Database

We describe each table in our database.

CustomUser
Each row is a user. The identifier is e-mail and the password is the hash password.

Research
This is the main table of our app. Each row define a research. For each research we have :

- a unique CustomUser associate with column «user». If the user is deleted, the research stays.

- a many-many relation with the table Article represented with intermediate table «Research_Article». These article are the fetched article from fetching article step of a research

- search: the search string that user gave for the research

- year_begin,year_end: two date that represent the interval the research fetch articles within.

- step: a string that indicate the current step of a research

- is_finish: a boolean indicate if the research terminated

- is_running: a boolean indicate if the research is running.

- max_article: the number of article there is at least.

- begining_date: the date when the research was created

- current_article_db: a string indicate the current article database where the research is fetching.

- number_neighbor: a integer that represents the number of neighbor

Keyword
This table keep all keywords of a research. We separate each keyword from the search string and save them here.

- word: the keyword
- research: the unique Research associated to the keyword.

Article
This table keeps all data of the article. Each row is an article.

- a many to many relation with table “Author” represented by Article_Author. The Author is only composed by a string for last name and a string for first name.

- title: a string for the title of the article

- doi: a string for the DOI of the article
- abstract : a string for the abstract of the article
- full_text: a string where there is the fulltext of a article that we extracted.
- publication: the date of the publication of this articles
- url_file: the url where we can fetch the pdf of the article.
- is_file_get: a boolean that check if the pdf of the article is in our machine.

Cluster
This table keep all information about the clustering of articles. Each row represent the clustering results for an article in a research.

- research: the unique Research associated to this clustering data.
- article: the unique Article associated to this clustering data.
- topic: a string that represent all topic of the cluster of the article. It identifie the cluster of the article
- pos_x,pos_y: represent the coordinate of article in the plot.

TableChoice
This table permit to manage the step of select manually the articles. Each row represent an article from a research and a user in the page with table and checkbox.

- user: the unique user that created the set of article to check manually. Each user may have a set of article. If he creates another set, the row that represents the ancient one are deleted.
- research: the unique Research associated to the set of article.
- article: the unique article of the row
- to_display: a boolean that checks if the article has to be displayed for the selection
- is_initial: a boolean that checks if this article is from the filtering and not from the neighbors of article checked by iteration
- is_check: a boolean that checks if an article was choosen and his checkbox was checked.

There are some table secondary.

Preprocess_text
during preprocesing of text of article, we save each word in this table beacause it is parallelized and this is the only way to keep result from parallelisation of preprocessing.

Number_trial et Number_preprocess
To keep the achievement progress of the number of trial done or the number of article done, we use these table. But, as we make parallelized processing and there are problem to write on the same row, we add for each trial/article preprocessed a new row with his research. We count the number of row with a research to obtain the progression.