Malformed UTF-8 & Spam

Matt Crane, Andrew Trotman & Richard O'Keefe

Department of Computer Science, University of Otago, New Zealand

Introduction

Malformed UTF-8 is a problem when processing data collected from uncontrolled domains, such as the web.

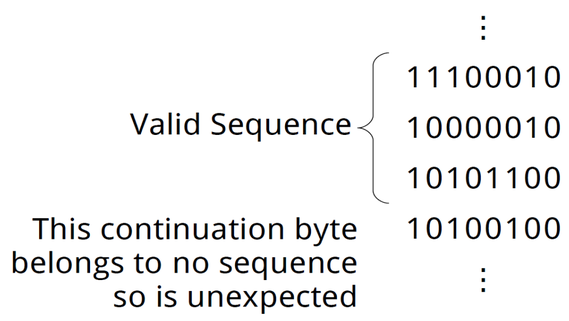
UTF-8 is one encoding of Unicode, a standard that defines a set of codepoints for characters and transformation rules for these codepoints.

Despite the assurances of distributors that "content [of the collection] is encoded in UTF-8 format (where proper UTF-8 character encodings apply)", we identify several types of encoding error present in the ClueWeb09 collection.

We then explore the potential relationship these errors have with another proven measure of document quality, the spamminess.

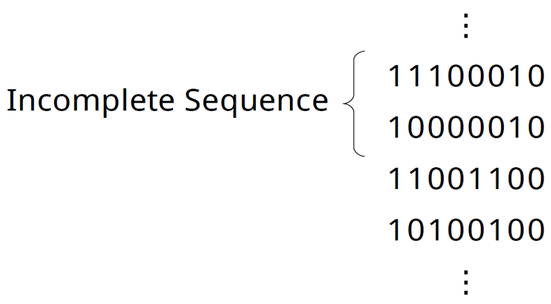
Encoding Errors

Several potential encoding errors must be protected against by standards conforming UTF-8 parsers:



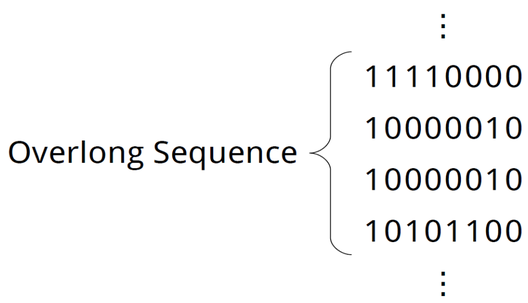
Unexpected continuation bytes: Continuation bytes should belong to a valid sequence, so seeing one out of sequence is invalid.

Incomplete sequence: A sequence that does not have enough continuation bytes is invalid.



Invalid surrogate halves: Some UTF-16 codepoints have been now been defined as invalid within Unicode.

Invalid 4-, 5- and 6-byte sequences: Codepoints above U+10FFFF are invalid, making all 5-, 6- and some 4- byte sequences invalid.



Overlong encodings: By left-padding with 0s it is possible to encode a codepoint in more bytes than necessary, such an encoding is invalid. This error is the only easily corrected error, replacing the encoding with the correct, shorter, encoding.

UTF-8

UTF-8 is a variable byte encoding of Unicode that retains backwards compatibility with ASCII.

Character (Codepoint)	Encoded Binary
$ (U+0024)	00100100
¢ (U+00A2)	11000010 10100010
€ (U+20AC)	11100010 10000010 10101100

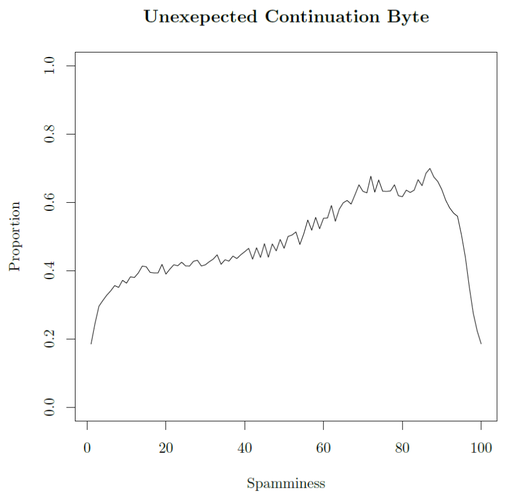
Header Sequence — Continuation Byte Header

Encoding Errors in ClueWeb09

Error	Count
Unexpected Continuation Byte	1,062,303,975
Incomplete Sequence	880,735,432
Invalid Surrogate Half	78,378,657
Invalid 4-, 5-, 6-byte Sequence	1,752,814,236
Overlong NUL	20,374,475
Overlong ASCII	11,496,910
Other Overlong	722,726

Relationship To Spam

We next investigated the relationship between the spamminess level of documents, and the proportion of documents at that spamminess level that contained each class of error.



The most common pattern of error was that typified by the unexpected continuation bytes, with a steady rise in proportion of documents containing the error, with a drop for the most spammiest documents, although the trend is not statistically significant.

The invalid 4-, 5- and 6-byte sequences show a different pattern. This error class is more prevalent among the least spammiest documents with a peak in the middle. This trend is statistically significant, and could be used to determine the quality of a document during indexing.