R programming – Assignment
M1 ROAD, M1 IREF
Fall 2022

LAURENT R. BERGÉ[*]

October 6, 2022

Due date. The 20^{th} of November 2022.

Notation. There are 22 points over 20.

Output expected. You shall return a single R file, containing a section for each question, well (but not too much) commented (so that anyone reading it can understand), and replicable (when I'll run it on my laptop, everything should work[1]).

The file should be named according to the persons in the group: `NAME1_NAME2.R`, etc.

If you do the bonus graph, please provide a JPEG file named in the same way as the R script.

1 The problem

The aim of this problem is to study whether specialized scientists perform better than diversified scientists (or vice versa). In other words, we'll study whether there is a correlation between specialization and citations (a common measure of publication performance). The challenge will be to create a measure of specialization.

The context. The study will focus on Bordeaux scientists (i.e. from Bordeaux institutions) from all fields.

A text-based specialization measure. To assess the level of specialization of a scientist, we will use the abstracts of their publications. The assumption is that the closer the abstracts are to each other in terms of word similarity, the closer they are in topic. The measure of specialization will be based on the Jaccard index.

Let W_a be the set of *unique* words of abstract a. The Jaccard index between two abstracts a and b is defined as:

$$Jaccard_{ab} = \frac{|W_a \cap W_b|}{|W_a \cup W_b|},$$

which is equal to the number of words the abstracts a and b have in common, divided by the total unique number of words across the two abstracts.

Using the Jaccard measure, the specialization measure for scientist i writes:

$$\text{Specialization}_i = \frac{1}{n_i \times (n_i - 1)} \sum_{a=1}^{n_i} \sum_{b=1, b\neq a}^{n_i} Jaccard_{ab}, \qquad (1)$$

which is simply the average Jaccard measure between all pairs of publications of scientist i. Of course, the higher this value, the more specialized the scientist is.

[*] B_xSE, UMR CNRS 6060, University of Bordeaux, *e-mail*: laurent.berge@u-bordeaux.fr

[1] Except file paths of course.

Data. For this problem we'll use data from five text files contained in the archive `R2022--DATA.zip`. The text files contain the following information:

1. `abstract.txt`: publication ID and abstracts
2. `author_name.txt`: auhtor ID and author name
3. `paper_author_instit.txt`: the paper ID, the author ID and the institution name
4. `paper_info.txt`: Paper information. Contains the paper ID, the type of document, and the date of publication.
5. `references.txt`: Contains all the references in a given paper, in the form of the paper ID citing and the paper ID cited.

The source of this data is the Microsoft Academic Graph (MAG), for which you can find more information here (Sinha et al., 2015). All the files you'll be using are small extractions (both in terms of rows and columns) of the larger MAG data set.

2 The exercises

1. **3 pts.** [*Importing and formatting the main database.*] Create the following main data set:[2]

author_id	author_name	institution	paper_id	year	abstract
2168626811	georges dooms	Centre [...] de Luxembourg...	1497678041	2016	Childhood obesity is...
2892250329	jeanclaude schmit	Lab [...] Retrovirology, [...] Luxembourg	2020087881	2012	Chemokines and their...

You will create the data with the following restrictions:

- **Keep only scientists working in Bordeaux institutions**: that is only institutions that contain "Bordeaux" in their name (forget about acronyms).
- Drop any missing values.

Tip: pen and paper planning before any action is taken will save time.

2. **1 pt.** Restrict the sample to relevant information only:

- A valid abstract must have at least 100 characters – drop non-valid abstracts,
- keep only scientists with at least 3 publications with valid abstracts.

This data set is the *focus sample*.

3. **2 pt.** Create a set of *unique* keywords for each publication.

(a) Clean the abstracts:

i. use the command `gsub("[^[:alpha:]]", " ", x)` to remove all the non alpha numeric characters,

ii. fix the case,

iii. drop all words with one or two letters (in other words: keep only words with 3+ letters).

(b) From these cleaned abstracts, create the keywords, knowing that keywords are defined as:

- not appearing in 10% or more abstracts (i.e. drop words appearing in 10% or more abstracts),
- appearing in at least 5 different abstracts (i.e. drop words appearing in less than 5 abstracts).

Each publication is then related to a set of *unique* keywords. You should end up with a data set containing two variables: the paper ID and the keyword.

4. **4 pts + 1 pt.** For each Bordeaux scientist, create the specialization measure defined in Equation (1).

Tip: For this question, think hard to the problem with pen and paper before dealing with it. There is a solution without loops: you get an extra point if you find it! (By the way, you may need to use the argument allow.cartesian = TRUE when merging.)

[2] This is an example of output. You can name the variables as you want.

5. **3 pts.** Using the file `references.txt`, for each publication of the focus sample, create the variable, say `nb_cites`, equal the number of citations a publication receives in a 5 years window. This means that for a publication published in 2004, call it the *origin* publication, `nb_cites` will be equal to the number of articles published between 2004 and 2009 that cite the *origin article.*[3]

6. **1 pt.** To the data set of question 3, add the number of citations obtained in a 5 years window. You should end with the following three variables:

   - keyword,
   - paper ID,
   - number of citations in a 5 years window.

7. **1 pt.** Create the table with the average number of citations per keyword, and display the 20 top keywords.

8. **2 pt.** Compute the average number of citation received for each scientist of the focus sample and merge this information to the specialization measure. Create the database containing the following information for each scientist (this is the aggregate information across all years):

   - author ID,
   - author name,
   - number of publications,
   - average number of citations,
   - total number of citations,
   - specialization measure.

9. **Bonus 1 pt.** If you do this question, please also provide a JPEG image of the graph (on top of the R script).

- **0.5 pt.** Using this data, plot the correlation (i.e. the scatterplot) between specialization and citations. You can use base R or ggplot2. Please try to make the graph look nice.

- **0.5 pt.** Show on the graph some noteworthy scientist names.

3 Notation

17 pts	Main questions
	You need a yes to the 3 Qs below to get full points to a question
	Is the answer conceptually correct? *Is the code correct?* *Is the code not too slow?*
3 pts	*General clarity and quality of the document*
+ 1 pt	*Bonus in Q4*
+ 1 pt	*Bonus graph*
+ 1 pt	The only packages used are data.table and ggplot2

References

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P., Wang, K., 2015. An overview of Microsoft academic service (MAG) and applications. ***Proceedings of the 24th international conference on world wide web***: 243–246.

[3] For late publications there is not enough information to create an accurate measure of the number of citations in a 5 years window: e.g. for publications in 2018, there is obviously no data up to 2023. Create the values for these late years too, just do as if they just received no citations.