**Character encoding detection**, **charset detection**, or **code page detection** is the process of [heuristically](https://en.wikipedia.org/wiki/Heuristic) guessing the [character encoding](https://en.wikipedia.org/wiki/Character_encoding) of a series of bytes that represent text. The technique is recognised to be unreliable and is only used when specific [metadata](https://en.wikipedia.org/wiki/Metadata), such as a HTTP Content-Type: header is either not available, or is assumed to be untrustworthy.

This algorithm usually involves statistical analysis of byte patterns, like frequency distribution of [trigraphs](https://en.wikipedia.org/wiki/Digraphs_and_trigraphs" \o "Digraphs and trigraphs) of various languages encoded in each code page that will be detected; such statistical analysis can also be used to perform [language detection](https://en.wikipedia.org/wiki/Language_detection). This process is not foolproof because it depends on statistical data.

**Universal**

We could easily detect if text was UTF-8, as there is a specific bit pattern in the top bits of bytes 2/3/etc. Once you found that pattern repeated a certain number of times you could be certain it was UTF-8.

If the file begins with a UTF-16 byte order mark, you can probably assume the rest of the text is that encoding. Otherwise, detecting UTF-16 isn't nearly as easy as UTF-8, unless you can detect the surrogate pairs pattern: but the use of surrogate pairs is rare, so that doesn't usually work. UTF-32 is similar, except there are no surrogate pairs to detect.

**Regional detection**

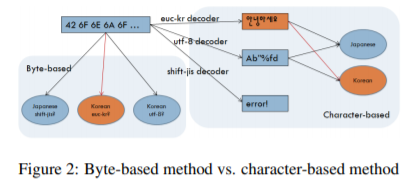
Next we would assume the reader was in a certain region. For instance, if the user was seeing the UI localized in Japanese, we could then attempt detection of the three main Japanese encodings. ISO-2022-JP is again east to detect with the escape sequences. If that fails, determining the difference between EUC-JP and Shift-JIS is not as straightforward. It's more likely that a user would receive Shift-JIS text, but there were characters in EUC-JP that didn't exist in Shift-JIS, and vice-versa, so sometimes you could get a good match.

The same procedure was used for Chinese encodings and other regions.

**User's choice**

If these didn't provide satisfactory results, the user must manually choose an encoding.

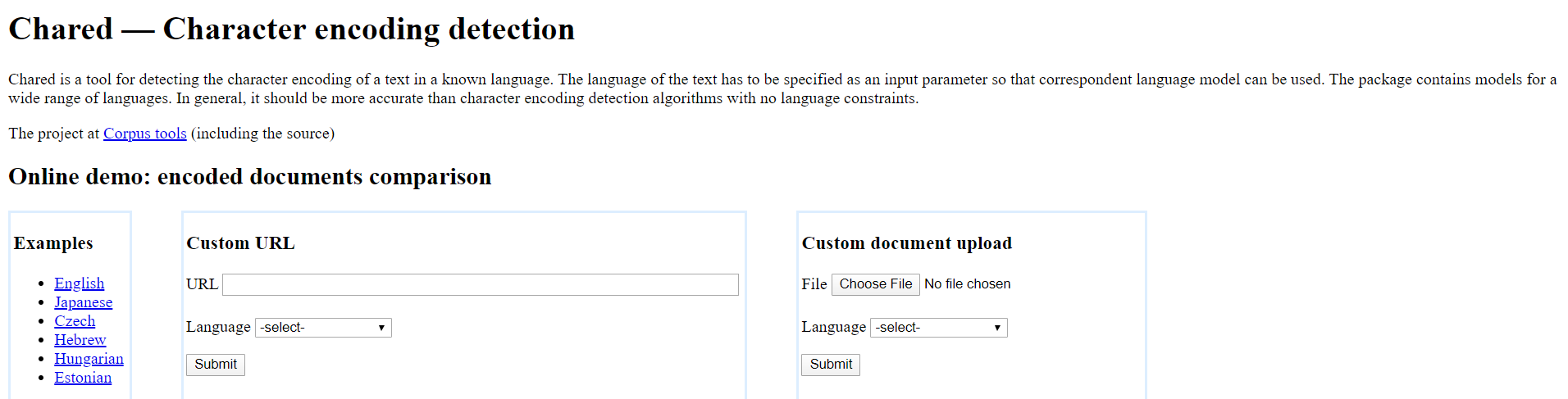
Due to its importance, automatic charset detection is already implemented in major Internet applications such as Mozilla or Internet Explorer. They are very accurate and fast, but the implementation applies many domain specific knowledges in case-by-case basis.



1. the character-based approach with the Na¨ıve Bayes algorithm works well for detection between charsets sharing a large code space (e.g. English US-ASCII, UTF-8, and ISO-8859-1 documents),

2. the byte-based approach with the Support Vector Machine works well with a smaller memory usage and at a faster speed, and

3. the hybrid method of character-based NB for charset detection and character-based SVM for language detection gives the best results.



Charset detection is particularly unreliable in Europe, in an environment of mixed [ISO-8859](https://en.wikipedia.org/wiki/ISO-8859) encodings. These are closely related eight-bit encodings that share an overlap in their lower half with [ASCII](https://en.wikipedia.org/wiki/ASCII). There is no technical way to tell these encodings apart and recognising them relies on identifying language features, such as letter frequencies or spellings.

Due to the unreliability of heuristic detection, it is better to properly label datasets with the correct encoding. HTML documents served across the web by [HTTP](https://en.wikipedia.org/wiki/HTTP) should have their encoding stated [out-of-band](https://en.wikipedia.org/wiki/Out-of-band_data) using the Content-Type: header.

Content-Type: text/html;charset=UTF-8

An isolated HTML document, such as one being edited as a file on disk, may imply such a header by a meta tag within the file:

<**meta** http-equiv="Content-Type" content="text/html;charset=UTF-8" >

or with a new meta type in HTML5[[1]](https://en.wikipedia.org/wiki/Charset_detection#cite_note-1)

<**meta** charset="utf-8" >

If the document is Unicode, then some UTF encodings explicitly label the document with an embedded initial [byte order mark](https://en.wikipedia.org/wiki/Byte_order_mark) (BOM).