cjklib Documentation

Release 0.3.2

Christoph Burgmer

July 02, 2013

CONTENTS

Cjklib provides language routines related to Han characters (characters based on Chinese characters named Hanzi, Kanji, Hanja and chu Han respectively) used in writing of the Chinese, the Japanese, infrequently the Korean and formerly the Vietnamese language(s). Functionality is included for character pronunciations, radicals, glyph components, stroke decomposition and variant information.

This document is about version 0.3.2, see http://cjklib.org/ for the newest and http://cjklib.org/current for the current development version. The project is hosted on http://code.google.com/p/cjklib. See http://characterdb.cjklib.org/ for a collaborative effort on gathering language data for cjklib.

Contents:

CHAPTER

ONE

DOWNLOADING & INSTALLING

cjklib has the following dependencies:

- Python 2.4 or above (currently no support for Python3)
- SQLite 3+
- SQLAlchemy 0.4.8+
- pysqlite2 (already ships with Python 2.5 and above)

Alternatively for MySQL as backend:

- MySQL 5+
- MySQL-Python

1.1 Windows

Download the `.exe` installer from the Python package index and run it.

Three scripts `cjknife.exe`, `buildcjkdb.exe`, and `installcjkdict.exe` will be added to the Python `Scripts` sub-directory. Make sure this directory is included in your `PATH` environment variable to access these programs from the command line.

CJK dictionaries are not included by default. If you want to install any of those run the following (with an Internet connection):

```
$ installcjkdict CEDICT
```

This will download CEDICT, create a SQLite database file and install it under the directory given by the `APPDATA` environment variable, e.g. `C:\windows\profiles\MY_USER\Application Data\cjklib`. Just substitute `CEDICT` for any other supported dictionary (i.e. EDICT, CEDICT, HanDeDict, CFDICT, CEDICTGR).

1.2 Unix

Get the source package from the Python package index and deploy the library on your system:

```
$ sudo python setup.py install
```

CJK dictionaries are not included by default. If you want to install any of those run the following (with an Internet connection):

```
$ sudo installcjkdict CEDICT
```

This will download CEDICT, create a SQLite database file and install it to `/usr/local/share/cjklib`. Just substitute `CEDICT` for any other supported dictionary (i.e. EDICT, CEDICT, HanDeDict, CFDICT, CEDICTGR).

1.3 Development version

The development version is available from svn:

```
$ git clone git://github.com/cburgmer/cjklib.git
```

You now need to generate the database. Download the Unihan database and call the build CLI (which is not yet installed as executable):

```
$ cd cjklib
$ wget ftp://ftp.unicode.org/Public/UNIDATA/Unihan.zip
$ python -m cjklib.build.cli build cjklibData --attach= \
    --database=sqlite:///cjklib/cjklib.db
$ sqlite3 cjklib/cjklib.db "VACUUM"
```

The last step is optional but will help to optimize the database file.

Install by running:

```
$ sudo python setup.py install
```

1.4 Database

Packaged versions of the library will ship with a pre-built SQLite database file. You can however easily rebuild the database yourself.

First download the newest Unihan file:

```
$ wget ftp://ftp.unicode.org/Public/UNIDATA/Unihan.zip
```

Then start the build process:

```
$ sudo buildcjkdb -r build cjklibData
```

1.4.1 SQLite

SQLite by default has no Unicode support for string operations. Optionally the ICU library can be compiled in for handling alphabetic non-ASCII characters. Cjklib can register own Unicode functions if ICU support is missing. Queries with `LIKE` will then use function `lower()`. This compatibility mode has negative impact on performance and as it is not needed for dictionaries like EDICT or CEDICT it is disabled by default. See `cjklib.conf` for enabling.

1.4.2 MySQL

With MySQL 5 the following `CREATE` command creates a database with `utf8` as character set using the general Unicode collation (MySQL from 5.5.3 on will support full Unicode given character set `utf8mb4` and collation `utf8mb4_bin`):

```
CREATE DATABASE cjklib DEFAULT CHARACTER SET utf8 COLLATE utf8_bin;
```

You might need to set access rights, too (substitute `user_name` and `host_name`):

```
GRANT ALL ON cjklib.* TO 'user_name'@'host_name';
```

Now update the settings in `cjklib.conf`.

MySQL < 5.5 doesn't support full UTF-8, and uses a version with max 3 bytes, so characters outside the Basic Multilingual Plane (BMP) can't be encoded. Building the Unihan database thus might result in warnings, characters above U+FFFF can't be built at all. You need to disable building the full character range by setting `wideBuild` to `False` in `cjklib.conf` before building. Alternatively pass `--wideBuild=False` to `buildcjkdb`.

CHAPTER

TWO

COMMAND LINE TOOLS

Contents:

2.1 cjknife — Command Line Interface

cjknife exposes most functions of the library to the command line.

2.1.1 Examples

Show character information:

```
$ cjknife -i
Information for character  (traditional locale, Unicode domain)
Unicode codepoint: 0x5468 (21608, character form)
Radical index: 30, radical form:
Stroke count: 8
Phonetic data (CantoneseYale): ju
Phonetic data (GR): jou
Phonetic data (Hangul):
Phonetic data (Jyutping): zau1
Phonetic data (MandarinBraille):
Phonetic data (MandarinIPA): tou
Phonetic data (Pinyin): zhu
Phonetic data (ShanghaineseIPA):
Phonetic data (WadeGiles): chou1
Semantic variants:
Glyph 0(*), stroke count: 8

Stroke order:  (SP-HZG H-S-H S-HZ-H)
```

Search the EDICT dictionary:

```
$ cjknife -w EDICT -x "knowledge"
 /(n) knowledge/
 /(n) knowledge/
 /(n) knowledge/
  /(n) learning/scholarship/erudition/knowledge/(P)/
  /(n) scholarship/learning/knowledge/
  /(n) scholarship/knowledge/literary ability/(P)/
  /(n) knowledge/information/(P)/
  /(n) human intellect/knowledge/
```

```
/(n) human intellect/knowledge/
/(n,vs) expertise/experience/knowledge/
/(n) knowledge/
/(n) knowledge/information/(P)/
/(n,vs) comprehension/knowledge/
/(n) sense/discretion/knowledge/
/(oK) (n) sense/discretion/knowledge/
```

See Also:

Screenshots Examples on the project's wiki.

2.1.2 Options

-i CHAR, **-information**=CHAR
: print information about the given char

-a READING, **-by-reading**=READING
: prints a list of characters for the given reading

-r CHARSTR, **-get-reading**=CHARSTR
: prints the reading for a given character string (for characters with multiple readings these are grouped in square brackets; shows the character itself if no reading information available)

-f CHARSTR, **-convert-form**=CHARSTR
: converts the given characters from/to Chinese simplified/traditional form (if ambiguous multiple characters are grouped in brackets)

-q CHARSTR
: performs commands -r and -f in one step

-k RADICALIDX, **-by-radicalidx**=RADICALIDX
: get all characters for a radical given by its index

-p CHARSTR, **-by-components**=CHARSTR
: get all characters that include all the chars contained in the given list as component

-m READING, **-convert-reading**=READING
: converts the given reading from the input reading to the output reading (compatibility needed)

-s SOURCE, **-source-reading**=SOURCE
: set given reading as input reading

-t TARGET, **-target-reading**=TARGET
: set given reading as output reading

-l LOCALE, **-locale**=LOCALE
: set locale, i.e. one character out of TCJKV

-d DOMAIN, **-domain**=DOMAIN
: set character domain, e.g. 'GB2312'

-L, -list-options
: list available options for parameters

-V, -version
: print version number and exit

-h, -help
: display this help and exit

-database=DATABASEURL
database url

-x SEARCHSTR
searches the dictionary (wildcards '_' and '%')

-w DICTIONARY, **-set-dictionary**=DICTIONARY
set dictionary

2.2 installcjkdict — Install dictionaries

installcjkdict downloads and installs a dictionary.

2.2.1 Examples

Download and install CEDICT to `$HOME/cjklib/` (Windows), `$HOME/.cjklib/` (Unix) or `$HOME/Library/Application Support/` (Mac OS X):

```
$ installcjkdict --local CEDICT
```

Download CFDICT:

```
$ installcjkdict --download CFDICT
Getting download page http://www.chinaboard.de/cfdict.php?mode=dl... done
Found version 2009-11-30
Downloading http://www.chinaboard.de/cfdict/cfdict-20091130.tar.bz2...
100% |################################################| Time: 00:00:00 193.85 B/s
Saved as cfdict-20091130.tar.bz2
```

2.2.2 Options

-version
show program's version number and exit

-h, -help
show this help message and exit

-f, -forceUpdate
install dictionary even if the version is older or equal

-prefix=PREFIX
installation prefix

-local
install to user directory

-download
download only

-targetName=TARGETNAME
target name of downloaded file (only with –download)

-targetPath=TARGETPATH
target directory of downloaded file (only with –download)

-q, -quiet
don't print anything on stdout

-database=URL
database url

-attach=URL
attachable databases

-registerUnicode=BOOL
register own Unicode functions if no ICU support available

Global builder options

-collation=VALUE
collation for dictionary entries

-enableFTS3=BOOL
enable SQLite full text search (FTS3)

-useCollation=BOOL
use collations for dictionary entries

2.3 buildcjkdb — Build database

buildcjkdb builds the database for the cjklib library. Example: `buildcjkdb build allAvail`.

Builders can be given specific options with format `--BuilderName-option` or `--TableName-option`, e.g. `--Unihan-wideBuild=yes`.

2.3.1 Options

-version
show program's version number and exit

-h, -help
show this help message and exit

-r, -rebuild
build tables even if they already exist

-d, -keepDepending
don't rebuild build-depends tables that are not given

-p BUILDER, **-prefer**=BUILDER
builder preferred where several provide the same table

-q, -quiet
don't print anything on stdout

-database=URL
database url

-attach=URL
attachable databases

-registerUnicode=BOOL
register own Unicode functions if no ICU support available

-ignoreConfig
ignore settings from cjklib.conf

Global builder options

-dataPath=VALUE
path to data files

-entrywise=BOOL
insert entries one at a time (for debugging)

-ignoreMissing=BOOL
ignore missing Unihan column and build empty table

-wideBuild=BOOL
include characters outside the Unicode BMP

-slimUnihanTable=BOOL
limit keys of Unihan table

-collation=VALUE
collation for dictionary entries

-enableFTS3=BOOL
enable SQLite full text search (FTS3)

-filePath=VALUE
file path including file name, overrides searching

-fileType=VALUE
file extension, overrides file type guessing

-useCollation=BOOL
use collations for dictionary entries

CHAPTER

THREE

REFERENCE

CHAPTER

FOUR

TO DO

CHAPTER

FIVE

EXAMPLES

Get characters by pronunciation (here: “” in Korean):

```
>>> from cjklib import characterlookup
>>> cjk = characterlookup.CharacterLookup('T')
>>> cjk.getCharactersForReading(u'', 'Hangul')
[u'', u'', u'', u'', u'', u'', u'', u'', u'', u'']
```

Get stroke order of characters:

```
>>> cjk.getStrokeOrder(u'')
[u'', u'', u'', u'', u'', u'', u'', u'', u'']
```

Convert pronunciation data (here from *Pinyin* to *IPA*):

```
>>> from cjklib.reading import ReadingFactory
>>> f = ReadingFactory()
>>> f.convert(u'losh', 'Pinyin', 'MandarinIPA')
u'lau.'
```

Access a dictionary (here using Jim Breen's EDICT):

```
>>> from cjklib.dictionary import EDICT
>>> d = EDICT()
>>> d.getForTranslation('Tokyo')
[EntryTuple(Headword=u'', Reading=u'', Translation=u'/(n) Tokyo (current capital of Japan)/(P)/'
```

CHAPTER

SIX

COPYRIGHT & LICENSE

CHAPTER

SEVEN

CONTACT

For help or discussions on cjklib, join cjklib-devel@googlegroups.com.

Please report bugs to the project's bug tracker.

CHAPTER

EIGHT

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*