chardet Documentation

Release 5.0.0

Mark Pilgrim, Dan Blanchard, Ian Cordasco

Jun 25, 2022

Contents

1 Documentation **3**
- 1.1 Frequently asked questions . . . 3
- 1.2 Supported encodings . . . 4
- 1.3 Usage . . . 5
- 1.4 How it works . . . 6
- 1.5 chardet . . . 8

2 Indices and tables **19**

Python Module Index **21**

Index **23**

Character encoding auto-detection in Python. As smart as your browser. Open source.

CHAPTER 1

Documentation

1.1 Frequently asked questions

1.1.1 What is character encoding?

When you think of "text", you probably think of "characters and symbols I see on my computer screen". But computers don't deal in characters and symbols; they deal in bits and bytes. Every piece of text you've ever seen on a computer screen is actually stored in a particular *character encoding*. There are many different character encodings, some optimized for particular languages like Russian or Chinese or English, and others that can be used for multiple languages. Very roughly speaking, the character encoding provides a mapping between the stuff you see on your screen and the stuff your computer actually stores in memory and on disk.

In reality, it's more complicated than that. Many characters are common to multiple encodings, but each encoding may use a different sequence of bytes to actually store those characters in memory or on disk. So you can think of the character encoding as a kind of decryption key for the text. Whenever someone gives you a sequence of bytes and claims it's "text", you need to know what character encoding they used so you can decode the bytes into characters and display them (or process them, or whatever).

1.1.2 What is character encoding auto-detection?

It means taking a sequence of bytes in an unknown character encoding, and attempting to determine the encoding so you can read the text. It's like cracking a code when you don't have the decryption key.

1.1.3 Isn't that impossible?

In general, yes. However, some encodings are optimized for specific languages, and languages are not random. Some character sequences pop up all the time, while other sequences make no sense. A person fluent in English who opens a newspaper and finds "txzqJv 2!dasd0a QqdKjvz" will instantly recognize that that isn't English (even though it is composed entirely of English letters). By studying lots of "typical" text, a computer algorithm can simulate this kind of fluency and make an educated guess about a text's language.

In other words, encoding detection is really language detection, combined with knowledge of which languages tend to use which character encodings.

1.1.4 Who wrote this detection algorithm?

This library is a port of the auto-detection code in Mozilla. I have attempted to maintain as much of the original structure as possible (mostly for selfish reasons, to make it easier to maintain the port as the original code evolves). I have also retained the original authors' comments, which are quite extensive and informative.

You may also be interested in the research paper which led to the Mozilla implementation, A composite approach to language/encoding detection.

1.1.5 Yippie! Screw the standards, I'll just auto-detect everything!

Don't do that. Virtually every format and protocol contains a method for specifying character encoding.

- HTTP can define a `charset` parameter in the `Content-type` header.
- HTML documents can define a `<meta http-equiv="content-type">` element in the `<head>` of a web page.
- XML documents can define an `encoding` attribute in the XML prolog.

If text comes with explicit character encoding information, you should use it. If the text has no explicit information, but the relevant standard defines a default encoding, you should use that. (This is harder than it sounds, because standards can overlap. If you fetch an XML document over HTTP, you need to support both standards *and* figure out which one wins if they give you conflicting information.)

Despite the complexity, it's worthwhile to follow standards and respect explicit character encoding information. It will almost certainly be faster and more accurate than trying to auto-detect the encoding. It will also make the world a better place, since your program will interoperate with other programs that follow the same standards.

1.1.6 Why bother with auto-detection if it's slow, inaccurate, and non-standard?

Sometimes you receive text with verifiably inaccurate encoding information. Or text without any encoding information, and the specified default encoding doesn't work. There are also some poorly designed standards that have no way to specify encoding at all.

If following the relevant standards gets you nowhere, *and* you decide that processing the text is more important than maintaining interoperability, then you can try to auto-detect the character encoding as a last resort. An example is my Universal Feed Parser, which calls this auto-detection library only after exhausting all other options.

1.2 Supported encodings

Universal Encoding Detector currently supports over two dozen character encodings.

- `Big5`, `GB2312`/`GB18030`, `EUC-TW`, `HZ-GB-2312`, and `ISO-2022-CN` (Traditional and Simplified Chinese)
- `EUC-JP`, `SHIFT_JIS`, and `ISO-2022-JP` (Japanese)
- `EUC-KR` and `ISO-2022-KR` (Korean)
- `KOI8-R`, `MacCyrillic`, `IBM855`, `IBM866`, `ISO-8859-5`, and `windows-1251` (Russian)
- `ISO-8859-2` and `windows-1250` (Hungarian)

- ISO-8859-5 and windows-1251 (Bulgarian)
- ISO-8859-1 and windows-1252 (Western European languages)
- ISO-8859-7 and windows-1253 (Greek)
- ISO-8859-8 and windows-1255 (Visual and Logical Hebrew)
- TIS-620 (Thai)
- UTF-32 BE, LE, 3412-ordered, or 2143-ordered (with a BOM)
- UTF-16 BE or LE (with a BOM)
- UTF-8 (with or without a BOM)
- ASCII

Warning: Due to inherent similarities between certain encodings, some encodings may be detected incorrectly. In my tests, the most problematic case was Hungarian text encoded as ISO-8859-2 or windows-1250 (encoded as one but reported as the other). Also, Greek text encoded as ISO-8859-7 was often mis-reported as ISO-8859-2. Your mileage may vary.

1.3 Usage

1.3.1 Basic usage

The easiest way to use the Universal Encoding Detector library is with the detect function.

1.3.2 Example: Using the detect function

The detect function takes one argument, a non-Unicode string. It returns a dictionary containing the auto-detected character encoding and a confidence level from 0 to 1.

```
>>> import urllib.request
>>> rawdata = urllib.request.urlopen('http://yahoo.co.jp/').read()
>>> import chardet
>>> chardet.detect(rawdata)
{'encoding': 'EUC-JP', 'confidence': 0.99}
```

1.3.3 Advanced usage

If you're dealing with a large amount of text, you can call the Universal Encoding Detector library incrementally, and it will stop as soon as it is confident enough to report its results.

Create a UniversalDetector object, then call its feed method repeatedly with each block of text. If the detector reaches a minimum threshold of confidence, it will set detector.done to True.

Once you've exhausted the source text, call detector.close(), which will do some final calculations in case the detector didn't hit its minimum confidence threshold earlier. Then detector.result will be a dictionary containing the auto-detected character encoding and confidence level (the same as the chardet.detect function returns).

1.3.4 Example: Detecting encoding incrementally

```
import urllib.request
from chardet.universaldetector import UniversalDetector

usock = urllib.request.urlopen('http://yahoo.co.jp/')
detector = UniversalDetector()
for line in usock.readlines():
    detector.feed(line)
    if detector.done: break
detector.close()
usock.close()
print(detector.result)
```

```
{'encoding': 'EUC-JP', 'confidence': 0.99}
```

If you want to detect the encoding of multiple texts (such as separate files), you can re-use a single `UniversalDetector` object. Just call `detector.reset()` at the start of each file, call `detector.feed` as many times as you like, and then call `detector.close()` and check the `detector.result` dictionary for the file's results.

1.3.5 Example: Detecting encodings of multiple files

```
import glob
from chardet.universaldetector import UniversalDetector

detector = UniversalDetector()
for filename in glob.glob('*.xml'):
    print(filename.ljust(60), end='')
    detector.reset()
    for line in open(filename, 'rb'):
        detector.feed(line)
        if detector.done: break
    detector.close()
    print(detector.result)
```

1.4 How it works

This is a brief guide to navigating the code itself.

First, you should read A composite approach to language/encoding detection, which explains the detection algorithm and how it was derived. This will help you later when you stumble across the huge character frequency distribution tables like `big5freq.py` and language models like `langcyrillicmodel.py`.

The main entry point for the detection algorithm is `universaldetector.py`, which has one class, `UniversalDetector`. (You might think the main entry point is the `detect` function in `chardet/__init__.py`, but that's really just a convenience function that creates a `UniversalDetector` object, calls it, and returns its result.)

There are 5 categories of encodings that `UniversalDetector` handles:

1. `UTF-n` with a BOM. This includes `UTF-8`, both BE and LE variants of `UTF-16`, and all 4 byte-order variants of `UTF-32`.

2. Escaped encodings, which are entirely 7-bit ASCII compatible, where non-ASCII characters start with an escape sequence. Examples: `ISO-2022-JP` (Japanese) and `HZ-GB-2312` (Chinese).
3. Multi-byte encodings, where each character is represented by a variable number of bytes. Examples: `Big5` (Chinese), `SHIFT_JIS` (Japanese), `EUC-KR` (Korean), and `UTF-8` without a BOM.
4. Single-byte encodings, where each character is represented by one byte. Examples: `KOI8-R` (Russian), `windows-1255` (Hebrew), and `TIS-620` (Thai).
5. `windows-1252`, which is used primarily on Microsoft Windows; its subset, `ISO-8859-1` is widely used for legacy 8-bit-encoded text. chardet, like many encoding detectors, defaults to guessing this encoding when no other can be reliably established.

1.4.1 `UTF-n` with a BOM

If the text starts with a BOM, we can reasonably assume that the text is encoded in `UTF-8`, `UTF-16`, or `UTF-32`. (The BOM will tell us exactly which one; that's what it's for.) This is handled inline in `UniversalDetector`, which returns the result immediately without any further processing.

1.4.2 Escaped encodings

If the text contains a recognizable escape sequence that might indicate an escaped encoding, `UniversalDetector` creates an `EscCharSetProber` (defined in `escprober.py`) and feeds it the text.

`EscCharSetProber` creates a series of state machines, based on models of `HZ-GB-2312`, `ISO-2022-CN`, `ISO-2022-JP`, and `ISO-2022-KR` (defined in `escsm.py`). `EscCharSetProber` feeds the text to each of these state machines, one byte at a time. If any state machine ends up uniquely identifying the encoding, `EscCharSetProber` immediately returns the positive result to `UniversalDetector`, which returns it to the caller. If any state machine hits an illegal sequence, it is dropped and processing continues with the other state machines.

1.4.3 Multi-byte encodings

Assuming no BOM, `UniversalDetector` checks whether the text contains any high-bit characters. If so, it creates a series of "probers" for detecting multi-byte encodings, single-byte encodings, and as a last resort, `windows-1252`.

The multi-byte encoding prober, `MBCSGroupProber` (defined in `mbcsgroupprober.py`), is really just a shell that manages a group of other probers, one for each multi-byte encoding: `Big5`, `GB2312`, `EUC-TW`, `EUC-KR`, `EUC-JP`, `SHIFT_JIS`, and `UTF-8`. `MBCSGroupProber` feeds the text to each of these encoding-specific probers and checks the results. If a prober reports that it has found an illegal byte sequence, it is dropped from further processing (so that, for instance, any subsequent calls to `UniversalDetector.feed` will skip that prober). If a prober reports that it is reasonably confident that it has detected the encoding, `MBCSGroupProber` reports this positive result to `UniversalDetector`, which reports the result to the caller.

Most of the multi-byte encoding probers are inherited from `MultiByteCharSetProber` (defined in `mbcharsetprober.py`), and simply hook up the appropriate state machine and distribution analyzer and let `MultiByteCharSetProber` do the rest of the work. `MultiByteCharSetProber` runs the text through the encoding-specific state machine, one byte at a time, to look for byte sequences that would indicate a conclusive positive or negative result. At the same time, `MultiByteCharSetProber` feeds the text to an encoding-specific distribution analyzer.

The distribution analyzers (each defined in `chardistribution.py`) use language-specific models of which characters are used most frequently. Once `MultiByteCharSetProber` has fed enough text to the distribution analyzer, it calculates a confidence rating based on the number of frequently-used characters, the total number of characters,

and a language-specific distribution ratio. If the confidence is high enough, `MultiByteCharSetProber` returns the result to `MBCSGroupProber`, which returns it to `UniversalDetector`, which returns it to the caller.

The case of Japanese is more difficult. Single-character distribution analysis is not always sufficient to distinguish between `EUC-JP` and `SHIFT_JIS`, so the `SJISProber` (defined in `sjisprober.py`) also uses 2-character distribution analysis. `SJISContextAnalysis` and `EUCJPContextAnalysis` (both defined in `jpcntx.py` and both inheriting from a common `JapaneseContextAnalysis` class) check the frequency of Hiragana syllabary characters within the text. Once enough text has been processed, they return a confidence level to `SJISProber`, which checks both analyzers and returns the higher confidence level to `MBCSGroupProber`.

1.4.4 Single-byte encodings

The single-byte encoding prober, `SBCSGroupProber` (defined in `sbcsgroupprober.py`), is also just a shell that manages a group of other probers, one for each combination of single-byte encoding and language: `windows-1251`, `KOI8-R`, `ISO-8859-5`, `MacCyrillic`, `IBM855`, and `IBM866` (Russian); `ISO-8859-7` and `windows-1253` (Greek); `ISO-8859-5` and `windows-1251` (Bulgarian); `ISO-8859-2` and `windows-1250` (Hungarian); `TIS-620` (Thai); `windows-1255` and `ISO-8859-8` (Hebrew).

`SBCSGroupProber` feeds the text to each of these encoding+language-specific probers and checks the results. These probers are all implemented as a single class, `SingleByteCharSetProber` (defined in `sbcharsetprober.py`), which takes a language model as an argument. The language model defines how frequently different 2-character sequences appear in typical text. `SingleByteCharSetProber` processes the text and tallies the most frequently used 2-character sequences. Once enough text has been processed, it calculates a confidence level based on the number of frequently-used sequences, the total number of characters, and a language-specific distribution ratio.

Hebrew is handled as a special case. If the text appears to be Hebrew based on 2-character distribution analysis, `HebrewProber` (defined in `hebrewprober.py`) tries to distinguish between Visual Hebrew (where the source text actually stored "backwards" line-by-line, and then displayed verbatim so it can be read from right to left) and Logical Hebrew (where the source text is stored in reading order and then rendered right-to-left by the client). Because certain characters are encoded differently based on whether they appear in the middle of or at the end of a word, we can make a reasonable guess about direction of the source text, and return the appropriate encoding (`windows-1255` for Logical Hebrew, or `ISO-8859-8` for Visual Hebrew).

1.4.5 windows-1252

If `UniversalDetector` detects a high-bit character in the text, but none of the other multi-byte or single-byte encoding probers return a confident result, it creates a `Latin1Prober` (defined in `latin1prober.py`) to try to detect English text in a `windows-1252` encoding. This detection is inherently unreliable, because English letters are encoded in the same way in many different encodings. The only way to distinguish `windows-1252` is through commonly used symbols like smart quotes, curly apostrophes, copyright symbols, and the like. `Latin1Prober` automatically reduces its confidence rating to allow more accurate probers to win if at all possible.

1.5 chardet

1.5.1 chardet package

Submodules

chardet.big5freq module

chardet.big5prober module

class chardet.big5prober.**Big5Prober**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

language

chardet.chardetect module

chardet.chardistribution module

class chardet.chardistribution.**Big5DistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**CharDistributionAnalysis**
Bases: object

ENOUGH_DATA_THRESHOLD = 1024

MINIMUM_DATA_THRESHOLD = 3

SURE_NO = 0.01

SURE_YES = 0.99

feed(*char*, *char_len*)
feed a character with known length

get_confidence()
return confidence based on existing data

get_order(_)

got_enough_data()

reset()
reset analyser, clear any state

class chardet.chardistribution.**EUCJPDistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**EUCKRDistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**EUCTWDistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**GB2312DistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**JOHABDistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

class chardet.chardistribution.**SJISDistributionAnalysis**
Bases: *chardet.chardistribution.CharDistributionAnalysis*

get_order(*byte_str*)

chardet.charsetgroupprober module

class chardet.charsetgroupprober.**CharSetGroupProber**(*lang_filter=None*)
Bases: *chardet.charsetprober.CharSetProber*

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.charsetprober module

class chardet.charsetprober.**CharSetProber**(*lang_filter=None*)
Bases: object

SHORTCUT_THRESHOLD = 0.95

charset_name

feed(*byte_str*)

static filter_high_byte_only(*buf*)

static filter_international_words(*buf*)
We define three types of bytes: alphabet: english alphabets [a-zA-Z] international: international characters [-ÿ] marker: everything else [^a-zA-Z-ÿ] The input buffer can be thought to contain a series of words delimited by markers. This function works to filter all words that contain at least one international character. All contiguous sequences of markers are replaced by a single space ascii character. This filter applies to all scripts which do not use English characters.

get_confidence()

static remove_xml_tags(*buf*)
Returns a copy of buf that retains only the sequences of English alphabet and high byte characters that are not between <> characters. This filter can be applied to all scripts which contain both English characters and extended ASCII characters, but is currently only used by Latin1Prober.

reset()

state

chardet.codingstatemachine module

class chardet.codingstatemachine.**CodingStateMachine**(*sm*)
Bases: object

A state machine to verify a byte sequence for a particular encoding. For each byte the detector receives, it will feed that byte to every active state machine available, one byte at a time. The state machine changes its state

based on its previous state and the byte it receives. There are 3 states in a state machine that are of interest to an auto-detector:

START state: This is the state to start with, or a legal byte sequence (i.e. a valid code point) for character has been identified.

ME state: This indicates that the state machine identified a byte sequence that is specific to the charset it is designed for and that there is no other possible encoding which can contain this byte sequence. This will to lead to an immediate positive answer for the detector.

ERROR state: This indicates the state machine identified an illegal byte sequence for that encoding. This will lead to an immediate negative answer for this encoding. Detector will exclude this encoding from consideration from here on.

get_coding_state_machine()

get_current_charlen()

language

next_state(*c*)

reset()

chardet.compat module

chardet.constants module

chardet.cp949prober module

class chardet.cp949prober.**CP949Prober**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

language

chardet.escprober module

class chardet.escprober.**EscCharSetProber**(*lang_filter=None*)
Bases: *chardet.charsetprober.CharSetProber*

This CharSetProber uses a "code scheme" approach for detecting encodings, whereby easily recognizable escape or shift sequences are relied on to identify these encodings.

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.escsm module

chardet.eucjpprober module

class chardet.eucjpprober.**EUCJPProber**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.euckrfreq module

chardet.euckrprober module

class chardet.euckrprober.**EUCKRProber**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

language

chardet.euctwfreq module

chardet.euctwprober module

class chardet.euctwprober.**EUCTWProber**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

language

chardet.gb2312freq module

chardet.gb2312prober module

class chardet.gb2312prober.**GB2312Prober**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

language

chardet.hebrewprober module

class chardet.hebrewprober.**HebrewProber**
Bases: *chardet.charsetprober.CharSetProber*

FINAL_KAF = 234

FINAL_MEM = 237

FINAL_NUN = 239

FINAL_PE = 243

FINAL_TSADI = 245

LOGICAL_HEBREW_NAME = 'windows-1255'

MIN_FINAL_CHAR_DISTANCE = 5

MIN_MODEL_DISTANCE = 0.01

NORMAL_KAF = 235

NORMAL_MEM = 238

NORMAL_NUN = 240

NORMAL_PE = 244

NORMAL_TSADI = 246

VISUAL_HEBREW_NAME = 'ISO-8859-8'

charset_name

feed(*byte_str*)

is_final(*c*)

is_non_final(*c*)

language

reset()

set_model_probers(*logical_prober*, *visual_prober*)

state

chardet.jisfreq module

chardet.jpcntx module

class chardet.jpcntx.**EUCJPContextAnalysis**

Bases: *chardet.jpcntx.JapaneseContextAnalysis*

get_order(*byte_str*)

class chardet.jpcntx.**JapaneseContextAnalysis**

Bases: object

DONT_KNOW = -1

ENOUGH_REL_THRESHOLD = 100

MAX_REL_THRESHOLD = 1000

MINIMUM_DATA_THRESHOLD = 4

NUM_OF_CATEGORY = 6

feed(*byte_str*, *num_bytes*)

get_confidence()

get_order(_)

got_enough_data()

reset()

class chardet.jpcntx.**SJISContextAnalysis**
Bases: *chardet.jpcntx.JapaneseContextAnalysis*

charset_name

get_order(*byte_str*)

chardet.langbulgarianmodel module

chardet.langcyrillicmodel module

chardet.langgreekmodel module

chardet.langhebrewmodel module

chardet.langhungarianmodel module

chardet.langthaimodel module

chardet.latin1prober module

class chardet.latin1prober.**Latin1Prober**
Bases: *chardet.charsetprober.CharSetProber*

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.mbcharsetprober module

class chardet.mbcharsetprober.**MultiByteCharSetProber**(*lang_filter=None*)
Bases: *chardet.charsetprober.CharSetProber*

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.mbcsgroupprober module

class chardet.mbcsgroupprober.**MBCSGroupProber**(*lang_filter=None*)
Bases: *chardet.charsetgroupprober.CharSetGroupProber*

chardet.mbcssm module

chardet.sbcharsetprober module

class chardet.sbcharsetprober.**SingleByteCharSetModel**(*charset_name*, *language*, *char_to_order_map*, *language_model*, *typical_positive_ratio*, *keep_ascii_letters*, *alphabet*)

Bases: tuple

alphabet
Alias for field number 6

char_to_order_map
Alias for field number 2

charset_name
Alias for field number 0

keep_ascii_letters
Alias for field number 5

language
Alias for field number 1

language_model
Alias for field number 3

typical_positive_ratio
Alias for field number 4

class chardet.sbcharsetprober.**SingleByteCharSetProber**(*model*, *is_reversed=False*, *name_prober=None*)

Bases: *chardet.charsetprober.CharSetProber*

NEGATIVE_SHORTCUT_THRESHOLD = 0.05

POSITIVE_SHORTCUT_THRESHOLD = 0.95

SAMPLE_SIZE = 64

SB_ENOUGH_REL_THRESHOLD = 1024

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.sbcsgroupprober module

class chardet.sbcsgroupprober.**SBCSGroupProber**

Bases: *chardet.charsetgroupprober.CharSetGroupProber*

chardet.sjisprober module

class chardet.sjisprober.**SJISProber**
Bases: *chardet.mbcharsetprober.MultiByteCharSetProber*

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

chardet.universaldetector module

Module containing the UniversalDetector detector class, which is the primary class a user of `chardet` should use.

author Mark Pilgrim (initial port to Python)

author Shy Shalom (original C code)

author Dan Blanchard (major refactoring for 3.0)

author Ian Cordasco

class chardet.universaldetector.**UniversalDetector**(*lang_filter=31*)
Bases: `object`

The `UniversalDetector` class underlies the `chardet.detect` function and coordinates all of the different charset probers.

To get a `dict` containing an encoding and its confidence, you can simply run:

```
u = UniversalDetector()
u.feed(some_bytes)
u.close()
detected = u.result
```

ESC_DETECTOR = re.compile(b'(\x1b|~{)')

HIGH_BYTE_DETECTOR = re.compile(b'[\x80-\xff]')

ISO_WIN_MAP = {'iso-8859-1': 'Windows-1252', 'iso-8859-13': 'Windows-1257', 'iso-885

MINIMUM_THRESHOLD = 0.2

WIN_BYTE_DETECTOR = re.compile(b'[\x80-\x9f]')

charset_probers

close()
Stop analyzing the current document and come up with a final prediction.

Returns The `result` attribute, a `dict` with the keys *encoding*, *confidence*, and *language*.

feed(*byte_str*)
Takes a chunk of a document and feeds it through all of the relevant charset probers.

After calling `feed`, you can check the value of the `done` attribute to see if you need to continue feeding the `UniversalDetector` more data, or if it has made a prediction (in the `result` attribute).

> **Note:** You should always call `close` when you're done feeding in your document if `done` is not already `True`.

has_win_bytes

input_state

reset()
Reset the UniversalDetector and all of its probers back to their initial states. This is called by `__init__`, so you only need to call this directly in between analyses of different documents.

chardet.utf8prober module

class chardet.utf8prober.**UTF8Prober**
Bases: *chardet.charsetprober.CharSetProber*

ONE_CHAR_PROB = 0.5

charset_name

feed(*byte_str*)

get_confidence()

language

reset()

Module contents

class chardet.**UniversalDetector**(*lang_filter=31*)
Bases: `object`

The `UniversalDetector` class underlies the `chardet.detect` function and coordinates all of the different charset probers.

To get a `dict` containing an encoding and its confidence, you can simply run:

```
u = UniversalDetector()
u.feed(some_bytes)
u.close()
detected = u.result
```

ESC_DETECTOR = re.compile(b'(\x1b|~{)')

HIGH_BYTE_DETECTOR = re.compile(b'[\x80-\xff]')

ISO_WIN_MAP = {'iso-8859-1': 'Windows-1252', 'iso-8859-13': 'Windows-1257', 'iso-885

MINIMUM_THRESHOLD = 0.2

WIN_BYTE_DETECTOR = re.compile(b'[\x80-\x9f]')

charset_probers

close()
Stop analyzing the current document and come up with a final prediction.

Returns The `result` attribute, a `dict` with the keys *encoding*, *confidence*, and *language*.

feed(*byte_str*)
Takes a chunk of a document and feeds it through all of the relevant charset probers.

After calling `feed`, you can check the value of the `done` attribute to see if you need to continue feeding the `UniversalDetector` more data, or if it has made a prediction (in the `result` attribute).

Note: You should always call `close` when you're done feeding in your document if `done` is not already `True`.

---

has_win_bytes

input_state

reset()
Reset the UniversalDetector and all of its probers back to their initial states. This is called by `__init__`, so you only need to call this directly in between analyses of different documents.

chardet.**detect**(*byte_str*)
Detect the encoding of the given byte string.

Parameters **byte_str** (`bytes` or `bytearray`) – The byte sequence to examine.

chardet.**detect_all**(*byte_str*, *ignore_threshold=False*)
Detect all the possible encodings of the given byte string.

Parameters

- **byte_str** (`bytes` or `bytearray`) – The byte sequence to examine.
- **ignore_threshold** (`bool`) – Include encodings that are below `UniversalDetector.MINIMUM_THRESHOLD` in results.

CHAPTER 2

Indices and tables

- genindex
- modindex
- search

Python Module Index

C

chardet, 17
chardet.big5freq, 8
chardet.big5prober, 9
chardet.chardistribution, 9
chardet.charsetgroupprober, 10
chardet.charsetprober, 10
chardet.codingstatemachine, 10
chardet.cp949prober, 11
chardet.escprober, 11
chardet.escsm, 12
chardet.eucjpprober, 12
chardet.euckrfreq, 12
chardet.euckrprober, 12
chardet.euctwfreq, 12
chardet.euctwprober, 12
chardet.gb2312freq, 12
chardet.gb2312prober, 12
chardet.hebrewprober, 12
chardet.jisfreq, 13
chardet.jpcntx, 13
chardet.langbulgarianmodel, 14
chardet.langgreekmodel, 14
chardet.langhebrewmodel, 14
chardet.langhungarianmodel, 14
chardet.langthaimodel, 14
chardet.latin1prober, 14
chardet.mbcharsetprober, 14
chardet.mbcsgroupprober, 14
chardet.mbcssm, 15
chardet.sbcharsetprober, 15
chardet.sbcsgroupprober, 15
chardet.sjisprober, 16
chardet.universaldetector, 16
chardet.utf8prober, 17

Index

A

alphabet (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15

B

Big5DistributionAnalysis (*class in chardet.chardistribution*), 9
Big5Prober (*class in chardet.big5prober*), 9

C

char_to_order_map (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15
chardet (*module*), 17
chardet.big5freq (*module*), 8
chardet.big5prober (*module*), 9
chardet.chardistribution (*module*), 9
chardet.charsetgroupprober (*module*), 10
chardet.charsetprober (*module*), 10
chardet.codingstatemachine (*module*), 10
chardet.cp949prober (*module*), 11
chardet.escprober (*module*), 11
chardet.escsm (*module*), 12
chardet.eucjpprober (*module*), 12
chardet.euckrfreq (*module*), 12
chardet.euckrprober (*module*), 12
chardet.euctwfreq (*module*), 12
chardet.euctwprober (*module*), 12
chardet.gb2312freq (*module*), 12
chardet.gb2312prober (*module*), 12
chardet.hebrewprober (*module*), 12
chardet.jisfreq (*module*), 13
chardet.jpcntx (*module*), 13
chardet.langbulgarianmodel (*module*), 14
chardet.langgreekmodel (*module*), 14
chardet.langhebrewmodel (*module*), 14
chardet.langhungarianmodel (*module*), 14
chardet.langthaimodel (*module*), 14
chardet.latin1prober (*module*), 14
chardet.mbcharsetprober (*module*), 14
chardet.mbcsgroupprober (*module*), 14
chardet.mbcssm (*module*), 15
chardet.sbcharsetprober (*module*), 15
chardet.sbcsgroupprober (*module*), 15
chardet.sjisprober (*module*), 16
chardet.universaldetector (*module*), 16
chardet.utf8prober (*module*), 17
CharDistributionAnalysis (*class in chardet.chardistribution*), 9
charset_name (*chardet.big5prober.Big5Prober attribute*), 9
charset_name (*chardet.charsetgroupprober.CharSetGroupProber attribute*), 10
charset_name (*chardet.charsetprober.CharSetProber attribute*), 10
charset_name (*chardet.cp949prober.CP949Prober attribute*), 11
charset_name (*chardet.escprober.EscCharSetProber attribute*), 11
charset_name (*chardet.eucjpprober.EUCJPProber attribute*), 12
charset_name (*chardet.euckrprober.EUCKRProber attribute*), 12
charset_name (*chardet.euctwprober.EUCTWProber attribute*), 12
charset_name (*chardet.gb2312prober.GB2312Prober attribute*), 12
charset_name (*chardet.hebrewprober.HebrewProber attribute*), 13
charset_name (*chardet.jpcntx.SJISContextAnalysis attribute*), 14
charset_name (*chardet.latin1prober.Latin1Prober attribute*), 14
charset_name (*chardet.mbcharsetprober.MultiByteCharSetProber attribute*), 14
charset_name (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15
charset_name (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15

charset_name (*chardet.sjisprober.SJISProber attribute*), 16
charset_name (*chardet.utf8prober.UTF8Prober attribute*), 17
charset_probers (*chardet.UniversalDetector attribute*), 17
charset_probers (*chardet.universaldetector.UniversalDetector attribute*), 16
CharSetGroupProber (*class in chardet.charsetgroupprober*), 10
CharSetProber (*class in chardet.charsetprober*), 10
close() (*chardet.UniversalDetector method*), 17
close() (*chardet.universaldetector.UniversalDetector method*), 16
CodingStateMachine (*class in chardet.codingstatemachine*), 10
CP949Prober (*class in chardet.cp949prober*), 11

D

detect() (*in module chardet*), 18
detect_all() (*in module chardet*), 18
DONT_KNOW (*chardet.jpcntx.JapaneseContextAnalysis attribute*), 13

E

ENOUGH_DATA_THRESHOLD (*chardet.chardistribution.CharDistributionAnalysis attribute*), 9
ENOUGH_REL_THRESHOLD (*chardet.jpcntx.JapaneseContextAnalysis attribute*), 13
ESC_DETECTOR (*chardet.UniversalDetector attribute*), 17
ESC_DETECTOR (*chardet.universaldetector.UniversalDetector attribute*), 16
EscCharSetProber (*class in chardet.escprober*), 11
EUCJPContextAnalysis (*class in chardet.jpcntx*), 13
EUCJPDistributionAnalysis (*class in chardet.chardistribution*), 9
EUCJPProber (*class in chardet.eucjpprober*), 12
EUCKRDistributionAnalysis (*class in chardet.chardistribution*), 9
EUCKRProber (*class in chardet.euckrprober*), 12
EUCTWDistributionAnalysis (*class in chardet.chardistribution*), 9
EUCTWProber (*class in chardet.euctwprober*), 12

F

feed() (*chardet.chardistribution.CharDistributionAnalysis method*), 9
feed() (*chardet.charsetgroupprober.CharSetGroupProber method*), 10
feed() (*chardet.charsetprober.CharSetProber method*), 10
feed() (*chardet.escprober.EscCharSetProber method*), 11
feed() (*chardet.eucjpprober.EUCJPProber method*), 12
feed() (*chardet.hebrewprober.HebrewProber method*), 13
feed() (*chardet.jpcntx.JapaneseContextAnalysis method*), 13
feed() (*chardet.latin1prober.Latin1Prober method*), 14
feed() (*chardet.mbcharsetprober.MultiByteCharSetProber method*), 14
feed() (*chardet.sbcharsetprober.SingleByteCharSetProber method*), 15
feed() (*chardet.sjisprober.SJISProber method*), 16
feed() (*chardet.UniversalDetector method*), 17
feed() (*chardet.universaldetector.UniversalDetector method*), 16
feed() (*chardet.utf8prober.UTF8Prober method*), 17
filter_high_byte_only() (*chardet.charsetprober.CharSetProber static method*), 10
filter_international_words() (*chardet.charsetprober.CharSetProber static method*), 10
FINAL_KAF (*chardet.hebrewprober.HebrewProber attribute*), 12
FINAL_MEM (*chardet.hebrewprober.HebrewProber attribute*), 12
FINAL_NUN (*chardet.hebrewprober.HebrewProber attribute*), 13
FINAL_PE (*chardet.hebrewprober.HebrewProber attribute*), 13
FINAL_TSADI (*chardet.hebrewprober.HebrewProber attribute*), 13

G

GB2312DistributionAnalysis (*class in chardet.chardistribution*), 9
GB2312Prober (*class in chardet.gb2312prober*), 12
get_coding_state_machine() (*chardet.codingstatemachine.CodingStateMachine method*), 11
get_confidence() (*chardet.chardistribution.CharDistributionAnalysis method*), 9
get_confidence() (*chardet.charsetgroupprober.CharSetGroupProber method*), 10
get_confidence() (*chardet.charsetprober.CharSetProber method*), 10
get_confidence() (*chardet.escprober.EscCharSetProber method*), 11
get_confidence() (*chardet.eucjpprober.EUCJPProber method*), 12

get_confidence() (*chardet.jpcntx.JapaneseContextAnalysis method*), 13
get_confidence() (*chardet.latin1prober.Latin1Prober method*), 14
get_confidence() (*chardet.mbcharsetprober.MultiByteCharSetProber method*), 14
get_confidence() (*chardet.sbcharsetprober.SingleByteCharSetProber method*), 15
get_confidence() (*chardet.sjisprober.SJISProber method*), 16
get_confidence() (*chardet.utf8prober.UTF8Prober method*), 17
get_current_charlen() (*chardet.codingstatemachine.CodingStateMachine method*), 11
get_order() (*chardet.chardistribution.Big5DistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.CharDistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.EUCJPDistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.EUCKRDistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.EUCTWDistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.GB2312DistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.JOHABDistributionAnalysis method*), 9
get_order() (*chardet.chardistribution.SJISDistributionAnalysis method*), 10
get_order() (*chardet.jpcntx.EUCJPContextAnalysis method*), 13
get_order() (*chardet.jpcntx.JapaneseContextAnalysis method*), 13
get_order() (*chardet.jpcntx.SJISContextAnalysis method*), 14
got_enough_data() (*chardet.chardistribution.CharDistributionAnalysis method*), 9
got_enough_data() (*chardet.jpcntx.JapaneseContextAnalysis method*), 14

H

has_win_bytes (*chardet.UniversalDetector attribute*), 18
has_win_bytes (*chardet.universaldetector.UniversalDetector attribute*), 17
HebrewProber (*class in chardet.hebrewprober*), 12
HIGH_BYTE_DETECTOR (*chardet.UniversalDetector attribute*), 17
HIGH_BYTE_DETECTOR (*chardet.universaldetector.UniversalDetector attribute*), 16

I

input_state (*chardet.UniversalDetector attribute*), 18
input_state (*chardet.universaldetector.UniversalDetector attribute*), 17
is_final() (*chardet.hebrewprober.HebrewProber method*), 13
is_non_final() (*chardet.hebrewprober.HebrewProber method*), 13
ISO_WIN_MAP (*chardet.UniversalDetector attribute*), 17
ISO_WIN_MAP (*chardet.universaldetector.UniversalDetector attribute*), 16

J

JapaneseContextAnalysis (*class in chardet.jpcntx*), 13
JOHABDistributionAnalysis (*class in chardet.chardistribution*), 9

K

keep_ascii_letters (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15

L

language (*chardet.big5prober.Big5Prober attribute*), 9
language (*chardet.charsetgroupprober.CharSetGroupProber attribute*), 10
language (*chardet.codingstatemachine.CodingStateMachine attribute*), 11
language (*chardet.cp949prober.CP949Prober attribute*), 11
language (*chardet.escprober.EscCharSetProber attribute*), 11
language (*chardet.eucjpprober.EUCJPProber attribute*), 12
language (*chardet.euckrprober.EUCKRProber attribute*), 12
language (*chardet.euctwprober.EUCTWProber attribute*), 12
language (*chardet.gb2312prober.GB2312Prober attribute*), 12
language (*chardet.hebrewprober.HebrewProber attribute*), 13
language (*chardet.latin1prober.Latin1Prober attribute*), 14
language (*chardet.mbcharsetprober.MultiByteCharSetProber attribute*), 14
language (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15

language (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15
language (*chardet.sjisprober.SJISProber attribute*), 16
language (*chardet.utf8prober.UTF8Prober attribute*), 17
language_model (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15
Latin1Prober (*class in chardet.latin1prober*), 14
LOGICAL_HEBREW_NAME (*chardet.hebrewprober.HebrewProber attribute*), 13

M

MAX_REL_THRESHOLD (*chardet.jpcntx.JapaneseContextAnalysis attribute*), 13
MBCSGroupProber (*class in chardet.mbcsgroupprober*), 14
MIN_FINAL_CHAR_DISTANCE (*chardet.hebrewprober.HebrewProber attribute*), 13
MIN_MODEL_DISTANCE (*chardet.hebrewprober.HebrewProber attribute*), 13
MINIMUM_DATA_THRESHOLD (*chardet.chardistribution.CharDistributionAnalysis attribute*), 9
MINIMUM_DATA_THRESHOLD (*chardet.jpcntx.JapaneseContextAnalysis attribute*), 13
MINIMUM_THRESHOLD (*chardet.UniversalDetector attribute*), 17
MINIMUM_THRESHOLD (*chardet.universaldetector.UniversalDetector attribute*), 16
MultiByteCharSetProber (*class in chardet.mbcharsetprober*), 14

N

NEGATIVE_SHORTCUT_THRESHOLD (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15
next_state() (*chardet.codingstatemachine.CodingStateMachine method*), 11
NORMAL_KAF (*chardet.hebrewprober.HebrewProber attribute*), 13
NORMAL_MEM (*chardet.hebrewprober.HebrewProber attribute*), 13
NORMAL_NUN (*chardet.hebrewprober.HebrewProber attribute*), 13
NORMAL_PE (*chardet.hebrewprober.HebrewProber attribute*), 13
NORMAL_TSADI (*chardet.hebrewprober.HebrewProber attribute*), 13
NUM_OF_CATEGORY (*chardet.jpcntx.JapaneseContextAnalysis attribute*), 13

O

ONE_CHAR_PROB (*chardet.utf8prober.UTF8Prober attribute*), 17

P

POSITIVE_SHORTCUT_THRESHOLD (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15

R

remove_xml_tags() (*chardet.charsetprober.CharSetProber static method*), 10
reset() (*chardet.chardistribution.CharDistributionAnalysis method*), 9
reset() (*chardet.charsetgroupprober.CharSetGroupProber method*), 10
reset() (*chardet.charsetprober.CharSetProber method*), 10
reset() (*chardet.codingstatemachine.CodingStateMachine method*), 11
reset() (*chardet.escprober.EscCharSetProber method*), 11
reset() (*chardet.eucjpprober.EUCJPProber method*), 12
reset() (*chardet.hebrewprober.HebrewProber method*), 13
reset() (*chardet.jpcntx.JapaneseContextAnalysis method*), 14
reset() (*chardet.latin1prober.Latin1Prober method*), 14
reset() (*chardet.mbcharsetprober.MultiByteCharSetProber method*), 14
reset() (*chardet.sbcharsetprober.SingleByteCharSetProber method*), 15
reset() (*chardet.sjisprober.SJISProber method*), 16
reset() (*chardet.UniversalDetector method*), 18
reset() (*chardet.universaldetector.UniversalDetector method*), 17
reset() (*chardet.utf8prober.UTF8Prober method*), 17

S

SAMPLE_SIZE (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15
SB_ENOUGH_REL_THRESHOLD (*chardet.sbcharsetprober.SingleByteCharSetProber attribute*), 15
SBCSGroupProber (*class in chardet.sbcsgroupprober*), 15

set_model_probers() (*chardet.hebrewprober.HebrewProber method*), 13
SHORTCUT_THRESHOLD (*chardet.charsetprober.CharSetProber attribute*), 10
SingleByteCharSetModel (*class in chardet.sbcharsetprober*), 15
SingleByteCharSetProber (*class in chardet.sbcharsetprober*), 15
SJISContextAnalysis (*class in chardet.jpcntx*), 14
SJISDistributionAnalysis (*class in chardet.chardistribution*), 10
SJISProber (*class in chardet.sjisprober*), 16
state (*chardet.charsetprober.CharSetProber attribute*), 10
state (*chardet.hebrewprober.HebrewProber attribute*), 13
SURE_NO (*chardet.chardistribution.CharDistributionAnalysis attribute*), 9
SURE_YES (*chardet.chardistribution.CharDistributionAnalysis attribute*), 9

T

typical_positive_ratio (*chardet.sbcharsetprober.SingleByteCharSetModel attribute*), 15

U

UniversalDetector (*class in chardet*), 17
UniversalDetector (*class in chardet.universaldetector*), 16
UTF8Prober (*class in chardet.utf8prober*), 17

V

VISUAL_HEBREW_NAME (*chardet.hebrewprober.HebrewProber attribute*), 13

W

WIN_BYTE_DETECTOR (*chardet.UniversalDetector attribute*), 17
WIN_BYTE_DETECTOR (*chardet.universaldetector.UniversalDetector attribute*), 16