|  |  |
| --- | --- |
| action | str\_replace\_all |
| remove all unicode characters (e.g. hair space <U+200A>) | pattern=\\s\*<U\\+\\w+>\\s\*|\\<\\S\\S\\S\\S\\S\\S\\>  replacement=" " |
| remove numbered equations/other centred typesetting | pattern=[\\s\*(\n)(.\*)(\n)\\s\*](file:///\\s*(\n)(.*)(\n)\\s*)  replacement = " " |
| Remove references e.g [37] | pattern = [\\[\\S\\S\\]|\\[\\S\\]](file:///\\[\\S\\S\\%5d|\\%5b\\S\\%5d)  replacement = "" |
| Remove (), keeping text within | pattern = "[()]"  replacement = "" |
| Standardise dashes | pattern=[\\s\*(–+)\\s\*](file:///\\s*(–+)\\s*)  replacement = "-" |
| Standardise references to p values | pattern = \\bp value\\b  replacement = "p-value"  pattern = \\bp less than\\b|\\bp less-than\\b  replacement = "p <"  pattern = \\bp equal to\\b|\\bp equal-to\\b  replacement = "p ="  pattern = \\bp greater than\\b|\\bp greater-than\\b  replacement = "p >" |
| Standardise spacing around <, >, = | pattern = \\s\*[<]\\s\*; replacement = “<”  pattern = \\s\*[>]\\s\*; replacement = “>”  pattern = [\\s\*[=]\\s\*](file:///\\s*[=]\\s*); replacement = “=” |
| Plus or minus symbol | pattern = [\\s\*(±)\\s\*|\\s\*(\\+/-+)\\s\*](file:///\\s*(±)\\s*|\\s*(\\+/-+)\\s*)  replacement = " plus-or-minus " |

Other cleaning steps

Common symbols ($,%,#,@,&,w/) converted to text, using textclean:: replace\_symbol()

Remove punctuation marks except for ‘.’ and ‘-‘, using textclean::strip()

Correct most common spelling errors (UK-EN, using ‘spelling’ package)

Standardised common statistical terms – see methods\_dictionary.xlsx

* For common methods (e.g. Kaplan-meier), all non-hyphenated (Kaplan meier) and combined (kaplanmeier) instances are standardised to include a hyphen
* All common terms expressed as plurals (e.g. t-tests) are converted to singular.

Removal of stopwords – based on list in textclean package

Outstanding issues

p-value: how best to standardise? E.g. p-value<0.05 vs p<0.05 vs p-value less-than 0.05, etc.

<= and >= symbols not accounted for (could standardise similar to ‘+/-‘ sign; changed to ‘plus-or-minus’

Validation of cleaning – take 1% random sample (~1100 rows) and compare original to cleaned text.