Steps of the analysis as implemented in R

by Raphaël Champeimont, OST, December 2014

raphael.champeimont@almacha.org

A. Extraction of the data from the Oracle OST database: computation of the similarity matrix between categories, computation of the number of references by category and by article.

These scripts have to be adapted to our own database, to your choice of the classes defining the Stirling index and to your choice of the corpora to study. We chose WoS categories as classes and WoS categories as corpora.

1. **compute_similarity_matrix_from_DB.R** computes the similarity matrix *simMat* between WoS categories as classes of references with whole counts (articles and references in journals attributed to many categories are counted with weight equal to 1 in each of these categories). This similarity matrix, as computed from the OST database for the period 2008-2012, is shown in directory ***data_integer***

1f. **compute_similarity_matrix_from_DB_frac.R** is the *Fractional count version* of the preceding script (articles and references in journals attributed to *n* categories are counted with weight *1/n* in each of these categories)

2. **compute_counts_from_DB_WoS.R** computes the counts of references by category for each article in the db. It provides separate tables for the sets of articles in each corpus. In these scripts, the corpora under study are defined as the publications in WoS categories, called category-corpus. Data not disclosed.

2f. **compute_counts_from_DB_WoS_frac.R** is the corresponding version for fractional counts of references in categories

3. **compute_counts_from_DB_WoS_institution.R** computes the counts of references by category for each article of an institution. For each institution, it provides separate tables for the sets of articles in each category-corpus where the institution has a *minimumNumberOfArticles.* As an example, counts for the corpora associated with WoS categories RU (Neurosciences) and RT (Clinical Neurology) for 8 universities are shown in directory ***data_integer/input***

3f. **compute_counts_from_DB_WoS_institution_frac.R** is the corresponding version for fractional counts of references in categories

B. Computation of the indicators for each institution and for each corpus of institution publications [1]

These scripts can be run on your own input data if they are displayed with the same filenames and format as ours in ***input/data_integer***[2]***.***

4. **run_all_stat_calc_step1.R** computes the different statistics for the whole WoS and for each category-corpus, using steps 1 and 2 above and the following scripts:

common.R which i) defines the impact of the *fracMode* parameter which selects the data corresponding to the choice of *whole* or *fractional* counts; ii) defines the function that normalizes the counts of references by article (EWA option in Cassi et al. 2014); iii) defines the function that computes the Stirling index *STa* for each single article *a;* iv) a few other useful tricks...

load_similarity_matrix.R

stat_calc_step1_world.R for the corpus associated with the publications in each category computes: i) the article indicators STa and the centiles of this distribution; ii) the within and between indexes; iii) the proportions (Q) and the contributions of the categories to the global index.

Results as computed from the OST database for the period 2008-2012 for the WoS categories RU (Neurosciences) and RT (Clinical Neurology) are presented in ***data_integer/stat_results_world***

4f. **run_all_stat_calc_step1_frac :** same as 4 with fractional counts of reference categories

1 We have implemented the EWA (Equal Weights by Article) option. See Cassi *et al.* 2014

2 Or in ***input/data_frac***

You can run the following scripts on the data we provide in **data_integer**. *Set as working directory (setwd) the directory where the scripts are*

5. **run_all_stat_calc_step2.R** computes i) the different statistics for all institutions and for each category-corpus ; ii) the contributions to the global index of each category-class for a category-corpus of each institution. It uses steps 1, 3 and 4 above and the following scripts:

common.R

load_similarity_matrix.R

stat_calc_step2_institutions.R computes, for the institutions selected in step A3 and for one category-corpus, the differences between the institution and the world indexes for the following indicators: global, within and between indexes and their z-scores and p-values, proportions (Q) and contributions of the categories (classes of references) to the global index.

stat_calc_step3_test_contributions.R computes each category contribution for the selected corpus and institutions and their the z-scores and p-values.

Results of these 2 scripts for the WoS category-corpora RU (Neurosciences) and RT (Clinical Neurology) for 8 universities are presented in ***data_integer/stat_results***

5f. **run_all_stat_calc_step2_frac.R :** same as 5 for fractional counts

C. Plots

6. **make_all_plots.R** launches the following five families of plots:

plot_by_corpus.R makes, for each corpus, a 2d plot of (STW, STB) for the selected institutions. The size of the points is proportional to the number of articles of the institutions.

plot_by_corpus_confidence.R makes, for each corpus, a 2d plot of (STW,STB) with the confidence areas of a chosen probability (parameter *alpha*).

plot_contributions.R makes, for each corpus, and for each institution, a figure with the contributions of each category (class of references). Differences between the institution and the world values are displayed. The size of each point is the proportion of references in the category.

plot_by_institution.R makes for each institution a 2d plot of (STW, STB) for the selected corpora. The size of the points is proportional to the number of articles in each corpus.

plot_by_institution_confidence.R makes for each institution a 2d plot of (STW, STB) with the confidence areas of a chosen probability (parameter *alpha*).

Results of these 6 scripts for the WoS category-corpus RU (Neurosciences) and RT (Clinical Neurology) for 8 universities are presented in ***data_integer/plots***

6f. **make_all_plots_frac.R** same as 6 for fractional counts