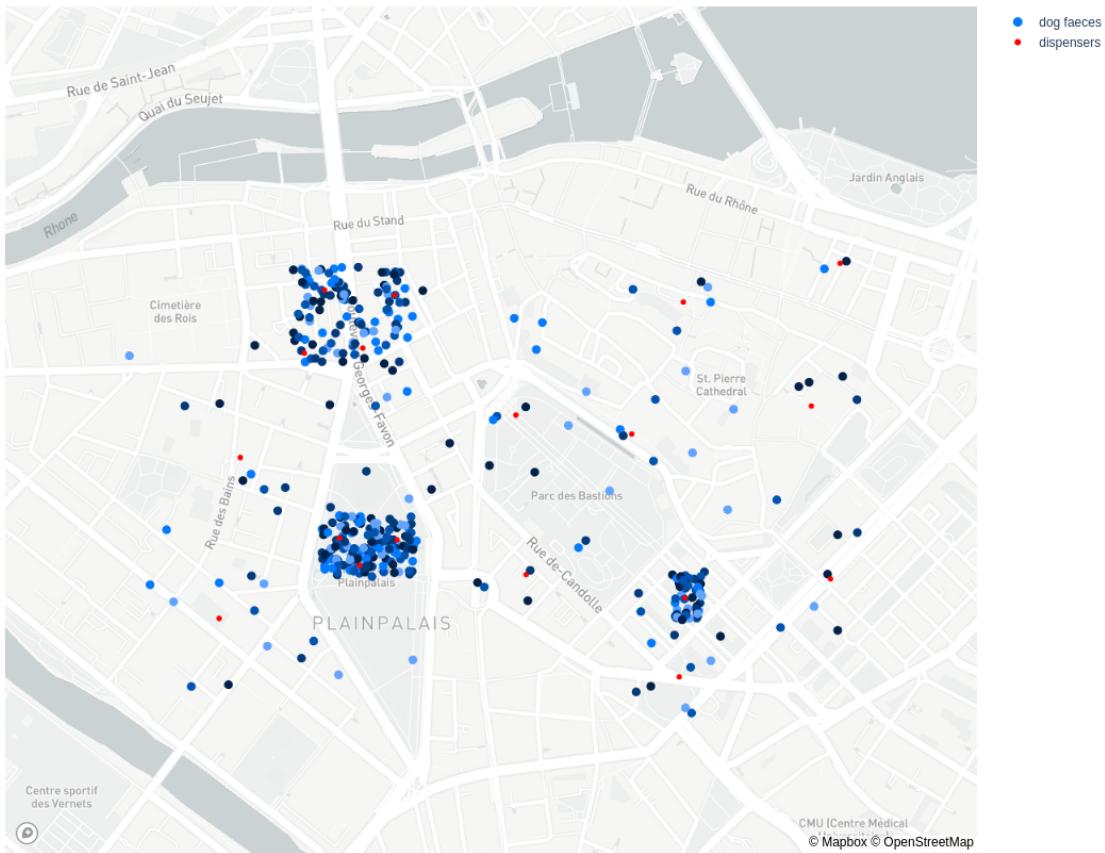
Mini-batch K-mean



Agglomerative clustering



Agglomerative clustering avec une distance en paramètre



Comme on peut le voir, pour les trois sets de données les résultats obtenus avec la méthode Mini-batch K-mean sont extrêmement similaires à ceux obtenus avec l'agglomerative clustering (sans distance en paramètre). On peut donc difficilement les départager à partir de ces observations.

Pour ce qui est du cas où on donne une distance maximale en paramètre, on peut voir que dans le cas où les données sont uniformément aléatoire la solution propose beaucoup plus de points et plus proches les uns des autres que pour le deuxième set de données aléatoire bien qu'on ait mis la même distance pour les deux. On peut en conclure que ce n'est probablement pas une solution idéale pour avoir une distance approximative entre les localisations proposées, car selon la répartition des données on obtient des distances et des quantités très variables. On pourrait, par exemple, utiliser la programmation par contrainte à la place.

Vu qu'on base nos observations principalement sur des fausses données, et que le choix dépend aussi de ce qu'on cherche à trouver (les meilleurs emplacements quel que soit le nombre de clusters ou pour un nombre défini de clusters) on ne peut pas choisir quel modèle est le plus adapté pour l'instant. Mais les trois peuvent être intéressants dans le cadre de notre problème.

Étapes suivantes

Comme expliqué plus tôt, nous avons implémenté une version simplifiée du problème au vu du temps à notre disposition. Il y aurait donc plusieurs points à améliorer pour avoir une solution qui s'adapte parfaitement à notre problème.

La première serait de faire en sorte que le modèle tienne compte de là où il est possible ou non de mettre un distributeur. Les données récoltées sont forcément situées dans la rue, ce n'est pas forcément le cas des résultats. En effet, comme le centroïde d'un cluster n'est pas forcément un des points présent dans les données, les positions suggérées pour ajouter des distributeurs peuvent être au milieu d'une route, dans un bâtiment, dans l'eau, etc. Il faudrait donc ajouter une contrainte pour que les modèles tiennent compte de ça.

Une autre chose dont il faudrait tenir compte, c'est les distributeurs déjà mis en place. Nous n'avons pas réussi à trouver la localisation des différents distributeurs dans Genève (même sur openstreetmap ils ne sont pas indiqués) c'est pour ça que les modèles implantés n'en tiennent pas compte, mais c'est un élément important pour pouvoir véritablement savoir où est-ce qu'il faut en ajouter. Il y aurait deux solutions pour ça, soit la ville de Genève à cette information et peut nous fournir les différentes localisations, auquel cas il suffit de modifier notre projet pour qu'il les prennent en compte, soit on peu nous même récolter cette information en utilisant le crowdsourcing (et potentiellement aussi de l'intelligence artificielle pour pouvoir les détecter automatiquement sur des photos, par exemple sur les réseaux sociaux, et ajouter leur localisations dans nos données).

Enfin, la dernière grosse amélioration à ajouter serait de faire en sorte de pouvoir reconnaître si deux personnes ont ajouté des données identiques et d'ainsi être capable de supprimer les doublons des données récoltées. Celà pourrait être fait relativement facilement en regardant la localisation et la date des différentes entrées, deux informations qu'on possède déjà.

Bien entendu il y a d'autres aspects qui pourraient être ajoutés ou améliorés suivant les besoins de la ville (par exemple améliorer le système de distance maximale entre les distributeurs, détecter quand une zone à déjà suffisamment de distributeurs mais qu'on y trouve toujours fréquemment des crottes afin de pouvoir essayer de mettre en place d'autres solutions, etc.) mais ce sont les trois points principaux qu'il faudrait ajouter.

Liens

<https://yukaichou.com/gamification-examples/five-examples-of-gamified-crowdsourcing-to-learn-from/>

<https://www.theguardian.com/cities/2018/may/04/poo-patrol-how-doodoowatch-could-solve-our-cities-dog-mess-problems>

<https://www.theguardian.com/world/2016/jun/13/paris-incivility-brigades-to-tackle-antisocial-behaviour>

<https://www.petbutler.com/blog-diseases-from-dog-poop/#:~:text=Both%20humans%20and%20canines%20can,shed%20eggs%20in%20dog%20feces>

Auteurs

Rose defossez



Dans ce projet je me suis occupée de la partie du rapport concernant la description de notre projet ainsi que de la communication avec la voirie. J'ai aussi participé aux différents brainstorming et j'ai apporté mon aide concernant l'IA.

Léa Heiniger



Une fois que nous avions décidé les trois ensembles de notre projet et de comment le mettre en place, j'ai principalement travaillé sur la partie technique et implémentation. Je me suis occupée de créer l'application sur Epicollect5, d'implémenter toute la partie pour intelligence artificielle (la récupération des données et la création de deux sets de fausses données, les différents modèles testés et l'affichage sur une carte des données et des résultats obtenus), de rédiger les sections "Implémentation" et "Étapes suivantes" du rapport.

Joao Filipe Costa da Quinta



Dans ce projet nous avons tous les trois participé au brainstorming. Ensuite je me suis consacré aux dangers du problème. J'ai récolté des données pour avoir un set de vraie données. Dans ce rapport j'ai écrit les parties introduction, problématique, et l'état de l'art. J'ai aussi fait toutes les recherches nécessaires pour bien compléter ces sections.