EECE 405/560 --- Detector programming assignment --- due October 14

Write a brief computer program that inputs text from `stdin`, computes the frequency `h[i]` of each letter of the alphabet, and then outputs the following 4 numbers on a single output line:

1. The Index of Coincidence:	$IC(s) = \sum_{i=1}^{26} h[i]^2$
2. The Shannon entropy formula:	$H(s) = \sum_{i=1}^{26} h[i] \cdot \log_2 \left(\frac{1}{h[i]} \right)$
3. The correlation of letter frequencies with English:	$h \cdot P_E = \sum_{i=1}^{26} h[i] \cdot P_E[i]$
4. The log-likelihood ratio of English versus Noise:	$LLR(s) \quad = \quad \log_{10} \frac{P_E[s]}{P_N[s]}$

Output these four numbers by themselves, each with 6 digits beyond the decimal point:

```
$ gcc -o freq frequency.c

$ echo "ABCDEFGHIJKLMNOPQRSTUVWXYZ" | ./freq
0.038462 4.700440 0.038462 -7.027473

$ cat EnglishMessage.txt | ./freq
0.072493 4.079364 0.067607 140.506455
```

We will then be able to use this program as a primitive to compare the four formulas, and compare their receiver operating characteristics (ROCs). This may seem like a complicated assignment, but really you are just writing a loop and computing four formulas that I just gave you. The program should be relatively short and easy to write.

Notes on the four formulas:

In the table above,

1. The index i goes from 1..26 for the letters A..Z.
2. The value `h[i]` is the fraction of text equal to that letter, so if the text is AaABBC, then `h[1]==0.5` and `h[4]==0`.
3. Compute the probability of a string as the product of the probabilities of their respective letters. Thus $P_E[AABC]==P_E[A]^2 \times P_E[B] \times P_E[C]$.
4. For all letters, $P_N[i]=1/26$. To the right is the table of values to use for P_E. Copy and paste these values into your source code.
5. For H(s), assume `0×log(1/0)==0`, so your computer doesn't explode when `h[i]==0`.

Pr[A]	0.081753
Pr[B]	0.017740
Pr[C]	0.023634
Pr[D]	0.040260
Pr[E]	0.122931
Pr[F]	0.021868
Pr[G]	0.021900
Pr[H]	0.066036
Pr[I]	0.068804
Pr[J]	0.001131
Pr[K]	0.008488
Pr[L]	0.044956
Pr[M]	0.024469
Pr[N]	0.068996
Pr[O]	0.072834
Pr[P]	0.018115
Pr[Q]	0.001650
Pr[R]	0.054670
Pr[S]	0.067462
Pr[T]	0.092509
Pr[U]	0.027998
Pr[V]	0.009004
Pr[W]	0.023286
Pr[X]	0.001079
Pr[Y]	0.017748
Pr[Z]	0.000679

A note on programming:

You can write this program in any language you want, except that it has to work as specified, and the TA must be able to run it on a typical Unix system (such as Bingsuns) without any special software installed. Thus you cannot do this assignment in Excel, or in Matlab (you'll thank me later for not letting you do this in Matlab,) or anything else that requires you to run some commercial software. If you want a suggestion for a language, I wrote the solution set in C while writing this handout.

Your program input will include uppercase and lowercase letters, and plenty of characters that are not letters. Make sure to ignore all nonletters and ignore case. Also, the input may be missing some letters of the alphabet, so if `h[i]==0` your program should not crash. You can assume the input will include at least one letter of the alphabet.

Your grade will depend on whether the program works on our input test sets, but you will also be graded on the size and readability of your code. A C program to perform this task should not be longer than 100 lines.

Hand in

The source file for your program, to me, in email (make sure the subject header includes EECE405 or EECE560, so I will not miss your submission.) If there are any special instructions for running the program, make sure to provide them.