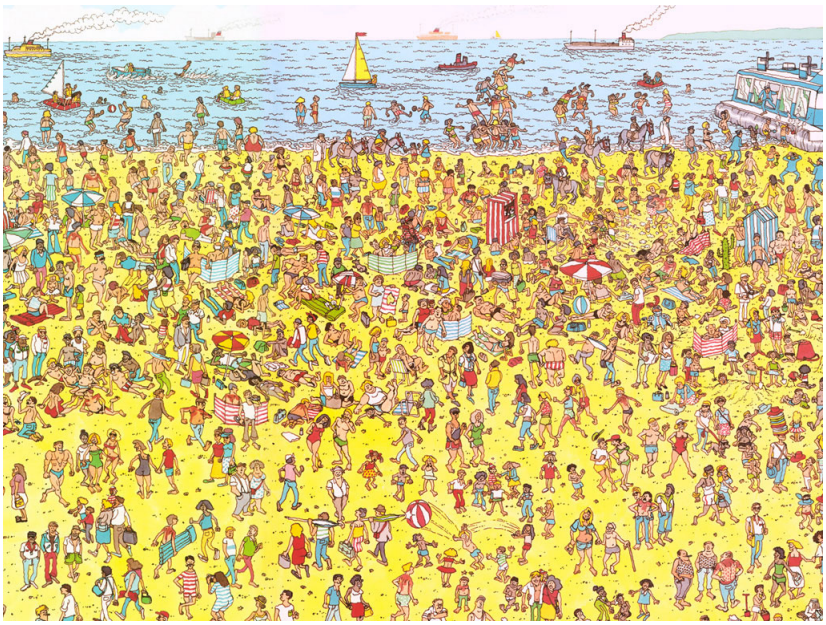
INSA Lyon
20, avenue Albert Einstein
69621 Villeurbanne Cedex

Livrable de projet

Fouille de données
« Flickr – Découverte de points d'intérêt »

du 24 février au 24 mars 2014



Les Martins :
Aline Martin
Martin Wetterwald

Enseignants :
Jean-François Boulicaut
Mehdi Kaytoue

Année scolaire 2013-2014

Sommaire

1. Préparation des données

2. Visualisation

2.1 Introduction

Pour la visualisation géographique des données, nous avons utilisé les technologies suivantes :

— Python ;
— Flask (serveur web) ;
— Scikit-learn (librairie de data mining) ;
— Google Maps API.

Avant de commencer toute tentative de représentation des données sous forme de cluster, nous avons commencé par afficher les points sur la carte. Étant donné qu'il y environ 80 000 points, nous ne pouvons en afficher qu'une partie, en raison de la lenteur observée sur les navigateurs lors du rendu de la carte.

Nous avons opté pour le rendu d'un pourcent des points. La figure 2.1 présente le rendu d'un échantillon des points.

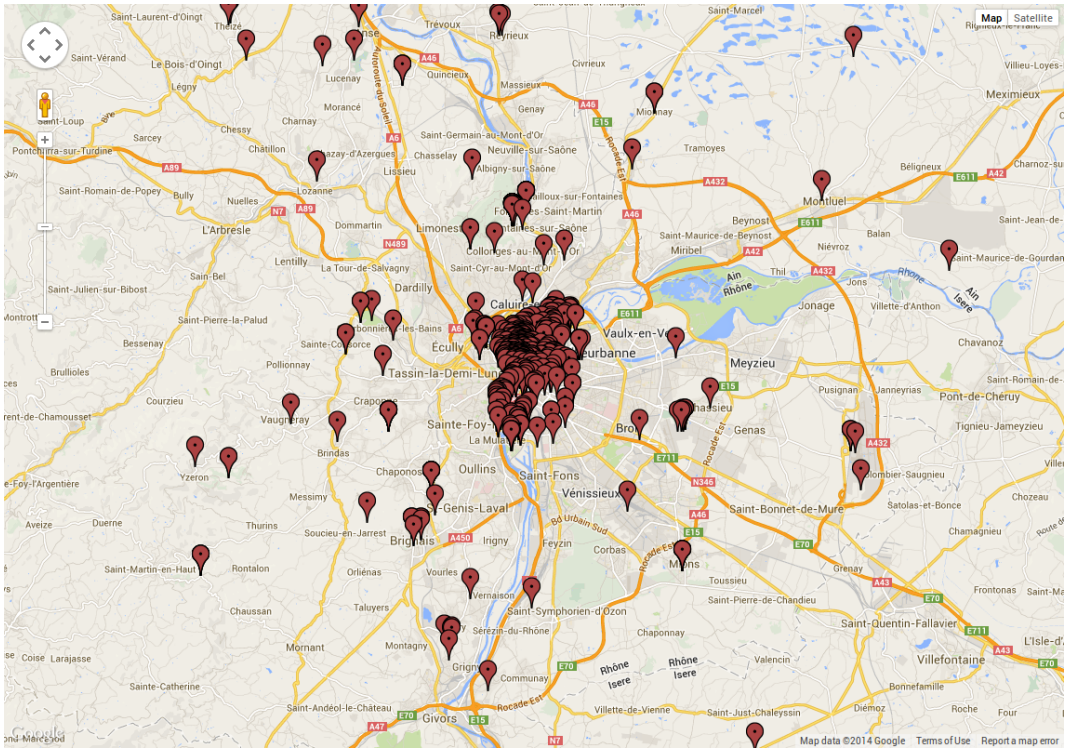


FIGURE 2.1 – Affichage d'un échantillon des points

Cela nous donne déjà un premier aperçu. Nos points sont répartis autour de la ville de Lyon, et on constate une forte densité de points dans la ville.

Mais zoomons entre les deux fleuves de Lyon (figure 2.2).

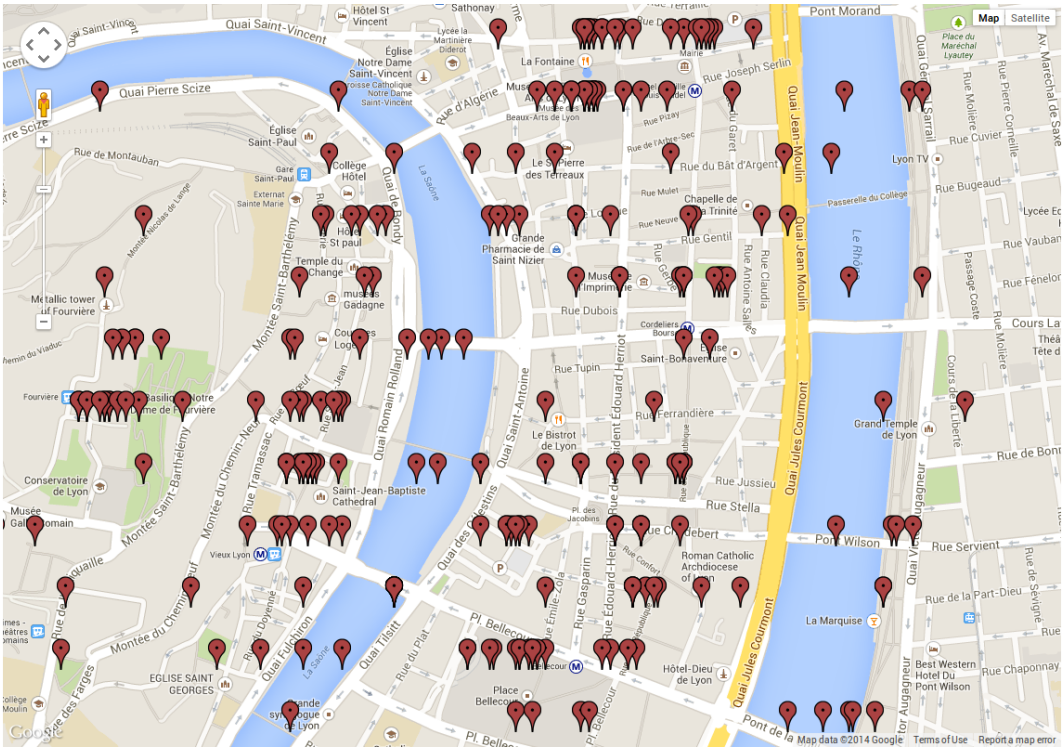


FIGURE 2.2 – Imprécision détectée lors du zoom

On constate que tous les points se trouvent sur des lignes dont la séparation est très nette. Cela est dû à un manque de précision dans nos données de départ. Dans le fichier CSV initial, nous remarquons en effet que les coordonnées GPS les plus précises que nous avons ne comportent que 4 chiffres après la virgule, ce qui explique la répartition étrange des points sur des lignes. Nous savons donc que cette erreur aura un impact sur la qualité de nos futurs clusters.

2.2 Meanshift

2.3 TODO