SUPINFO Academic Dept.

SUPINFO International University

INSTITUTE OF INFORMATION TECHNOLOGY

5DAT – BIG DATA

Mini-Projet

Version 1.1
Use: Students/Staff
Author: SAD

SOMMAIRE

1 PREAMBULE

Ce mini-projet est à résoudre en groupe de trois au maximum.

Toute forme de plagiat, même de manière partielle, est strictement interdite et se verra sanctionnée.

Vous devrez envoyer une unique archive ZIP contenant tous vos scripts PIG commentés.

Un barème indicatif est donné dans la dernière partie de ce sujet.

Vous pouvez utiliser la VM du cours ou tout autre environnement hadoop (ex :hortonworks sandbox, cloudera quickstart VM, …).

2 PARTIE 1

2.1 CONTEXTE

Vous travaillerez dans un premier temps sur les données des émissions de CO2 et de polluants des véhicules commercialisés en France. Celles-ci sont contenues dans le fichier mars-2014-complete.csv fourni en annexe et dont le dictionnaire de données est le suivant :

nom-colonne	typerubrique	longueur	légende
lib_mrq_utac	varchar	12	la marque
lib_mod_doss	varchar	20	le modele du dossier
lib_mod	varchar	20	le modèle commercial
dscom	varchar	91	la désignation commerciale
cnit	varchar	15	le Code National d'Identification du Type (CNIT)
tvv	varchar	36	le Type-Variante-Version (TVV) ou le type Mines
cod_cbr	varchar	5	le type de carburant
hybride	varchar	3	une information permettant d'identifier les véhicules hybrides (O/N)
puiss_admin_98	varnb	2	la puissance administrative
puiss_max	varnb	11	la puissance maximale (en kW)
typ_boite_nb_rapp	varchar	3	le type de boîte de vitesse et le nombre de rapports,
conso_urb	varnb	11	consommation urbaine de carburant (en l/100km),
conso_exurb	varnb	11	consommation extra urbaine de carburant (en l/100km),
conso_mixte	varnb	11	consommation mixte de carburant (en l/100km),
co2	varnb	3	l'émission de CO2 (en g/km),
co_typ_1	varnb	11	le résultat d'essai de CO type I
hc	varnb	11	les résultats d'essai HC
nox	varnb	11	les résultats d'essai NOx
hcnox	varnb	11	les résultats d'essai HC+NOX
ptcl	varnb	5	le résultat d'essai de particules
masse_ordma_min	varnb	4	la masse en ordre de marche mini
masse_ordma_max	varnb	4	la masse en ordre de marche maxi
champ_v9	varchar	23	le champ V9 du certificat d'immatriculation qui contient la norme EURO
date_maj	varchar	6	la date de la dernière mise à jour.
Carrosserie	varchar	19	Carrosserie
gamme	varchar	14	gamme

Une fois le fichier copié sur la VM, dans HDFS, utilisez PIG pour réaliser les traitements demandés.

2.2 QUESTIONS

1. Créer un script permettant de nettoyer le fichier source pour ne garder que les données suivantes en sortie :
    a. la marque
    b. le modèle du dossier
    c. le modèle commercial
    d. la désignation commerciale
    e. le type de carburant
    f. une information permettant d'identifier les véhicules hybrides (O/N)
    g. la puissance administrative
    h. la puissance maximale (en kW)
    i. consommation urbaine de carburant (en l/100km)
    j. consommation mixte de carburant (en l/100km)
    k. consommation extra urbaine de carburant (en l/100km)
    l. l'émission de CO2 (en g/km)
    m. la masse en ordre de marche mini
    n. carrosserie
    o. gamme

Stocker le résultat dans un fichier texte que vous utiliserez comme source dans les questions suivantes.

2. Créer un script affichant à l'écran, pour chaque véhicule :
    - La désignation commerciale
    - le type de boîte de vitesse et le nombre de rapports
    - le rapport poids(en kg)/puissance (en CV).
        - utiliser puiss_max et masse_ordma_min.
        - se baser sur le ratio 1kW = 1,35 CV.

3. Créer un script affichant à l'écran le nombre de modèles distincts (lib_mod) pour chaque marque et chaque gamme.

4. Créer un script retournant à l'écran la moyenne de la consommation mixte par marque. Tirer le résultat par ordre alphabétique en fonction de la marque.

5. Créer un script retournant dans un fichier la moyenne de l'émission de CO2 en fonction de la puissance administrative. Trier le résultat par puissance administrative croissante. Ne retourner que les 10 moyennes les plus élevées.

3 PARTIE 2

3.1 CONTEXTE

Vous travaillerez ici sur un ouvrage choisi dans la collection mise à disposition par le projet Gutenberg : www.gutenberg.org.

Il n'y a pas de contrainte de langue ni de sujet. Seul le format texte brut doit être respecté.

Une fois le fichier .txt récupéré, le copier sur HDFS.

Il s'agit ici de compter le nombre de fois qu'une lettre est la première dans un mot et ce sur l'ensemble de l'ouvrage.

Il est recommandé de valider chaque étape avant de passer à la suivante en comparant le retour obtenu avec l'exemple proposé.

3.2 QUESTIONS

1. Charger le texte en renseignant le format suivant : ligne:chararray

Exemple :

```
(The two most memorable barricades which the observer of social diseases)
(can mention do not belong to the period in which the action of this)
(book is laid. These two barricades, both symbols under different)
(aspects of a formidable situation, emerged from the earth during the)
(fatal insurrection of June, 1848, the greatest street-war which history)
…
```

2. Découper chaque ligne source pour n'avoir qu'un mot par ligne en sortie avec le format suivant : mot:chararray

Exemple :

```
(The)
(two)
(most)
(memorable)
(barricades)
(which)
(the)
(observer)
(of)
(social)
(diseases)
…
```

3. Récupérer ensuite la première lettre de chaque mot

Exemple :

```
(T)
(t)
(m)
(m)
(b)
(w)
(t)
(o)
(o)
(s)
(d)
…
```

4. Créer maintenant un bag pour chaque lettre distincte en restant sensible à la casse.

Exemple :

```
(T, {(T)})
(t, {(t), (t)})
(m, {(m), (m)})
(b, {(b)})
(w, {(w)})
(o, {(o), (o)})
…
```

5. Pour chaque lettre, compter le nombre d'occurrences.

Exemple :

```
(T, 1)
(t, 2)
(m, 2)
(b, 1)
(w, 1)
(o, 2)
…
```

6. Pour terminer, trier le résultat par ordre alphabétique et le stocker dans un fichier texte.

4 BAREME INDICATIF

Partie 1 :

1. 1 point
2. 1,5 points
3. 2,5 points
4. 2,5 points
5. 2,5 points

Partie 2 :

1. 1 point
2. 1 points
3. 1 points
4. 2 points
5. 2 points
6. 1 points

Documentation : 2 points