|  |  |  |
| --- | --- | --- |
| **FieldName** | **Type** | **Description** |
| url | string | Url of the webpage to be classified |
| urlid | integer | StumbleUpon's unique identifier for each url |
| boilerplate | json | Boilerplate text |
| alchemy\_category | string | Alchemy category (per the publicly available Alchemy API found at www.alchemyapi.com) |
| alchemy\_category\_score | double | Alchemy category score (per the publicly available Alchemy API found at www.alchemyapi.com) |
| avglinksize | double | Average number of words in each link |
| commonLinkRatio\_1 | double | # of links sharing at least 1 word with 1 other links / # of links |
| commonLinkRatio\_2 | double | # of links sharing at least 1 word with 2 other links / # of links |
| commonLinkRatio\_3 | double | # of links sharing at least 1 word with 3 other links / # of links |
| commonLinkRatio\_4 | double | # of links sharing at least 1 word with 4 other links / # of links |
| compression\_ratio | double | Compression achieved on this page via gzip (measure of redundancy) |
| embed\_ratio | double | Count of number of <embed>  usage |
| frameBased | integer (0 or 1) | A page is frame-based (1) if it has no body markup but have a frameset markup |
| frameTagRatio | double | Ratio of iframe markups over total number of markups |
| hasDomainLink | integer (0 or 1) | True (1) if it contains an <a>  with an url with domain |
| html\_ratio | double | Ratio of tags vs text in the page |
| image\_ratio | double | Ratio of <img> tags vs text in the page |
| is\_news | integer (0 or 1) | True (1) if StumbleUpon's news classifier determines that this webpage is news |
| lengthyLinkDomain | integer (0 or 1) | True (1) if at least 3 <a> 's text contains more than 30 alphanumeric characters |
| linkwordscore | double | Percentage of words on the page that are in hyperlink's text |
| news\_front\_page | integer (0 or 1) | True (1) if StumbleUpon's news classifier determines that this webpage is front-page news |
| non\_markup\_alphanum\_characters | integer | Page's text's number of alphanumeric characters |
| numberOfLinks | integer | Number of <a>  markups |
| numwords\_in\_url | double | Number of words in url |
| parametrizedLinkRatio | double | A link is parametrized if it's url contains parameters  or has an attached onClick event |
| spelling\_errors\_ratio | double | Ratio of words not found in wiki (considered to be a spelling mistake) |
| label | integer (0 or 1) | User-determined label. Either evergreen (1) or non-evergreen (0); available for train.tsv only |