bernardo

```
## Loading required package: DBI
```

```
## Loading required package: RSQLite
```

```
## Loading required package: ff
```

```
## Loading required package: bit
```

```
##
## Attaching package: 'bit'
```

```
## The following object is masked from 'package:base':
##
##     xor
```

```
## Attaching package ff
```

```
## - getOption("fftempdir")=="/var/folders/bv/xhclmwh90zg08bvwnjvxtrz80000gn/T//RtmpSfLq5J/ff"
```

```
## - getOption("ffextension")=="ff"
```

```
## - getOption("ffdrop")==TRUE
```

```
## - getOption("fffinonexit")==TRUE
```

```
## - getOption("ffpagesize")==65536
```

```
## - getOption("ffcaching")=="mmnoflush"  -- consider "ffeachflush" if your system stalls on large write
```

```
## - getOption("ffbatchbytes")==16777216 -- consider a different value for tuning your system
```

```
## - getOption("ffmaxbytes")==536870912 -- consider a different value for tuning your system
```

```
##
## Attaching package: 'ff'
```

```
## The following objects are masked from 'package:utils':
##
##     write.csv, write.csv2
```

```
## The following objects are masked from 'package:base':
##
##     is.factor, is.ordered
```

```
## RecordLinkage library
```

```
## [c] IMBEI Mainz
```

```
##
## Attaching package: 'RecordLinkage'
```

```
## The following object is masked from 'package:bit':
##
##     clone
```

```
## The following object is masked from 'package:base':
##
##     isFALSE
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: stringdist

## Loading required package: plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

##
## Attaching package: 'blink'

## The following object is masked from 'package:RecordLinkage':
##
##     RLdata500

##
## Attaching package: 'tokenizers'

## The following objects are masked from 'package:textreuse':
##
##     tokenize_ngrams, tokenize_sentences, tokenize_skip_ngrams,
##     tokenize_words

## Loading required package: usethis

##
## Attaching package: 'igraph'

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.1.0     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x plyr::arrange()         masks dplyr::arrange()
## x tibble::as_data_frame() masks igraph::as_data_frame(), dplyr::as_data_frame()
## x purrr::compact()        masks plyr::compact()
## x purrr::compose()        masks igraph::compose()
## x plyr::count()           masks dplyr::count()
## x tidyr::crossing()       masks igraph::crossing()
## x tidyr::extract()        masks stringdist::extract()
## x plyr::failwith()        masks dplyr::failwith()
## x dplyr::filter()         masks stats::filter()
## x igraph::groups()        masks dplyr::groups()
## x plyr::id()              masks dplyr::id()
## x dplyr::lag()            masks stats::lag()
## x plyr::mutate()          masks dplyr::mutate()
## x plyr::rename()          masks dplyr::rename()
## x purrr::simplify()       masks igraph::simplify()
## x plyr::summarise()       masks dplyr::summarise()
## x plyr::summarize()       masks dplyr::summarize()
## x readr::tokenize()       masks textreuse::tokenize()
```

```
b = 10
m = 100
minhash = minhash_generator(n = m, seed = 1234)

dat = read.csv("sv-mauricio.csv")
dat = dat %>%
  filter(!is.na(HandID))



docs <- apply(dat, 1, function(x) paste(x[-c(1, 2, 5:11)], collapse = " ")) # get strings
head(docs)
```

```
## [1] "ALEMAN SOLIS ALFREDO"    "CRUS CARMEN"
## [3] "MONTOYA CARMEN"          "PAS SINGUENSA JUAN JOSE"
## [5] "GUIYEN TEODORO"          "MANOQUIN JULIA"
```

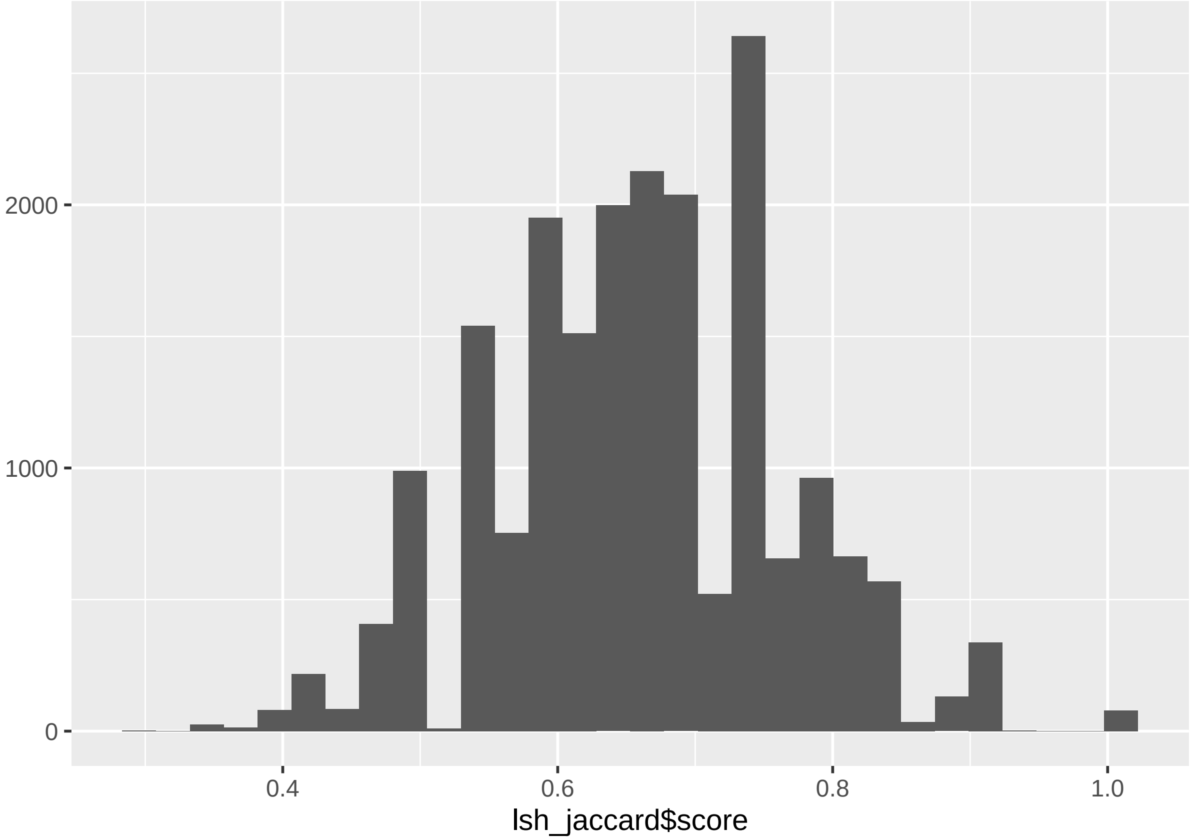
```
#docs <- apply(dat, 1, function(x) paste(x[-c(1, 2, 9)], collapse = " ")) # get strings


names(docs) <- dat$id # add id as names in vector
corpus <- TextReuseCorpus(text = docs, # dataset
                          tokenizer = tokenize_character_shingles, n = 1, simplify = TRUE, # shingles
                          progress = FALSE, # quietly
                          keep_tokens = TRUE, # store shingles
                          minhash_func = minhash) # use minhash

buckets <- lsh(corpus, bands = b, progress = FALSE)
candidates <- lsh_candidates(buckets)
lsh_jaccard <- lsh_compare(candidates, corpus,
                                        jaccard_similarity, progress = FALSE)


qplot(lsh_jaccard$score)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
g <- make_empty_graph(nrow(dat), directed = FALSE) # empty graph
#edges <- as.integer(as.vector(t(as.data.frame(candidates[, 1:2]))))
#g <- add_edges(g, edges)
#g <- set_vertex_attr(g, "id", value = dat$id) # add id

#clust <- components(g, "weak") # get clusters
#blocks <- data.frame(id = V(g)$id, # record id
                     #block = clust$membership) # block number
```