DAT 560M Final Presentation Report

Group: Excel Fan Club
Shuyun Wan-489555
Wei Xu-491987
Lingjia Wang-491854

Description of Data

The data set we chose for the final presentation is called “Gender predict from email”. It was collected by Kaggle user Maksim Drobchak in 2020. The set contains a database of gender-marked email addresses with a total of over 100 million email addresses from around the world. Out of performance concerns, we selected a subset of the database with a size of over 1.5G to balance the data variety and server availability. The subset of the data contains 72,525,659 records and each record has two features: Email Address and Gender. Email Address column contains real email but the part of the address after @ is hidden to protect privacy. The data we downloaded from Kaggle.com is a CSV file, thus it should be classified as semi-structured data.

Problem Statement

We would like to discover whether male and female have different preferences on email addresses. Our assumption is that male and female have different patterns when it comes to naming an email address and we want to use the data set we chose to verify our assumption. We are going to test the following patterns:

- Which gender tends to use signs (-, _,.) in their email address more?
- What’s the ratio of each gender that includes numbers in their email address?
- What about the ratio of each gender that starts their email address with a number?
- Among those that have numbers in the email address, how many of them end the email with a number?
- Which gender is more likely to have a longer email address? What about a shorter email address?

By discovering whether there’s a correlation between each of the factors and gender, we are hoping to find the factors that significantly correlate to the gender so that we will be able to make some preliminary predictions on the user’s gender given his or her email address.

Why is this big data?

The reason why we chose this big data is that the data is big enough as it contains more than 70,000,000 rows so that it has sufficient records for us to come up with some reliable data analysis. Furthermore, the features are clearly defined and easy to interpret. The only two columns in the data are Email Address and Gender. So, the result of the analysis can be both insightful and readable.

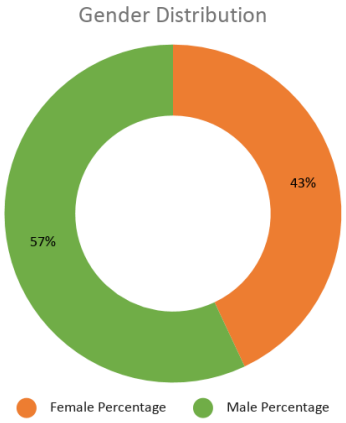
Methods & Results

Gender Distribution

We decided to use Hadoop MapReduce to analyze the big data. Our first step is to map each email address that has removed the part after "@" with the corresponding gender and omit the records that have gender values other than "M" and "F". Then, get the total number of each gender and the percentage of them. The number of females is 31709208, which takes 43% of the total number. And the number of males is 41829413, which takes 57% of the total number. The result is shown in the chart below. According to the result, the gender distribution of our data is relatively even. Therefore, it is feasible for us to continue our data analysis.

Table 1 - Number of Each Gender

Gender	Total Number	Percentage
Female	31,709,208	43%
Male	41,829,413	57%



Plot 1 - Gender Distribution

Which gender tends to use signs in their email address more?

We are now interested in whether there are some patterns in email addresses that are preferred by a certain gender. The first problem we raise is which gender tends to use signs more including dash line, underline and dot in the email address. In order to calculate the difference, we made some adjustments to the reduced file. We first created two lists with three

items to count the number of emails that used the designated signs for each gender [0,0,0] and also defined a list of signs [“_”, “-”, “.”]. For each email address, we ran a for loop to see if it contains any of the signs in the sign list. If it does, we classify it by its corresponding gender and add 1 to the count for that gender. As a result, the percentage of females using signs is 43%, which is 4% larger than the percentage of males using signs.

Table 2 - Percentage of Using Signs for Each Gender

Gender	Using Signs	Not Using
Female	43%	57%
Male	39%	61%

Also, according to a more specific result, the percentage of females using “_”, “-”, “.” is 19%, 6%, 17% separately, while the percentage of males using “_”, “-”, “.” is 20%, 5%, 14% separately. We can find that both females and males like to use “-” and “.” more often than “-”, whereas “-” and “.” are more preferred by females than by males.

Table 3 - Percentage of Using Each Sign for Each Gender

Gender	"_"	"-"	"."
Female	19.47%	6.09%	17.44%
Male	19.65%	5.21%	14.18%

What’s the ratio of each gender that includes numbers in their email address?

In this problem, we first defined a list in the reduce file which contains 10 numbers from 0 to 9. Then, we ran a for loop to check if the email address has any of the numbers in it. If it does, we add 1 for the count of the corresponding gender. The result shows that although the percentage of males using numbers is slightly larger than that of females, there’s no significant difference in male and female’s likelihood of using numbers in email addresses.

Table 4 - Percentage of Including Numbers for Each Gender

Gender	Using Numbers	Not Using
Female	44%	56%
Male	45%	55%

What about the ratio of each gender that starts their email address with a number?

We obtained the number of emails that contain numbers for each gender. In this problem, we made a few adjustments to the previous reduce file to add a new factor “head” which is the first character of the email address, and then run a boolean check to see if that head character is a number or not. If so, add 1 to the corresponding gender count. From what we got from the

analysis, less than 0.5% of the email addresses start with a number and the ratio of that for each gender is close to the overall percentage with female higher by an insignificant advantage.

Table 5 - Percentage of Starting with a Number for Each Gender

Gender	Using Numbers	Not Using
Female Start	2%	98%
Male Start	1%	99%

Among those that have numbers in the email address, how many of them end the email with a number?

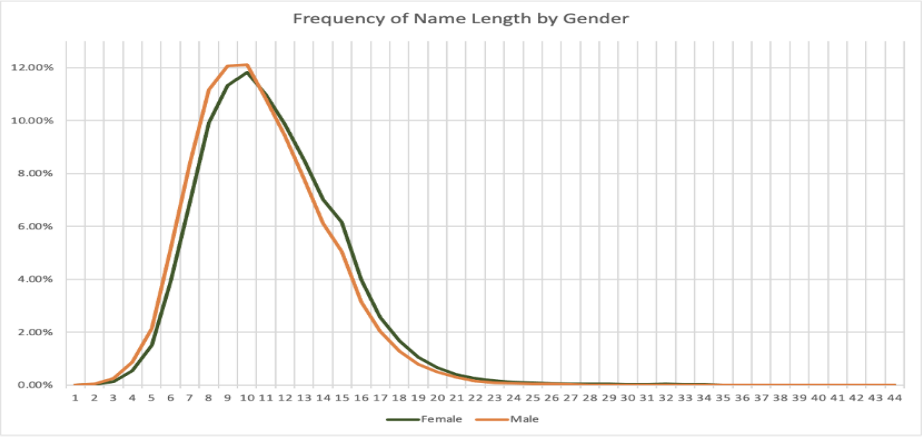
Similar to the previous question, we added a factor "tail" representing the last character of the email address and ran the boolean check. This time, the overall percentage of emails ending with number and the two ratios for each gender are almost identical, meaning it is indifferent for male and female to end the email with a number.

Table 6 - Percentage of Starting with a Number for Each Gender

Gender	Using Numbers	Not Using
Female End	41%	59%
Male End	42%	58%

Which gender is more likely to have a longer email address? What about a shorter email address?

For this question, we created two dictionaries in the reduce file where the keys are the length of the email addresses and the values are the count of the length for each gender. The shortest email address has only one character while the longest contains 44. In order to better reflect the difference of email address length between male and female, we created a graph to show the comparison with more readability.



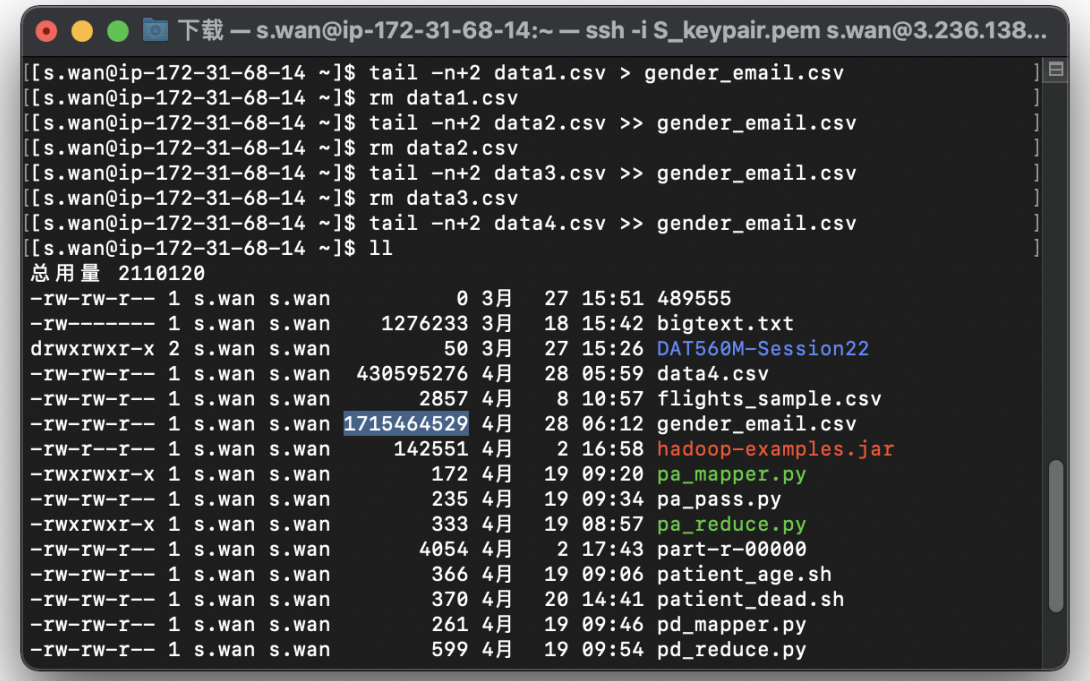
Plot 2 - Frequency of Name Length

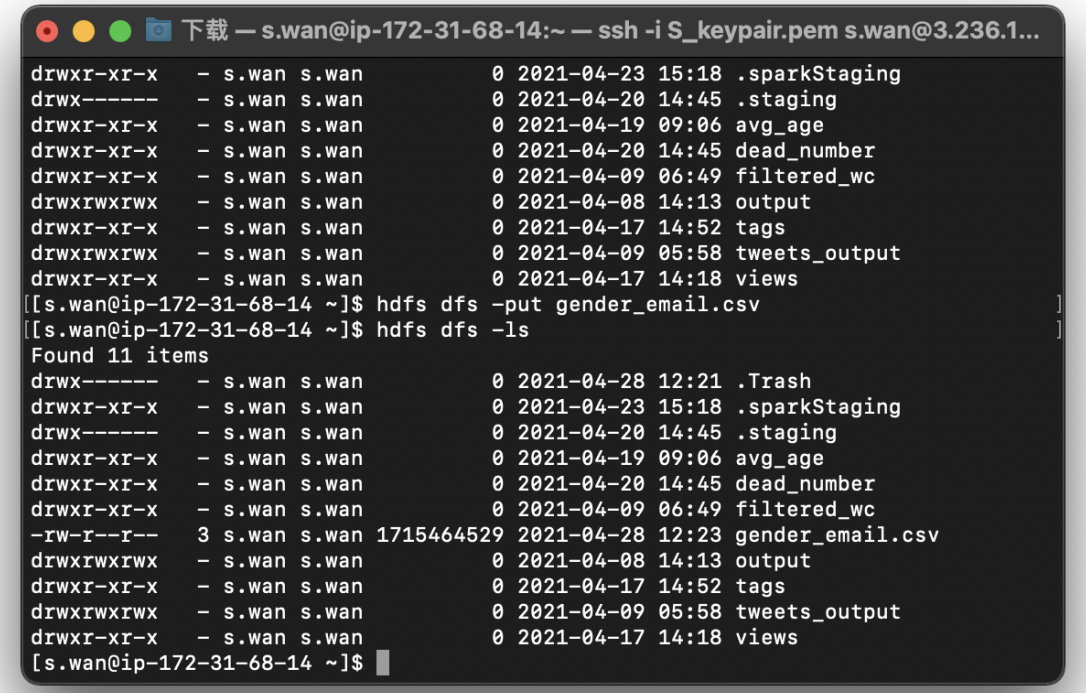
Conclusion

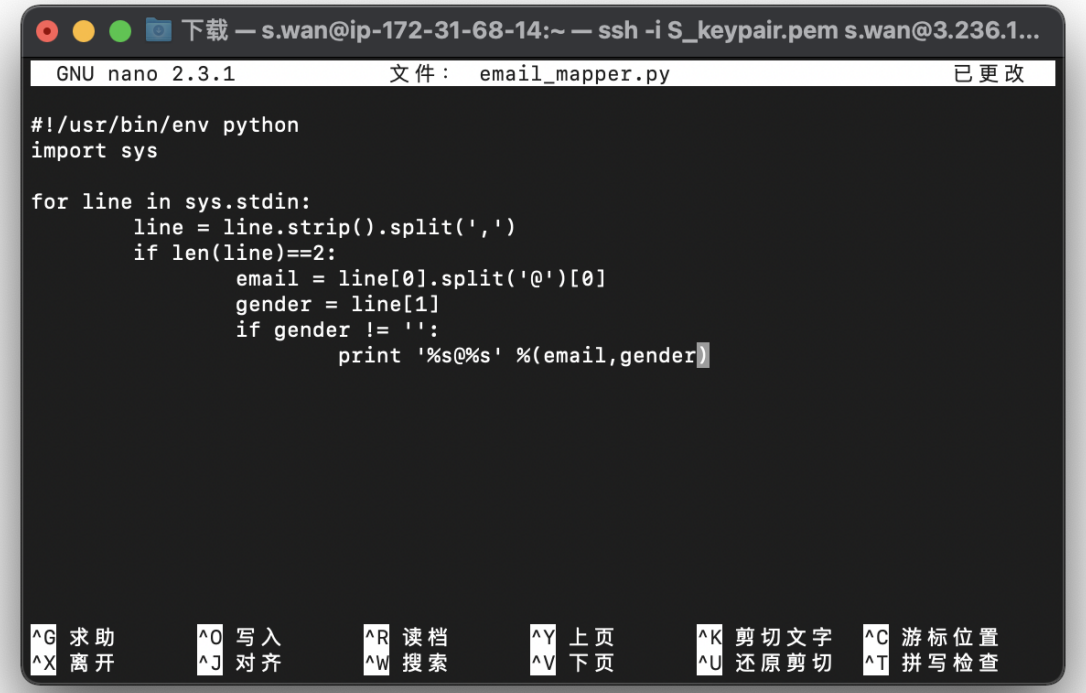
From our discoveries, we find that compared to male, female are more likely to include signs in the email address. Furthermore, among the three signs, the use of dot has the biggest difference between the two genders. Male and female are indifferent on the use of numbers in email addresses whether the number shows up in the beginning or at the end. Lastly, it seems male and female have different preferences on the length of email address. When the length is below 10, there's more male than female. As the length goes beyond 10, the email addresses are more likely to be owned by female.

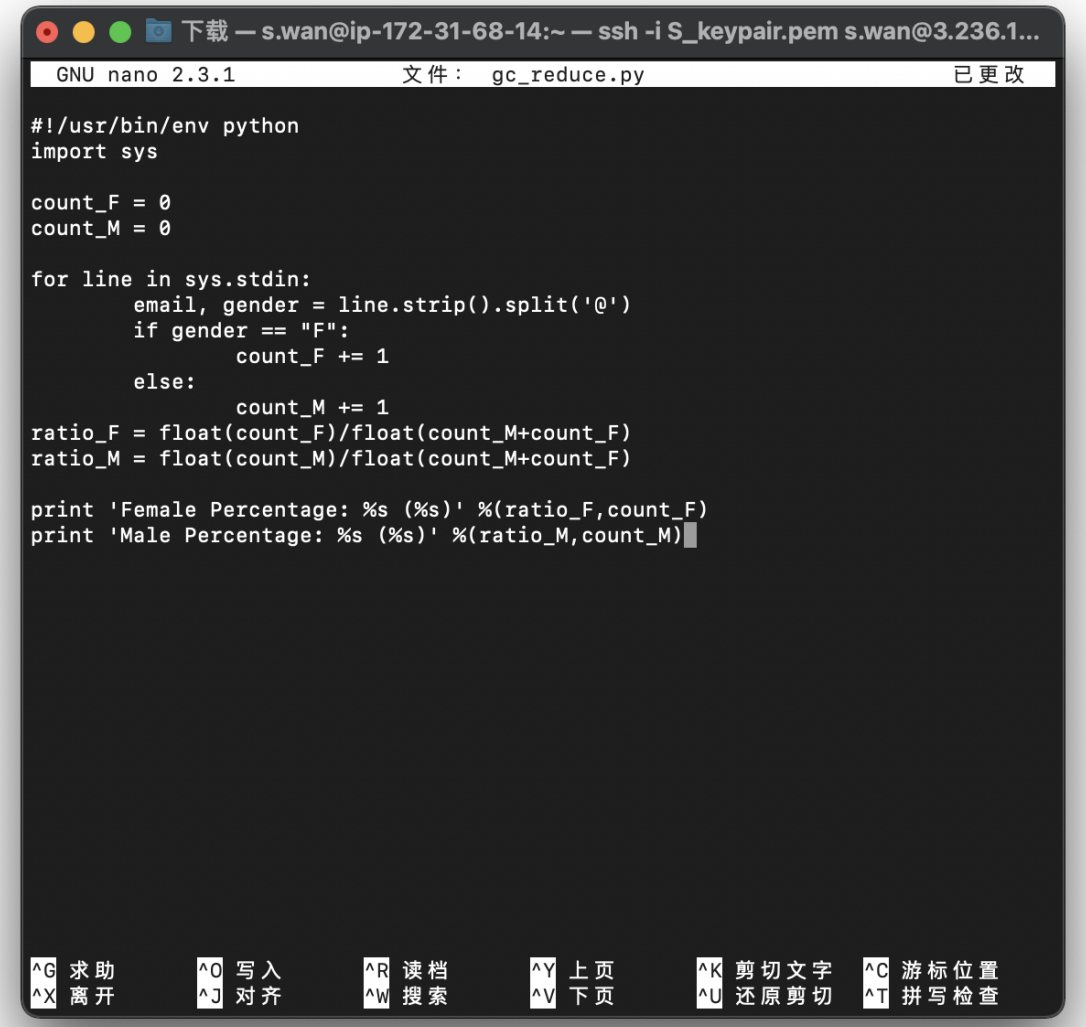
Appendix

Code




下载 — s.wan@ip-172-31-68-14:~ — ssh -i S_keypair.pem s.wan@3.236.1...
drwxr-xr-x   - s.wan s.wan   0 2021-04-23 15:18 .sparkStaging
drwx------   - s.wan s.wan   0 2021-04-20 14:45 .staging
drwxr-xr-x   - s.wan s.wan   0 2021-04-19 09:06 avg_age
drwxr-xr-x   - s.wan s.wan   0 2021-04-20 14:45 dead_number
drwxr-xr-x   - s.wan s.wan   0 2021-04-09 06:49 filtered_wc
drwxrwxrwx   - s.wan s.wan   0 2021-04-08 14:13 output
drwxr-xr-x   - s.wan s.wan   0 2021-04-17 14:52 tags
drwxrwxrwx   - s.wan s.wan   0 2021-04-09 05:58 tweets_output
drwxr-xr-x   - s.wan s.wan   0 2021-04-17 14:18 views
[s.wan@ip-172-31-68-14 ~]$ hdfs dfs -put gender_email.csv
[s.wan@ip-172-31-68-14 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - s.wan s.wan   0 2021-04-28 12:21 .Trash
drwxr-xr-x   - s.wan s.wan   0 2021-04-23 15:18 .sparkStaging
drwx------   - s.wan s.wan   0 2021-04-20 14:45 .staging
drwxr-xr-x   - s.wan s.wan   0 2021-04-19 09:06 avg_age
drwxr-xr-x   - s.wan s.wan   0 2021-04-20 14:45 dead_number
drwxr-xr-x   - s.wan s.wan   0 2021-04-09 06:49 filtered_wc
-rw-r--r--   3 s.wan s.wan 1715464529 2021-04-28 12:23 gender_email.csv
drwxrwxrwx   - s.wan s.wan   0 2021-04-08 14:13 output
drwxr-xr-x   - s.wan s.wan   0 2021-04-17 14:52 tags
drwxrwxrwx   - s.wan s.wan   0 2021-04-09 05:58 tweets_output
drwxr-xr-x   - s.wan s.wan   0 2021-04-17 14:18 views
[s.wan@ip-172-31-68-14 ~]$

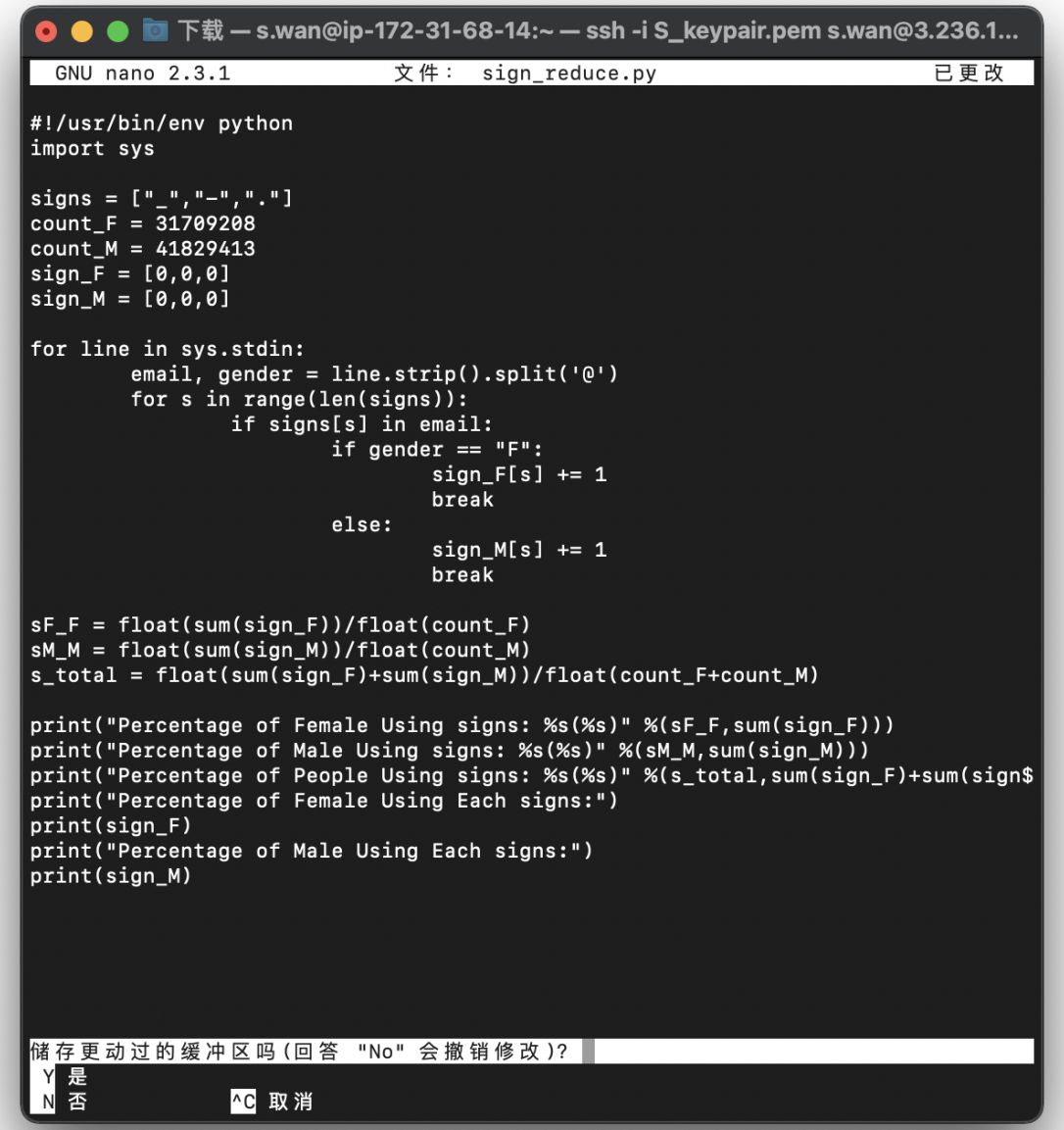

下载 — s.wan@ip-172-31-68-14:~ — ssh -i S_keypair.pem s.wan@3.236.1...
GNU nano 2.3.1 文件： email_mapper.py 已更改
#!/usr/bin/env python
import sys
for line in sys.stdin:
line = line.strip().split(',')
if len(line)==2:
email = line[0].split('@')[0]
gender = line[1]
if gender != '':
print '%s@%s' %(email,gender)
^G 求助 ^O 写入 ^R 读档 ^Y 上页 ^K 剪切文字 ^C 游标位置
^X 离开 ^J 对齐 ^W 搜索 ^V 下页 ^U 还原剪切 ^T 拼写检查



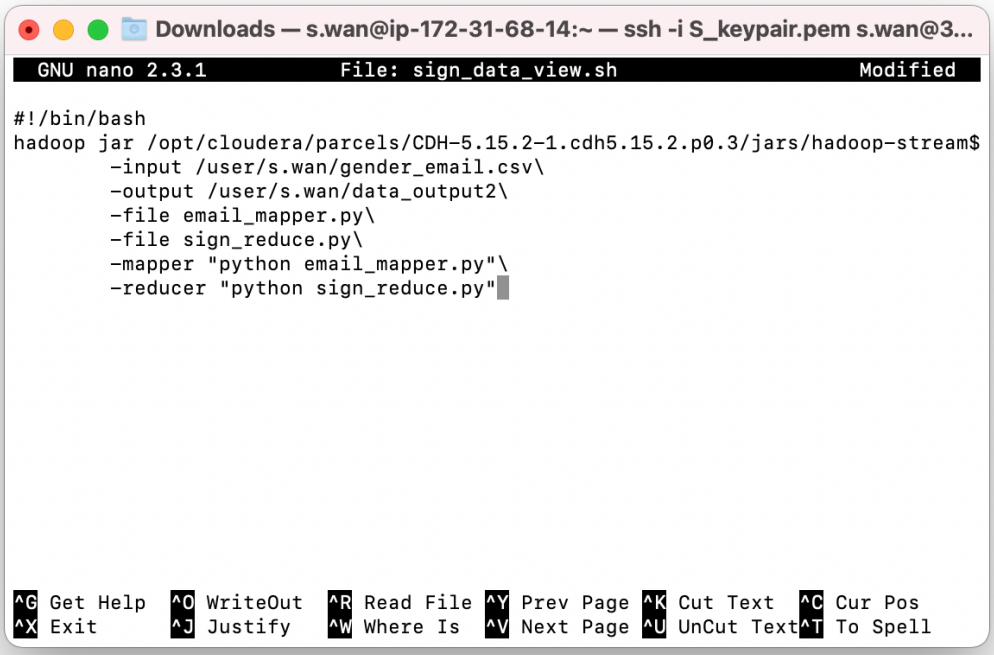

```
[s.wan@ip-172-31-68-14 ~]$ cat gender_email.csv | python email_mapper.py | pytho
n gc_reduce.py
Female Percentage: 0.431191223996 (31709208)
Male Percentage: 0.568808776004 (41829413)
```

```
  GNU nano 2.3.1            File: gc_data_view.sh

#!/bin/bash
hadoop jar /opt/cloudera/parcels/CDH-5.15.2-1.cdh5.15.2.p0.3/jars/hadoop-stream$
        -input /user/s.wan/gender_email.csv\
        -output /user/s.wan/data_output1\
        -file email_mapper.py\
        -file gc_reduce.py\
        -mapper "python email_mapper.py"\
        -reducer "python gc_reduce.py"









                          [ Read 8 lines ]
^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^V Next Page ^U UnCut Text^T To Spell
```

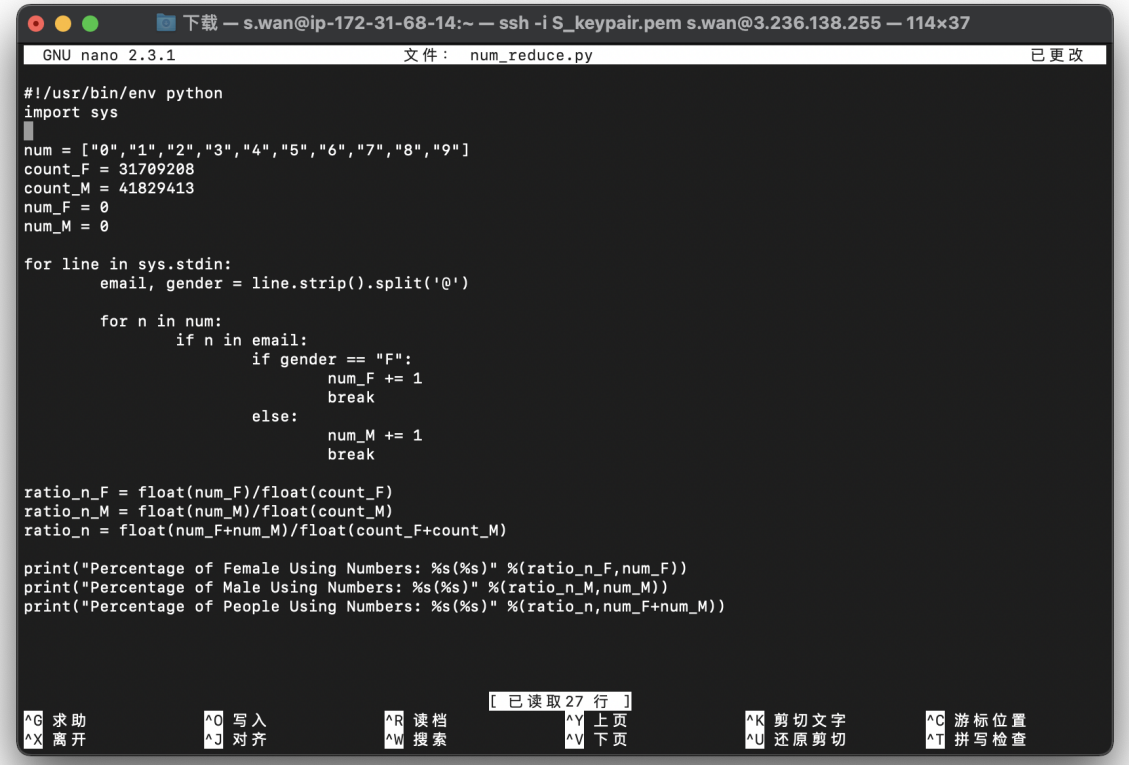
```
[s.wan@ip-172-31-68-14 ~]$ bash gc_data_view.sh
```

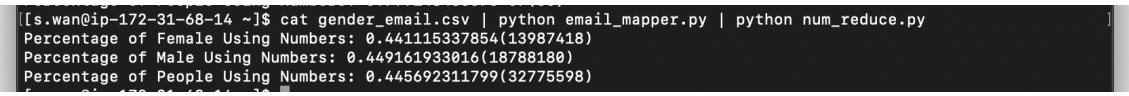


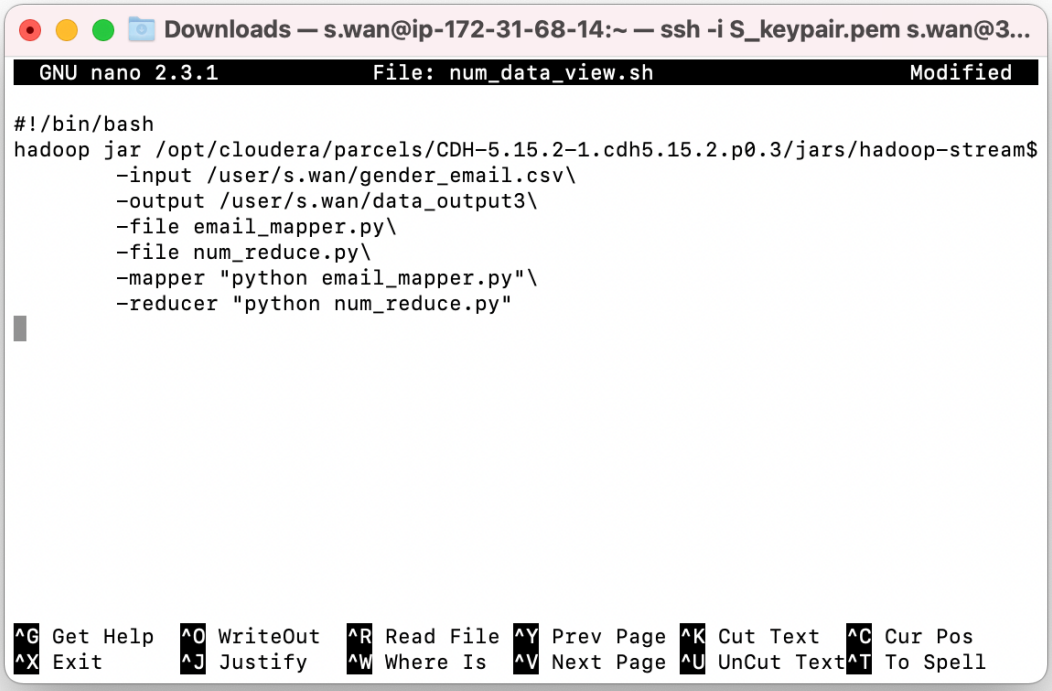

```
[s.wan@ip-172-31-68-14 ~]$ cat gender_email.csv | python email_mapper.py | pytho
n sign_reduce.py
Percentage of Female Using signs: 0.43003811385(13636168)
Percentage of Male Using signs: 0.390440477852(16331896)
Percentage of People Using signs: 0.407514630986(29968064)
Percentage of Female Using Each signs:
[6172242, 1932313, 5531613]
Percentage of Male Using Each signs:
[8220199, 2178783, 5932914]
```



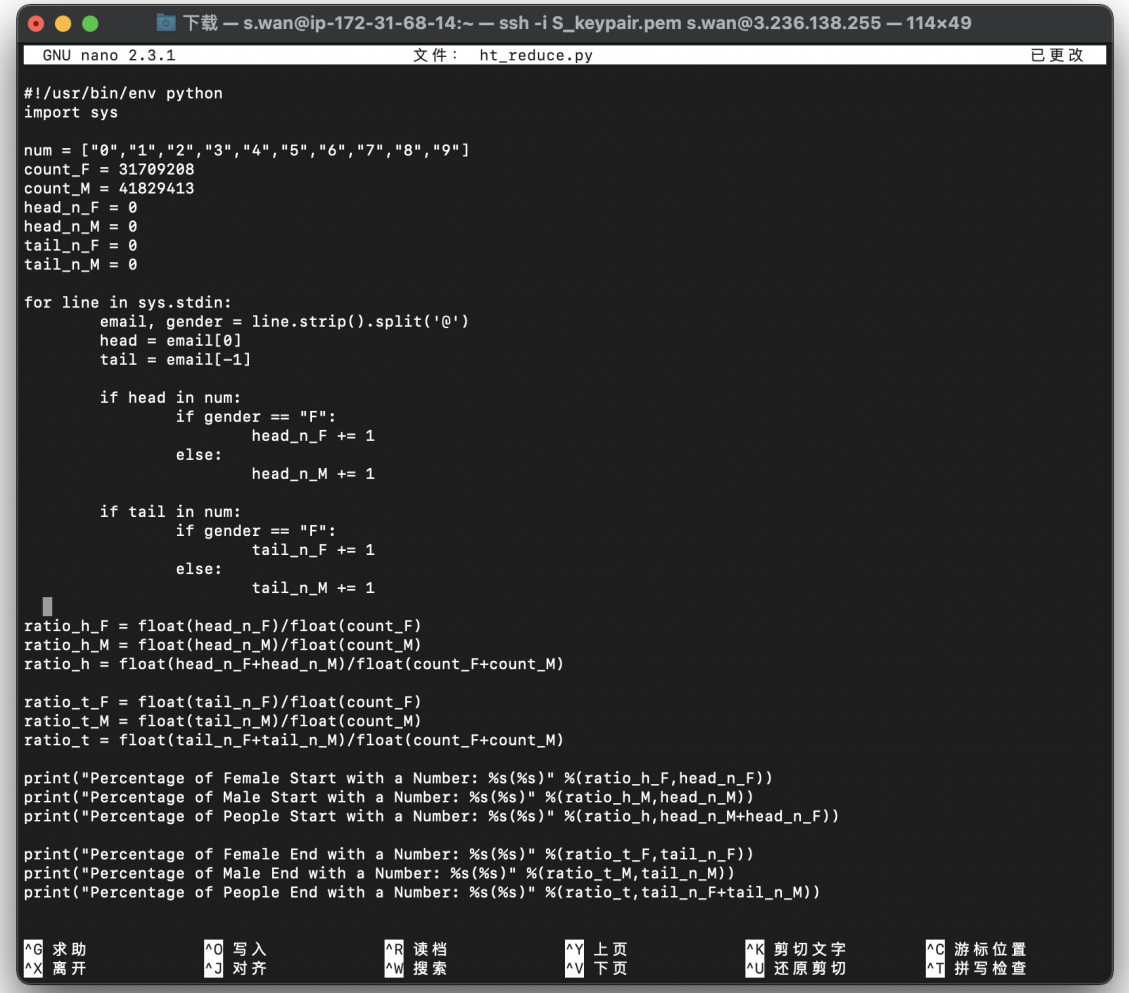

```
[s.wan@ip-172-31-68-14 ~]$ nano sign_data_view.sh
[s.wan@ip-172-31-68-14 ~]$ bash sign_data_view.sh
21/04/30 11:48:28 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJarJar: [email_mapper.py, sign_reduce.py] [/opt/cloudera/parcels/CDH-5.15
.2-1.cdh5.15.2.p0.3/jars/hadoop-streaming-2.6.0-cdh5.15.2.jar] /tmp/streamjob427
8161499646218623.jar tmpDir=null
21/04/30 11:48:30 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-68-14.ec2.internal/172.31.68.14:8032
21/04/30 11:48:30 INFO client.RMProxy: Connecting to ResourceManager at ip-172-3
1-68-14.ec2.internal/172.31.68.14:8032
21/04/30 11:48:31 INFO mapred.FileInputFormat: Total input paths to process : 1
21/04/30 11:48:31 INFO mapreduce.JobSubmitter: number of splits:4
21/04/30 11:48:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
19707237029_0023
21/04/30 11:48:31 INFO impl.YarnClientImpl: Submitted application application_16
19707237029_0023
21/04/30 11:48:31 INFO mapreduce.Job: The url to track the job: http://ip-172-31
-68-14.ec2.internal:8088/proxy/application_1619707237029_0023/
21/04/30 11:48:31 INFO mapreduce.Job: Running job: job_1619707237029_0023
```


下载 — s.wan@ip-172-31-68-14:~ — ssh -i S_keypair.pem s.wan@3.236.138.255 — 114×37
GNU nano 2.3.1 文件： num_reduce.py 已更改
#!/usr/bin/env python
import sys
num = ["0","1","2","3","4","5","6","7","8","9"]
count_F = 31709208
count_M = 41829413
num_F = 0
num_M = 0
for line in sys.stdin:
email, gender = line.strip().split('@')
for n in num:
if n in email:
if gender == "F":
num_F += 1
break
else:
num_M += 1
break
ratio_n_F = float(num_F)/float(count_F)
ratio_n_M = float(num_M)/float(count_M)
ratio_n = float(num_F+num_M)/float(count_F+count_M)
print("Percentage of Female Using Numbers: %s(%s)" %(ratio_n_F,num_F))
print("Percentage of Male Using Numbers: %s(%s)" %(ratio_n_M,num_M))
print("Percentage of People Using Numbers: %s(%s)" %(ratio_n,num_F+num_M))
[已读取27 行]
^G 求助 ^O 写入 ^R 读档 ^Y 上页 ^K 剪切文字 ^C 游标位置
^X 离开 ^J 对齐 ^W 搜索 ^V 下页 ^U 还原剪切 ^T 拼写检查


[s.wan@ip-172-31-68-14 ~]$ cat gender_email.csv | python email_mapper.py | python num_reduce.py
Percentage of Female Using Numbers: 0.441115337854(13987418)
Percentage of Male Using Numbers: 0.449161933016(18788180)
Percentage of People Using Numbers: 0.445692311799(32775598)

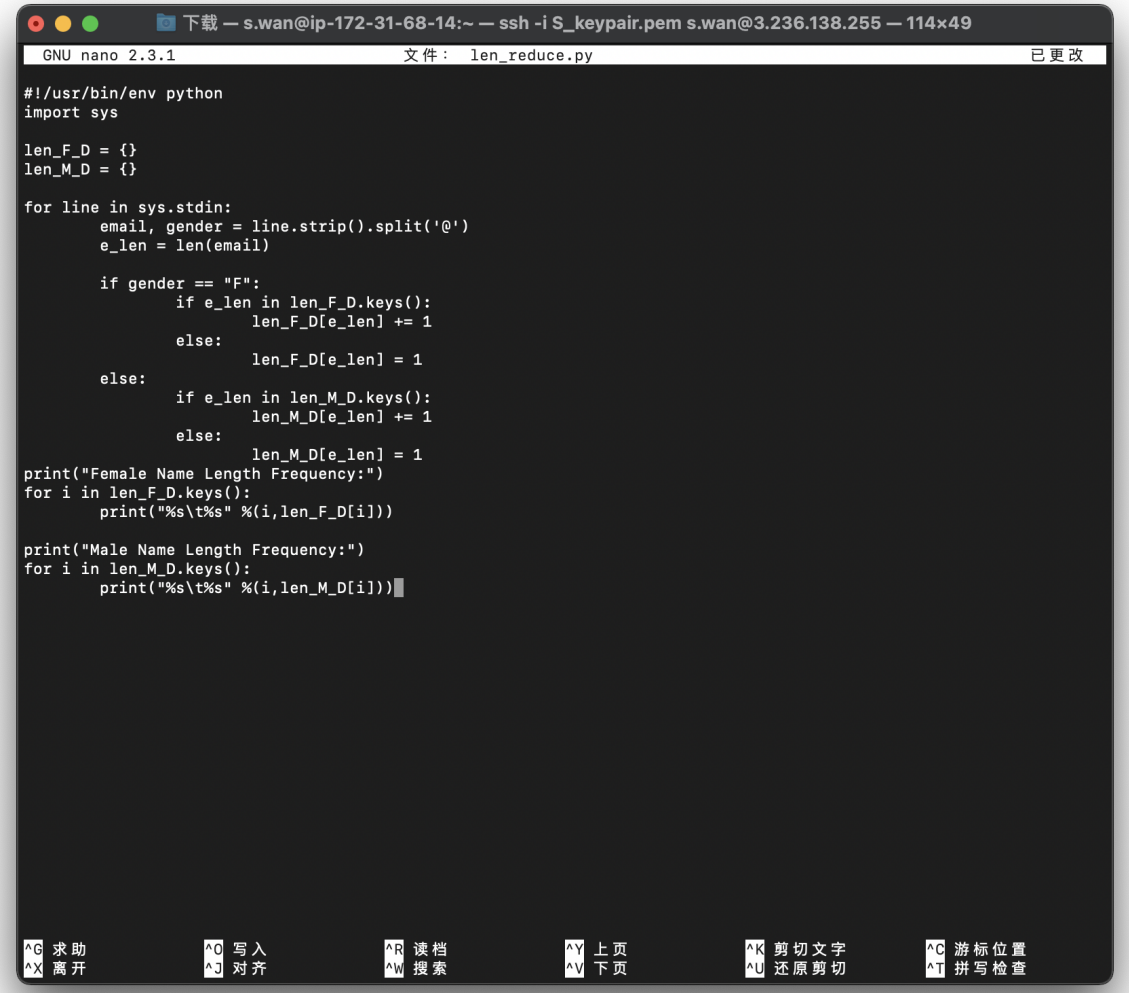


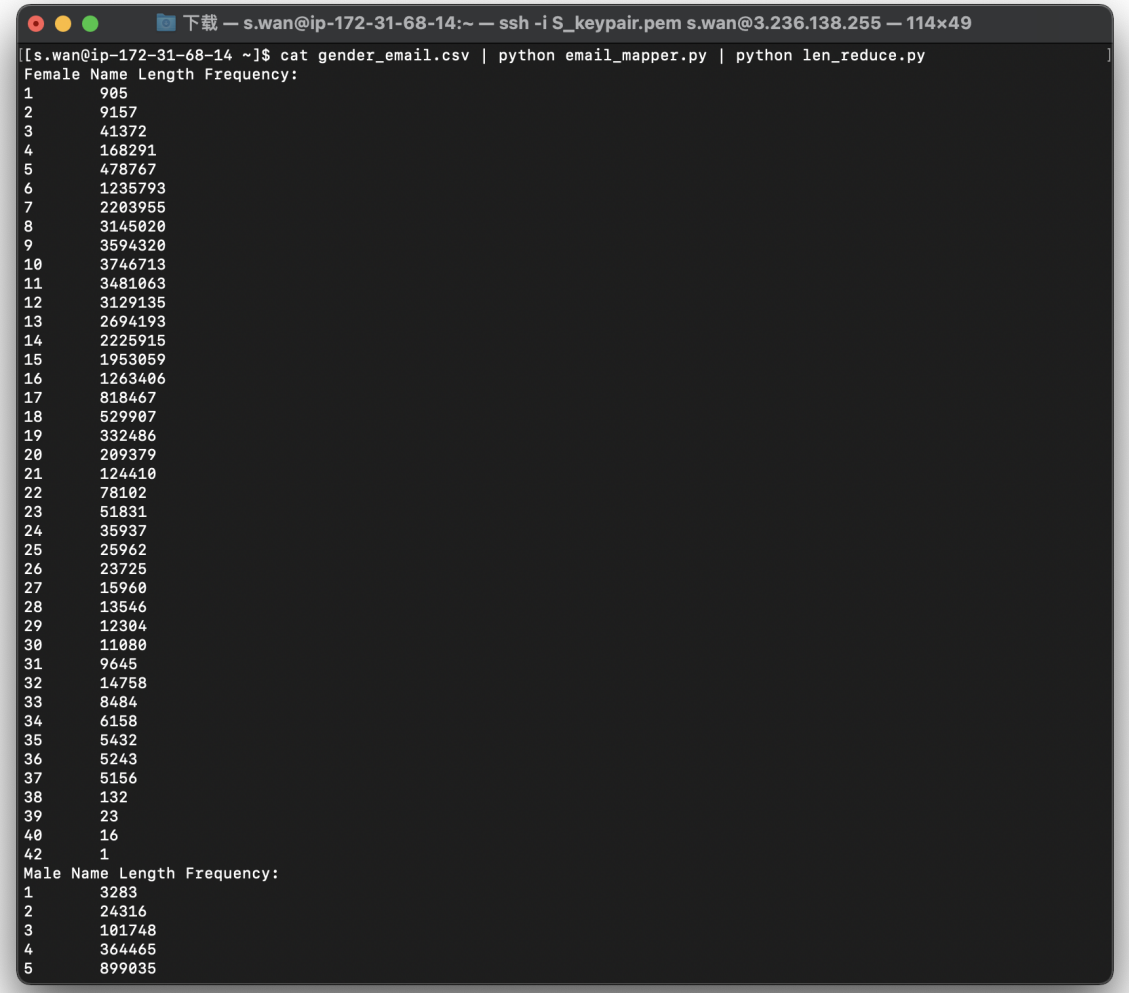

```
[s.wan@ip-172-31-68-14 ~]$ bash num_data_view.sh
```

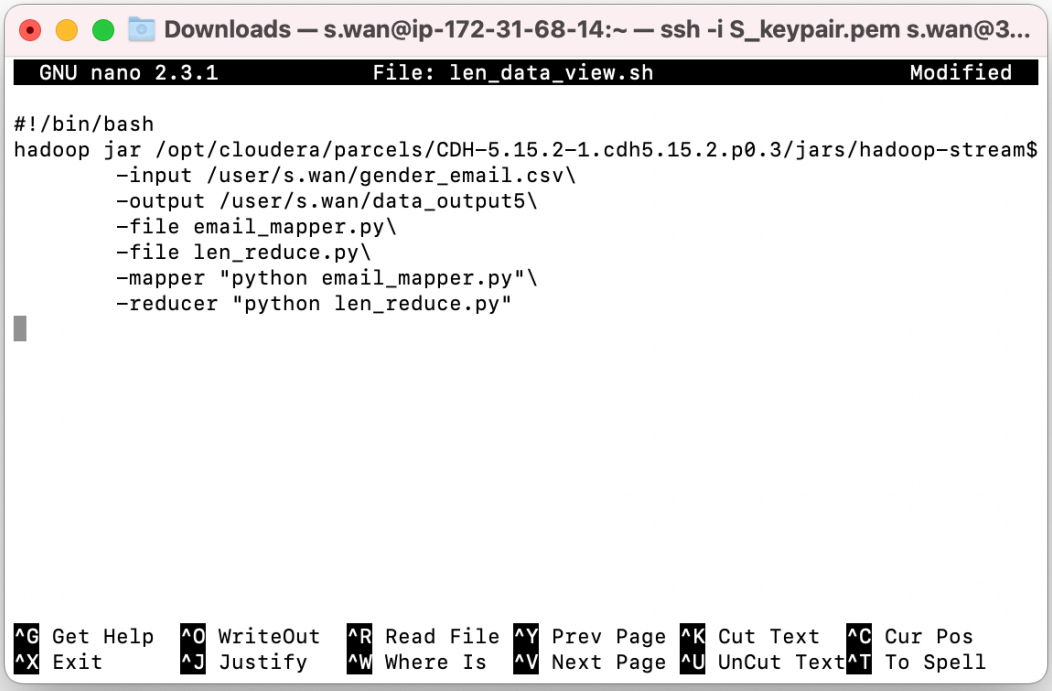



```
[s.wan@ip-172-31-68-14 ~]$ cat gender_email.csv | python email_mapper.py | python ht_reduce.py
Percentage of Female Start with a Number: 0.015837797021(502204)
Percentage of Male Start with a Number: 0.0143244898034(599185)
Percentage of People Start with a Number: 0.0149770145948(1101389)
Percentage of Female End with a Number: 0.412703811461(13086511)
Percentage of Male End with a Number: 0.416277703921(17412652)
Percentage of People End with a Number: 0.414736672857(30499163)
```


```
[s.wan@ip-172-31-68-14 ~]$ bash ht_data_view.sh
```


下载 — s.wan@ip-172-31-68-14:~ — ssh -i S_keypair.pem s.wan@3.236.138.255 — 114×49
GNU nano 2.3.1 文件： len_reduce.py 已更改
#!/usr/bin/env python
import sys
len_F_D = {}
len_M_D = {}
for line in sys.stdin:
email, gender = line.strip().split('@')
e_len = len(email)
if gender == "F":
if e_len in len_F_D.keys():
len_F_D[e_len] += 1
else:
len_F_D[e_len] = 1
else:
if e_len in len_M_D.keys():
len_M_D[e_len] += 1
else:
len_M_D[e_len] = 1
print("Female Name Length Frequency:")
for i in len_F_D.keys():
print("%s\t%s" %(i,len_F_D[i]))
print("Male Name Length Frequency:")
for i in len_M_D.keys():
print("%s\t%s" %(i,len_M_D[i]))
^G 求助 ^O 写入 ^R 读档 ^Y 上页 ^K 剪切文字 ^C 游标位置
^X 离开 ^J 对齐 ^W 搜索 ^V 下页 ^U 还原剪切 ^T 拼写检查


下载 — s.wan@ip-172-31-68-14:~ — ssh -i S_keypair.pem s.wan@3.236.138.255 — 114×49
[s.wan@ip-172-31-68-14 ~]$ cat gender_email.csv | python email_mapper.py | python len_reduce.py
Female Name Length Frequency:
1 905
2 9157
3 41372
4 168291
5 478767
6 1235793
7 2203955
8 3145020
9 3594320
10 3746713
11 3481063
12 3129135
13 2694193
14 2225915
15 1953059
16 1263406
17 818467
18 529907
19 332486
20 209379
21 124410
22 78102
23 51831
24 35937
25 25962
26 23725
27 15960
28 13546
29 12304
30 11080
31 9645
32 14758
33 8484
34 6158
35 5432
36 5243
37 5156
38 132
39 23
40 16
42 1
Male Name Length Frequency:
1 3283
2 24316
3 101748
4 364465
5 899035




```
[s.wan@ip-172-31-68-14 ~]$ bash len_data_view.sh
```