7402 Assignment 1

Aing Ragunathan A00765949

Task 1:

Alice’s Adventures in Wonderland:

```
a :  9846         distribution:  0.0797828376954866
b :  1757         distribution:  0.014237095859330686
c :  3028         distribution:  0.024536099181589822
d :  5491         distribution:  0.04449396321205737
e :  15441        distribution:  0.125119520298193
f :  2385         distribution:  0.019325824487480756
g :  2948         distribution:  0.023887853496475164
h :  7915         distribution:  0.06413580747103152
i :  8669         distribution:  0.07024552305323718
j :  235          distribution:  0.0019042217000243091
k :  1291         distribution:  0.010461064743537802
l :  5227         distribution:  0.042354752451179
m :  2469         distribution:  0.020006482456851146
n :  8066         distribution:  0.06535937120168545
o :  9496         distribution:  0.07694676282310996
p :  1988         distribution:  0.016108905275099263
q :  223          distribution:  0.0018069848472571104
r :  6648         distribution:  0.05386921643302812
s :  7280         distribution:  0.05899035734543392
t :  12241        distribution:  0.09918969289360667
u :  3990         distribution:  0.03233125354509359
v :  972          distribution:  0.0078761850741431
w :  2956         distribution:  0.02395267806498663
x :  179          distribution:  0.0014504497204440483
y :  2589         distribution:  0.020978850984523133
z :  80  distribution:  0.0006482456851146585
total =  123410
total distribution =  1.0
```

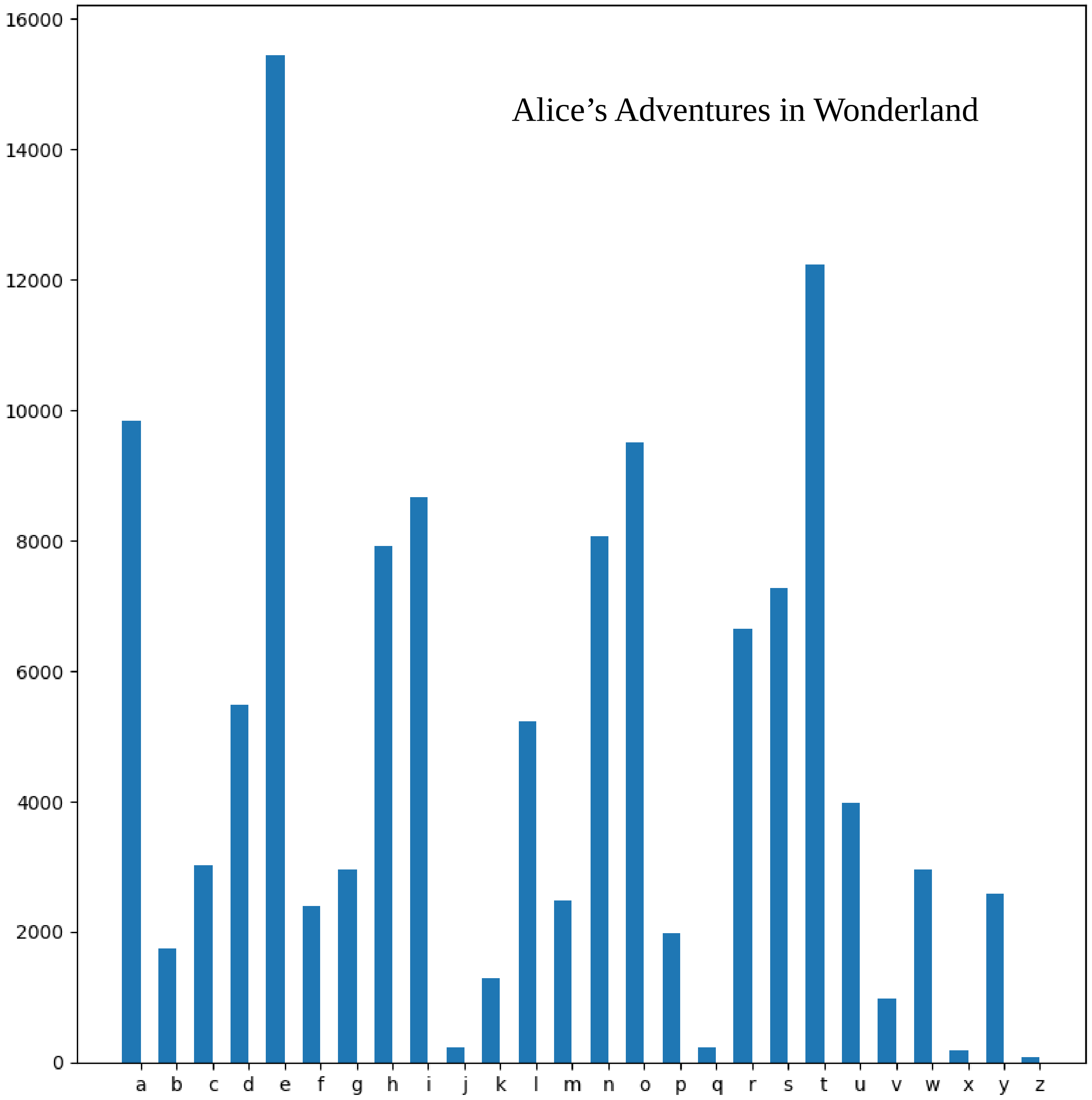
Alice’s Adventures in Wonderland
16000
14000
12000
10000
8000
6000
4000
2000
0
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z

Figure 1.0

Moby Dick:

```
a :  79235        distribution:  0.08163978140094627
b :  17212        distribution:  0.017734384015562408
c :  23319        distribution:  0.024026731400121994
d :  38853        distribution:  0.04003218813366524
e :  119333       distribution:  0.12295475527127055
f :  21261        distribution:  0.02190627112217478
g :  21285        distribution:  0.021930999521917606
h :  63768        distribution:  0.06570335811668508
i :  66702        distribution:  0.06872640498524539
j :  1176         distribution:  0.0012116915873984075
k :  8223         distribution:  0.008472567961885294
l :  43369        distribution:  0.04468524868527341
m :  23697        distribution:  0.024416203696071483
n :  66781        distribution:  0.06880780263439885
o :  70790        distribution:  0.07293847574143986
p :  17886        distribution:  0.018428839908340065
q :  1581         distribution:  0.0016289833330585734
r :  53586        distribution:  0.05521233452579172
s :  65145        distribution:  0.06712215005192963
t :  89894        distribution:  0.09262228193672827
u :  27203        distribution:  0.028028610758502447
v :  8730         distribution:  0.008994955406452464
w :  22540        distribution:  0.02322408875846948
x :  1063         distribution:  0.001095262038609275
y :  17230        distribution:  0.017752930315369526
z :  638          distribution:  0.0006573632931634217
total =  970544
total distribution =  1.0
```

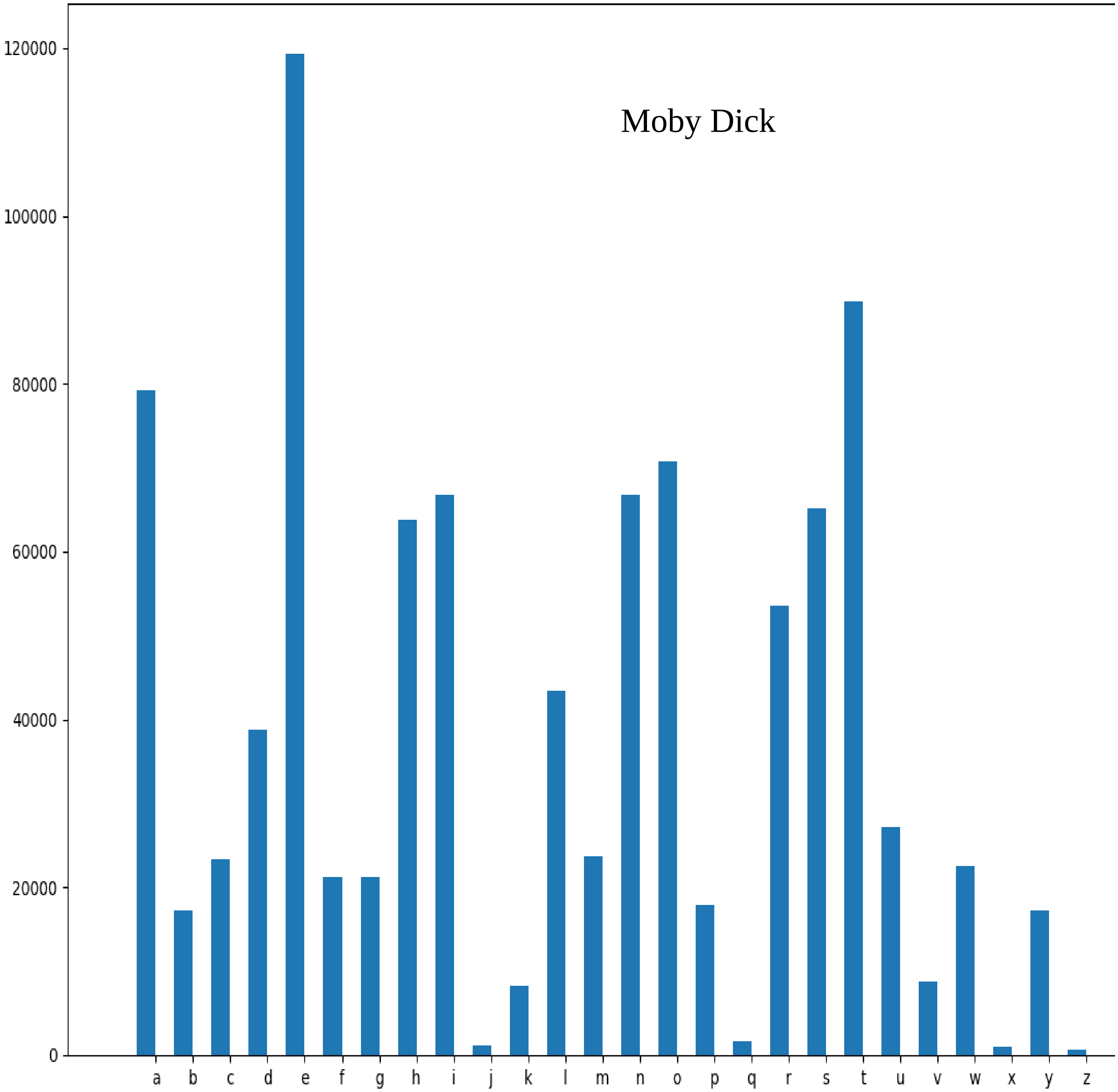
Moby Dick
120000
100000
80000
60000
40000
20000
0
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z

Figure 1.1

Comparison:

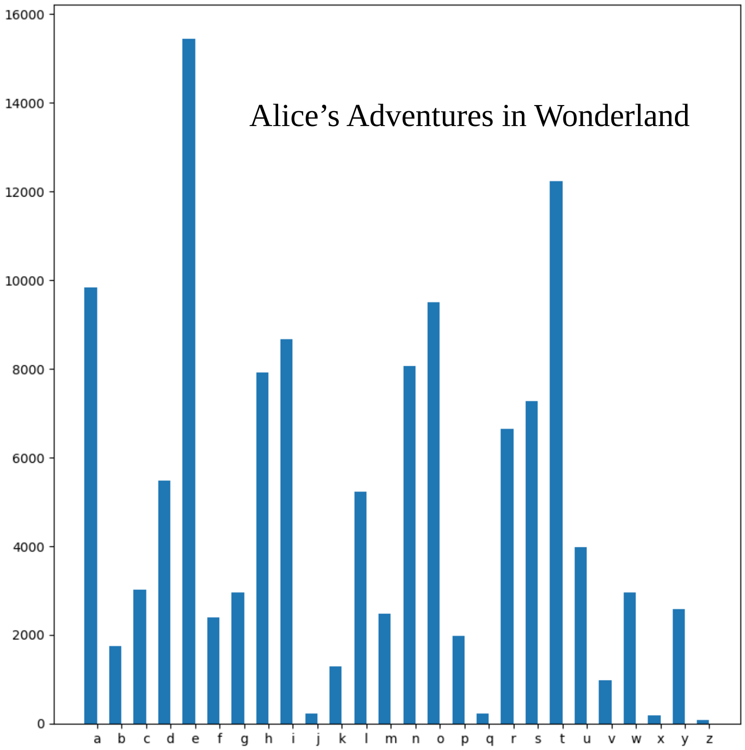

Figure 1.2

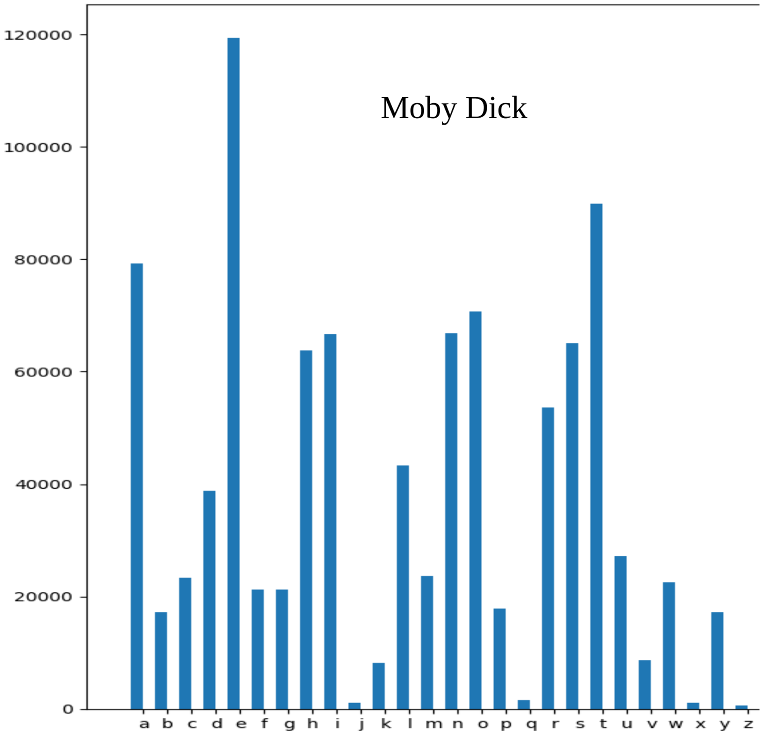

Figure 1.3

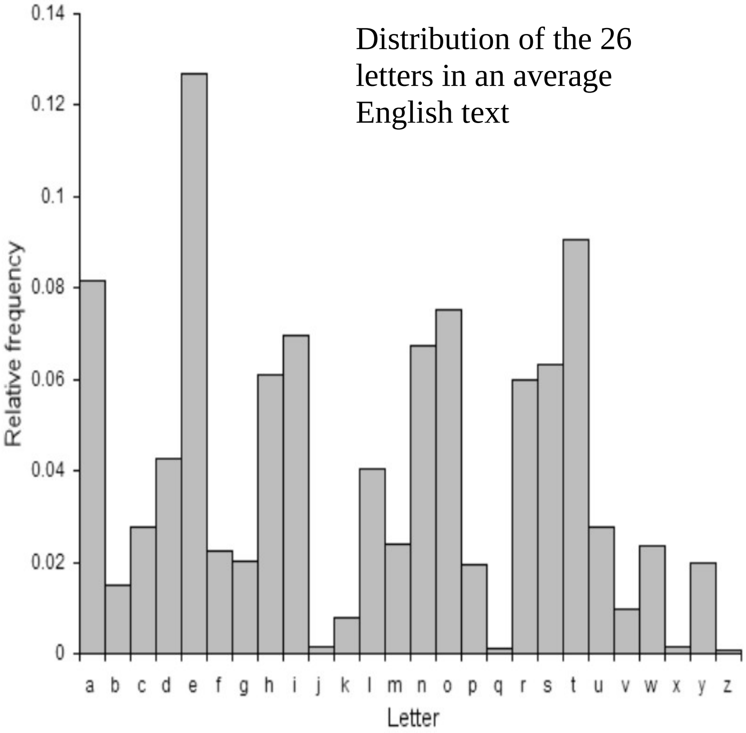

Figure 1.4

The two graphs representing the distribution of characters between Alice’s Adventures in Wonderland, Moby Dick and the 26 letters in an average English text graph look almost identical. This distribution shows that there could be a very strong relationship between the frequency of letters used in a piece of text regardless of the author. Furthermore, if the text were to be encrypted by something like the Caesar cipher, it could be trivial to decrypt it with a simple analysis of letter frequencies to find the number of shifts in a key

Task 2:

Alice’s Adventures in Wonderland Encrypted

```
a :  223        distribution:  0.0018069848472571104
b :  6648       distribution:  0.05386921643302812
c :  7280       distribution:  0.05899035734543392
d :  12241      distribution:  0.09918969289360667
e :  3990       distribution:  0.03233125354509359
f :  972        distribution:  0.0078761850741431
g :  2956       distribution:  0.02395267806498663
h :  179        distribution:  0.0014504497204440483
i :  2589       distribution:  0.020978850984523133
j :  80         distribution:  0.0006482456851146585
k :  9846       distribution:  0.0797828376954866
l :  1757       distribution:  0.014237095859330686
m :  3028       distribution:  0.024536099181589822
n :  5491       distribution:  0.04449396321205737
o :  15441      distribution:  0.125119520298193
p :  2385       distribution:  0.019325824487480756
q :  2948       distribution:  0.023887853496475164
r :  7915       distribution:  0.06413580747103152
s :  8669       distribution:  0.07024552305323718
t :  235        distribution:  0.0019042217000243091
u :  1291       distribution:  0.010461064743537802
v :  5227       distribution:  0.042354752451179
w :  2469       distribution:  0.020006482456851146
x :  8066       distribution:  0.06535937120168545
y :  9496       distribution:  0.07694676282310996
z :  1988       distribution:  0.016108905275099263
total =  123410
total distribution =  1.0
```

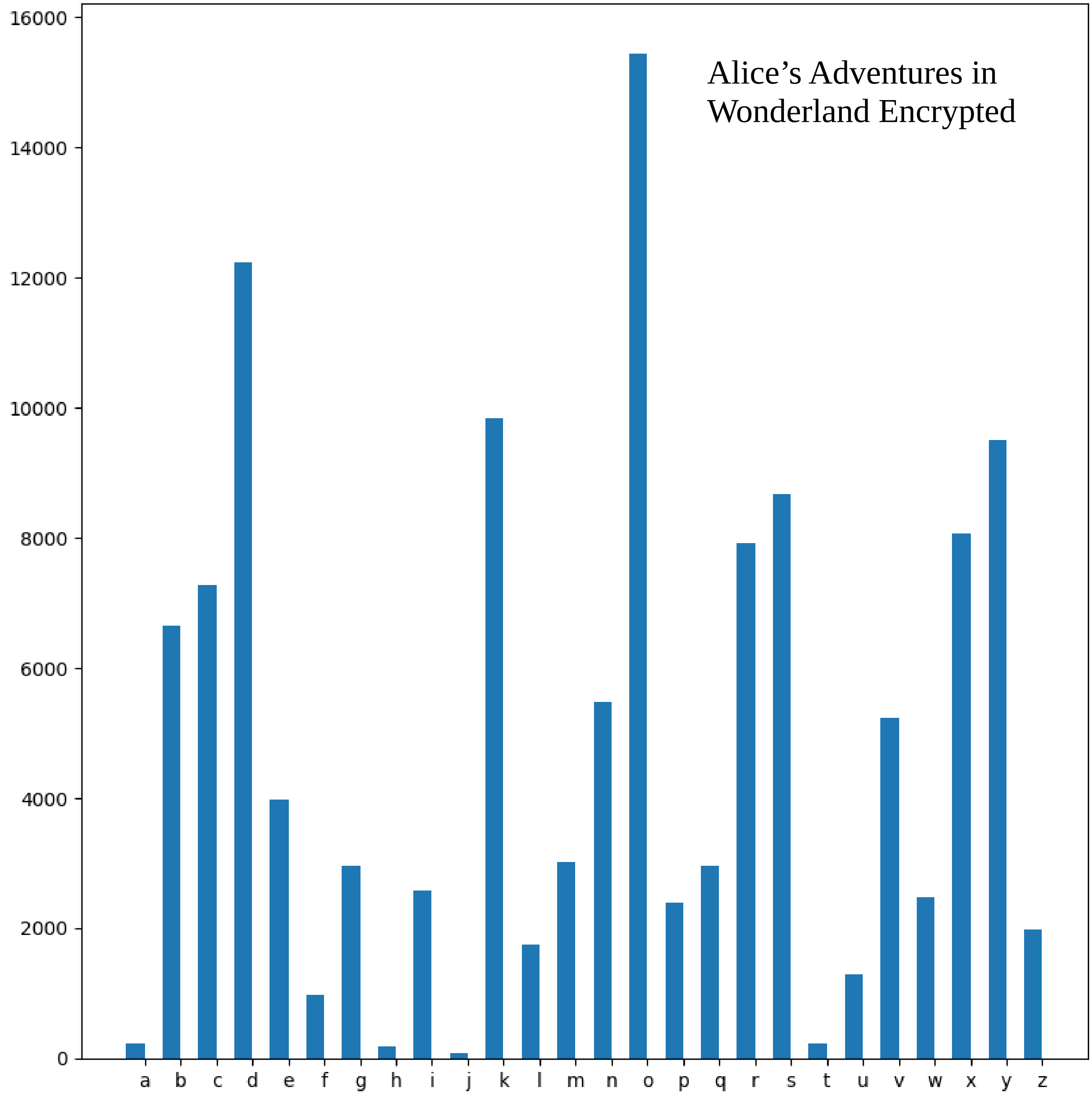
16000
14000
12000
10000
8000
6000
4000
2000
0
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
Alice's Adventures in
Wonderland Encrypted

Figure 2.0

Moby Dick Encrypted:

```
a :  1581         distribution:  0.0016289833330585734
b :  53586        distribution:  0.05521233452579172
c :  65145        distribution:  0.06712215005192963
d :  89894        distribution:  0.09262228193672827
e :  27203        distribution:  0.028028610758502447
f :  8730         distribution:  0.008994955406452464
g :  22540        distribution:  0.02322408875846948
h :  1063         distribution:  0.001095262038609275
i :  17230        distribution:  0.017752930315369526
j :  638          distribution:  0.0006573632931634217
k :  79235        distribution:  0.08163978140094627
l :  17212        distribution:  0.017734384015562408
m :  23319        distribution:  0.024026731400121994
n :  38853        distribution:  0.04003218813366524
o :  119333       distribution:  0.12295475527127055
p :  21261        distribution:  0.02190627112217478
q :  21285        distribution:  0.021930999521917606
r :  63768        distribution:  0.06570335811668508
s :  66702        distribution:  0.06872640498524539
t :  1176         distribution:  0.0012116915873984075
u :  8223         distribution:  0.008472567961885294
v :  43369        distribution:  0.04468524868527341
w :  23697        distribution:  0.024416203696071483
x :  66781        distribution:  0.06880780263439885
y :  70790        distribution:  0.07293847574143986
z :  17886        distribution:  0.018428839908340065
total =  970544
total distribution =  1.0
```

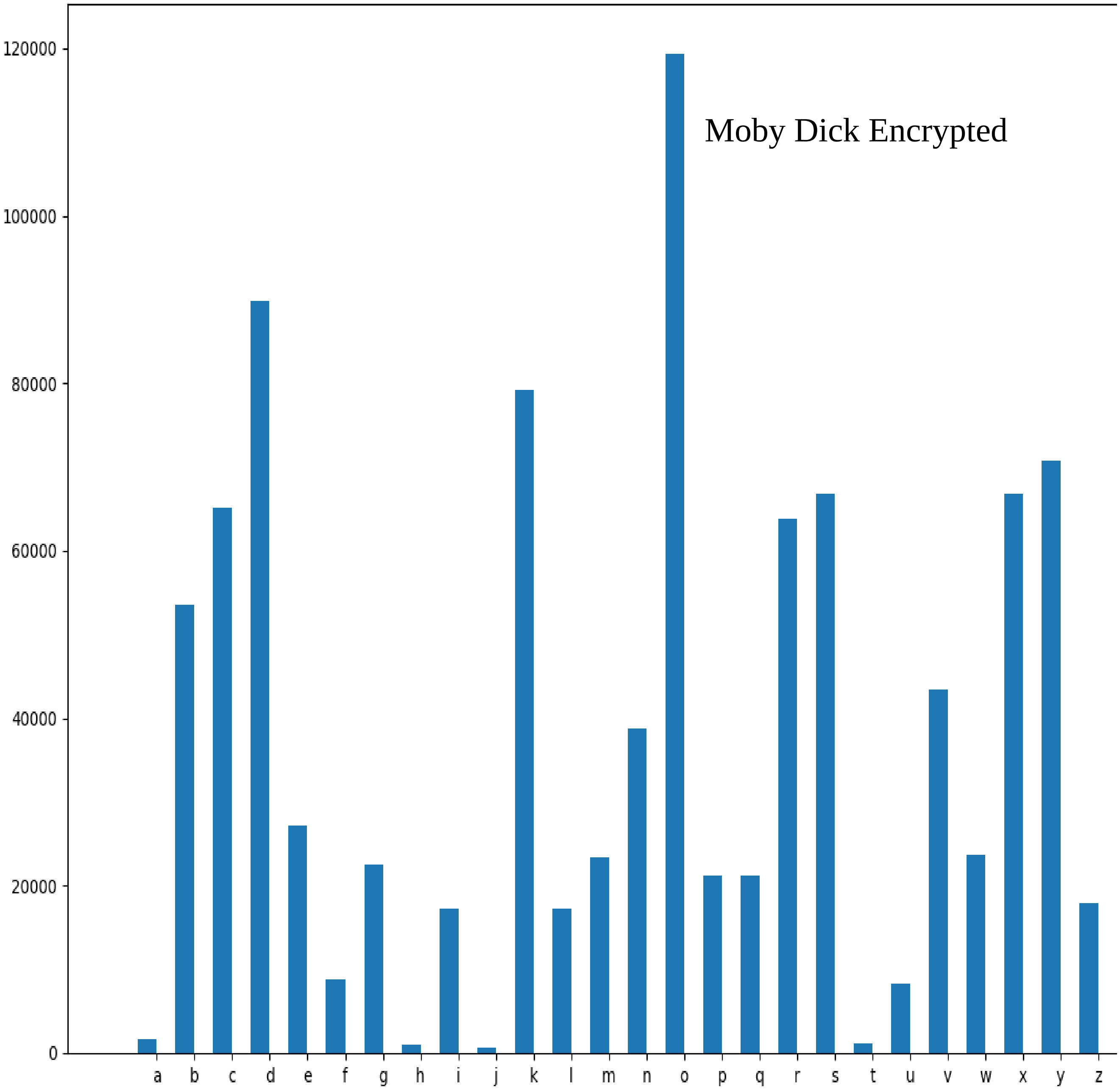
Moby Dick Encrypted
0
20000
40000
60000
80000
100000
120000
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z

Figure 2.1

Calculate conditional probabilities:

$$P(M = m|C = c) = \frac{\sum_{\{k:m=d_k(c)\}} P(K = k) \cdot P(M = m)}{\sum_{\{k:c\in C(k)\}} P(K = k) \cdot P(M = d_k(c))}$$

Figure 2.2

Alice’s Adventures in Wonderland

Character analysis	Conditional Probability
P(M=e|c_i) $c_i \in C$	0.004812289242238193
P(M=t|c_i) $c_i \in C$	0.003814988188215641
P(M=a|c_i) $c_i \in C$	0.003068570680595638
P(M=i|c_i) $c_i \in C$	0.0027017508866629685
P(M=o|c_i) $c_i \in C$	0.0029594908778119216
P(M=n|c_i) $c_i \in C$	0.002513821969295594

Table 2.0

Moby Dick

Character analysis	Conditional Probability
P(M=e|c_i) $c_i \in C$	0.004729029048895021
P(M=t|c_i) $c_i \in C$	0.003562395459104933
P(M=a|c_i) $c_i \in C$	0.0031399915923440874
P(M=i|c_i) $c_i \in C$	0.0026433232686632843
P(M=o|c_i) $c_i \in C$	0.0028053259900553793
P(M=n|c_i) $c_i \in C$	0.0010780234907116326

Table 2.1

The conditional probabilities (see Table 2.0) represent the chances of any given encrypted value to be a certain character, the specific formula used is stated above in figure 2.2. There is a 0.5% chance that any given character in the cipher text file is a an ‘e’, 0.4% chance that it is a ‘t’ and so on given a key with a 1/26 chance and a valid cipher. As expected, the ranking of the conditional probability of each of these characters follows their frequency rankings in the plain text files. This is clear evidence that the specific Caesar cipher algorithm used, reveals a significant amount of information about the plain text. The conditional probabilities determined above could be very useful in identifying the plain text and cipher text pairing since a frequency analysis could be made on the cipher text (see figure 2.0) to easily determine the number of shifts made in the Caesar cipher or key.