20180313_three_function_overview

references<-read_references()

authors<-read_authors(references, sim_score=0.9)

At this stage you go through and change any incorrectly grouped authors groupID to their authorID.

This function takes those changes and makes the names all the same, filters again, and makes some small changes and deletes some columns for input to your functions

refine_authors is doing what merge_records and remove_duplicates did before

authors_final<-refine_authors(authors, sim_score=0.94)

```
eb_references <- read_references("./data/EBpubs.txt", dir=FALSE,
                                 filename_root="./output/eb")

eb_authors <- read_authors(eb_references, filename_root="./output/eb")

eb_refined <- refine_authors(authors=eb_authors)

save(eb_refined, file="./output/eb_refined.Rdata")
```

```
load("./output/eb_refined.Rdata")

dat <- separate(data=eb_refined, col = address,
         into=c("university","department","short_address"),
         sep=",",extra = "merge", remove=FALSE) %>%
       mutate(country=stri_extract_last_words(short_address),
        zip = str_extract(string=short_address,
          pattern="[:digit:][:digit:][:digit:][:digit:][:digit:]"),
        city_state = str_extract(string=short_address,
                pattern="[:alnum:]{1,20}[,][ ][A-Z][A-Z]") ) %>%
      select(address, short_address, city_state, zip, country, university, department)
```

```r
head(dat[,c("address")])
```

```
## [1] Univ Florida, Dept Wildlife Ecol & Conservat, Gainesville, FL 32611 USA.
## [2] Univ Fed Pernambuco, Dept Bot, BR-50372970 Recife, PE, Brazil.
## [3] Univ Fed Pernambuco, Dept Bot, BR-50372970 Recife, PE, Brazil.
## [4] Univ Fed Pernambuco, Dept Bot, BR-50372970 Recife, PE, Brazil.
## [5] Univ Fed Pernambuco, Programa Posgrad Biol Vegetal, BR-50372970 Recife, PE, Brazil.
## [6] Univ Fed Pernambuco, Dept Bot, BR-50372970 Recife, PE, Brazil.
## 94 Levels: Univ Fed Pernambuco, Dept Bot, BR-50372970 Recife, PE, Brazil. ...
```

```r
head(dat[,c("university","department")])
```

```
##            university                      department
## 1        Univ Florida  Dept Wildlife Ecol & Conservat
## 2 Univ Fed Pernambuco                        Dept Bot
## 3 Univ Fed Pernambuco                        Dept Bot
## 4 Univ Fed Pernambuco                        Dept Bot
## 5 Univ Fed Pernambuco   Programa Posgrad Biol Vegetal
## 6 Univ Fed Pernambuco                        Dept Bot
```

```r
head(dat[,c("short_address","city_state","zip","country")])
```

```
##                      short_address      city_state   zip country
## 1       Gainesville, FL 32611 USA. Gainesville, FL 32611     USA
## 2  BR-50372970 Recife, PE, Brazil.      Recife, PE 50372  Brazil
## 3  BR-50372970 Recife, PE, Brazil.      Recife, PE 50372  Brazil
## 4  BR-50372970 Recife, PE, Brazil.      Recife, PE 50372  Brazil
## 5  BR-50372970 Recife, PE, Brazil.      Recife, PE 50372  Brazil
## 6  BR-50372970 Recife, PE, Brazil.      Recife, PE 50372  Brazil
```