2002/2003 SOUTHERN CALIFORNIA REGIONAL
ACM INTERNATIONAL COLLEGIATE PROGRAMMING CONTEST

Problem 1
Find Crypt

The Swamp Count Sheriff's Department has captured several computers which they think contain important evidence. But there are thousands of files on them and the data in question may be encrypted. The task of your team is to write a program that will analyze each file and produce a report of the distribution of byte values in the file. This, along with other information, will be used to select files to be investigated further.

Consider a file to be made up of 8-bit bytes (0..255). A compact way of representing a group of byte values is a bitmap. That is, for a given character, say 'A' which has decimal value 65, bit 65 in the bit map will be on. It can be printed conveniently in hexadecimal with 64 characters. For example, a group consisting only of the so-called printing characters (32..126) plus newline (10) looks like:

```
00200000FFFFFFFFFFFFFFFFFFFFFFFE0000000000000000000000000000000000
```

A Chi^2 test is a measure of how a group of numbers are distributed. In this case, the numbers are the byte values.

$$\mathrm{Chi}^2 = \frac{\sum(f_i - (N/r))^2}{(N/r)}$$

where i = 0..255, f_i is the count of bytes with value i, N is the number of bytes in the file, and r is 256. Also of interest is the distribution using only the values found in the file. In this case, i is a list consisting of all j where $f_j \neq 0$ and r is the length of that list.

Also calculate a measure of the 'compressed' length of the file. Considering each byte value to be a symbol, the number of bits needed can be measured by

$$-\sum f_i \log_2(f_i/N)$$

where $\log_2$ is $\log$ to the base 2. Report the answer in bytes, which is the least integer greater or equal to the above divided by 8.

Input

The input is standard input terminated by end-of-file. Note that the file is *binary*, that is, all characters, including newlines, are significant.

Output

The report has two fields. The first field is of fixed length 15 and contains a label or an integer, each left adjusted. Any integer should have no leading zeros. The second field can be an integer (again with no leading zeros), a bitmap, or a floating point number, all left adjusted. Use uppercase when printing bitmaps. The format of a floating point number is one digit before the decimal point and three after, with a power-of-ten exponent using a lower case 'e', a sign, and two digits.

The first line has the label 'length' and the length of the file. The second line has the label 'all' and the bitmap of all byte values that appear in the file.

The next 1-to-5 lines detail the most commonly occuring byte values in the file. The first field is an integer n and the second is the bitmap of all byte values that occur n times in the file. Print the five highest values of n in decending order. If a file has fewer than five values of n, print only those. Note that there will always be at least one line, because you will not be given a zero length file.

Next print a line with the label 'chisq' and the floating point value of the Chi^2 test, followed by a line with the label 'nzchisq' and the floating point value of the Chi^2 test using only the values found in the file. Finally, print the label 'minlen' and the estimate of the compressed length of the file.

Problem 1
Find Crypt (continued)

The Sample Input is a hex dump of the sample binary file. Use the command **`od -A x -t x1 filename`** *to display the contents of a binary file as hexadecimal byte values.*

Sample Input

```
000000 20 60 31 32 33 34 35 36 37 38 39 30 2d 3d 5c 7e
000010 21 40 23 24 25 5e 26 2a 28 29 5f 2b 7c 0a 71 77
000020 65 72 74 79 75 69 6f 70 5b 5d 51 57 45 52 54 59
000030 55 49 4f 50 7b 7d 0a 61 73 64 66 67 68 6a 6b 6c
000040 3b 27 41 53 44 46 47 48 4a 4b 4c 3a 22 0a 7a 78
000050 63 76 62 6e 6d 2c 2e 2f 5a 58 43 56 42 4e 4d 3c
000060 3e 3f 0a 65 65 65 65 65 65 65 65 65 65 0a
00006e
```

Sample Output

```
length          110
all             00200000FFFFFFFFFFFFFFFFFFFFFE0000000000000000000000000000000000
11              00000000000000000000000040000000000000000000000000000000000000000
5               00200000000000000000000000000000000000000000000000000000000000000
1               00000000FFFFFFFFFFFFFFFFBFFFFFE0000000000000000000000000000000000
chisq           4.485e+02
nzchisq         9.945e+01
minlen          88
```