Remorquages Ville de Montréal

Opérations déneigement | Remorquages

TP Final

| Étude de cas données massives |

Python MRJob et Hive

Cours 420-T16-BB

Professeur Jocelyn Cloutier

Jean-François Pedneault | 2210037

Collège Multihexa

Introduction

Ce travail présente la démarche suivie pour répondre à cinq questions ayant trait au lien entre les remorquages commandés par la ville de Montréal lors des opérations de déneigement qui suivent les précipitations de neige. Le but est de démontrer l’incidence du déneigement sur l’activité économique du remorquage et les pôles de localisation où elle a lieu. Dans cette recherche j’investigue l’année 2019 seulement.

N.B. Lorsque je fait allusion au nombre total de remorquages dans ce texte ou les résultats, je sous-entends le nombre total de remorquages commandé par la ville de Montréal ayant comme motif un constat d’infraction qui comprend les remorquages dus aux opérations de déneigement.

Environnement de test

Section présentant les détails en ce qui a trait aux répertoires, fichiers et scripts linux de la démarche.

1 Arborenscence

Sous linux, dans le compte cloudera : [cloudera@quickstart ~]$

Dans le répertoire : [cloudera@quickstart Documents]$ mkdir TP\_remorquages

2 Fichiers sources

2.1 Fichier source : *Remorquages.csv*

Le fichier est disponible sur le site de la ville de Montréal :

@ <http://donnees.ville.montreal.qc.ca/dataset/remorquages-de-vehicules-genants>

Décompresser le fichier remorquages.csv : [cloudera@quickstart Documents]$ cd TP\_remorquages

2.1.1 Le fichier contient les colonnes suivantes :

1. DATE\_ORIGINE
2. LONGITUDE\_ORIGINE
3. LATITUDE\_ORIGINE
4. RUE\_ORIGINE
5. SECTEUR\_ORIGINE
6. ARRONDISSEMENT\_ORIGINE
7. DATE\_DESTINATION
8. LONGITUDE\_DESTINATION
9. LATITUDE\_DESTINATION
10. RUE\_DESTINATION
11. SECTEUR\_DESTINATION
12. ARRONDISSEMENT\_DESTINATION
13. MOTIF\_REMORQUAGE

2.1.2 Le fichier a deux champs vides :

1. SECTEUR\_ORIGINE
2. SECTEUR\_DESTINATION

2.1.3 Trie et utilisation du fichier *Remorquages.csv* :

Pour le traitement j’utiliserai seulement les colonnes suivantes :

1. DATE\_ORIGINE
2. RUE\_ORIGINE
3. ARRONDISSEMENT\_ORIGINE
4. MOTIF\_REMORQUAGE

2.2 Fichier source : *export-montreal2019.csv*

Le fichier est disponible sur le site historique-météo.net :

@ https://www.historique-meteo.net/amerique-du-nord/quebec/montreal/2019/01

Décompresser le fichier export-montreal2019.csv : [cloudera@quickstart Documents]$ cd TP\_remorquages

2.2.1 Le fichier contient les colonnes suivantes :

1. DATE
2. MAX\_TEMPERATURE\_C
3. MIN\_TEMPERATURE\_C
4. WINDSPEED\_MAX\_KMH
5. TEMPERATURE\_MORNING\_C
6. TEMPERATURE\_NOON\_C
7. TEMPERATURE\_EVENING\_C
8. PRECIP\_TOTAL\_DAY\_MM
9. HUMIDITY\_MAX\_PERCENT
10. VISIBILITY\_AVG\_KM
11. PRESSURE\_MAX\_MB
12. CLOUDCOVER\_AVG\_PERCEN
13. HEATINDEX\_MAX\_C
14. DEWPOINT\_MAX\_C
15. WINDTEMP\_MAX\_C
16. WEATHER\_CODE\_MORNING
17. WEATHER\_CODE\_NOON
18. WEATHER\_CODE\_EVENING
19. TOTAL\_SNOW\_MM
20. UV\_INDEX
21. SUNHOUR
22. OPINION

2.2.2 N.B. Le fichier ne contient aucun champs vides :

2.2.3 Trie et utilisation du fichier *export-montreal201901.csv*:

Pour le traitement j’utiliserai seulement les colonnes suivantes :

1. DATE
2. TOTAL\_SNOW\_MM
3. MAX\_TEMPERATURE\_C
4. N.B. Il semble y avoir une erreur dans le fichier de données puisque le champ TOTAL\_SNOW\_MM semble donner des centimètres (CM) plutôt que des millimètres (MM). J’ai considéré les précipitations en centimètres pour des raisons logiques.

Les cinq questions posées par la démarche

1 Quelles sont les journées de l’année ou il y a eu des chutes de neige?

La première information à extirper des données, afin de trouver les résultats recherches sur la fréquence, la localisation et la somme de l’activité économique générée par les remorquages en période de déneigement, est de trouver les journées où il y a eu des précipitations de neige. Trouver les journées consécutives de chute de neige afin de délimiter des épisodes de chute de neige qui nécessiteront une opération de déneigement.

* 1. Description du code

Trouver les journées lorsqu’il y a eu plus de 3 centimètres de précipitation de neige. Deux champs concernés dans *export-montreal201901.csv*:

DATE et TOTAL\_SNOW\_MM.

1.2. MRJob et Hive:

J’ai sorti les journées de chute de neige de plus de 3 centimètres. J’ai repéré les épisodes où un total de 12 centimètres a été atteint en 2 à 3 jours. Le lendemain et les jours suivant sont le moment où se déroule les opérations de déneigement. J’ai trouvé les totaux de remorquages lors des opérations de déneigement mais aussi lors des journées qui précèdent et des journées qui suivent pour faire une comparaison. J’ai fait une impression du Mapper pour avoir toutes les journées et fait ressortir le maximum du Reducer.

1.2.1 Exécution du code MRJob

[cloudera@quickstart TP\_remorquages]$ python ChuteDeNeigeMax.py Data/export-montreal2019.csv

1.5 Résultats

J’ai extirpé 26 dates avec des chutes de neige de plus de 3 centimètres. J'ai ajouté la température maximum au requêtes Hive pour m'assurer qu'il n'y avait pas eu un dégèle, donc une fonte de neige pendant les épisodes de chute de neige.

*N.B. Voir les codes python MRJob et Hive ainsi que les résultats en annexe.*

2 Quelles sont les journées où il y a eu le plus d’activités de remorquages?

C'est la deuxième information à extirper des données pour trouver les corrélations existantes entre ce résultat et celui des épisodes de tempête de neige.

* 1. Description du code :

Trouver les journées ou il y a eu les plus d’activités de remorquages. 2 champs concernés dans *Remorquages.csv*: DATE\_ORIGINE, MOTIF\_REMORQUAGE

N.B. Plusieurs manipulations de split pour arriver à manipuler la date de façon efficace tant dans Hive que MRJob. Voir les commentaires dans le code.

2.2 MRJob et Hive :

J’ai trouvé les totaux de remorquages coïncidant avec les dates d’opérations de déneigement. Je n’ai pas eu à utiliser MRStep pour extirper ces résultats. J'ai choisi les dates en fonction de la première épisode de chute de neige en 2019.

2.2.1 Exécution du code :

[cloudera@quickstart TP\_remorquages]$ python RemorquagesParDateMax.py Data/remorquages.csv

[cloudera@quickstart TP\_remorquages]$ python RemorquagesParDate.py Data/remorquages.csv

* 1. Résultats :

On remarque l’incidence exponentielle des opérations de déneigement sur le nombre de remorquages. Prenons par exemple le [1jan=1rem, 2jan=1rem, 3jan=18rem] pour un total de 20 remorquages. Le 3 janvier fut la journée ou l’on a atteint plus de 12 centimètres avec 14 centimètres. Ensuite les 4 jours suivants [4jan=828rem, 5jan=1122rem, 6jan=571rem, 7jan=304rem] là où a eu lieu l’opération de déneigement. Il y a donc eu un total 2825 remorquages durant cette période. Le lendemain de l’opération soit le [8jan=75rem] donc 75 remorquages seulement, ce qui pourrait correspondre à la dernière journée de déneigement. La journée avec le maximum de remorquages en 2019 fût le 10 janvier 2019 avec 3068 remorquages. Il y a eu une autre chute de neige de 8.2 centimètres le 9 janvier 2019.

*N.B. Voir les codes python MRJob et* *Hive ainsi que les résultats en annexe.*

3 Quelles sont les arrondissements qui sont le plus touchés par les activités de remorquages?

3.1 Description du code

Ensuite faire ressortir dans quels arrondissements de la ville ces remorquages ont tendances à avoir lieu. 3 champs concernés dans *Remorquages.csv*: DATE\_ORIGINE, ARRONDISSEMENT\_ORIGINE, MOTIF\_REMORQUAGE

3.2 MRJob et Hive :

J’ai utilisé MRStep pour extirper le maximum. Le résultat de Côte-des-Neiges - Notre-Dame-de-Grâce ne veut pas sortir du à des caractères ASCII non reconnaissable. Je pourrais probablement régler ce problème en intégrant le WORD\_REGEXP = re.compile(r"[\w']+") au code. J’ai essayé mais je n’ai pas réussi. Hive n'a pas posé ce problème.

3.2.1 Exécution du code :

[cloudera@quickstart TP\_remorquages]$ python RemorquagesParArrondissementParAnMax.py Data/remorquages.csv

[cloudera@quickstart TP\_remorquages]$ python RemorquagesParArrondissementParAn.py Data/remorquages.csv

3.3 Résultats :

Dans le centre de la ville, tous les arrondissements ont des totaux semblables de 8176 à 4087 remorquages, à l’exception d’Outremont qui voit ses totaux divisés par 6 en moyenne avec un total de 1067. Ceci s’explique probablement par le fait que le pourcentage d’habitations avec stationnements privés dans cet arrondissement est beaucoup plus élevé.

*N.B. Voir les codes python MRJob et* *Hive ainsi que les résultats en annexe.*

4 Quelle est l’activité économique générée par la ville de Montréal avec la commande de ces remorquages lors des épisodes de chute de neige et de déneigement qui s’en suivent.

4.1 Description du code

Faire une estimation de l’activité économique générée par les remorquages de la ville de Montréal. Trouver les totaux de nombre de remorquages par année et pas mois. Deux champs concernés dans *Remorquages.csv*: DATE\_ORIGINE et MOTIF\_REMORQUAGE.

4.2 MRJob :

4.2.1 Exécution du code

[cloudera@quickstart TP\_remorquages]$ python RemorquagesTotal.py Data/remorquages.csv

4.3 Résultats

Selon la comparaison faite des nombres de remorquages avec la première opération de déneigement en janvier 2019 fait précédemment au point 1.3, on peut considérer qu’un pourcentage négligeable (0.707%) des remorquages sont effectués pour des raisons autre que les opérations de déneigement à ce moment de l’année.

De ce fait, les 60 911 remorquages qui ont été facturés à 87$, effectués en janvier, février, mars, novembre et décembre 2019, mois ou il y a eu des chutes de neiges, ont générés un chiffre d’affaire total de $5 299 257.00!

Pour toutes les années du fichier et les mois concernés (jan-fév-mars-nov-déc) j’obtiens un total de 213 069 remorquages qui ont donc générés un chiffre d’affaire total de $18 537 003!

Il est à noter que le nombre total de remorquages pour tous les mois est de 223 023 une différence de seulement 10 000 remorquages. Donc plus de 99% des remorquages ont pour cause les opérations de déneigement.

*N.B. Voir les codes python MRJob et* *Hive ainsi que les résultats en annexe.*