Eerste programmeeropdracht: UNIX Shell Usage

Andrew Huang s1913999

24 februari 2017

1 Introduction

Dit is het eerste opdracht voor het vak Programmeer Technieken. Deze opdracht bestaat uit 4 analyses die met behulp van de UNIX environment werden gedaan. Samen met het een en ander programmeren in Python kon belangrijke informatie uit grote hoeveelheden data worden gehaald. Hieronder de volgend 4 analyses.

2 Analyse 1

Voor deze analyse werd alle HTTP reply codes genegeerd die een code hadden van 200. Om dit te bereiken werd de volgende bash code gebruikt:

```
bzcat $src/wc_day*.out.bz2 | grep -E '200|3??' | awk '{print $7}' | sed 's%/[^/]*$%/%'
| egrep -v '\.|\*' | egrep -v '\/\/' | sort | uniq -c | sort -k2 > ../temp/tree_data.txt
```

De verschillende logs worden uitgepakt met bzcat en met grep -v "200"worden de logs met het getal 200 eruit gehaald. Daarna worden de overige logs uitgeprint en opgeslagen in een text bestand.

3 Analyse 2

In deze analyse werd het ip-adres van alle logs onderzocht en eruit gehaald. Vervolgens werd deze door de GeoIP Python module gehaald om de landcodes op te halen. Met deze landcodes kon weer gesorteerd worden totdat een lijst van de 10 landen waaruit de meeste bezoekers tevoorschijn kwam. Er zijn 2 queries die gedaan werden 1 om het ip-adres te isoleren en 1 om de gekregen landcodes te sorteren, optellen en de eerste 10 te laten zien. Het werd gedaan als volgt:

```
bzcat $src/wc_day*.out.bz2 | awk -F' ' '{print $1}' > $temp/ipaddress.txt
```

Weer met bzcat de logs vrijmaken envervolgens met awk de eerste kolom uit printen en in een text bestand toevoegen zodat deze in Python kan worden omgezet in een landcode:

```
#!/usr/bin/python
import GeoIP
gi = GeoIP.newGeoIP.GEOIP_MEMORY_CACHE
out = open"../temp/landcodes.txt", "w"

with open"../temp/ipaddress.txt","r" as f:
    for line in f:
        line = line.rstrip'\n'
        land = gi.country_code_by_addrline
        temp = strland
        out.writetemp+'\n'
f.close
out.close
```

Vervolgens werd de output text bestand door deze pipeline heengehaald:

```
cat $temp/landcodes.txt | sort | uniq -c | head > $result/p2_result.txt
```

Hier werd het tekst bestand gelezen, gesorteerd met sort en de landcodes bij elkaar opgeteld die hetzelfde waren met uniq -c. Vervolgens werd de eerste 10 landcodes getoond door head en dit alles werd in een tekst bestand gestopt.

4 Analyse 3

Voor deze analyse werden de datum waarop de HTTP Reply was vestuurd en het aantal bytes dat was verstuurd geisoleerd en vervolgens bij elkaar opgeteld, zodat het gemiddelde aantal bytes per uur kon worden geplot met behulp van een Python plot module.
De bash code voor het verzamelen en optellen van de data en bytes is als volgt:

```
    for i in $src/*.bz2; do
echo "Processing: " $i;
bzcat $i | awk '{print $4 $5 " " $10}'| tr  []/:+ " "
| awk '{h[$4] += $NF} END{for (i in h) print i, h[i]}' >> $temp/byte_data.txt;
```

Hier werx de code gerund voor het aantal bestanden dat in de folder zit. De data werd geisoleerd met behulp van awk en vervolgens werden de speciale karakters eruit gehaald. Weer met awk werd het aantal bytes van alle uren die hetzelfde zijn bij elkaar opgeteld en in een tekst bestand gezet. Daarna werd de tekst bestand gelezen om het gemiddelde van de bytes per uur te bereken:

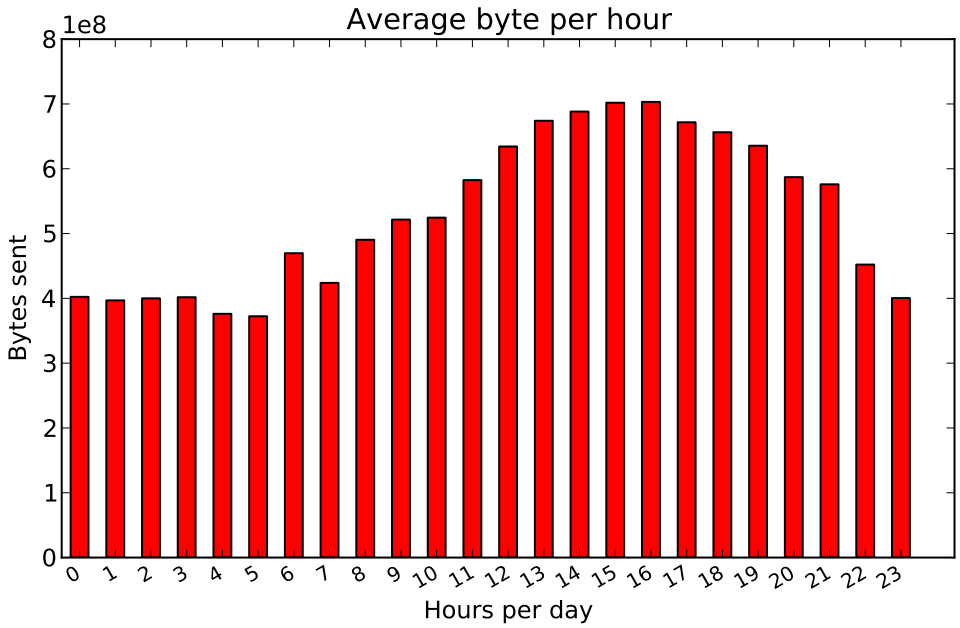
```
cat $temp/byte_data.txt | awk -v c="$num" '{h[$1] += $NF} END{for (i in h) print i, h[i]/c}'
| sort > $temp/plot_byte.txt
```

De deling werd gedaan door een constante c, wat gelijk is aan het aantal files in de folder waar alle data staat, omdat de files per dag zijn gesorteerd. Vervolgens werd deze in een tekst bestand gezet en kon gelijk worden geplot door een Python module:

```
#!/usr/bin/env python
import pandas
import matplotlib.pyplot as plt
import sys

D = pandas.read_csv"../temp/plot_byte.txt", sep=" ", header=None, index_col=0
D.plotkind="bar", rot=30, legend=False
plt.xlabel'Hours per day'
plt.ylabel'Bytes sent'
plt.title"Average byte per hour"
plt.savefig"../Result/avg_byte_per_hour.pdf"
exit0
```

Deze plot werd dan opgeslagen in een pdf.bestand:



Figuur 1: Gemiddelde bytes per uur

Zoals te zien waren er veel meer bytes die verzonden waren gedurende de avonden dit kan te maken hebben met het feit dat wedstrijden vaker in de avond plaatsvinden, waardoor er meer bezoek is op de website.

5 Analyse 4

In deze analyse werd de bezochte directory van alle bezoekers geisoleerd en vervolgens werd hiervan door Python een ascii boom van gemaakt waar goed te zien is hoevaak een bepaalde directory is bezocht. De directories werden geisoleerd met deze code:

```
bzcat $src/wc_day*.out.bz2 | grep -E '200|3??'
| awk '{print $7}' | sed 's%/[^/]*$%/%' | egrep -v '\.|\*'
| egrep -v '\/\/' | sort | uniq -c | sort -k2 > ../temp/tree_data.txt
```

Met grep werden alleen logs die 200 of een 3 nummer getal dat met een 3 begint bevatten gepakt. Vervolgens werden de directories geisoleerd met behulp van awk en met sed en egrep werden speciale karakters eruit gehaald. De uitkomst werd gesorteerd, bij elkaar gestopt en weer gesorteerd op tweede naam. Deze werd dan opgeslagen in een tekst bestand om door Python te kunnen halen.
De Python programma zet deze in een ascii boom met het aantal bezoeken per directory links afgebeeld en de boom rechts afgebeeld. Hoe verder naar rechts een directory is, hoe hoger het niveau in de boom de directory is:

```
Hits:     Directory:
15053     /
1         x----=>/
1         x----="/
4            x----=images/
12           x----=pics/
28        x----=art/
26           x----=romania/
1         x----=cgi-bin/
3            x----=hall_of_fame/
670          x----=member/
4133         x----=trivia/
3         x----=china/
2         x----=chinese/
2         x----=cities/
1         x----=cn/
1         x----=com/
1            x----=english/
1               x----=mainnav/
1         x----=deutsch/
1         x----=download/
1         x----=engilsh/
2         x----=engish/
1         x----=engli"/
1            x----=competition/
50396     x----=english/
1         x----=English/
6            x----=00frames/
0               x----=images/
2                  x----=body/
6                  x----=nav/
6                  x----=temp-nav/
2                     x----=digits/
1                     x----=cup/
1                     x----=movies/
2                     x----=infos/
8182           x----=competition/
518            x----=help/
1330              x----=image/
1                 x----=images/
796            x----=history/
457               x----=history_of/
7                    x----=images/
3046                    x----=cup/
1074                    x----=football/
1                          x----=theballisroundtx/
279                     x----=france/
6                       x----=movies/
7461              x----=images/
1              x----=history_of/
1513              x----=past_cups/
11623                x----=images/
1365                    x----=posters/
64                x----=reading/
58                   x----=images/
2              x----=hosts/
530               x----=cfo/
36                   x----=images/
4                       x----=anim/
481                     x----=cfo/
2                       x----=fin/
3                       x----=info/
3                       x----=logis/
2                       x----=marketing/
3                       x----=media/
2                       x----=press/
2                       x----=sec/
1                       x----=special/
4                       x----=sports/
4                       x----=tickets/
```

```
1                    x----=venues/
62                x----=fff/
462                  x----=images/
5                    x----=98cup/
6                    x----=competent/
5                    x----=compititions/
5                    x----=future/
4                    x----=history/
3                    x----=international/
165               x----=fifa/
362                  x----=images/
3842              x----=images/
143               x----=sponsors/
386                  x----=images/
503               x----=suppliers/
156                  x----=images/
14                   x----=credit/
53                   x----=danone/
162                  x----=eds/
272                  x----=francetele/
155                  x----=hp/
12                   x----=laposte/
16                   x----=manpower/
129                  x----=sybase/
1              x----=i/
243819         x----=images/
13                x----=anim/
5                 x----=black_nav/
1                 x----=co/
5                 x----=homepage/
11                x----=jscriptlnav/
1                 x----=nav_venue_ol/
3              x----=index/
1              x----=index,html/
3876           x----=individuals/
2                 x----=player111722"/
2                 x----=player115188"/
1                 x----=player2862/
1                 x----=player33482/
1                 x----=player908"/
1                 x----=awt/
1                 x----=download/
1033           x----=member/
3610              x----=images/
10697          x----=news/
1                 x----=english/
2                 x----=><HR><H3>Transfer/
1                 x----=mroul_images/
1                 x----=newsletters/
2              x----=past_cups/
2              x----=pastcups/
2919           x----=playing/
1391              x----=download/
11132                x----=images/
0                      x----=screen/
71                         x----=mac/
120                        x----=win_3x/
2519                       x----=win_95/
29812             x----=images/
12231                x----=anim/
3                    x----=play/
5625                 x----=trivia/
1935              x----=mascot/
8062                 x----=images/
10                x----=rules/
2              x----=programme/
6              x----=ProScroll/
2              x----=site/
1                 x----=image/
2              x----=splash_inet/
```

```
0              x----=beans/
2                 x----=infos/
1           x----=teambio/
10736       x----=teams/
1              x----=brazil/
1              x----=groupa/
1              x----=images/
3663        x----=tickets/
10574          x----=images/
1528        x----=venues/
778            x----=cities/
0               x----=images/
830                 x----=bordeaux/
1712                x----=denis/
360                 x----=etienne/
815                 x----=lens/
712                 x----=lyon/
1525                x----=marseille/
265                 x----=montpellier/
481                 x----=nantes/
1718                x----=paris/
1                   x----=tou><HR><H3>Overdracht/
459                 x----=toulouse/
8336           x----=images/
1164              x----=france/
1              x----=Lyon/
1              x----=Marseille/
1              x----=Nantes/
1                 x----=St-Denis/
537            x----=venues/
3              x----=news/
49                x----=briefs/
2              x----=suppliers/
2              x----=travel/
0                 x----=images/
1                    x----=bordeaux/
1         x----=esp/
1         x----=fr_/
1         x----=fr/
3         x----=francais/
1            x----=mascot/
10389     x----=french/
1            x----=0/
1            x----=2/
1               x----=history_of/
1            x----=BGCOLOR/
1            x----=bienvenue/
0               x----=news/
1                  x----=briefs/
1               x----=suppliers/
1            x----=billets/
1               x----=tck_/
0                  x----=beans/
1                     x----=infos/
1418         x----=competition/
85           x----=help/
211             x----=image/
155          x----=history/
100             x----=history_of/
2                  x----=images/
500                   x----=cup/
234                   x----=football/
129                   x----=france/
1319           x----=images/
206            x----=past_cups/
1694              x----=images/
209                  x----=posters/
19             x----=reading/
14                x----=images/
117            x----=cfo/
```

```
22                x----=images/
86                   x----=cfo/
5                    x----=info/
4                    x----=marketing/
6                    x----=media/
5                    x----=sec/
1                    x----=special/
5                    x----=sports/
6                    x----=venues/
27             x----=fff/
207               x----=images/
8                    x----=98cup/
1                    x----=competent/
1                    x----=compititions/
1                    x----=history/
21             x----=fifa/
53                x----=images/
671            x----=images/
28             x----=sponsors/
58                x----=images/
36             x----=suppliers/
43                x----=images/
6                    x----=credit/
25                   x----=danone/
2                    x----=eds/
10                   x----=francetele/
4                    x----=hp/
11                   x----=sybase/
44360       x----=images/
7              x----=anim/
2           x----=index/
591         x----=individuals/
1              x----=player111722/
1              x----=player111760/
1              x----=player2846/
1           x----=marseille/
138         x----=member/
451            x----=images/
1           x----=monde_qui_joue/
1              x----=events/
5              x----=travel/
2023        x----=news/
427         x-----=playing/
133            x----=download/
994               x----=images/
4480           x----=images/
1473              x----=anim/
932               x----=trivia/
410            x----=mascot/
1552              x----=images/
3              x----=rules/
2              x----=venues/
1           x-----=ProScroll/
0              x----=beans/
2                 x----=infos/
1594        x----=teams/
0              x----=mascot/
2                 x----=images/
976         x----=tickets/
2781           x----=images/
356         x----=venues/
1              x----=404/
1              x----=439/
1              x----=507/
234            x----=cities/
0               x----=images/
265                  x----=bordeaux/
432                  x----=denis/
91                   x----=etienne/
231                  x----=lens/
```

```
189                  x----=lyon/
454                  x----=marseille/
113                  x----=montpellier/
226                  x----=nantes/
379                  x----=paris/
191                  x----=toulouse/
2389            x----=images/
373                x----=france/
1               x----=javascript:finddist'Bordeaux'/
1               x----=javascript:finddist'Nantes'/
1               x----=javascript:finddist'St-Etienne'/
1               x----=javascript:finddist'Toulouse'/
174             x----=venues/
2          x----=german/
0               x----=images/
1                  x----=movies/
7          x----=hk/
1             x----=images/
2          x----=hongkong/
3          x----=><HR><H3>Transfer/
604062     x----=images/
2          x----=Images/
1          x----=images2/
2             x----=anim/
3             x----=bordeaux/
2             x----=cal_t><HR><H3>Transfer/
1             x----=denis/
4             x----=etienne/
1             x----=home_fr_phra><HR><H3>/
3             x----=marseille/
1             x----=montpellier/
1             x----=nantes/
1             x----=toulouse/
1          x----=ima><HR><H3>Transfer/
1          x----=img/
2          x----=index/
1          x----=italian/
3          x----=jp/
1          x----=js/
1          x----=l/
3          x----=martes/
1          x----=mascot/
1          x----=maxweb/
1          x----=platini/
5             x----=images/
1          x----=portuguese/
7          x----=spanish/
1             x----=images/
1          x----=stadiums/
0                x----=shockwave/
1                   x----=images/
1          x----=team/
1          x----=uk/
104        x----=welcome/
```