Extraire un corpus à partir du Web

MorDev 23 octobre 2018

Quelques rappels sur la ligne de commande

Navigation

- ls
- cd *dossier*

Affichage

- echo "*Un texte à afficher*"
- cat *unFichier*
- curl *unLien*

Quelques rappels sur la ligne de commande

Filtrage

- grep *UnTrucÀChercher*
  - cat ***unFichier*** | grep **"<"**
    -> affiche toutes les lignes qui contiennent **<** dans ***unFichier***
- grep *UnTrucÀEnlever*
  - cat ***unFichier*** | grep -v **"<"**
    -> affiche toutes les lignes qui ne contiennent pas **<** dans ***unFichier***
- perl -pe 's/.*BordGauche(.*?)BordDroit.*/$1/'
  - cat ***unFichier*** | grep **"<"** | perl -pe 's/.***<**(.*?)**>**.*/$1/'
    -> comme ci-dessus, mais on ne garde ***que*** ce qui est entre <chevrons>
- sort -u

Quelques rappels sur la ligne de commande

Sortie

- > *unFichier*
  - echo "*toto*" > *unFichier*
    -> crée un fichier *unFichier* et écrit *toto* dedans
- >> *unFichier*
  - echo "*toto*" >> *unFichier*
    -> écrit *toto* à la fin du fichier *unFichier*

Quelques rappels sur la ligne de commande

Contrôle

- sleep *n*
- exit

- for *variable* in $(*affichage*)
  do
  …
  done

for *toto* in $(*cat monFichier*)
do
    echo "La ligne contient $toto"
done

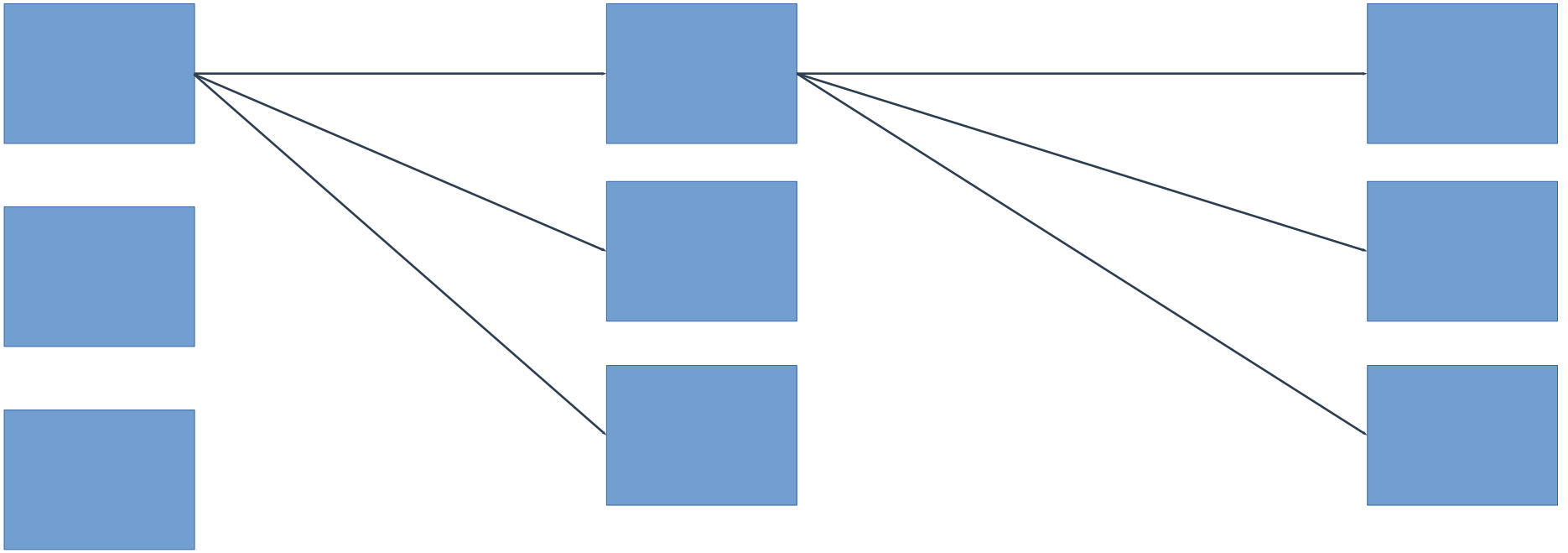
Le Web comme corpus

« en Paris » dans *Google*

- hôtel en <nomDeLieu> – hôtel à <nomDeLieu>
- hôtel en Paris – hôtel à Paris
- emploi en Paris – emploi à Paris
- … il faut cibler !

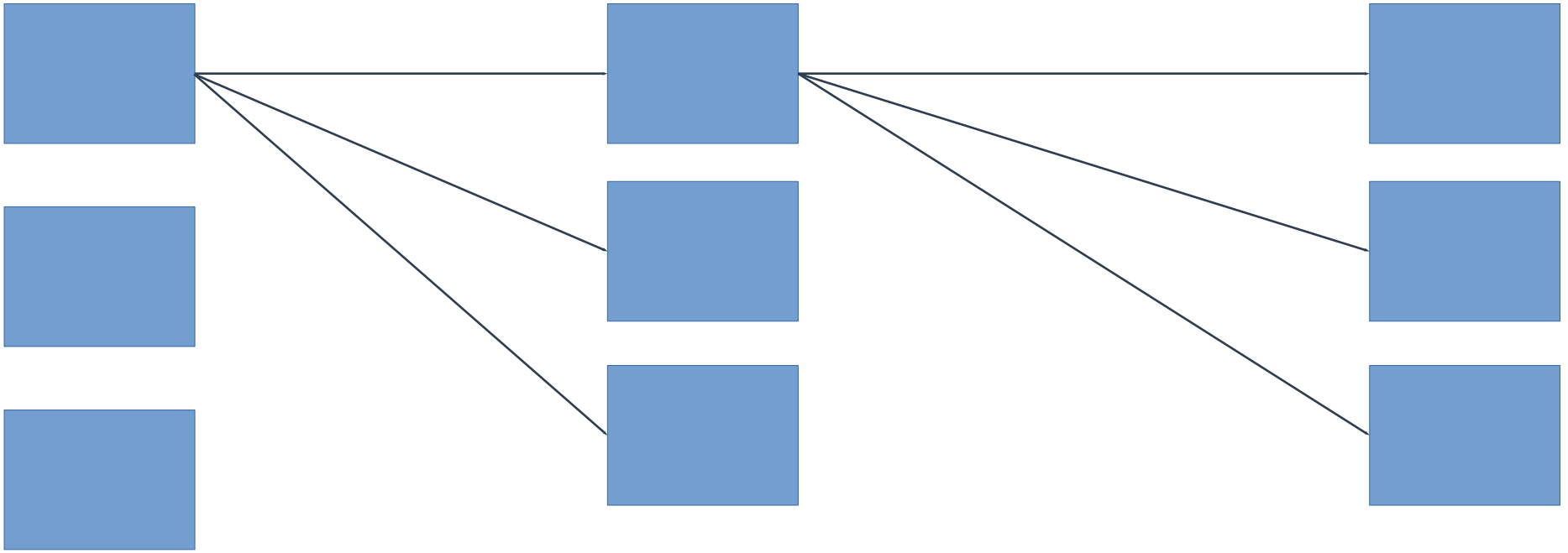
Connaître ses données

Structure d’un site Web



Connaître ses données

Structure d’un site Web (*TripAdvisor*)



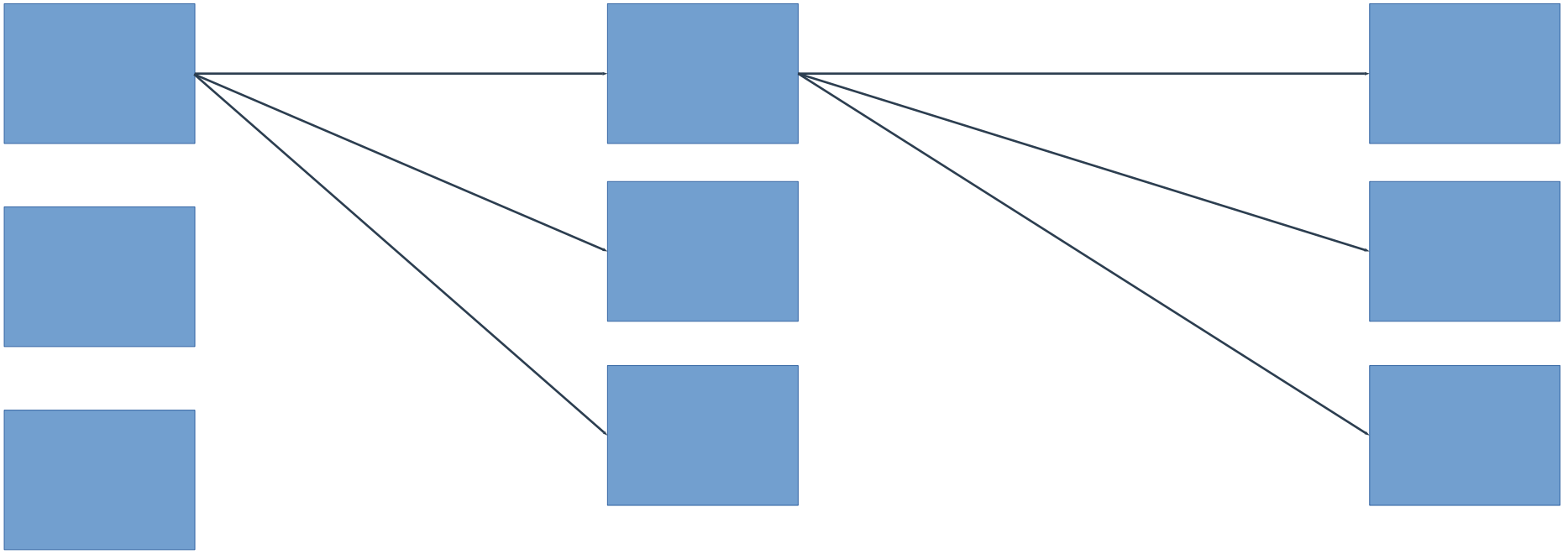
Sommaire

Sous-sommaire

Page de commentaires

Connaître ses données

Structure d’un site Web (*TripAdvisor*)



Lieu

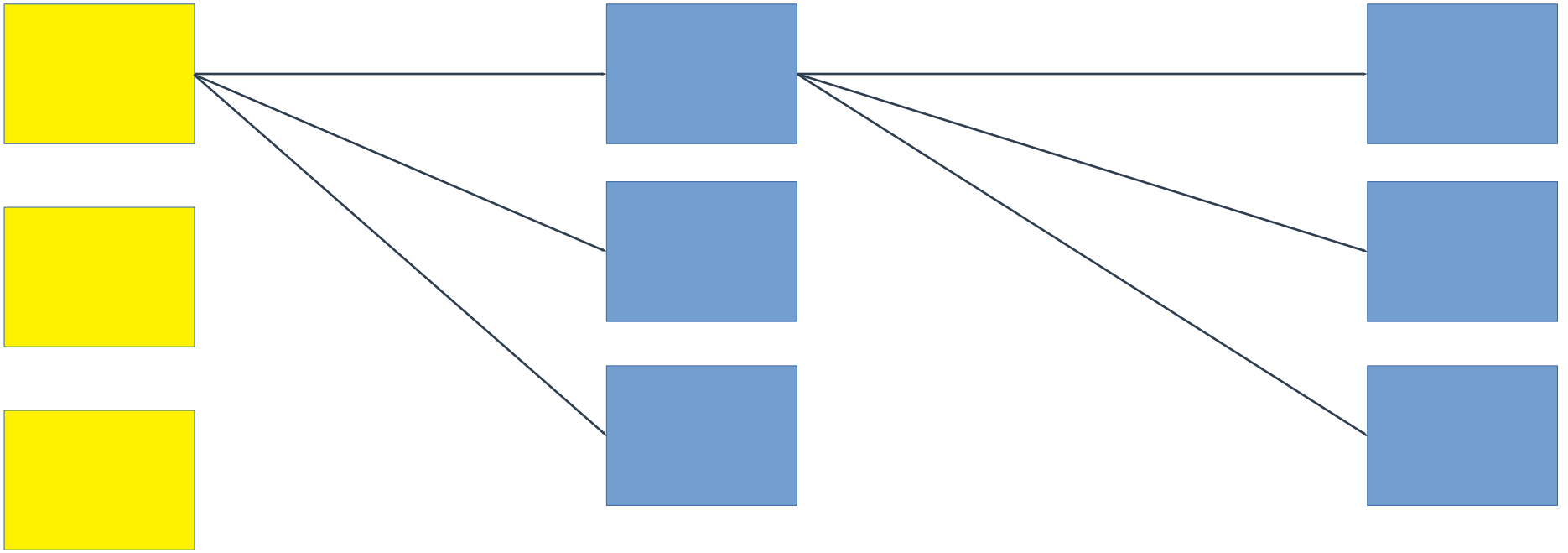
Attraction

Page de commentaires

Préparer une liste de liens

Listes d’URL de lieux

- À écrire dans un fichier *lieux.txt* (un par ligne)



Lieu

Attraction

Page de commentaires

Recherche des listes d’attractions

Créer un fichier programme.sh

- On va écrire notre programme dans ce fichier
- *bash programme.sh* pour exécuter ce programme

1 – Afficher le fichier lieux.txt

- cat lieux.txt

Recherche des listes d’attractions

2 – Afficher chaque URL du fichier lieux.txt

- for URL in $(**cat lieux.txt**)
  do
    echo $URL
    sleep 1
  done

Recherche des listes d’attractions

3 – Télécharger chaque URL

- for URL in $(cat lieux.txt)
  do
    echo $URL
    curl $URL
    sleep 1
  done

Recherche des listes d’attractions

4 – Attraper les liens qui nous intéressent

- for URL in $(cat lieux.txt)
  do
    echo $URL
    curl $URL | grep '<a href="/Attraction_Review'
    sleep 1
  done

Recherche des listes d’attractions

5 – Attraper les liens qui nous intéressent

- for URL in $(cat lieux.txt)
  do
    echo $URL
    curl $URL | grep '<a href="/Attraction_Review' |
      perl -pe 's/.*<a href="(.*?)".*/$1/' perl -pe 's/.*<a href="(.*?)".*/$1/'
    sleep 1
  done

Recherche des listes d’attractions

6 – Attraper les liens qui nous intéressent

- for URL in $(cat lieux.txt)
  do
    echo $URL
    curl $URL | grep '<a href="/Attraction_Review' |
      perl -pe 's/.*<a href="(.*?)".*/$1/' perl -pe 's/.*<a href="(.*?)".*/$1/'
    sleep 1
  done

Recherche des listes d’attractions

7 – Enlever les doublons

- for URL in $(cat lieux.txt)
  do
    echo $URL
    curl $URL | grep '<a href="/Attraction_Review' |
      perl -pe 's/.*<a href="(.*?)".*/$1/' perl -pe 's/.*<a href="(.*?)".*/$1/' |
      sort -u | grep -v "#REVIEWS"
    sleep 1
  done

Recherche des listes d’attractions

8 – Sauvegarder la liste

- **echo "" > attractions.txt**

```
for URL in $(cat lieux.txt)
do
  echo $URL
  curl $URL | grep '<a href="/Attraction_Review' |
    perl -pe 's/.*<a href="(.*?)".*/$1/'  perl -pe 's/.*<a href="(.*?)".*/$1/' |
    sort -u  | grep -v "#REVIEWS"  >> attractions.txt
  sleep 1
done
```

Recherche des attractions

Afficher chaque URL du fichier attractions.txt

- for URL in $(**cat attractions.txt**)
  do
    echo $URL
    sleep 1
  done

Recherche des commentaires

Afficher chaque URL du fichier attractions.txt

- for URL in $(**cat attractions.txt**)
  do
    echo "**https://tripadvisor.fr/**$URL"
    sleep 1
  done

Recherche des commentaires

- **echo "" > activites.txt**
  for URL in $(cat attractions.txt)
  do
    echo $URL
    curl $URL | grep '<a href="/Attraction_Review' |
      perl -pe 's/.*<a href="(.*?)".*/$1/' perl -pe 's/.*<a href="(.*?)".*/$1/' |
      sort -u | grep -v "#REVIEWS" **>> activites.txt**
    sleep 1
  done

Recherche des commentaires

Afficher chaque URL du fichier attractions.txt

- for URL in $(**cat attractions.txt**)
  do
    echo $URL
    sleep 1
  done

Recherche des commentaires

Afficher chaque URL du fichier attractions.txt

- for URL in $(**cat attractions.txt**)
  do
    echo $URL
    curl $URL | grep '<p class="partial_entry" >' |
    perl -pe 's/.*<p class="partial_entry".*?>(.*?)<\/p>.*/$1/' |
    grep -v '<'
    sleep 1
  done

Recherche des commentaires

Afficher chaque URL du fichier attractions.txt

- for URL in $(**cat attractions.txt**)
  do
    echo $URL
    sleep 1
  done