2017 年 5 月 8 日　 情報理論　 第 4 回　実習 1　　担当：間野

情報理論のシェアホルダにデータ（半角英数字及び半角の記号からなるテキストデータ、ファイル名 "alice29.txt"）がある。このデータについて以下を計算しなさい。
(Linux 上で C 言語で計算することを想定しているが、他の OS/言語でもよい。)

問題 1 (必ずやること) #アルファベットのエントロピーの計算

(1) データより，アルファベット（半角英字で大文字('A'—'Z')と小文字('a'—'z')の区別なしで、'a'—'z'の 26 文字)の生起確率 p(X)(=p('a'), p('b'), ..., p('z'))をそれぞれ計算しなさい。ただし、アルファベット以外の数字やその他の記号・空白は無視して詰めること。これを情報源 S とする。

(2) アルファベット情報源 S のエントロピーH(S)を計算しなさい。また、26 文字の生起確率を等確率(=1/26)と仮定した場合のエントロピーと値を比較しなさい。

問題 2 （必ずやること） #2 次の拡大情報源のエントロピーの計算

(1) 問題 1 の情報源 S（大文字小文字の区別なし）の 2 次の拡大情報源 S^2 (2 文字を 1 アルファベットとした情報源）を考える。データより，最も生起確率の大きい 2 次のアルファベット情報源の文字列とその生起確率を求めなさい。(つまり、データファイルからアルファベット 2 文字の組を取り出してそれを S^2 の情報源記号として計算し，その最大生起確率の組を求めればよい． 全組の印刷は大きくなるので不要.）

(2) (3)の 2 次の拡大情報源のエントロピーH(S^2)，その 1 文字当たりのエントロピー H(S^2)/2 を求めなさい。また、問題 1(2)の H(S)と値を比較しなさい。

問題 3 （できる人） #条件付きエントロピーの計算

問題 1 のアルファベット情報源 S（大文字小文字の区別なし）について、i 番目の文字がわかった時の i+1 番目の文字の条件付きエントロピー$H(X_{i+1}|X_i)$を求めなさい。(つまり、データファイルからアルファベット 1 文字ずつを順に取り出して、X_i 、X_{i+1} として 1 文字ずつシフトしながら $P(X_{i+1}|X_i)$を計算し、そこから、条件付きエントロピーを求めればよい.）

〇提出

A4 のレポートの書き出しに、授業名，実習 1, 授業日，提出日，学籍番号，氏名 を書く.
各問題の答えとソースプログラムを添付し，ファイル名= “学籍番号”_実習 1.pdf として，
提出先：シェアフォルダ ¥2017¥大宮月曜２限情報理論 - 1311101130¥提出用¥実習 1
提出締切：5/23(金) 18:00〆切

以上

○参考： プログラムに関する説明

（１）データの読み込み方

・問題１は、例えば、入力が「 This is .　OK. (Please read page 10.)」であるとすると、英字のみをすべて小文字にしてつなげた「thisisokpleasereadpage」を 1 文字ずつ、't' 'h' 'i' 's' 'i' 's' 'o' 'k' 'p' 'l' 'e' 'a' 's' 'e' 'r' 'e' 'a' 'd' 'p' 'a' 'g' 'e'　読み込んで、カウントしていく。

・問題２は、問題１の入力「thisisokpleasereadpage」を 2 文字ずつ、'th' 'is' 'is' 'ok' 'pl' 'ea' 'se' 're' 'ad' 'pa' 'ge' ' と読み込んで、カウントしていく。

・問題３は、問題１の入力「thisisokpleasereadpage」を 1 文字ずつずらしながら、'th' hi' 'is' 'si' 'is' 'so' 'ok' 'kp' 'pl' 'le' 'ea' ... と読み込んで、カウントしていく。

（２）使うかもしれない c 言語の関数

・double log10( )　常用対数　　(`#include <math.h>` を忘れない。)
・FILE *fopen( )　ファイルオープン　(`#include <stdio.h>`, `#include <stdlib.h>`)
・ic = fgetc(fin)　　char　1byte 入力　(`#include <stdio.h>`, `#include <stdlib.h>`)
・isalpha(ic)　　ic が英字アルファベットなら真　　(`#include <ctype.h>`)
・ic2 = tolower(ic)　ic が英大文字なら英小文字を出力，小文字ならそのまま出力
(`#include <ctype.h>`)

（３）配列のインデックス計算のテクニック

ASCII の'a','b',...,'z'は、バイナリ数字に変換すると　97, 98,..., 122 となるので、例えば、'a','b',...,'z'のカウンタとして、`count[0],count[1],...,count[25]`を対応させるには、入力 `char` を `ic` とすると、`count[ic - 'a']`とすれば、`count[0]`が'a'のカウント，`count[1]`が'b'のカウント,... となる。

これで、配列のインデックスを求めるのに

```
if (ic == 'a') count[0]++;
if (ic == 'b') count[1]++;
  ...
```

を繰り返さなくてもよい.（もちろん，こうしてもいい.）

以上