Sanjana Agarwal
Username: sa14593

ADVANCED NLP ASSIGNMENT 2

1) To find the entropy of the given distribution, we can apply the formula of entropy, which is,

$$H(x) = -\Sigma p(i)\ log p(i)$$

Therefore,

$$H(x) = -[\frac{1}{8}log(\frac{1}{8}) + \frac{1}{16}log(\frac{1}{16}) + \frac{1}{4}log(\frac{1}{4}) + \frac{1}{8}log(\frac{1}{8}) + \frac{1}{16}log(\frac{1}{16}) + \frac{1}{16}log(\frac{1}{16}) + \frac{1}{4}log(\frac{1}{4}) + \frac{1}{16}log(\frac{1}{16})]$$

$$= -[\frac{2}{8}log(\frac{1}{8}) + \frac{4}{16}log(\frac{1}{16}) + \frac{2}{4}log(\frac{1}{4})]$$

$$= -[\frac{1}{4}log(\frac{1}{8}) + \frac{1}{4}log(\frac{1}{16}) + \frac{1}{2}log(\frac{1}{4})]$$

$$= -[-\frac{1}{4}log({}_2 3) - \frac{1}{4}log({}_2 4) - \frac{1}{2}log({}_2 2)]$$

$$= [\frac{3}{4} + \frac{4}{4} + \frac{2}{2}]$$

$$= 2.75$$

2 a) Using fp1.py, we find the frequencies of the words. They have been tabulated below.

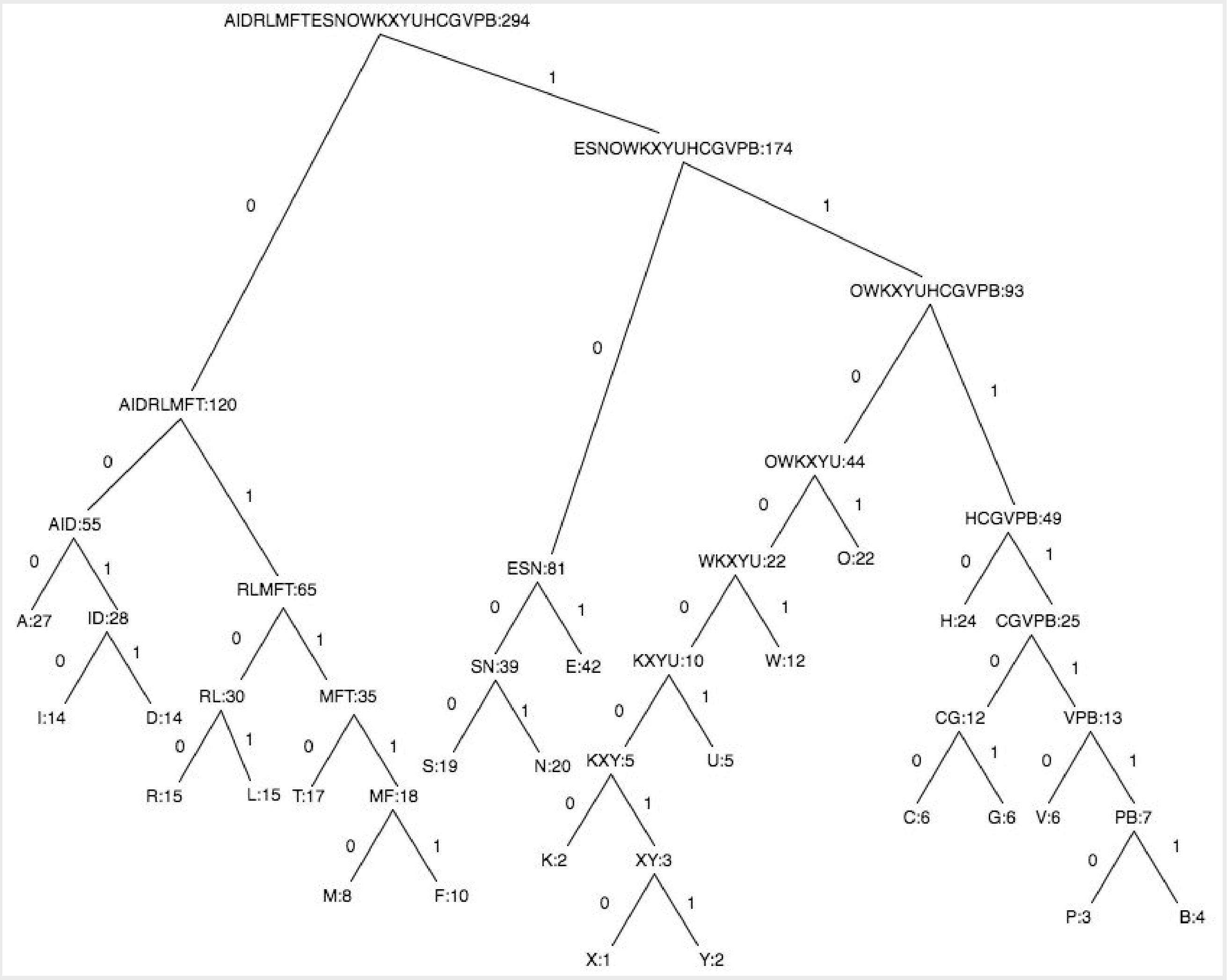
A 27	L 15	W 12
B 4	M 8	X 1
C 6	N 20	Y 2
D 14	O 22	
E 42	P 3	
F 10	R 15	
G 6	S 19	
H 24	T 17	
I 14	U 5	
K 2	V 6	

Total sum of the frequencies of each letter is 294
The probability of occurrence of each letter is equal to division of each character frequency by total number of characters.

A	27	0.091836735	N	20	0.068027211
B	4	0.013605442	O	22	0.074829932
C	6	0.020408163	P	3	0.010204082
D	14	0.047619048	R	15	0.051020408
E	42	0.142857143	S	19	0.06462585
F	10	0.034013605	T	17	0.057823129
G	6	0.020408163	U	5	0.017006803
H	24	0.081632653	V	6	0.020408163
I	14	0.047619048	W	12	0.040816327
K	2	0.006802721	X	1	0.003401361
L	15	0.051020408	Y	2	0.006802721
M	8	0.027210884			

2 b) We can use the famous Huffman Encoding scheme for the letters.
In the Huffman encoding scheme, we arrange all the characters in the decreasing order of their occurrences and start building a tree in a bottom up fashion with the least frequent characters. Once that is done, we can use a binary encoding like, we assign 0 for every left side character and 1 for every right side. The Huffman tree generated is:
Thus, values of each character are:

A	000	G	111110	N	1001	U	011110
B	1111111	H	1110	O	1101	V	110010
C	110011	I	0010	P	1111110	W	11110
D	0011	K	01111111	R	0100	X	01111110
E	101	L	0101	S	1000	Y	0111110
F	11000	M	01110	T	0110		



Thus, according to Huffman Encoding, we need 8 digits.

2 c) To calculate the variance:

Formula is: $V(X) = \sum_{X=x} (X - \mu)^2 p(X)$

Here, X is the decimal value of each of the characters. We have their binary formats, we just need to convert it to decimal values.

Character	Binary Value	Decimal Value	Mean	$(x-\mu)^2$	Variance
A	0	0	0	255.6735688	23.4802257
B	1111111	127	1.727891156	12323.26544	167.6634754
C	110011	51	1.040816327	1225.7144	25.0145796
D	11	3	0.142857143	168.7347941	8.034990197
E	101	5	0.714285714	120.7756111	17.25365872
F	11000	24	0.816326531	64.16337181	2.182427612
G	111110	62	1.265306122	2116.938893	43.20283456
H	1110	14	1.142857143	3.959287206	0.323207119
I	10	2	0.095238095	195.7143857	9.319732652
K	1111111	127	0.863945578	12323.26544	83.83173771
L	101	5	0.255102041	120.7756111	6.162020973
M	1110	14	0.380952381	3.959287206	0.107735706
N	1001	9	0.612244898	48.85724491	3.323622102
O	1101	13	0.972789116	8.938878746	0.668895688
P	1111110	126	1.285714286	12102.24503	123.4922963
R	100	4	0.204081633	143.7552026	7.334449113
S	1000	8	0.517006803	63.83683645	4.125509838
T	110	6	0.346938776	99.79601953	5.770518136
U	11110	30	0.510204082	196.2858226	3.338194261
V	110010	50	1.020408163	1156.693992	23.60599983
W	11110	30	1.224489796	196.2858226	8.011666227
X	1111110	126	0.428571429	12102.24503	41.16409876
Y	111110	62	0.421768707	2116.938893	14.40094485

To calculate μ, we need to find the expected value, which is nothing but μ

$$E(X) = \sum_{X=x} Xp(X)$$

We calculated p(X) in part a. Hence the expected value of each character has been summarized in the table above.
Thus, the mean is 15.98979592.

The variance is calculated and summarized in the table above.
Thus the total variance is: 621.812821

3) I wrote my code in Java. The code has been attached with the assignment called Entropy.java
To run the code, change the location of the input file (only the string part). Let the variable ‘i’ remain.
Also, the output has been attached, called output.txt

The total number of POS tags are: 342
Entropy of POS for the sub-corpora A is: 65.96079