Basic analysis

Jinxi_Hu-48528608

2025-11-09

```
library(readr)
```

```
# this file is for do some most basic analysis to the data we collected.
data <- read_csv("data/nodes.csv")
```

```
## Rows: 60152 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): local_id, paper_id, title, first_author, institution, country, venu...
## dbl (5): year, citations, references, n_authors, author_share
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# head of data
head(data)
```

```
## # A tibble: 6 x 13
##   local_id paper_id  title  year first_author institution country venue subtopic
##   <chr>    <chr>     <chr> <dbl> <chr>        <chr>       <chr>   <chr> <chr>
## 1 P0001    https://~ Foun~  2023 Michael Moor Stanford U~ US      Natu~ Machine~
## 2 P0002    https://~ AI i~  2022 Pranav Rajp~ Harvard Un~ US      Natu~ Artific~
## 3 P0003    https://~ Mult~  2022 Julián N. A~ Yale Unive~ US      Natu~ Artific~
## 4 P0004    https://~ Mole~  2019 Philippe Sc~ Ibm Resear~ CH      ACS ~ Machine~
## 5 P0005    https://~ The ~  2020 Jessica Mor~ University~ GB      Soci~ Artific~
## 6 P0006    https://~ Prot~  2021 Gary S Coll~ John Radcl~ GB      BMJ ~ Artific~
## # i 4 more variables: citations <dbl>, references <dbl>, n_authors <dbl>,
## #   author_share <dbl>
```

```
# row and colum number
dim(data)
```

```
## [1] 60152    13
```

```
# colum names
names(data)
```

```
##  [1] "local_id"     "paper_id"     "title"        "year"         "first_author"
##  [6] "institution"  "country"      "venue"        "subtopic"     "citations"
## [11] "references"   "n_authors"    "author_share"
```

```r
# type of colums
str(data)
```

```
## spc_tbl_ [60,152 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ local_id    : chr [1:60152] "P0001" "P0002" "P0003" "P0004" ...
##  $ paper_id    : chr [1:60152] "https://openalex.org/W4365143687" "https://openalex.org/W4205164650"
##  $ title       : chr [1:60152] "Foundation models for generalist medical artificial intelligence" "Al
##  $ year        : num [1:60152] 2023 2022 2022 2019 2020 ...
##  $ first_author: chr [1:60152] "Michael Moor" "Pranav Rajpurkar" "Julián N. Acosta" "Philippe Schwall
##  $ institution : chr [1:60152] "Stanford University" "Harvard University" "Yale University" "Ibm Rese
##  $ country     : chr [1:60152] "US" "US" "US" "CH" ...
##  $ venue       : chr [1:60152] "Nature" "Nature Medicine" "Nature Medicine" "ACS Central Science" ..
##  $ subtopic    : chr [1:60152] "Machine Learning in Healthcare" "Artificial Intelligence in Healthcar
##  $ citations   : num [1:60152] 1155 1931 810 722 671 ...
##  $ references  : num [1:60152] 50 127 180 40 174 33 22 42 60 99 ...
##  $ n_authors   : num [1:60152] 7 4 4 7 7 13 4 9 44 34 ...
##  $ author_share: num [1:60152] 0.143 0.25 0.25 0.143 0.143 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   local_id = col_character(),
##   ..   paper_id = col_character(),
##   ..   title = col_character(),
##   ..   year = col_double(),
##   ..   first_author = col_character(),
##   ..   institution = col_character(),
##   ..   country = col_character(),
##   ..   venue = col_character(),
##   ..   subtopic = col_character(),
##   ..   citations = col_double(),
##   ..   references = col_double(),
##   ..   n_authors = col_double(),
##   ..   author_share = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
# summary
summary(data)
```

```
##    local_id           paper_id            title                year
##  Length:60152       Length:60152       Length:60152       Min.   :2015
##  Class :character   Class :character   Class :character   1st Qu.:2022
##  Mode  :character   Mode  :character   Mode  :character   Median :2023
##                                                           Mean   :2023
##                                                           3rd Qu.:2024
##                                                           Max.   :2025
##                                                           NA's   :57542
##  first_author       institution          country            venue
##  Length:60152       Length:60152       Length:60152       Length:60152
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##    subtopic           citations         references      n_authors
##  Length:60152       Min.   :   0.00   Min.   :  0.0   Min.   :  1.00
##  Class :character   1st Qu.:   1.00   1st Qu.: 10.0   1st Qu.:  3.00
##  Mode  :character   Median :   6.00   Median : 27.0   Median :  5.00
##                     Mean   :  34.77   Mean   : 36.2   Mean   :  7.65
##                     3rd Qu.:  25.00   3rd Qu.: 48.0   3rd Qu.:  9.00
##                     Max.   :2383.00   Max.   :629.0   Max.   :100.00
##                     NA's   :57542     NA's   :57542   NA's   :57542
##   author_share
##  Min.   :0.01
##  1st Qu.:0.11
##  Median :0.20
##  Mean   :0.26
##  3rd Qu.:0.33
##  Max.   :1.00
##  NA's   :57542
```