Probabilities and entropy

Input: Nehru speech.txt

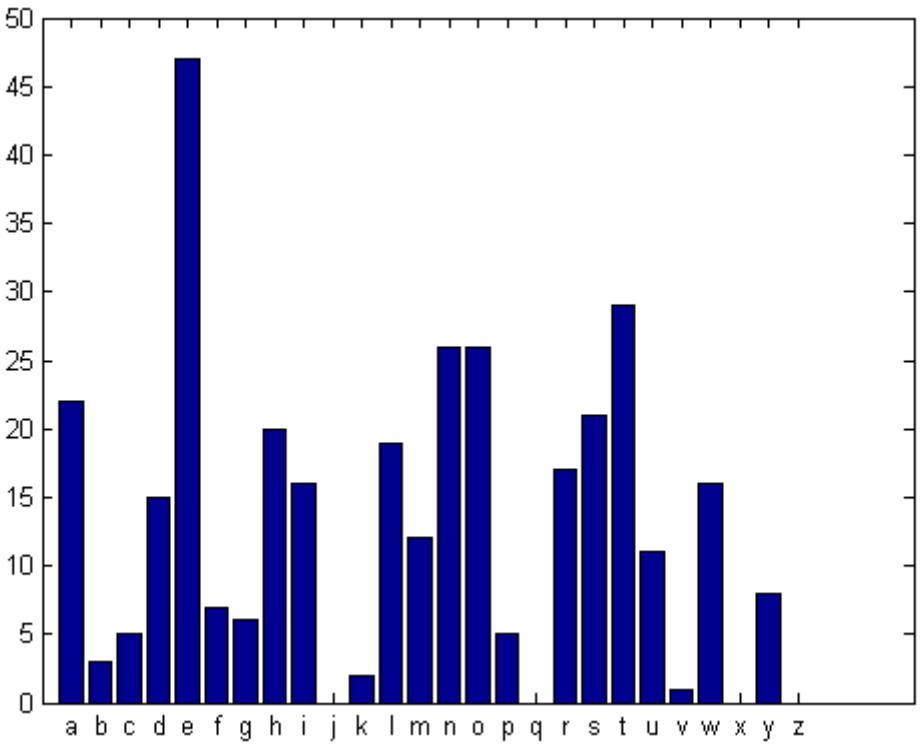


Figure 1: Frequency of letters in Nehru speech.txt

a) The probability of occurrence of each letter based on the frequency is:

a	0.065868
b	0.008982
c	0.01497
d	0.04491
e	0.140719
f	0.020958
g	0.017964
h	0.05988
i	0.047904
j	0
k	0.005988
l	0.056886
m	0.035928
n	0.077844
o	0.077844
p	0.01497
q	0

r	0.050898
s	0.062874
t	0.086826
u	0.032934
v	0.002994
w	0.047904
x	0
y	0.023952
z	0

b) The entropy of the text is 4.101775.

c) Six speeches were chosen for computing frequency in longer texts: The unabridged version of Tryst with Destiny (Speech English), I Have a Dream (Speech English), The Daffodils (Poem English), Une certaine idée de la France (Speech French), Le Printemps (Poem French) and Sur le Pont D'Avignon (Poem French).

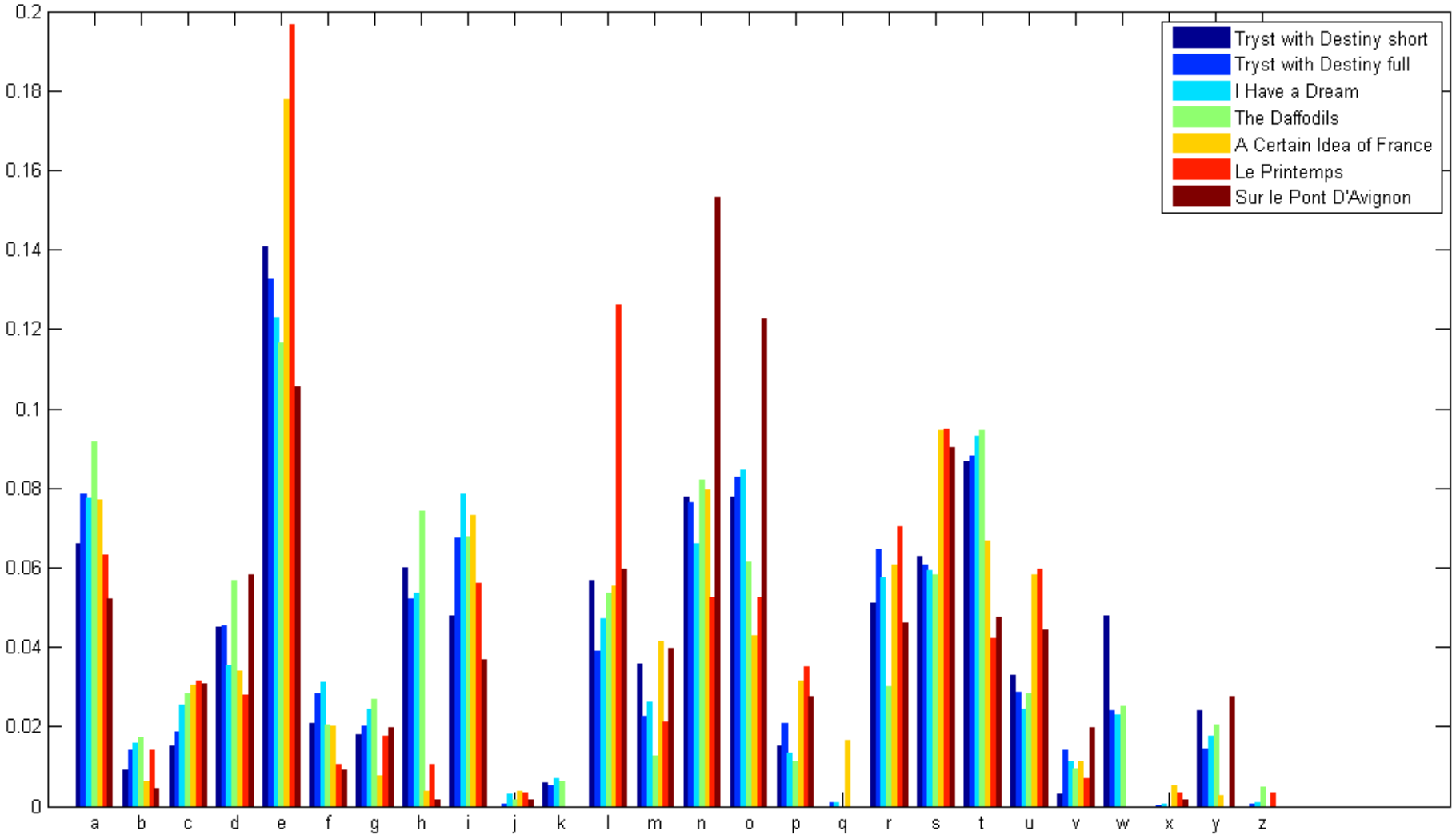
It was observed that although in English, there is not much difference in the porbability of the occurrence of letters based on frequncy, but in French there seems to be a large difference. In case of Sur le Pont D'Avignon, a festival song, 'n' and 'o' seems to occur more frequently due to its rhythm. In case of Le Printemps, the was an unusually high occurrence of 'e' and 'I'.

The entropies were

Tryst with Destiny	4.138875
I Have a Dream	4.173899
The Daffodils	4.145939
Une certaine idée de la France	3.957631
Le Printemps	3.848505
Sur le Pont D'Avignon	3.923404

In general, the French texts had lesser entropies than the English ones. Due to their similar letter distribution, the English texts had similar entropies.

[See Next Page for the Graph]


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
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
Tryst with Destiny short
Tryst with Destiny full
I Have a Dream
The Daffodils
A Certain Idea of France
Le Printemps
Sur le Pont D'Avignon