Homework 2
Fall 2016
Advanced Natural Language Processing

Anusha Ramamurthy
anuramam@iu.edu

September 15, 2016

All the work herein is solely mine.

Question 1

1. 1 2 3 4 5 6 7 8
   $\frac{1}{8}\ \frac{1}{16}\ \frac{1}{4}\ \frac{1}{8}\ \frac{1}{16}\ \frac{1}{16}\ \frac{1}{4}\ \frac{1}{16}$
   What is the entropy of this distribution?
   Solution:

   $H(x) = -\Sigma_{i=1}^{8} p(1) \log_2 p(i) = \{\frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{16}\log_2\frac{1}{16} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{16}\log_2\frac{1}{16}\}$=2.75

Question 2.1 - Probability of
A : 0.0918
C : 0.0204
B : 0.0136
E : 0.1429
D : 0.0476
G : 0.0204
F : 0.034
I : 0.0476
H : 0.0816
K : 0.0068
M : 0.0272
L : 0.051
O : 0.0748
N : 0.068
P : 0.0102
S : 0.0646
R : 0.051
U : 0.017
T : 0.0578
W : 0.0408
V : 0.0204
Y : 0.0068
X : 0.0034

Question 2.2 - Using Huffman Coding Algorithm
('Max No of Bits required:', 8)
('Min No of Bits required:', 3)
[['A', '000'], ['E', '101'], ['D', '0010'], ['H', '1110'], ['I', '0011'], ['L', '0100'], ['N', '1001'], ['O', '1100'], ['R', '0101'], ['S', '1000'], ['T', '0110'], ['F', '01111'], ['M', '01110'], ['W', '11110'], ['C', '110110'], ['G', '110111'], ['U', '110100'], ['V', '111110'], ['B', '1111111'], ['P', '1101011'], ['Y', '1101010'], ['K', '11111101'], ['X', '11111100']]

Question 2.3 - Keeping the Text (not including punctuations and spaces) as is and converting everything to Upper Case
('Variance is ', 93.996219281663528)

Question 2.3 - Keeping the Text (includes punctuations and spaces) as is and converting everything to Upper Case
('Variance is ', 199.20110192837458)

Question 2.3 - Keeping the Text (includes punctuations and spaces as is and not converting everything to Upper Case
('Variance is ', 182.39737034331631)

Process finished with exit code 0

```
Enter the subcorpora name: or DONE"A"
('Entropy of Corpa A is ', 6.269610053587205)
Enter the subcorpora name or DONE:"G"
('Entropy of Corpa G is ', 6.244113821853599)
Enter the subcorpora name or DONE:"J"
('Entropy of Corpa J is ', 5.873465203669868)
Enter the subcorpora name or DONE:"N"
('Entropy of Corpa N is ', 6.290773611483493)
Enter the subcorpora name or DONE:n]
Please enter wihin quotes either A,G,J or N:
Please enter wihin quotes either A,G,J or N:

Process finished with exit code 0
```