Rufai Balogun e2008985

Shell scripting assignment

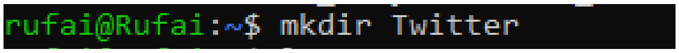
A basic introduction to Shell scripting for Msc GeoData science students at UBS as part of the Copernicus Master in Digital Earth.
Facilitated by: Dr. Charlotte Petellier

The following highlights below shows the steps used to achieve certain tasks using the UNIX-based shell. The shell scripts and the rationale for using them were also mentioned.

1. Create a new Directory

   As a first step, it is important to create a singular working folder for organizing the files, data and other output of the task. Here, a twitter folder was created using the shell command:

   mkdir Twitter



```
rufai@Rufai:~$ mkdir Twitter
```

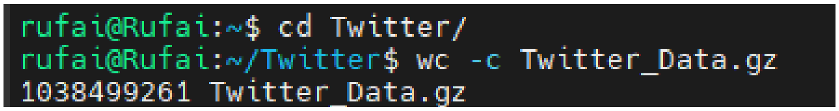
Fig 1: Shell command for creating a directory.

2. Download the file. How big is it?

   The twitter data was uploaded on the Unix environment using a third party software -- MobaXterm and decompressed afterwards using the shell command:

   gzip - d Twitter.gz

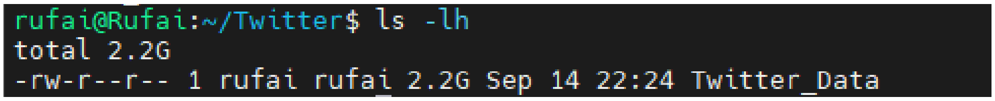
   To get an idea of the size of the file, the ***wc -c*** command was used to check the file size prior to decompressing it.



```
rufai@Rufai:~$ cd Twitter/
rufai@Rufai:~/Twitter$ wc -c Twitter_Data.gz
1038499261 Twitter_Data.gz
```

Fig 2: Shell command for extracting the number of bytes occupied by a file.

   From the information displayed, we can tell that the twitter data in its compressed format is about **1038499261 bytes (~ 1 GB) in** size. After decompression, the file size was checked using the ***ls -lh*** command



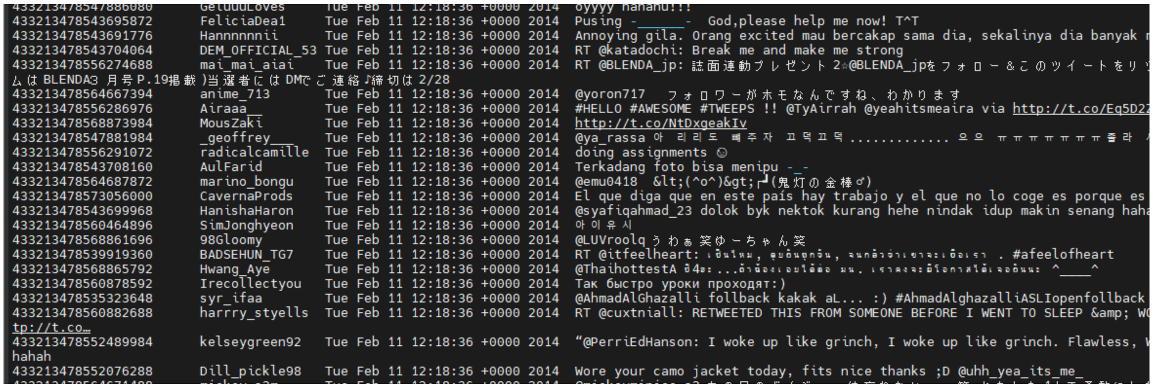
```
rufai@Rufai:~/Twitter$ ls -lh
total 2.2G
-rw-r--r-- 1 rufai rufai 2.2G Sep 14 22:24 Twitter_Data
```

Fig 3: Shell command for showing human-readable information (including the size) about the files in a folder.

   This shows that the twitter file in its decompressed format **is approximately 2.2GB.**

3. What delimiter is used to separate the columns in the file?

   A quick visual inspection of the twitter data, shows that the columns are tab-delimited; separated by tabs. A See a sample of the data in the image shown below.

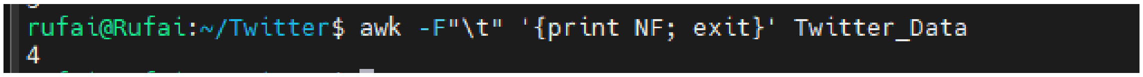


```
433213478547886080      GetuuuLoves     Tue Feb 11 12:18:36 +0000 2014  oyyyy hahahu!!!
433213478543695872      FeliciaDea1     Tue Feb 11 12:18:36 +0000 2014  Pusing -______-  God,please help me now! T^T
433213478543691776      Hannnnnnii      Tue Feb 11 12:18:36 +0000 2014  Annoying gila. Orang excited mau bercakap sama dia, sekalinya dia banyak m
433213478543704064      DEM_OFFICIAL_53 Tue Feb 11 12:18:36 +0000 2014  RT @katadochi: Break me and make me strong
433213478556274688      mai_mai_aiai    Tue Feb 11 12:18:36 +0000 2014  RT @BLENDA_jp: 誌面連動プレゼント2☆@BLENDA_jpをフォロー＆このツイートをリツ
ムはBLENDA3 月号P.19掲載)当選者にはDMでご連絡♪締切は2/28
433213478564667394      anime_713       Tue Feb 11 12:18:36 +0000 2014  @yoron717  フォロワーがホモなんですね、わかります
433213478556286976      Airaaa__        Tue Feb 11 12:18:36 +0000 2014  #HELLO #AWESOME #TWEEPS !! @TyAirrah @yeahitsmeaira via http://t.co/Eq5D2Z
433213478568873984      MousZaki        Tue Feb 11 12:18:36 +0000 2014  http://t.co/NtDxgeakIv
433213478547881984      _geoffrey___    Tue Feb 11 12:18:36 +0000 2014  @ya_rassa 아 리리도 빼주자 끄덕끄덕............ 으으 ㅠㅠㅠㅠㅠㅠㅠ좋라 ㅅ
433213478556291072      radicalcamille  Tue Feb 11 12:18:36 +0000 2014  doing assignments ☹
433213478543708160      AulFarid        Tue Feb 11 12:18:36 +0000 2014  Terkadang foto bisa menipu -_-
433213478564687872      marino_bongu    Tue Feb 11 12:18:36 +0000 2014  @emu0418  &lt;(^o^)&gt;┌┛(鬼灯の金棒♂)
433213478573056000      CavernaProds    Tue Feb 11 12:18:36 +0000 2014  El que diga que en este país hay trabajo y el que no lo coge es porque es
433213478543699968      HanishaHaron    Tue Feb 11 12:18:36 +0000 2014  @syafiqahmad_23 dolok byk nektok kurang hehe nindak idup makin senang haha
433213478560464896      SimJonghyeon    Tue Feb 11 12:18:36 +0000 2014  아 이 유 시
433213478568861696      98Gloomy        Tue Feb 11 12:18:36 +0000 2014  @LUVroolq うわぁ笑ゆーちゃん笑
433213478539919360      BADSEHUN_TG7    Tue Feb 11 12:18:36 +0000 2014  RT @itfeelheart: เป็นไหม, คุยอินดูกอิน, จนกล้าจำเขาจะเชื่อเรา . #afeelofheart
433213478568865792      Hwang_Aye       Tue Feb 11 12:18:36 +0000 2014  @ThaihottestA ปี4อะ...ถ้าน้องเอยได้ต่อ มน. เราคงจะมีโอกาสได้เจออันนะ ^____^
433213478560878592      Irecollectyou   Tue Feb 11 12:18:36 +0000 2014  Так быстро уроки проходят:)
433213478535323648      syr_ifaa        Tue Feb 11 12:18:36 +0000 2014  @AhmadAlGhazalli follback kakak aL... :) #AhmadAlghazalliASLIopenfollback
433213478560882688      harrry_styells  Tue Feb 11 12:18:36 +0000 2014  RT @cuxtniall: RETWEETED THIS FROM SOMEONE BEFORE I WENT TO SLEEP &amp; WO
tp://t.co…
433213478552489984      kelseygreen92   Tue Feb 11 12:18:36 +0000 2014  “@PerriEdHanson: I woke up like grinch, I woke up like grinch. Flawless, W
hahah
433213478552076288      Dill_pickle98   Tue Feb 11 12:18:36 +0000 2014  Wore your camo jacket today, fits nice thanks ;D @uhh_yea_its_me_
```

4. How many columns are there? What do the columns describe?

   To extract the number of columns in the data, the awk shell command shown below was used:

   awk -F "\t"'{print NF; exit}' Twitter_Data

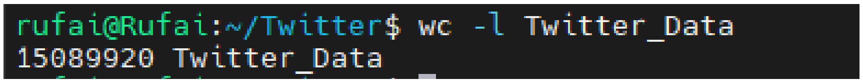
   

   ```
   rufai@Rufai:~/Twitter$ awk -F"\t" '{print NF; exit}' Twitter_Data
   4
   ```

   The output shows that there are **4 columns i**n the data. These columns describe the individual identity of the tweets (tweet id), the username, data of the tweets and the content of the tweets respectively.

5. How many Tweets are there in the file?

   The data contains at least 15,089,920 tweets as extracted by the shell command shown below.

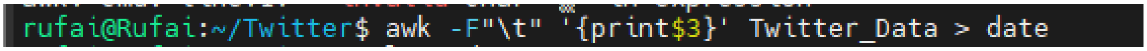
   

   ```
   rufai@Rufai:~/Twitter$ wc -l Twitter_Data
   15089920 Twitter_Data
   ```

6. What is the date range for Tweets in this file?

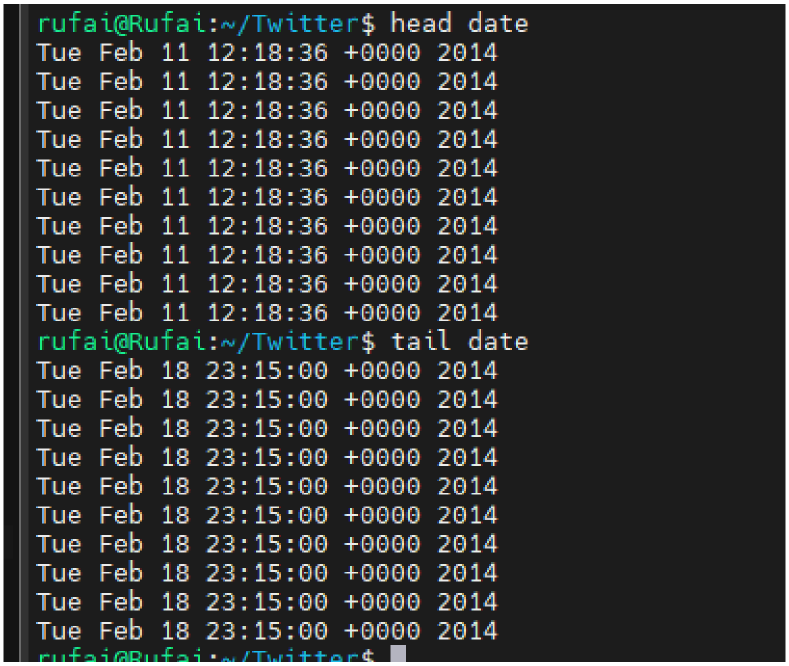
   To extract the date range from the twitter data, the awk command was also used but this time, the output was redirected to a date file.

   awk -F'\t' '{print$3}' Twitter_Data > date

   ```
   rufai@Rufai:~/Twitter$ awk -F"\t" '{print$3}' Twitter_Data > date
   ```

   The top and last 10 content of this date file was examined to get an idea of the tweet date range. A sample of these two sets of information are shown below:

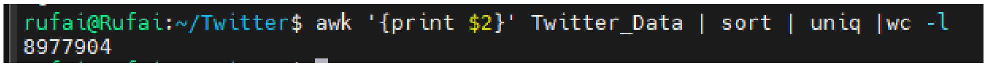


```
rufai@Rufai:~/Twitter$ head date
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
rufai@Rufai:~/Twitter$ tail date
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
rufai@Rufai:~/Twitter$
```

Based on this, we can conclude that the tweet date ranges from **Tue, Feb 11 12:18:36 +0000 2014** to **Tue, Feb 18 23:15:00 +0000 2014.**

7. How many unique users are there?
   The number of unique users in the file was computed using the shell command:

   awk '{print $2}' Twitter_Data | sort | uniq |wc -l

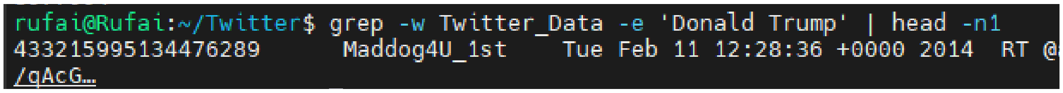


```
rufai@Rufai:~/Twitter$ awk '{print $2}' Twitter_Data | sort | uniq |wc -l
8977904
```

   From this output we can conclude that there are **8977904 unique users.**

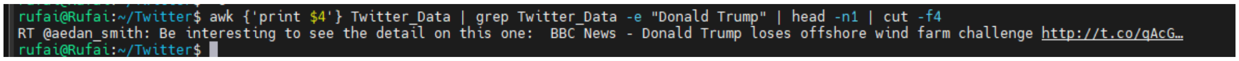
8. When was the first mention in the file of Donald Trump and what was the tweet?
   To extract the first mention of Donald Trump in the twitter file, the shell command shown below was used.

   grep -w Twitter_Data -e 'Donald Trump' | head -n1



```
rufai@Rufai:~/Twitter$ grep -w Twitter_Data -e 'Donald Trump' | head -n1
433215995134476289      Maddog4U_1st    Tue Feb 11 12:28:36 +0000 2014  RT @
/qAcG…
```

   The tweet that first mention Donald Trump was:



```
rufai@Rufai:~/Twitter$ awk {'print $4'} Twitter_Data | grep Twitter_Data -e "Donald Trump" | head -n1 | cut -f4
RT @aedan_smith: Be interesting to see the detail on this one:  BBC News - Donald Trump loses offshore wind farm challenge http://t.co/qAcG…
rufai@Rufai:~/Twitter$
```

   RT @aedan_smith: Be interesting to see the details on this one: BBC News - Donald Trump loses offshore wind farm challenge http://t.co/qAcG…

   The command used to extract this tweet is:

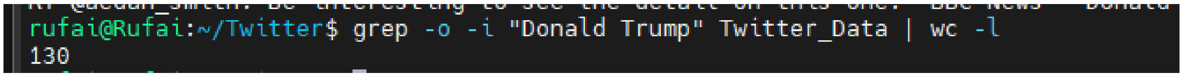
   awk {'print $4'} Twitter_Data | grep Twitter_Data -e "Donald Trump" | head -n1 | cut -f4

9. How many times has Donald Trump been mentioned? What about Barack Obama? Hillary Clinton?

   The respective shell commands shown below were used to extract the number of times each aforementioned string was found in the twitter file.

   I. Donald Trump: A total of 130 references were found as shown in the graphical display below.

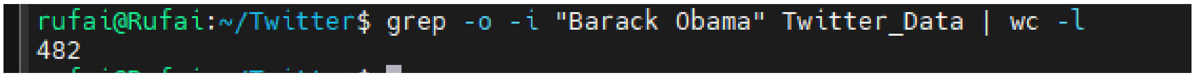
   grep -o -i Twitter_Data -o -i "Donald Trump" | wc -l



```
rufai@Rufai:~/Twitter$ grep -o -i "Donald Trump" Twitter_Data | wc -l
130
```

   II. Barack Obama: A total of 482 references were found as shown in the graphical display below.

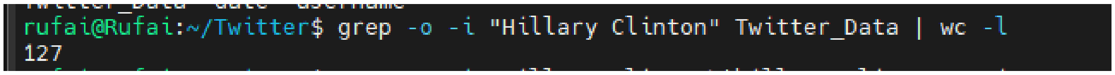
   grep -o -i Twitter_Data -o -i "Barack Obama" | wc -l



```
rufai@Rufai:~/Twitter$ grep -o -i "Barack Obama" Twitter_Data | wc -l
482
```

   III. Hillary Clinton: A total of 127 references were found as shown in the graphical display below.

   grep -o -i Twitter_Data -o -i "Hillary Clinton" | wc -l



```
rufai@Rufai:~/Twitter$ grep -o -i "Hillary Clinton" Twitter_Data | wc -l
127
```

10. Do you think we have captured all the references to Donald Trump, Barack Obama, and Hilary Clinton? What other strings might we need to try? What problems might we face?

    No, we haven't captured all the references to the aforementioned names in the field. Largely because of different sentence cases, incomplete and incorrect spellings in the file. For instance, there is still a reference to Barack Obama using ***Barack obam***, ***Barack Obam,*** for ***Hillary Clinton*** we still have **Hilary Clinton**. **Donald trump, and donald trump.** There are also possibilities for people to make tweets about these people using their nicknames, which might be specific to certain regions or groups of people.