File analysis tool

John Taylor (onefouronefour limited) 2015-07-30

Years ago I wrote a simple utility to analyse a file into its byte components. I find it very useful and always bring it to each new system I work on, together with some elementary test files; for example a file '256' containing 256 bytes, one of each bit pattern. Where necessary I have rewritten it in C, Java, Perl and awk (making use of od to scan the file to accommodate null characters, which awk cannot handle properly).

As a first examination of a file, it easy to tell whether a text file is DOS- or Unix-formatted by looking at the counts of new line (0x0a) and carriage return (0x0d) characters. Often, problems are resolved by checking whether a file contains tabs, nulls, delete and other unprintables. In text files, characters above 0x7f may represent accented characters or other special values.

However, nowadays many files are encoded in UTF-8 (and most web pages too). Simply adding counts of all bit patterns is not sufficient to understand the large “alphabet” of Unicode characters which are represented in 1, 2, 3 or 4 bytes.

I have rewritten the program again, now in python (utfa.py), this time including support for all UTF-8 characters and adding a mechanism to read web pages. Since the two widely used versions of python (2.x and 3.x) handle unicode differently, I wrote the program to detect which version of python is in use and to handle unicode characters appropriately.

By default, the program assumes files/web pages are UTF-8 encoded. If the input cannot be decoded correctly it suggests two alternatives to handle this.

To illustrate, here is the same '256' test file:

```
utfa.py 256
file 256 is not valid utf-8, try analysing file as bytes using flag -b or enable error replacement with flag -e
```

using the first suggestion:

```
utfa.py 256 -b

byte analysis for 256

        0 1 2 3 4 5 6 7 8 9 a b c d e f       -- characters --

     0  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ................
     1  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ................
     2  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1        !"#$%&'()*+,-./
     3  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       0123456789:;<=>?
     4  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       @ABCDEFGHIJKLMNO
     5  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       PQRSTUVWXYZ[\]^_
     6  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       `abcdefghijklmno
     7  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       pqrstuvwxyz{|}~
     8  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ................
     9  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ................
     a  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1        ¡¢£¤¥¦§¨©ª«¬®¯
     b  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       °±²³´µ¶·¸¹º»¼½¾¿
     c  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ
     d  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
     e  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       àáâãäåæçèéêëìíîï
     f  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1       ðñòóôõö÷øùúûüýþÿ

total bytes : 256
```

shows that there are, indeed, one of each byte value. The total shown is a count of bytes.

By allowing python's decode() method to replace invalid UTF-8 bytes with the standard U_FFFD replacement character (flag -e):

```
utfa.py 256 -e

unicode analysis for 256  (running Python 3)

          0   1   2   3   4   5   6   7   8   9   a   b   c   d   e   f -- unicode block --         -- characters --

     0    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               ................
     1    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               ................
     2    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin                !"#$%&'()*+,-./
     3    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               0123456789:;<=>?
     4    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               @ABCDEFGHIJKLMNO
     5    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               PQRSTUVWXYZ[\]^_
     6    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               `abcdefghijklmno
     7    1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1 Basic Latin               pqrstuvwxyz{|}~
 0xfff0   0   0   0   0   0   0   0   0   0   0   0   0   0 128   0   0 Specials                  ￿￿

total bytes      : 256
total characters : 256
```

Note that the count spacing is adjusted to match the width of the largest character count (here 128). Various flags can be enabled to include or suppress parts of the output, and the table of individual counts can be replaced with summary totals for each block of sixteen consecutive characters (flag -x):

```
utfa.py 256 -ex

unicode analysis for 256  (running Python 3)

        total -- unicode block --        -- characters --

     0   16 Basic Latin                ................
     1   16 Basic Latin                ................
     2   16 Basic Latin                 !"#$%&'()*+,-./
     3   16 Basic Latin                0123456789:;<=>?
     4   16 Basic Latin                @ABCDEFGHIJKLMNO
     5   16 Basic Latin                PQRSTUVWXYZ[\]^_
     6   16 Basic Latin                `abcdefghijklmno
     7   16 Basic Latin                pqrstuvwxyz{|}~
 0xfff0  128 Specials                  ￿￿

total bytes      : 256
total characters : 256
```

Depending on operating system character handling, normally most Unicode characters are displayed correctly, but a few special characters may be unprintable, reverse video, or have unusual vertical spacing or print directions (e.g. Hebrew, Arabic). The 'Specials' here show this. It is possible to display only the characters actually present in the file/web page rather than the full set of sixteen values (flag -m):

```
utfa.py 256 -exm

unicode analysis for 256  (running Python 3)
```

```
       total -- unicode block --       -- characters --

     0   16 Basic Latin                ................
     1   16 Basic Latin                ................
     2   16 Basic Latin                 !"#$%&'()*+,-./
     3   16 Basic Latin                0123456789:;<=>?
     4   16 Basic Latin                @ABCDEFGHIJKLMNO
     5   16 Basic Latin                PQRSTUVWXYZ[\]^_
     6   16 Basic Latin                `abcdefghijklmno
     7   16 Basic Latin                pqrstuvwxyz{|}~
 0xfff0  128 Specials                                 

total bytes      : 256
total characters : 256
```

Here, flag -m masks all but the 0xfffd in the last line, replacing all others with a space.

Since '256' is not valid UTF-8, here it has been converted to 'u256':

```
utfa.py u256

unicode analysis for u256  (running Python 3)

        0 1 2 3 4 5 6 7 8 9 a b c d e f      -- unicode block --       -- characters --

     0  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               ................
     1  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               ................
     2  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin                !"#$%&'()*+,-./
     3  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               0123456789:;<=>?
     4  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               @ABCDEFGHIJKLMNO
     5  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               PQRSTUVWXYZ[\]^_
     6  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               `abcdefghijklmno
     7  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Basic Latin               pqrstuvwxyz{|}~
   0x80 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        ................
   0x90 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        ................
   0xa0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement         ¡¢£¤¥¦§¨©ª«¬®¯
   0xb0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        °±²³´µ¶·¸¹º»¼½¾¿
   0xc0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ
   0xd0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
   0xe0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        àáâãäåæçèéêëìíîï
   0xf0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1      Latin-1 Supplement        ðñòóôõö÷øùúûüýþÿ

total bytes      : 384
total characters : 256
```

Note that for characters above 0x7f rows are shown only where at least one example of a character in each group of sixteen is present in the input. Now the two totals at the end show bytes and characters. Each byte value above 0x7f in the original file requires two bytes in UTF-8. This can be seen by analysing 'u256' as bytes:

```
utfa.py u256 -b

byte analysis for u256

        0  1  2  3  4  5  6  7  8  9  a  b  c  d  e  f           -- characters --
```

```
    0   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          ................
    1   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          ................
    2   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          !"#$%&'()*+,-./
    3   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          0123456789:;<=>?
    4   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          @ABCDEFGHIJKLMNO
    5   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          PQRSTUVWXYZ[\]^_
    6   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          `abcdefghijklmno
    7   1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1          pqrstuvwxyz{|}~
    8   2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2          ................
    9   2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2          ................
    a   2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2           ¡¢£¤¥¦§¨©ª«¬®¯
    b   2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2          °±²³´µ¶·¸¹º»¼½¾¿
    c   0  0 64 64  0  0  0  0  0  0  0  0  0  0  0  0          ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ
    d   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0          ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
    e   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0          àáâãäåæçèéêëìíîï
    f   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0          ðñòóôõö÷øùúûüýþÿ

total bytes : 384
```

Of course, although the legend is headed 'characters' the characters displayed here are not true UTF-8 characters, but those same glyphs are part of ISO/IEC 8859, so it is helpful to make the values distinct.

Example usage

Having shown '256' many different ways, more interesting inputs can show the program in other useful cases.

Here is the Wikipedia web page on UTF-8, referencing the web page using flag -u:

```
utfa.py -u http://en.wikipedia.org/wiki/UTF-8

unicode analysis for http://en.wikipedia.org/wiki/UTF-8     (running Python 3)

            0     1     2     3     4     5     6     7     8     9     a     b     c     d     e     f      -- unicode block --      -- characters --

      0     0     0     0     0     0     0     0     0     0  1827  3029     0     0     0     0     0      Basic Latin              ................
      1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0      Basic Latin              ................
      2 15449    21 10545   495     6   801   761    60   486   486    33   384   441  2871  1686  9036      Basic Latin               !"#$%&'()*+,-./
      3  2679  2023  1731  1154   573   607   720   614  1430   776  1431  1614 11509  5926 11509   202      Basic Latin              0123456789:;<=>?
      4     3   644   350  1335   328   573  1383   134   161   612   129   140   229   376   164   257      Basic Latin              @ABCDEFGHIJKLMNO
      5   263    28   226   704   897   937    76   176    50    23    92   119    28   119    38  1941      Basic Latin              PQRSTUVWXYZ[\]^_
      6     1 16224  3487  6586  6581 16639  3880  3266  4254 14765   118  2439 11454  3892 10620  8188      Basic Latin              `abcdefghijklmno
      7  6585   110  8941 11721 13555  2391  1295  2824  1419  1983   236    36     1    36     6     0      Basic Latin              pqrstuvwxyz{|}~
   0xa0     0     0     2     1     0     0     0     9     0     0     0     0     0     0     1     0      Latin-1 Supplement        ¡¢£¤¥¦§¨©ª«¬®¯
   0xb0     0     0     0     0     0     0     0     0     0     0     0     1     0     2     0     1      Latin-1 Supplement       °±²³´µ¶·¸¹º»¼½¾¿
   0xc0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0      Latin-1 Supplement       ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏ
   0xd0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0      Latin-1 Supplement       ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
   0xe0     1     0     0     0     0     2     0     2     0     2     1     0     0     0     0     1      Latin-1 Supplement       àáâãäåæçèéêëìíîï
   0xf0     0     1     0     2     0     0     0     0     0     0     0     0     3     2     0     0      Latin-1 Supplement       ðñòóôõö÷øùúûüýþÿ
  0x100     0     8     0     0     0     0     0     0     0     0     0     0     1     2     0     0      Latin Extended-A         ĀāĂăĄąĆćĈĉĊċČčĎď
  0x120     0     0     0     0     0     0     0     0     0     0     0     3     0     0     0     0      Latin Extended-A         ĠġĢģĤĥĦħĨĩĪīĬĭĮį
  0x150     0     0     0     0     0     0     0     0     0     0     0     2     0     0     0     0      Latin Extended-A         ŐőŒœŔŕŖŗŘřŚśŜŝŞş
```

```
  0x160     0    3    0    0    0    0    0    0    0    0    0    0    0    0    0    0    Latin Extended-A          ŠšŢţŤťŦŧŨũŪūŬŭŮů
  0x170     0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    Latin Extended-A          ŰűŲųŴŵŶŷŸŹźŻżŽžſ
  0x250     0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    IPA Extensions            ɐɑɒɔɕɖɗɘəɚɛɜɞɟɠɡɢɨ
  0x390     0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    Greek and Coptic          ΆΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟ
  0x3a0     0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    Greek and Coptic          ΠΡ□ΣΤΥΦΧΨΩΪΫάέήί
  0x3b0     0    0    0    0    0    0    0    1    0    1    1    2    0    1    0    0    Greek and Coptic          ΰαβγδεζηθικλμνξο
  0x410     0    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    Cyrillic                  АБВГДЕЖЗИЙКЛМНОП
  0x420     1    1    0    1    0    0    0    1    0    0    0    0    0    0    0    0    Cyrillic                  РСТУФХЦЧШЩЪЫЬЭЮЯ
  0x430     5    0    1    1    0    0    0    0    3    1    5    2    0    1    0    1    Cyrillic                  абвгдежзийклмноп
  0x440     3    5    0    1    0    0    0    0    1    0    1    0    1    0    0    0    Cyrillic                  рстуфхцчшщъыьэюя
  0x450     0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    Cyrillic                  ѐёђѓєѕіїјљњћќѝўџ
  0x4d0     0    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    Cyrillic                  ĂăÄäÆæĔĕƏəÖöЖжZ̈z̈
  0x5d0     0    1    0    0    0    0    0    0    0    0    1    0    0    0    0    0    Hebrew                    אבגדהוחטיךכלמן
  0x5e0     0    0    1    0    0    0    0    0    0    1    0    1    0    0    0    0    Hebrew                    □□□□□נסעפץצקרשת
  0x620     0    0    0    0    0    0    0    5    1    2    2    0    0    2    0    1    Arabic                    ،ةحخدذآأؤإئابةتثجخ
  0x630     0    2    0    1    0    1    0    0    0    1    1    0    0    0    0    0    Arabic                    ذرزسشصضطظعغڪڪتټ
  0x640     0    2    0    0    4    1    0    0    3    0    3    0    0    0    0    0    Arabic                    ـفقكلمنهوىيً
  0x6c0     0    0    0    0    0    0    0    0    0    0    0    0    3    0    0    0    Arabic                    ۀہۂۃۄۅۆۇۈۉۊۋیۍۏ
  0x6f0     0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    Arabic                    ۰۱۲۳۴۵۶۷۸۹ۺۻۼ
  0x910     0    0    0    0    0    4    0    0    0    0    0    0    0    0    0    0    Devanagari                ऐऑओऔकखगघङचछजझञट
  0x920     0    4    0    0    0    0    0    0    4    0    0    0    0    0    0    4    Devanagari                ठडढणतथदधनऩपफबभमय
  0x930     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    4    Devanagari                ररलळऴवशषसहॉ◌ऽ◌ि
  0x940     0    0    4    0    0    0    0    0    0    0    0    4    0    0    0    0    Devanagari                ◌ी◌ु◌ू◌ृ◌ॄ◌ॅ◌ॆ◌े◌ै◌ॉ◌ॊ◌ो◌ौ
  0xd00     0    0    1    0    0    0    0    0    0    0    0    0    0    0    1    0    Malayalam                 □□ംഃ□അആഇഈഉഊഋഌ□എഏ
  0xd10     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    Malayalam                 ഐ□ഒഓഔകഖഗഘങചഛജഝഞടഠ
  0xd20     0    0    0    0    0    0    0    0    0    0    0    1    0    0    1    2    Malayalam                 ഡഢണതഥദധന□പഫബഭമയ
  0xd30     0    0    1    1    0    0    0    0    0    0    0    0    0    0    1    1    Malayalam                 രറലളഴവശഷസഹ□□□ാിീ
  0xd40     0    1    0    0    0    0    0    0    0    0    0    0    0    1    0    0    Malayalam                 ീുൂൃൄ□െേൈ□ൊോൌ്□□
 0x1e40     0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    Latin Extended Additional ẀẁẂẃẄẅỲỳỴỵỸỹỾỿ
 0x1eb0     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    Latin Extended Additional ẰằẲẳẴẵẶặẸẹẺẻẼẽ
 0x1ec0     0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    Latin Extended Additional ỀềỂểỄễỆệỈỉỊịỌọỎỏ
 0x2000     0    0    0    0    0    0    0    0    0    1    0    0    1    0    0    0    General Punctuation
 0x2010     1    1    0   49    4    0    0    0    0    0    0    0    0    0    0    0    General Punctuation       ------‖_‘’,‘“”„‟
 0x2020     0    0   64    0    0    0    0    0    0    0    0    0    0    0    0    0    General Punctuation       †‡●‣‥….·
 0x20a0     0    0    0    0    0    0    0    0    0    0    0    0    4    0    0    0    Currency Symbols          ₲₡₵₣₤₥₦₧₨₩₪₫€₭₮₱₽₹₺₼₾₿
 0x4e20     0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    CJK Unified Ideographs    丄丏丐丙严並丧丨丩个丫丬中乳乹
 0x6580     0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    CJK Unified Ideographs    斀斁斂斒斢斨斪斲文孛斉斊斋斌斠斎斛
 0x65e0     0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    CJK Unified Ideographs    无旡既旣日旦旧旨早旪旫旬旭旮旻
 0x6720     0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    CJK Unified Ideographs    朘朡望朧朋勝朦朧木朩未末本札朮术
 0x8a90     0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    CJK Unified Ideographs    誀誑誒誓誕誕誶誷誘誣誚誤諓諂語誶
 0xad60     0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    Hangul Syllables          곰곱곲곳공곶곷곸곹곺곻구국굮굯
 0xc5b0     0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    Hangul Syllables          얠얱얤얥어억얶얻얽얾얿얻얽얾얿
 0xd550     0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    Hangul Syllables          핑핒핓핔핕핖핗하학핚핛한핝핞핟
 0xfff0     0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    Specials                  ￰￱￲
0x10340     0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    Gothic                    𐍀𐍁𐍂𐍃𐍄𐍅𐍆𐍇𐍈𐍉𐍊𐍋𐍌𐍍𐍎𐍏

total bytes      : 269811
total characters : 269307
```

which shows a wide range of unicode blocks. Wikipedia pages comprise very good test data for the tool because most pages include a list of links to translations in other languages, shown in the alphabets of those languages, thus contributing at least a few Unicode characters from each. Notice that the last character line shows 'Gothic' characters from a code block with a start address greater than 0xffff.

The standard Unicode Basic Multilingual Plane (BMP) of 65,312 characters (out of 65,636 possible values) is insufficient to include all Unicode characters. Python3 on OSX is compiled to support 32-bit Unicode characters internally, so can handle characters from the Supplementary Multilingual Plane (SMP) and Supplementary Ideographic Plane (SIP) correctly. However, Python2 supports only the Basic Multilingual Plane and converts any out-of-range unicode to pairs of 16-bit values (surrogate pairs). Running the same

version of the analysis program with the python 2.7 interpreter shows (the unchanged middle part of the output is removed here for brevity):

```
python2 utfa.py -u http://en.wikipedia.org/wiki/UTF-8

unicode analysis for http://en.wikipedia.org/wiki/UTF-8     (running Python 2 : showing Basic Multilingual Plane including surrogate pairs)

            0     1     2     3     4     5     6     7     8     9     a     b     c     d     e     f      -- unicode block --      -- characters --

      0     0     0     0     0     0     0     0     0     0  1827  3029     0     0     0     0     0     Basic Latin              ................
      1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     Basic Latin              ................
 ...
 0x20a0     0     0     0     0     0     0     0     0     0     0     0     0     4     0     0     0     Currency Symbols         ₠₡₢₣₤₥₦₧₨₩₪₫€₭₮₯₰
 0x4e20     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     CJK Unified Ideographs   丠両丢丣两严並丧丨丩个丫丬中丮丯
 0x6580     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     CJK Unified Ideographs   歀歁歂歃歄歅歆歇歈歉歊歋文歍歎歏歐歑歒歓歔歕歖歗歘歙歚歛歜歝歞歟歠歡止正歨歩歪歫歬歭歮歯歰歱歲歳歴歵歶歷歸歹歺
 0x65e0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     CJK Unified Ideographs   无旡既旣旦旧旨早旪旫旬旭旮旯
 0x6720     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     CJK Unified Ideographs   朠朡朢朣朤朥朦朧木朩未末本札朮术
 0x8a90     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     CJK Unified Ideographs   誀誁誂誃誄誅誆誇誈誉誊誋誌認誎誏誐誑誒誓誔誕誖誗誘誙誚誛誜誝語誟誠誡誢誣誤誥誦誧誨誩說誫説読誮誯誰誱課
 0xad60     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     Hangul Syllables         곰곱곲곳곴공곶곷곸곹곺곻과곽곾곿관괁괂괃괄괅괆괇괈괉괊괋괌괍괎괏괐광괒괓괔괕괖괗괘괙괚괛괜괝괞괟괠괡괢괣괤괥괦괧괨괩괪괫괬괭괮괯괰괱괲괳괴괵괶괷괸괹괺괻괼괽괾괿굀굁굂굃굄굅굆굇굈굉굊굋굌굍굎굏교굑굒굓굔굕굖굗굘굙
 0xc5b0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     Hangul Syllables         얰얱얲얳어억얶얷언얹얺얻얼얽얾엀엁엂엃엄업없엇었엉엊엋엌엍엎엏에엑엒엓엔엕엖엗엘엙엚엛엜엝엞엟엠엡엢엣엤엥엦엧엨엩엪엫엮엯연엱엲엳열엵엶엷엸엹엺엻염엽엾엿였
 0xd550     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     Hangul Syllables         핑핒핓핔필핖핗하학핚핛한핝할
 0xd800     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     High Surrogates          ????????????????????????????????????????????????
 0xdf40     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     Low Surrogates           ????????????????????????????????????????????????
 0xfff0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     Specials                 ￰￱￲

total bytes      : 269811
total characters : 269308
```

Now all characters are part of BMP and there are 'High Surrogates' and 'Low Surrogates' which were not present in the python3 version.

Note that the first line of output displays the changed python version and character support.

It is possible to ask the program to show relevant version information (flag -v). Compare:

```
python3 utfa.py -v

sys.version_info(major=3, minor=4, micro=3, releaselevel='final', serial=0)
deutfault encoding utf-8
maximum Unicode code point 0x10ffff
no filename given


python2 utfa.py -v

sys.version_info(major=2, minor=7, micro=9, releaselevel='final', serial=0)
deutfault encoding ascii
maximum Unicode code point 0xffff
no filename given
```

Another Wikipedia example, showing many character sets:

```
utfa.py -u http://en.wikipedia.org/wiki/Armenian_alphabet
```

unicode analysis for http://en.wikipedia.org/wiki/Armenian_alphabet (running Python 3)

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f	-- unicode block --	-- characters --
0	0	0	0	0	0	0	0	0	0	1918	2995	0	0	0	0	0	Basic Latin	.....
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Basic Latin	.....
2	12751	13	9734	182	2	1582	174	76	359	357	11	174	404	1783	1316	7342	Basic Latin	!"#$%&'()*+,-./
3	1215	821	852	481	272	745	312	160	501	335	1413	724	8289	4840	8289	44	Basic Latin	0123456789:;<=>?
4	0	2233	793	680	650	718	282	221	226	836	42	165	450	449	404	164	Basic Latin	@ABCDEFGHIJKLMNO
5	763	19	493	511	573	309	68	174	9	50	45	85	25	85	23	1598	Basic Latin	PQRSTUVWXYZ[\]^_
6	0	14164	2236	4326	5141	14647	2605	2616	4854	15173	142	2296	10033	2858	9921	6484	Basic Latin	`abcdefghijklmno
7	4587	51	8614	8017	14096	2179	1133	2768	962	1663	223	22	0	22	0	0	Basic Latin	pqrstuvwxyz{|}~
0xa0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	Latin-1 Supplement	¡¢£¤¥¦§¨©ª«¬®¯
0xb0	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	Latin-1 Supplement	°±²³´µ¶·¸¹º»¼½¾¿
0xd0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Latin-1 Supplement	ÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
0xe0	2	1	1	0	0	2	0	6	0	15	5	3	0	3	0	0	Latin-1 Supplement	àáâãäåæçèéêëìíîï
0xf0	0	3	6	2	2	0	0	0	0	1	1	0	12	0	0	0	Latin-1 Supplement	ðñòóôõö÷øùúûüýþÿ
0x100	0	20	0	1	0	0	0	0	0	0	0	0	2	18	0	0	Latin Extended-A	ĀāĂăĄąĆćĈĉĊċČčĎď
0x110	0	0	0	15	0	0	0	1	0	0	0	0	0	0	0	2	Latin Extended-A	ĐđĒēĔĕĖėĘęĚěĜĝĞğ
0x120	0	3	0	0	0	0	0	0	0	0	0	11	0	0	0	0	Latin Extended-A	ĠġĢģĤĥĦħĨĩĪīĬĭĮį
0x130	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Latin Extended-A	İıIJijĴĵĶķĸĹĺĻļĽľĿ
0x140	0	0	2	0	1	0	1	0	0	0	0	0	0	6	0	0	Latin Extended-A	ŀŁłŃńŅņŇňŉŊŋŌōŎŏ
0x150	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Latin Extended-A	ŐőŒœŔŕŖŗŘřŚśŜŝŞş
0x160	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Latin Extended-A	ŠšŢţŤťŦŧŨũŪūŬŭŮů
0x170	0	0	0	2	0	0	0	0	0	0	0	0	0	0	5	0	Latin Extended-A	ŰűŲųŴŵŶŷŸŹźŻżŽžſ
0x1f0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Latin Extended-B	ǰDZDzdzǴǵHƕPǸǹÅåÆæØø
0x250	0	16	0	0	0	0	0	0	0	9	0	34	0	0	0	0	IPA Extensions	ɐɑɒɔɕɖɗɘəɚɛɜɝɞɟ
0x260	0	4	0	0	0	0	0	0	0	0	0	2	0	0	0	0	IPA Extensions	ɠɡɢɣɤɥɦɧɨɪɫɬɭɮ
0x270	0	0	0	0	0	0	0	0	0	6	0	0	0	0	6	0	IPA Extensions	ɯɰɱɲɳɴɵɶɷɸɹɺɻɼɽ
0x280	0	3	0	10	0	0	0	0	0	0	0	0	0	0	0	7	IPA Extensions	ʀʁʂʃʄʅʆʇʈʉʊʋʌʍʎʏ
0x290	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	IPA Extensions	ʐʑʒʓʔʕʖʗʘʙʚʛʜʝʞʟ
0x2b0	27	0	0	0	0	0	0	0	0	0	0	1	0	0	1	7	Spacing Modifier Letters	ʰʱʲʳʴʵʶʷʸʹʺʻʼʽʾʿˀˁ
0x2d0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Spacing Modifier Letters	ːˑ˒˓˔˕˖˗˘˙˚˛˜˝˞˟ˠˡˢˣˤ
0x300	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Combining Diacritical Marks	◌̀◌́◌̂◌̃◌̄◌̅◌̆◌̇◌̈◌̉◌̊◌̋◌̌◌̍◌̎◌̏
0x320	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	Combining Diacritical Marks	◌̠◌̡◌̢◌̣◌̤◌̥◌̦◌̧◌̨◌̩◌̪◌̫◌̬◌̭◌̮◌̯
0x3c0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	Greek and Coptic	πρςστυφχψωϊϋόύώϗ
0x400	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Cyrillic	ЀЁЂЃЄЅІЇЈЉЊЋЌЍЎЏ
0x410	4	1	1	0	0	1	0	0	0	0	1	0	1	0	0	0	Cyrillic	АБВГДЕЖЗИЙКЛМНОП
0x420	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	Cyrillic	РСТУФХЦЧШЩЪЫЬЭЮЯ
0x430	26	5	6	3	2	13	0	4	18	3	24	4	10	17	17	4	Cyrillic	абвгдежзийклмноп
0x440	16	26	4	5	3	1	1	1	0	0	1	2	3	0	0	3	Cyrillic	рстуфхцчшщъыьэюя
0x450	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	Cyrillic	ѐёђѓєѕіїјљњћќѝўџ
0x4b0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Cyrillic	ҰұҲҳҴҵҶҷҸҹҺһҼҽҾҿ
0x4e0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	Cyrillic	ӡӢӣӤӥӦӧӨөӪӫӬӭӮӯ
0x530	0	7	5	4	4	4	3	3	3	3	3	3	4	3	3	3	Armenian	ՀԱԲԳԴԵԶԷԸԹԺԻԼԽԾԿ
0x540	10	3	3	4	4	3	3	3	8	3	3	3	3	3	4	4	Armenian	ՁՂՃՄՅՆՇՈՉՊՋՌՍՎՏ
0x550	3	3	7	4	4	3	3	0	0	1	2	3	3	2	3	4	Armenian	ՐՑՒՓՔՕՖ◌՚՛՜՝՞
0x560	0	70	20	9	8	50	13	16	9	15	7	31	12	10	9	10	Armenian	◌աբգդեզէըթժիլխծկ
0x570	11	7	8	10	18	26	47	7	47	9	9	8	7	18	13	18	Armenian	հձղճմյնշոչպջռսվտ
0x580	32	12	60	7	13	15	9	9	0	3	3	0	0	2	2	2	Armenian	րցւփքօֆև։
0x620	0	0	0	2	0	0	0	5	3	3	0	0	1	0	0	1	Arabic	ؠءآأؤإئابةتثجحخدذ
0x630	0	4	0	1	0	0	0	0	0	1	0	0	0	0	0	0	Arabic	رزسشصضطظعغـفقكلمن
0x640	0	2	0	0	2	2	2	0	0	0	3	0	0	0	0	0	Arabic	هوىيًٌٍَُِّْ
0x6c0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	Arabic	ۀہۂۃۄۅۆۇۈۉۊۋیۍێۏ
0xc80	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Kannada	◌ಀ◌ಁಂಃ◌ಅಆಇಈಉಊಋಌ◌ಎಏ
0xc90	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	Kannada	ಐ◌ಒಓಔಕಖಗಘಙಚಛಜಝಞಟಠ
0xca0	0	1	0	0	0	0	0	0	2	0	1	0	0	0	1	1	Kannada	ಡಢಣತಥದಧನ◌ಪಫಬಭಮಯ

0xcb0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	3	Kannada	ರಱಲಳ□ವಶಷಸಹ□□಼ಽಾಿ
0xcc0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	Kannada	ೀುೂೃೄ□ೆೇೈ□ೊೋೌ್□□
0xe00	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Thai	กขฃคฅฆงจฉชซฌญฎฏ
0xe10	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	Thai	ฐฑฒณดตถทธนบปผฝพฟ
0xe20	0	1	2	2	0	0	0	0	0	1	0	0	0	2	0	0	Thai	ภมยรฤลฦวศษสหฬอฮฯ
0xe30	0	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0	Thai	ะัาำิีึืฺุู□□□□฿
0xe40	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Thai	เแโใไๅๆ็่้๊๋์ํ๎๏
0x10d0	3	1	0	1	1	0	0	1	2	0	2	2	0	2	0	0	Georgian	აბგდევზთიკლმნოპჟ
0x10e0	3	1	0	2	0	1	0	0	0	0	0	0	1	0	1	0	Georgian	რსტუფქღყშჩცძწჭხჯ
0x1e00	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	Latin Extended Additional	ḀḁḂḃḄḅḆḇḈḉḊḋḌḍḎḏ
0x1e20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Latin Extended Additional	ḠḡḢḣḤḥḦḧḨḩḪḫḬḭḮḯ
0x1e40	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	0	Latin Extended Additional	ṀṁṂṃṄṅṆṇṈṉṊṋṌṍṎṏ
0x1e50	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	Latin Extended Additional	ṐṑṒṓṔṕṖṗṘṙṚṛṜṝṞṟ
0x1e60	0	0	0	2	0	0	0	0	0	0	0	0	0	4	0	0	Latin Extended Additional	ṠṡṢṣṤṥṦṧṨṩṪṫṬṭṮṯ
0x1ee0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	Latin Extended Additional	ỠỡỢợỤụỦủỨứỪừỬửỮữ
0x2010	1	0	0	77	3	0	0	0	1	10	0	5	6	6	0	0	General Punctuation	‐‑‒–—―‖‗‘’‚‛“”„‟
0x2020	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	General Punctuation	†‡•‣․‥…‧
0x27e0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	Miscellaneous Mathematical Symbols-A	⟠⟡⟢⟣⟤⟥⟦⟧⟨⟩⟪⟫⟬⟭⟮⟯
0x2800	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	Braille Patterns	⠀⠁⠂⠃⠄⠅⠆⠇⠈⠉⠊⠋⠌⠍⠎⠏
0x2810	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Braille Patterns	⠐⠑⠒⠓⠔⠕⠖⠗⠘⠙⠚⠛⠜⠝⠞⠟
0x30a0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	Katakana	゠ァアィイゥウェエォオカガキギク
0x30c0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Katakana	ダチヂッツヅテデトドナニヌネノハ
0x30e0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Katakana	ムメモャヤュユョヨラリルレロヮワ
0x4e20	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	CJK Unified Ideographs	丠両丢丣两严並丧丨丩个丫丬中丮丯
0x4e90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	CJK Unified Ideographs	亐云互亓五井亖亗亘亙亚些亜亝亞亟
0x5b50	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	CJK Unified Ideographs	子孑孒孓孔孕孖字存孙孚孛孜孝孞孟
0x5c30	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	CJK Unified Ideographs	尰就尲尳尴尵尶尷尸尹尺尻尼尽尾尿
0x6580	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	CJK Unified Ideographs	斀斁斂斃斄斅斆文斈斉斊斋斌斍斎斏
0x65e0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	CJK Unified Ideographs	无旡既旣旤日旦旧旨早旪旫旬旭旮旯
0x6720	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	CJK Unified Ideographs	朠朡朢朣朤朥朦朧木朩未末本札朮术
0x6bc0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	CJK Unified Ideographs	毀毁毂毃毄毅毆毇毈毉毊毋毌母毎每
0x7f80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	CJK Unified Ideographs	羀羁羂羃羄羅羆羇羈羉羊羋羌羍美羏
0x8a90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	CJK Unified Ideographs	誐誑誒誓誔誕誖誗誘誙誚誛誜誝語誟
0xad60	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	Hangul Syllables	굠굡굢굣굤굥굦굧굨굩굪굫구국굮굯
0xb2c0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Hangul Syllables	늸늹늺늻늼늽늾늿니닉닊닋닌닍닎닏
0xb970	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Hangul Syllables	륔륕륖륗르륵륶륷른륹륺륻를륽륾륿
0xba50	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Hangul Syllables	멐멑멒멓메멕멖멗멘멙멚멛멜멝멞멟
0xbb30	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Hangul Syllables	묔묕묖묗무묵묶묷문묹묺묻물묽묾묿
0xc540	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	Hangul Syllables	씤씥씦씧아악앆앇안앉않앋알앍앎앏
0xc5b0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Hangul Syllables	얰얱얲얳어억얶얷언얹얺얻얼얽얾얿
0xc790	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hangul Syllables	자작잒잓잔잕잖잗잘잙잚잛잜잝잞잟
0xd550	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Hangul Syllables	핐핑핒핓핔핕핖핗하학핚핛한핝핞핟
0xfb10	0	0	0	2	2	2	2	2	0	0	0	0	0	0	0	0	Alphabetic Presentation Forms	□□□ﬓﬔﬕﬖﬗ□□□□□יִﬞײַ

```
total bytes      : 229138
total characters : 227173
```

Here there are three Braille characters (for 'b', 'r' and 'l'). They are a little hard to discover on the web page but can be found by expanding 'Type of writing systems' at the foot of the page. <brl> is a frequently-used braille contraction of the name 'braille' itself.

A quick summary of characters on a BBC web page:

```
utfa.py -u http://bbc.co.uk -xm
```

```
unicode analysis for http://bbc.co.uk   (running Python 3)

       total -- unicode block --       -- characters --

      0    241      Basic Latin                         .  .
      1      0      Basic Latin
      2  25365      Basic Latin               !"#$%&'()*+,-./
      3  11273      Basic Latin              0123456789:;<=>?
      4   1767      Basic Latin               ABCDEFGHIJKLMNO
      5   1605      Basic Latin              PQRSTUVWXY [\]^_
      6  41106      Basic Latin               abcdefghijklmno
      7  24356      Basic Latin              pqrstuvwxyz{|}
   0xa0      2      Latin-1 Supplement          £
 0x2010      4      General Punctuation               ’
 0x2020      3      General Punctuation            …

total bytes      : 105738
total characters : 105722
```

shows that almost all characters are 'Basic Latin' except for a few pound signs and some 'General punctuation' marks.

There are clearly both new line and carriage return characters, but not in equal numbers (241 is odd). The full breakdown:

```
utfa.py -u http://bbc.co.uk

unicode analysis for http://bbc.co.uk   (running Python 3)

             0     1     2     3     4     5     6     7     8     9     a     b     c     d     e     f      -- unicode block --      -- characters --

      0      0     0     0     0     0     0     0     0     0     0   179     0     0    62     0     0      Basic Latin              .................
      1      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0      Basic Latin              .................
      2  10491   170  2131     6     8    20  2434   872   771   771   169   172   435  1187  2372  3356      Basic Latin               !"#$%&'()*+,-./
      3    519   487   554   486   431   271   220   271   226   289   811  2711  1090  1737  1096    74      Basic Latin              0123456789:;<=>?
      4      0   208   211   285   128   221    73    47    48   105    17    40    73   130   105    76      Basic Latin              @ABCDEFGHIJKLMNO
      5     99     2    76   147   184    81    40    48     4    25     0   198    97   198    15   391      Basic Latin              PQRSTUVWXYZ[\]^_
      6      0  4383  2035  3573  2479  5323  1179  1867  1911  4466   198   839  3359  1371  3197  4926      Basic Latin              `abcdefghijklmno
      7   1969   892  3443  3907  7180  2294   987  1606   281   733   187   390    95   392     0     0      Basic Latin              pqrstuvwxyz{|}~
   0xa0      0     0     0     2     0     0     0     0     0     0     0     0     0     0     0     0      Latin-1 Supplement        ¡¢£¤¥¦§¨©ª«¬®¯
 0x2010      0     0     0     0     0     0     0     0     0     4     0     0     0     0     0     0      General Punctuation      ------‖‗‘’‚‛“”„‟
 0x2020      0     0     0     0     0     0     3     0     0     0     0     0     0     0     0     0      General Punctuation      †‡•‣‥…‧·


total bytes      : 105738
total characters : 105722
```

shows many more new line characters (179) than carriage returns (62). This is a little unusual because most DOS-format files contain these in pairs.

Command line options

Here are the command line controls for utfa.py:

```
utfa.py -h
usage: utfa [-h] [-u URL] [-b] [-e] [-t] [-l] [-n] [-m] [-x] [-o OFFSET]
          [-s SIZE] [-v]
          [file]

file analysis

positional arguments:
  file

optional arguments:
  -h, --help            show this help message and exit
  -u URL, --url URL     url e.g. -u http://bbc.co.uk
  -b, --bytes           analyse 256 possible bit patterns
  -e, --errors          enable unicode error replacement
  -t, --time            show run time
  -l, --legend          disable legend display
  -n, --name            disable unicode block names
  -m, --mask            show only characters present in file/url
  -x, --number          disable individual counts
  -o OFFSET, --offset OFFSET
  -s SIZE, --size SIZE
  -v, --version         python version, encoding and max code point
```