An Upper-Bound on Information Contained Within a Tweet

Karl Koscher
University of Washington

ABSTRACT

While tweets (and this paper) are limited to 140 characters, not all characters are created equal. This paper explores abuses of character encoding schemes to maximize the number of bits that can be conveyed by a tweet. In particular, since Twitter supports Unicode, we examine how we can abuse UTF8. For example, while people equate a Unicode codepoint with a character, some can be combined to form a single character. Does Twitter count these as one or two characters? Furthermore, some encodings (such as UTF8) allow more codepoints than are specified by Unicode – does Twitter accept these too? We ignore external links, embedded media, Twitter entities, and geotags, which are not universally supported.

BODY

Max bits/tweet? UTF8=31b/chr Can use chrs forbidden by RFC3629 Composing chars count as distinct codepts Max bits is 4339 (-1 for ctrl chrs)

REFERENCES

[1] UTF-8. http://en.wikipedia.org/wiki/UTF-8, July 2012.

[2] Twitter, Inc. Counting characters. https://dev.twitter.com/docs/counting-characters, Apr. 2012.

[3] F. Yergeau. UTF-8, a transformation format of ISO 10646. RFC 3629 (Standard), Nov. 2003.

Volume 1 of Tiny Transactions on Computer Science