Table retrieval project report [CM-UCL I]

Contentmine

2017-04-23

The current project relates to information extraction from tables in scholarly articles for reuse in systematic reviews. Scholarly articles frequently contain vital information/statistics in tabular format, but given that the PDF format is a visually oriented tool instead of also being machine-readable, the information from these tables is not readily exportable. For example, simply copy-pasting a table into a spreadsheet is virtually impossible. Tools such as 't

This document is the final report for the project that ran from circa March 2017 through April 2017, contracted by the EPPI-centre.

We developed software for table extraction in two stages: development and testing.

The current report showcases some of the vital metrics of the resulting software.

Corpus collection

Together with members of the UCL

Table sections

Table structure

Limitations

```
library(ggplot2)
library(plyr)

dat <- read.csv('data/metrics-test.csv')

# Add normalized
dat$normal_disc <- dat$discrepancy_cell_count / (dat$man_cols * dat$man_rows)

# Select only the in scope
sum(dat$scope_ucl == 0 | is.na(dat$table_nr))
```

```
## [1] 56
```

```
dat <- dat[dat$scope_ucl == 1 & !is.na(dat$table_nr), ]

# Failure rate
sum(is.na(dat$discrepancy_cell_count))
```

```
## [1] 36
```

```
# split p complexity
table(is.na(dat$discrepancy_cell_count), dat$table_complexity)
```

```
##
##         messy tidy untidy
##   FALSE     2   33     41
##   TRUE     10   12     14
```

```
dat <- dat[!is.na(dat$discrepancy_cell_count), ]

# Perfect
sum(dat$normal_disc == 0 &
      dat$man_cols == dat$cols_retrieved &
      dat$man_rows == dat$rows_retrieved, na.rm = TRUE)
```

```
## [1] 27
```

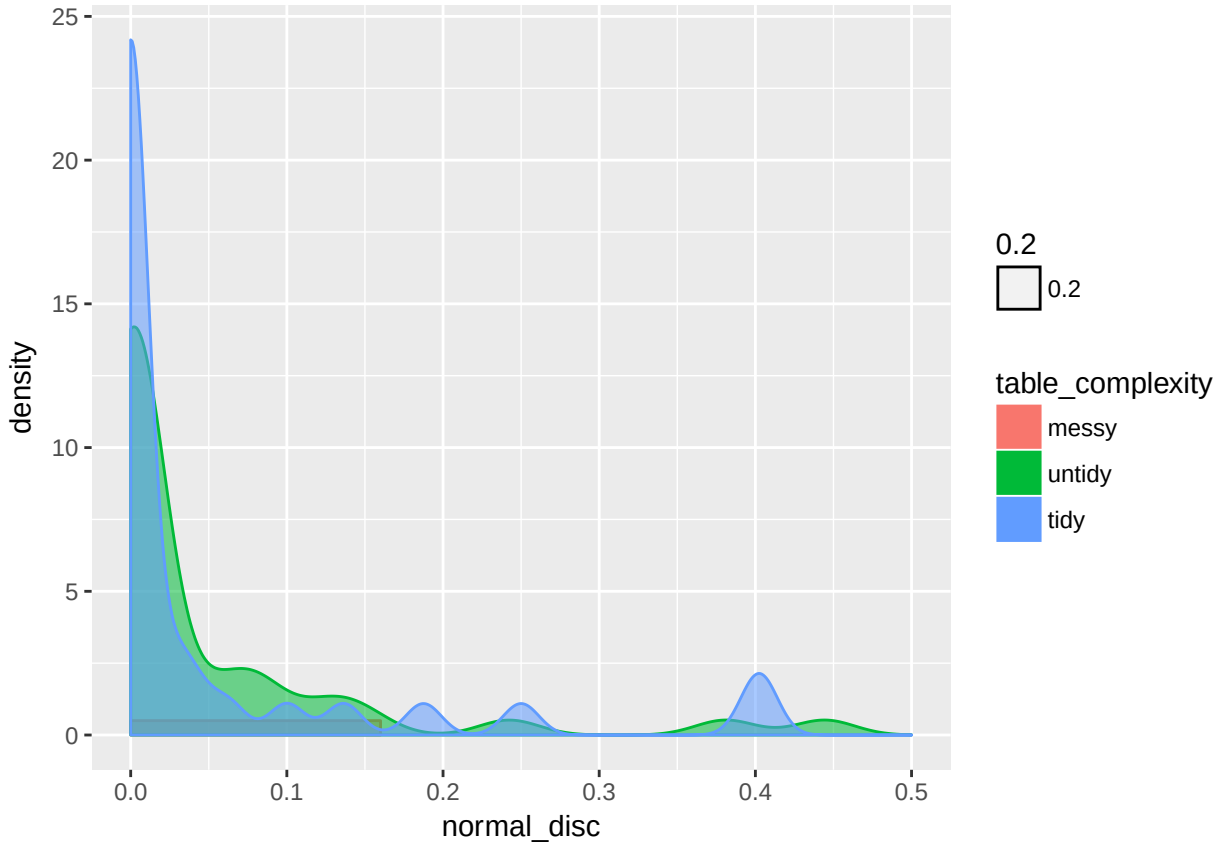
```
# Split p complexity
table(round(dat$normal_disc, 2) == 0 & dat$man_cols == dat$cols_retrieved & dat$man_rows == dat$rows_re
```

```
##
##         messy tidy untidy
##   FALSE     2   13     34
##   TRUE      0   20      7
```

```
#
tmp <- abs(dat$man_cols - dat$cols_retrieved) + abs(dat$man_rows - dat$rows_retrieved)

dat$structure_retrieved <- ifelse(tmp == 0, 'perfect structure',
       ifelse(tmp == 1, 'close to perfect structure',
              ifelse(tmp > 1 & tmp < 4,
                     'reasonable structure',
                     'bad structure')))
dat$discrepancy_factor <- ifelse(dat$discrepancy_cell_count == 0,
                                 'perfect contents',
                                 ifelse(dat$discrepancy_cell_count == 1,
                                        'close to perfect contents',
                                        ifelse(dat$discrepancy_cell_count > 1  & dat$discrepancy_cell_c
                                               'reasonable contents',
                                               'bad contents')))
dat$table_complexity <- factor(dat$table_complexity,
                               levels = c('messy',
                                          'untidy',
                                          'tidy'))
dat$structure_retrieved <- factor(dat$structure_retrieved,
                          levels = c('bad structure',
                                     'reasonable structure',
                                     'close to perfect structure',
                                     'perfect structure'))
dat$discrepancy_factor <- factor(dat$discrepancy_factor,
                                 levels = c('bad contents',
                                            'reasonable contents',
                                            'close to perfect contents',
                                            'perfect contents'))

write.csv(table(dat$structure_retrieved, dat$discrepancy_factor), 'tmp2.csv')


ggplot(dat, aes(x = normal_disc)) +
```

```
  geom_density(aes(color = table_complexity, fill = table_complexity,
                   alpha = .2)) +
  xlim(0, 0.50)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```




```
sum(dat$man_cols == dat$cols_retrieved, na.rm = TRUE)
```

```
## [1] 62
```