NAME

unicharambigs − Tesseract unicharset ambiguities

DESCRIPTION

The unicharambigs file (a component of traineddata, see combine_tessdata(1)) is used by Tesseract to represent possible ambiguities between characters, or groups of characters.

The file contains a number of lines, laid out as follow:

[num] <TAB> [char(s)] <TAB> [num] <TAB> [char(s)] <TAB> [num]

Field one   the number of characters contained in field two

Field two   the character sequence to be replaced

Field three   the number of characters contained in field four

Field four   the character sequence used to replace field two

Field five   contains either 1 or 0. 1 denotes a mandatory replacement, 0 denotes an optional replacement.

Characters appearing in fields two and four should appear in unicharset. The numbers in fields one and three refer to the number of unichars (not bytes).

EXAMPLE

```
v1
2     ' '    1      "     1
1     m      2      r n   0
3     i i i  1      m     0
```

The first line is a version identifier. In this example, all instances of the *2* character sequence *''* will **always** be replaced by the *1* character sequence *"*; a *1* character sequence *m* **may** be replaced by the *2* character sequence *rn*, and the *3* character sequence **may** be replaced by the *1* character sequence *m*.

Version 3.03 and on supports a new, simpler format for the unicharambigs file:

```
v2
" " 1
m rn 0
iii m 0
```

In this format, the "error" and "correction" are simple UTF−8 strings separated by a space, and, after another space, the same type specifier as v1 (0 for optional and 1 for mandatory substitution). Note the downside of this simpler format is that Tesseract has to encode the UTF−8 strings into the components of the unicharset. In complex scripts, this encoding may be ambiguous. In this case, the encoding is chosen such as to use the least UTF−8 characters for each component, ie the shortest unicharset components will make up the encoding.

HISTORY

The unicharambigs file first appeared in Tesseract 3.00; prior to that, a similar format, called DangAmbigs (*dangerous ambiguities*) was used: the format was almost identical, except only mandatory replacements could be specified, and field 5 was absent.

BUGS

This is a documentation "bug": it’s not currently clear what should be done in the case of ligatures (such as *fi*) which may also appear as regular letters in the unicharset.

SEE ALSO

tesseract(1), unicharset(5)
https://github.com/tesseract−ocr/tesseract/wiki/Training−Tesseract−3.03%E2%80%933.05#the−unicharambigs−fi

AUTHOR

The Tesseract OCR engine was written by Ray Smith and his research groups at Hewlett Packard (1985−1995) and Google (2006−present).