CSC555 Mining Big Data Assignment 1
Sarah Cummings

Part 1

A) Exercise 1.3.1 : Suppose there is a repository of ten million documents. What (to the nearest integer) is the IDF for a word that appears in (a) 40 documents (b) 10,000 documents?

B) Compute:

i. 2^10 = 2*2*2*2*2*2*2*2*2*2 = **1024**

ii. 2^20 = 2*2*2*2*2*2*2*2*2*2* 2*2*2*2*2*2*2*2*2*2 = **1048567**

iii. 101 MOD 3 -> 101 = 99 + 2. Since 99 is the closest multiple of three to 101, we have remainder 2 and 101 MOD 3 = **2**

iv. 37 MOD 13 -> 37 = 26 + 11. 37 MOD 13 = **11**

v. 42 MOD 7= **0**

C) Given vectors V1 = (3, 2, 1) and V2 = (6, 5, 4) and a 3x3 matrix M = [(0, 1, 2), (1, 2, 0), (3, 0, 1)], compute:

i. V1 + V2= (3+6, 2+5, 1+4) = **(9, 7, 5)**

ii. V1- V2= (3-6, 2-5, 1-4) = **(-3, -3, -3)**

iii. |V1| = sqrt (3^2 + 2^2 + 1^2) = sqrt(14) = **~3.7416**

iv. M * V1 = [(0*3 + 1*2 + 2*1), (1*3 + 2*2 + 0*1), (3*3 + 0*2 + 1*1)]

= **(4, 7, 10)**

v. M*M = **[(7, 2, 2), (2, 5, 2,), (3, 3, 7)]**

vi. M^4 = **[(59, 30, 32), (30, 35, 28), (48, 42, 610)]**

D) Suppose we are flipping a coin with Head (H) and Tail (T) sides. The coin is not balanced with 0.4 probability of H coming up (and 0.6 of T). Compute the probabilities of getting:

i. HTT = (0.4)(0.6)(0.6)= **0.144**

ii. TTH= (0.6)(0.6)(0.4)= **0.144**

iii. Either one head or three heads of three flips:

P(HTT)+ P(TTH) + P(THT) + P(HHH)=

(0.4)(0.6)(0.6) + (0.6)(0.6)(0.4) + (0.6)(0.4)(0.6) + (0.4)(0.4)(0.4) = 0.432 + 0.064 = **0.496**

iv. Exactly two tails: P(TTH) + P(THT) + P(HTT) = 3(0.6)(0.6)(0.4)= **0.432**

E) Consider a database schema consisting of two tables, Employee(ID, Name, Address, Position) and a table that stores employee certifications, Certificates(EID, CertName, Date). (ID) and (EID, CertName) are the primary keys for each respective table. EID is the foreign key referencing Employee(ID). Write SQL queries for:

i. Find all employees whose position is “Level-2 Manager”

```
SELECT Name
FROM Employee
WHERE Position = ‘Level-2 Manager’;
```

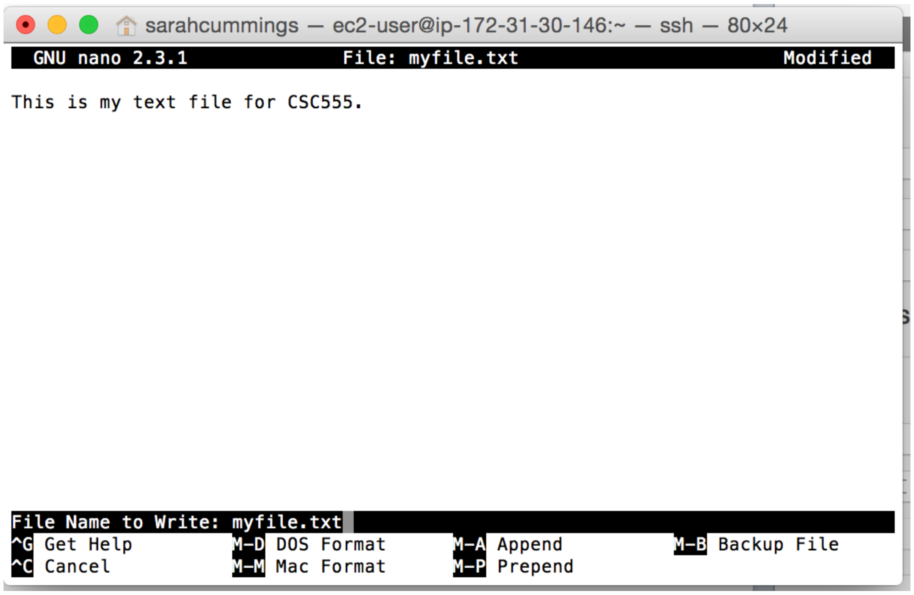
i. Find out how many different certifications (CertName) are stored in the database

```
SELECT Count( DISTINCT CertName) FROM Certificates;
```

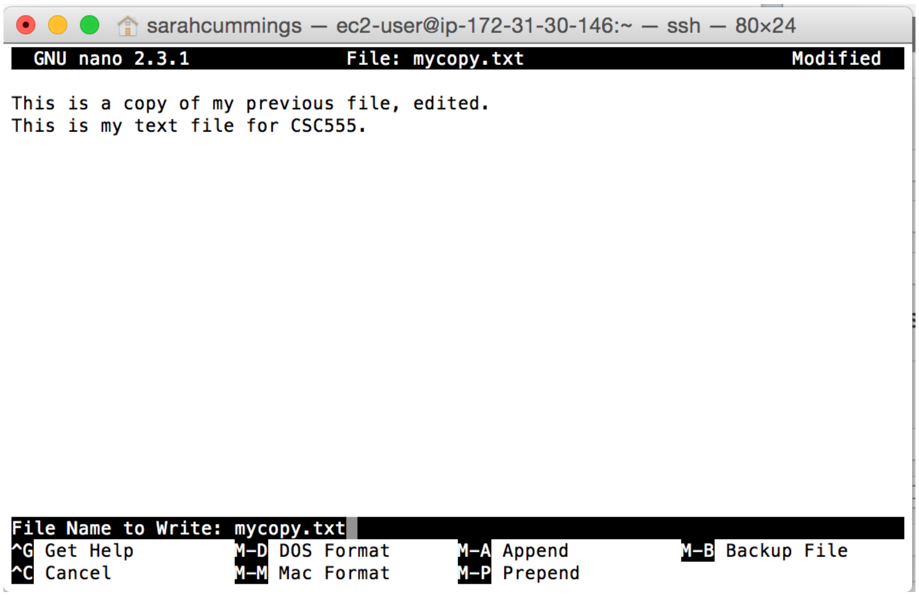
iii. Find all employees that have fewer than 2 certifications (note that this should include 0 and 1 certifications)

```
SELECT Name FROM Employee
WHERE (SELECT Count(*) FROM Certificates WHERE ID=EID) <=2;
```

Part 2

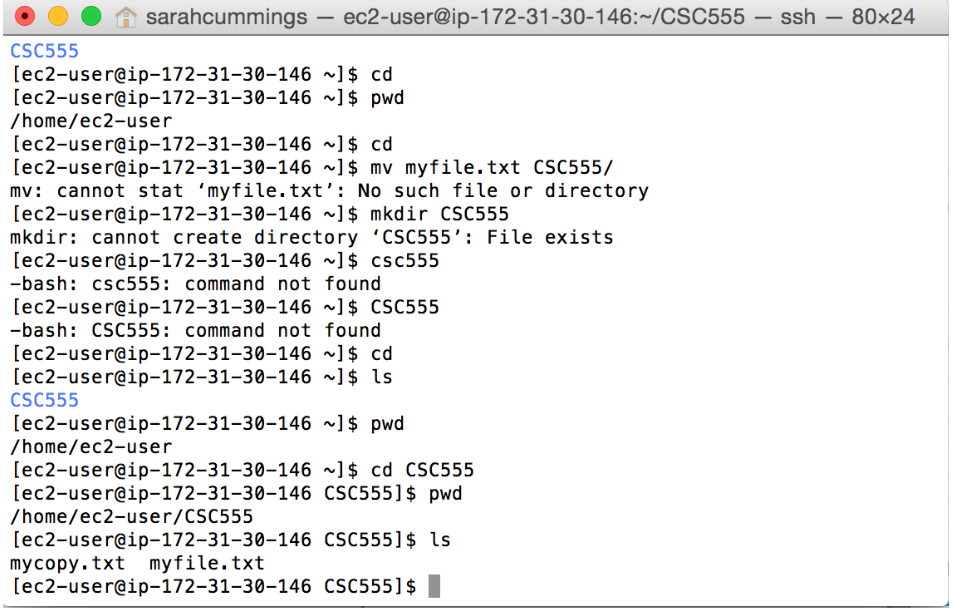


```
  GNU nano 2.3.1                File: myfile.txt                        Modified

This is my text file for CSC555.

File Name to Write: myfile.txt
^G Get Help          M-D DOS Format       M-A Append           M-B Backup File
^C Cancel            M-M Mac Format       M-P Prepend
```



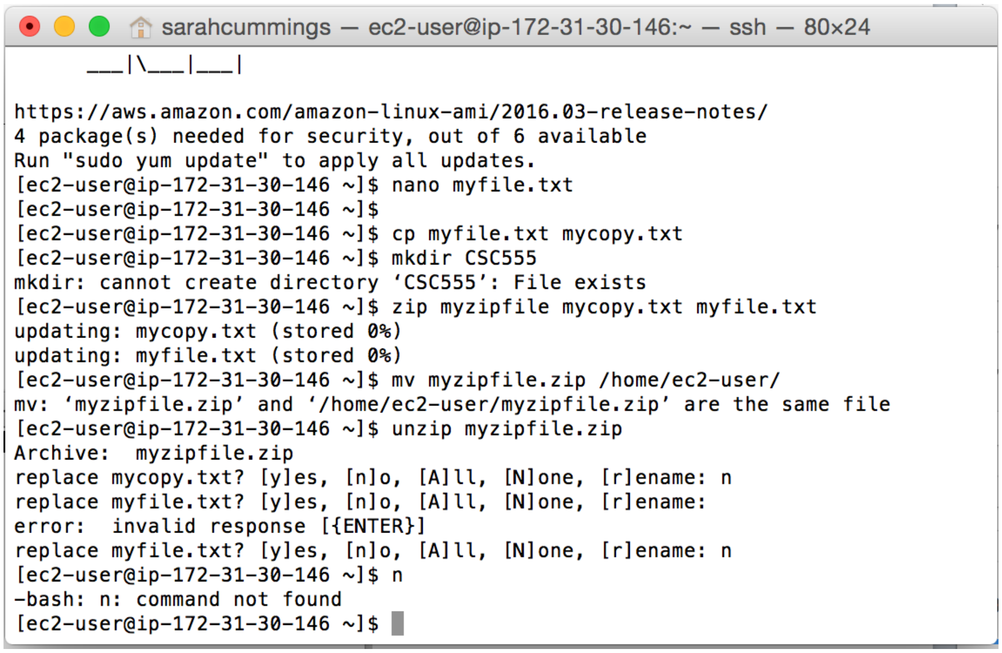
```
  GNU nano 2.3.1                File: mycopy.txt                        Modified

This is a copy of my previous file, edited.
This is my text file for CSC555.

File Name to Write: mycopy.txt
^G Get Help          M-D DOS Format       M-A Append           M-B Backup File
^C Cancel            M-M Mac Format       M-P Prepend
```

Directory contents— below you can see I’ve created the files


sarahcummings — ec2-user@ip-172-31-30-146:~/CSC555 — ssh — 80×24

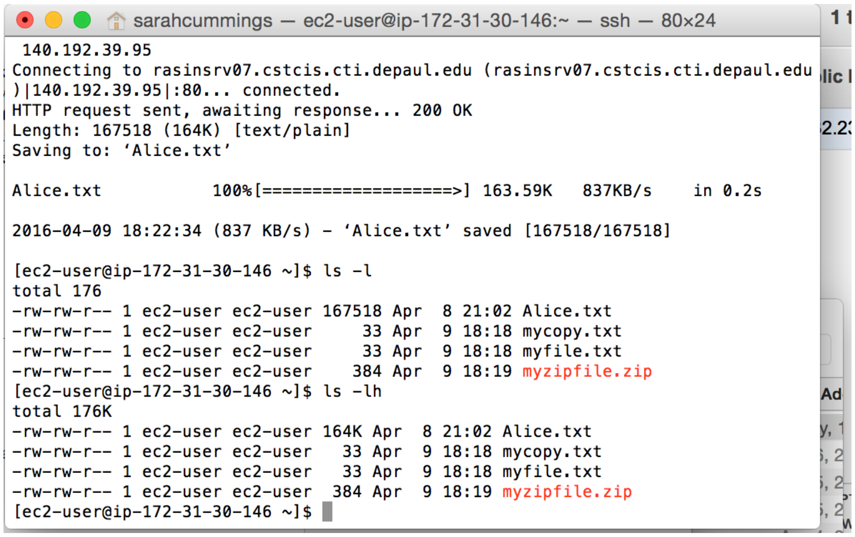
```
CSC555
[ec2-user@ip-172-31-30-146 ~]$ cd
[ec2-user@ip-172-31-30-146 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-30-146 ~]$ cd
[ec2-user@ip-172-31-30-146 ~]$ mv myfile.txt CSC555/
mv: cannot stat ‘myfile.txt’: No such file or directory
[ec2-user@ip-172-31-30-146 ~]$ mkdir CSC555
mkdir: cannot create directory ‘CSC555’: File exists
[ec2-user@ip-172-31-30-146 ~]$ csc555
-bash: csc555: command not found
[ec2-user@ip-172-31-30-146 ~]$ CSC555
-bash: CSC555: command not found
[ec2-user@ip-172-31-30-146 ~]$ cd
[ec2-user@ip-172-31-30-146 ~]$ ls
CSC555
[ec2-user@ip-172-31-30-146 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-30-146 ~]$ cd CSC555
[ec2-user@ip-172-31-30-146 CSC555]$ pwd
/home/ec2-user/CSC555
[ec2-user@ip-172-31-30-146 CSC555]$ ls
mycopy.txt  myfile.txt
[ec2-user@ip-172-31-30-146 CSC555]$ 
```

Screenshot after I did the $ unzip myzipfile.zip:



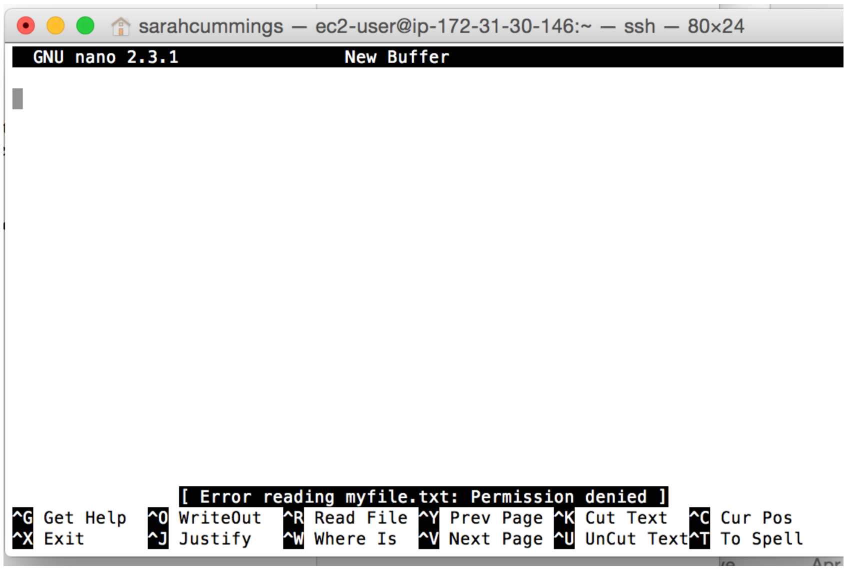
```
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2016.03-release-notes/
4 package(s) needed for security, out of 6 available
Run "sudo yum update" to apply all updates.
[ec2-user@ip-172-31-30-146 ~]$ nano myfile.txt
[ec2-user@ip-172-31-30-146 ~]$
[ec2-user@ip-172-31-30-146 ~]$ cp myfile.txt mycopy.txt
[ec2-user@ip-172-31-30-146 ~]$ mkdir CSC555
mkdir: cannot create directory ‘CSC555’: File exists
[ec2-user@ip-172-31-30-146 ~]$ zip myzipfile mycopy.txt myfile.txt
updating: mycopy.txt (stored 0%)
updating: myfile.txt (stored 0%)
[ec2-user@ip-172-31-30-146 ~]$ mv myzipfile.zip /home/ec2-user/
mv: ‘myzipfile.zip’ and ‘/home/ec2-user/myzipfile.zip’ are the same file
[ec2-user@ip-172-31-30-146 ~]$ unzip myzipfile.zip
Archive:  myzipfile.zip
replace mycopy.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace myfile.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename:
error:  invalid response [{ENTER}]
replace myfile.txt? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
[ec2-user@ip-172-31-30-146 ~]$ n
-bash: n: command not found
[ec2-user@ip-172-31-30-146 ~]$ 
```

After downloading the Alice in wonderland file, I found the size is 164k



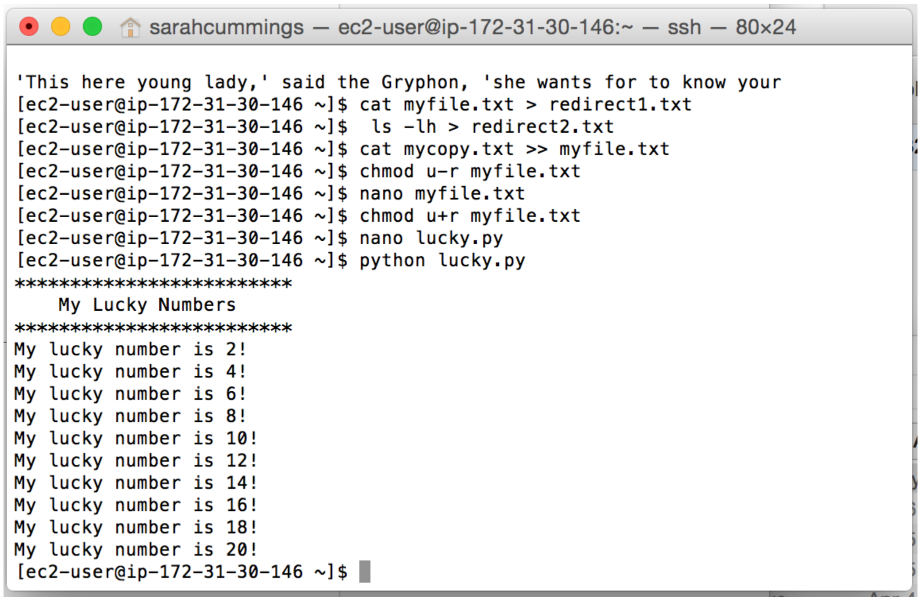
```
 140.192.39.95
Connecting to rasinsrv07.cstcis.cti.depaul.edu (rasinsrv07.cstcis.cti.depaul.edu
)|140.192.39.95|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 167518 (164K) [text/plain]
Saving to: ‘Alice.txt’

Alice.txt           100%[===================>] 163.59K   837KB/s    in 0.2s

2016-04-09 18:22:34 (837 KB/s) - ‘Alice.txt’ saved [167518/167518]

[ec2-user@ip-172-31-30-146 ~]$ ls -l
total 176
-rw-rw-r-- 1 ec2-user ec2-user 167518 Apr  8 21:02 Alice.txt
-rw-rw-r-- 1 ec2-user ec2-user     33 Apr  9 18:18 mycopy.txt
-rw-rw-r-- 1 ec2-user ec2-user     33 Apr  9 18:18 myfile.txt
-rw-rw-r-- 1 ec2-user ec2-user    384 Apr  9 18:19 myzipfile.zip
[ec2-user@ip-172-31-30-146 ~]$ ls -lh
total 176K
-rw-rw-r-- 1 ec2-user ec2-user 164K Apr  8 21:02 Alice.txt
-rw-rw-r-- 1 ec2-user ec2-user   33 Apr  9 18:18 mycopy.txt
-rw-rw-r-- 1 ec2-user ec2-user   33 Apr  9 18:18 myfile.txt
-rw-rw-r-- 1 ec2-user ec2-user  384 Apr  9 18:19 myzipfile.zip
[ec2-user@ip-172-31-30-146 ~]$
```

step 11 permission denied error:

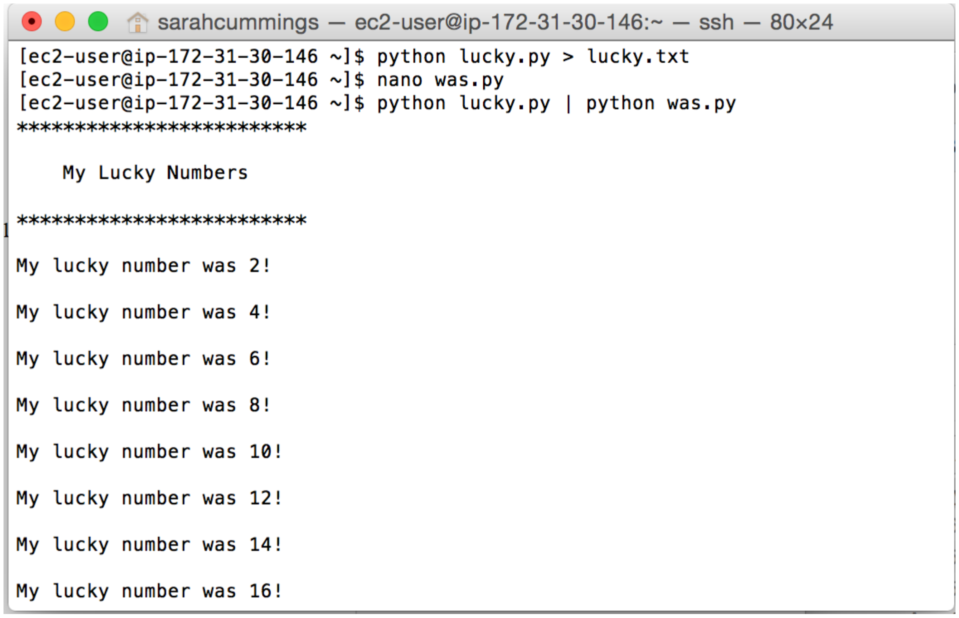


```
  GNU nano 2.3.1                 New Buffer



                 [ Error reading myfile.txt: Permission denied ]
^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
```

Proof I made it through the lucky number python part:

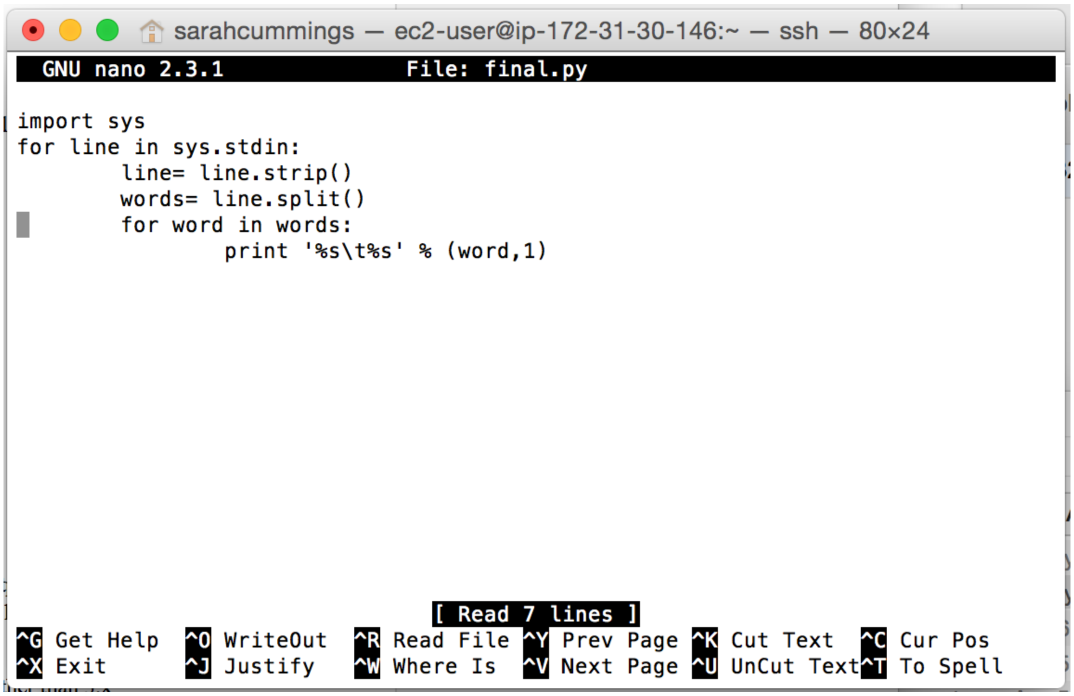


```
'This here young lady,' said the Gryphon, 'she wants for to know your
[ec2-user@ip-172-31-30-146 ~]$ cat myfile.txt > redirect1.txt
[ec2-user@ip-172-31-30-146 ~]$  ls -lh > redirect2.txt
[ec2-user@ip-172-31-30-146 ~]$ cat mycopy.txt >> myfile.txt
[ec2-user@ip-172-31-30-146 ~]$ chmod u-r myfile.txt
[ec2-user@ip-172-31-30-146 ~]$ nano myfile.txt
[ec2-user@ip-172-31-30-146 ~]$ chmod u+r myfile.txt
[ec2-user@ip-172-31-30-146 ~]$ nano lucky.py
[ec2-user@ip-172-31-30-146 ~]$ python lucky.py
*************************
    My Lucky Numbers
*************************
My lucky number is 2!
My lucky number is 4!
My lucky number is 6!
My lucky number is 8!
My lucky number is 10!
My lucky number is 12!
My lucky number is 14!
My lucky number is 16!
My lucky number is 18!
My lucky number is 20!
[ec2-user@ip-172-31-30-146 ~]$
```



```
[ec2-user@ip-172-31-30-146 ~]$ python lucky.py > lucky.txt
[ec2-user@ip-172-31-30-146 ~]$ nano was.py
[ec2-user@ip-172-31-30-146 ~]$ python lucky.py | python was.py
*************************

    My Lucky Numbers

*************************

My lucky number was 2!

My lucky number was 4!

My lucky number was 6!

My lucky number was 8!

My lucky number was 10!

My lucky number was 12!

My lucky number was 14!

My lucky number was 16!
```

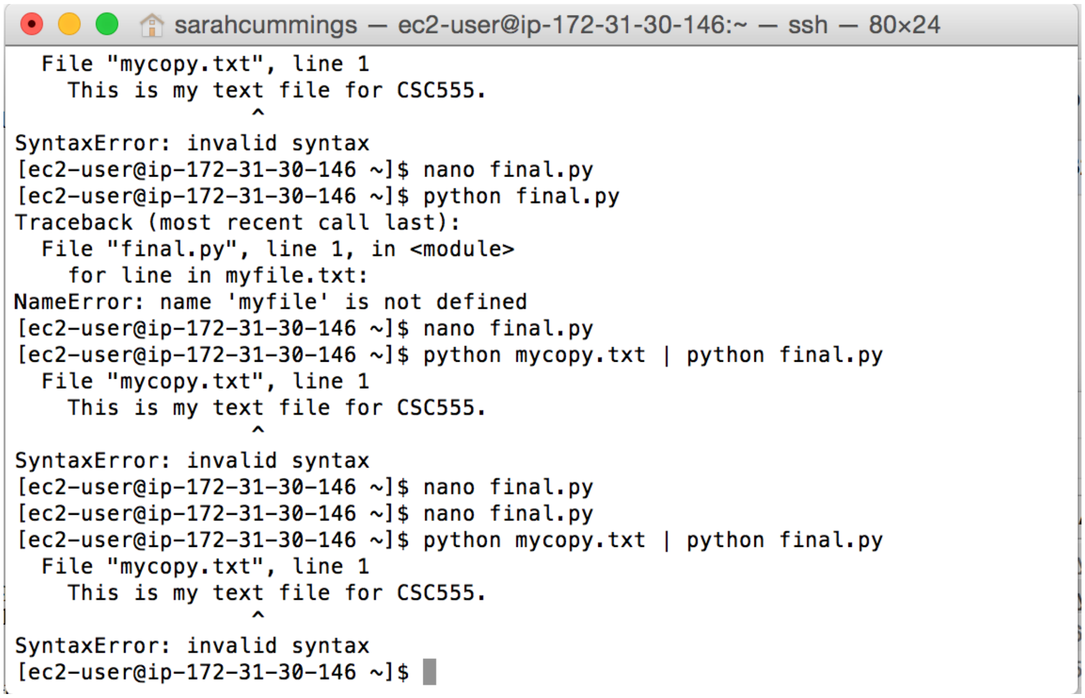
I can’t get the word counter to work but here’s proof that I tried:



sarahcummings — ec2-user@ip-172-31-30-146:~ — ssh — 80×24

```
  GNU nano 2.3.1                File: final.py

import sys
for line in sys.stdin:
        line= line.strip()
        words= line.split()
        for word in words:
                print '%s\t%s' % (word,1)
```

```
                                [ Read 7 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
```



sarahcummings — ec2-user@ip-172-31-30-146:~ — ssh — 80×24

```
  File "mycopy.txt", line 1
    This is my text file for CSC555.
                  ^
SyntaxError: invalid syntax
[ec2-user@ip-172-31-30-146 ~]$ nano final.py
[ec2-user@ip-172-31-30-146 ~]$ python final.py
Traceback (most recent call last):
  File "final.py", line 1, in <module>
    for line in myfile.txt:
NameError: name 'myfile' is not defined
[ec2-user@ip-172-31-30-146 ~]$ nano final.py
[ec2-user@ip-172-31-30-146 ~]$ python mycopy.txt | python final.py
  File "mycopy.txt", line 1
    This is my text file for CSC555.
                  ^
SyntaxError: invalid syntax
[ec2-user@ip-172-31-30-146 ~]$ nano final.py
[ec2-user@ip-172-31-30-146 ~]$ nano final.py
[ec2-user@ip-172-31-30-146 ~]$ python mycopy.txt | python final.py
  File "mycopy.txt", line 1
    This is my text file for CSC555.
                  ^
SyntaxError: invalid syntax
[ec2-user@ip-172-31-30-146 ~]$
```