Année universitaire 2019-2020
M2 Data Science - Intelligence Artificielle
Université Claude Bernard Lyon 1

Data Visualization
Document de Cadrage

Anthony BARDOU - 11808020
Hugo POLLOLI - 11707049
Tristan SYRZISKO - 11809087
Théo RABUT - 11307400

Contents

1 Problème abordé

Dans le métro, à l'université ou à notre domicile, nous utilisons notre smartphone, nous naviguons parmi les différentes applications que nous possédons. Cependant, nous ne le faisons pas tous de la même manière. Nous développons certaines habitudes qui nous sont propres et qui varient, entre autres, fonction du jour de la semaine, de l'heure et de l'endroit où nous nous trouvons. Nous souhaitons fournir un outil capable d'apporter et de visualiser de l'information synthétique et pertinente sur ces habitudes de consommation.

2 Public et tâches

2.1 Public ciblé

Les visualisations proposées étant simples à comprendre et riches en informations, le public cible est assez large, s'étendant de l'utilisateur curieux à l'analyste souhaitant visualiser les habitudes de consommation d'une population, en passant par un concepteur d'application désireux de savoir quelles sont les applications renvoyant le plus vers son produit pour conclure d'éventuels partenariats.

2.2 Tâches rendues possibles

Un coup d'oeil aux visualisations que nous proposons permettra ainsi de répondre à des questions relatives à ces sujets impliquant plusieurs critères :

- **Temps d'utilisation** : quelle est l'application sur laquelle je passe *le plus de temps* ?
- **Localisation** : quelle est la suite d'applications la plus fréquente entre le moment où je déverrouille mon téléphone et celui où je le verrouille *quand je suis à l'université* ?
- **Temporalité** : sur quelle application vais-je le plus souvent après avoir utilisé l'application X *le samedi entre 8h et 10h* ?
- **Fréquence** : ai-je *plus de 20% de chance* d'aller sur Y quand je suis sur X ?
- **Agrégation de personnes** : quelles sont les habitudes de consommation *de* P_i *et* P_j ?
- **Toutes les combinaisons possibles de ces critères**

Les réponses à ces questions nous semblent importantes, car elles peuvent alimenter des réflexions diverses et variées (curieuses, marketing, sociologiques) sur les habitudes de consommation d'un objet qui est à présent profondément ancré dans le quotidien de l'écrasante majorité des populations des sociétés développées.

3 Source de données

Les données dont nous disposons ont été générées par 3 des 4 membres du groupe. Pour chacun d'entre eux, les données d'activité de l'application App Usage sont collectées sur une période prédifinie. Sur la même période, nous récupérons l'historique des positions fourni par Google et nous le lions avec les données d'activité. Enfin, nous agrégeons les

données des différents membres pour former le dataset final sur lequel se formeront nos visualisations.

Pour des raisons esthétiques, nous présentons les applications grâce à leurs icônes. Afin de récupérer automatiquement les icônes des applications présentes dans notre dataset, nous avons écrit un site de webscrapping capable de récupérer l'icône d'une application dans le Google Play Store.

Nous avons construit notre propre dataset après nous être rendus compte que peu de datasets de ce genre existaient en ligne, mais également parce que nous étions curieux de nos propres habitudes de consommation. Cette construction offre avantages et inconvénients :

- **Imprévu** : un mauvais paramétrage (autorisations non accordées, GPS désactivé) au début de la période de collecte des données peut rendre toute une partie des données inutilisables
- **Diversité** : étant donné que seuls les membres du groupe participent à la collecte de données, celles-ci ne sont pas aussi diversifiées que celles d'un dataset qui rassemblerait des données issues d'un plus grand nombres d'individus appartenant à des populations différentes
- **Génération à la demande** : les données d'une ou deux semaines d'utilisation sont largement suffisantes pour nourrir nos visualisations, et il est facile de recommencer / continuer la collecte s'il le faut, de manière totalement gratuite
- **Facilités d'interprétation** : les données étant issues des membres qui les analyseront par la suite, il est facile d'expliquer un phénomène mis en valeur par les visualisations

3.1 Données d'activité

Au coeur de notre projet, cette source nous permet de récupérer chaque application visitée par l'utilisateur, le moment où l'application a été démarrée et le temps pendant laquelle elle a été utilisée. Ceci nous permet de retracer aisément l'historique complet des applications visitées et de déduire la plupart des informations que nous présentons visuellement.

Ces données présentent cependant certaines limites, en particulier les données récupérées sont dépendantes de la langue du smartphone de l'utilisateur ainsi que de son modèle. Ceci rend obligatoire une étape de preprocessing que nous réalisons en Python.

3.2 Historique des positions

Bien que les positions jouent un rôle plus secondaire dans notre projet, elles restent un élément important. Elles donnent la possibilité à l'utilisateur d'effectuer un filtrage spatial sur les données et d'explorer ainsi des visualisations spécifiques à des lieux.

Ce sont ces données qui possèdent le plus de limites, car leur granularité temporelle est assez grosse et leur collecte est dépendante des autorisations données à Google par l'utilisateur. Certaines activités ne peuvent donc pas être rattachées à des données spatiales. Enfin, elles sont également dépendantes de la calibration GPS du mobile qui effectue les mesures, ce qui peut donner lieu à quelques imprécisions.

4 Travaux liés au projet

5 Organisation

Nous communiquons principalement par Slack et avons mis en place une liste d'issues sur GitHub représentant chaque tâche à accomplir pour la réalisation du projet. Les issues sont réparties selon plusieurs catégories :

- data: relatif au dataset c'est-à-dire collecter des données ou les transformer;
- filter: composant filtre, par exemple le filtre temporel;
- visu: principales visualisations des données à réaliser.

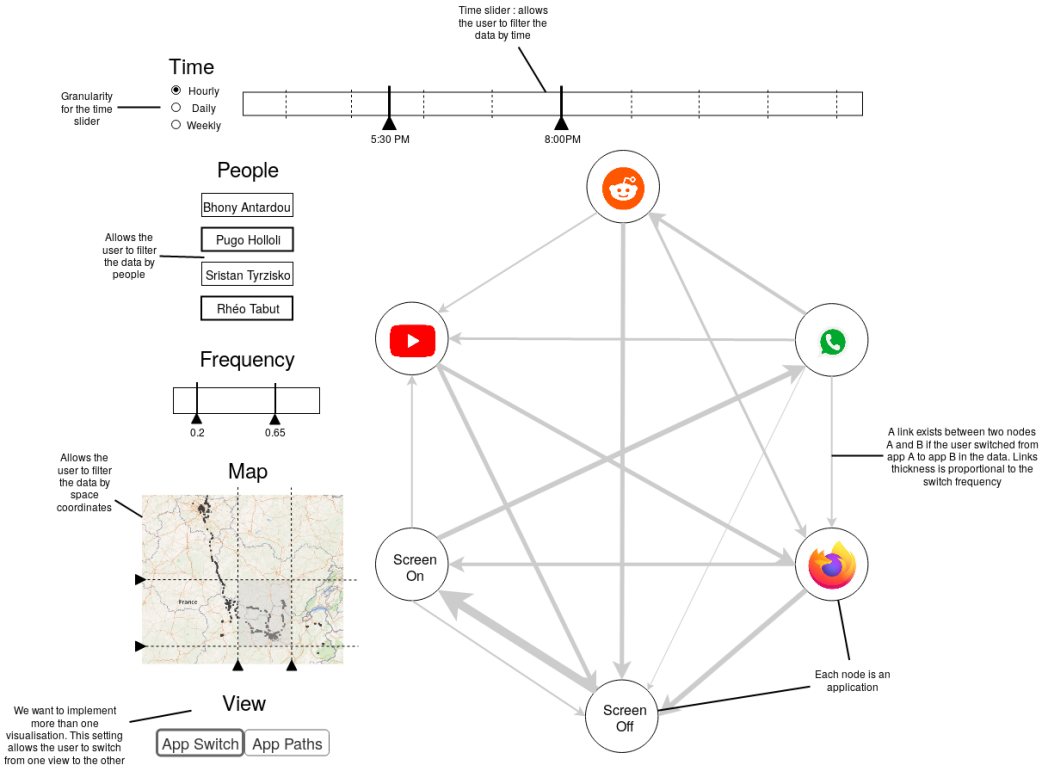
Nous n'avons pas réalisé ni prévu de séances de travail particulières, la coordination se réalise par Slack et chacun réalise les tâches qui lui sont assignées.

Rôles :

- Pré-traitement des données: Anthony;
- Développement D3 (filtres): Anthony et Hugo;
- Développement D3 (visus): Anthony et Tristan;
- Traduction de JS vers VueJS: Hugo.

6 Esquisses finales

6.1 Force directed Graph



6.2 Radial Tree