A tutorial for vosviewer

Clément Levallois

Version 1.6.5, 2017-03-29

Table of Contents

Presentation of this tutorial 1
Importing a dataset 2
    1. Importing a bibliometric dataset (WoK, Scopus, etc.) 2
    1. Importing a text (of a bibliometric nature or not) 7
    3. Importing a network (gml format or pajek) 15
(to be continued) 15
More tutorials on vosviewer 15
the end 15

last modified: 2017-03-30

VOSviewer

Visualizing scientific landscapes

Nees Jan van Eck and Ludo Waltman
Centre for Science and Technology Studies
Leiden University

Presentation of this tutorial

This tutorial explains how to create a semantic map like this one:

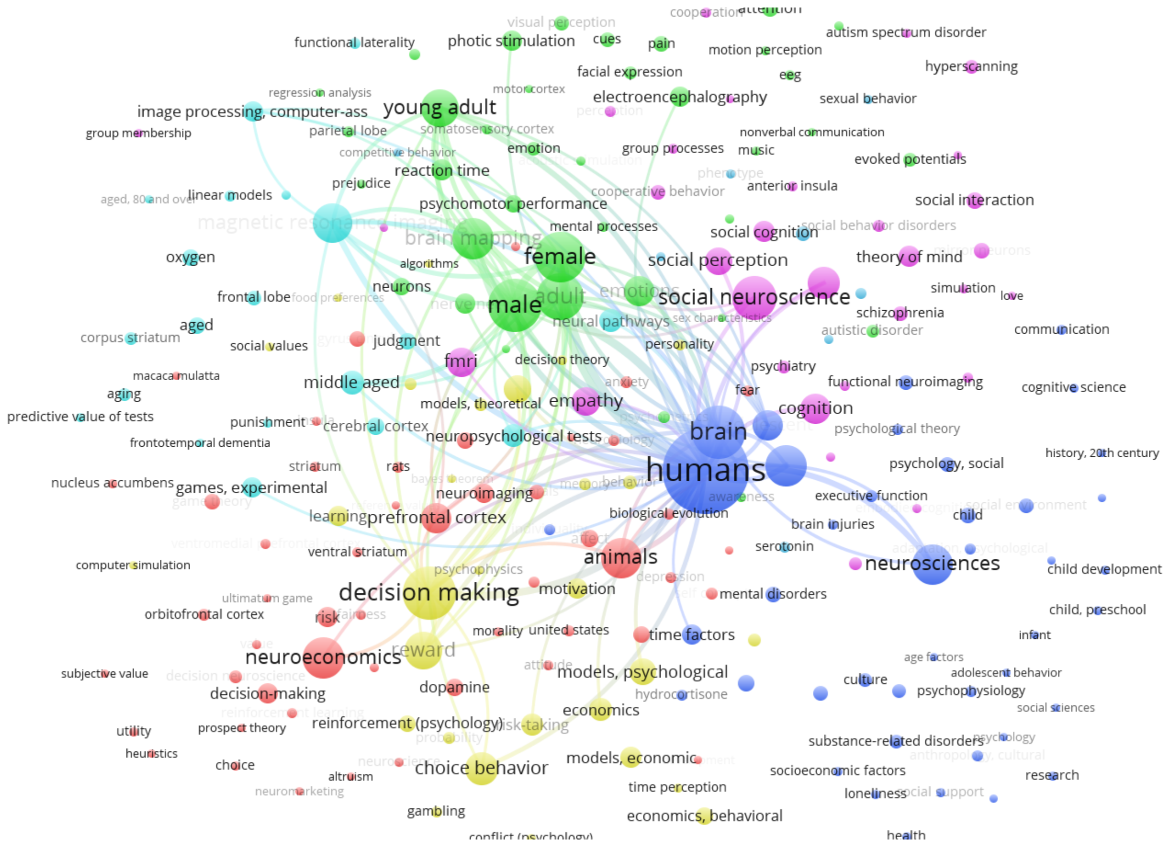


Figure 1. a map with VosViewer

VosViewer is developed by Nees Jan van Eck and Ludo Waltman, researchers at the CWTS Leiden of Leiden University in Leiden.

The software was built for the analysis of scientometric data (the scientific study of the scientific activity), but the software has a broader relevance.

In particular, VosViewer is **particularly good at producing textual maps** of any sorts, not just from scientometric datasets.

A 2.0 version of VosViewer is announced for 2017, capable of handling larger datasets and broadening its focus (will explicitly target non-scientometricians).

Importing a dataset

We can use several methods:

- directly importing datasets of a bibliometric nature (Web of Knowledge, Scopus, Medline or RIS)
- importing a text, where each paragraph will be considered a "unit of text" to compute co-occurrences.
- importing your own networks (created for example with Gephi, etc.)

Let’s see each of these methods in turn.

1. Importing a bibliometric dataset (WoK, Scopus, etc.)

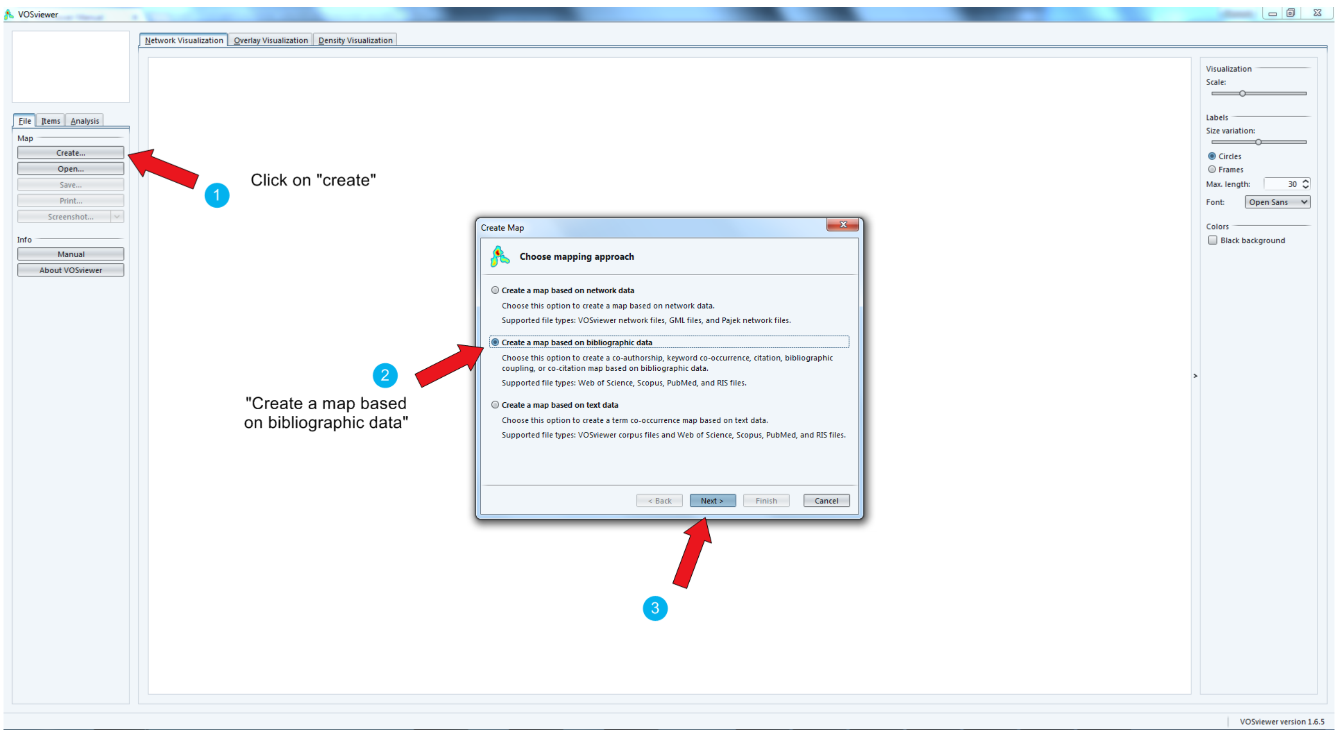
We need a bibliometric dataset to practice.

download this zip file and unzip it on your computer.

The file `pubmed_medline.txt` inside the zip contains 1484 research articles from the PubMed database of scientific reports, retrieved by conducting this query:

"social neuroscience" OR "neuroeco*" OR "decision neuroscience"

→ The query can be seen at online here.



Create Map

Choose mapping approach

( ) **Create a map based on network data**

Choose this option to create a map based on network data.

Supported file types: VOSviewer network files, GML files, and Pajek network files.

(●) **Create a map based on bibliographic data**

Choose this option to create a co-authorship, keyword co-occurrence, citation, bibliographic coupling, or co-citation map based on bibliographic data.

Supported file types: Web of Science, Scopus, PubMed, and RIS files.

( ) **Create a map based on text data**

Choose this option to create a term co-occurrence map based on text data.

Supported file types: VOSviewer corpus files and Web of Science, Scopus, PubMed, and RIS files.

< Back | Next > | Finish | Cancel

Figure 2. Follow the wizard for biblio data - 1

Get your file `pubmed_medline.txt` ready as we need it now:

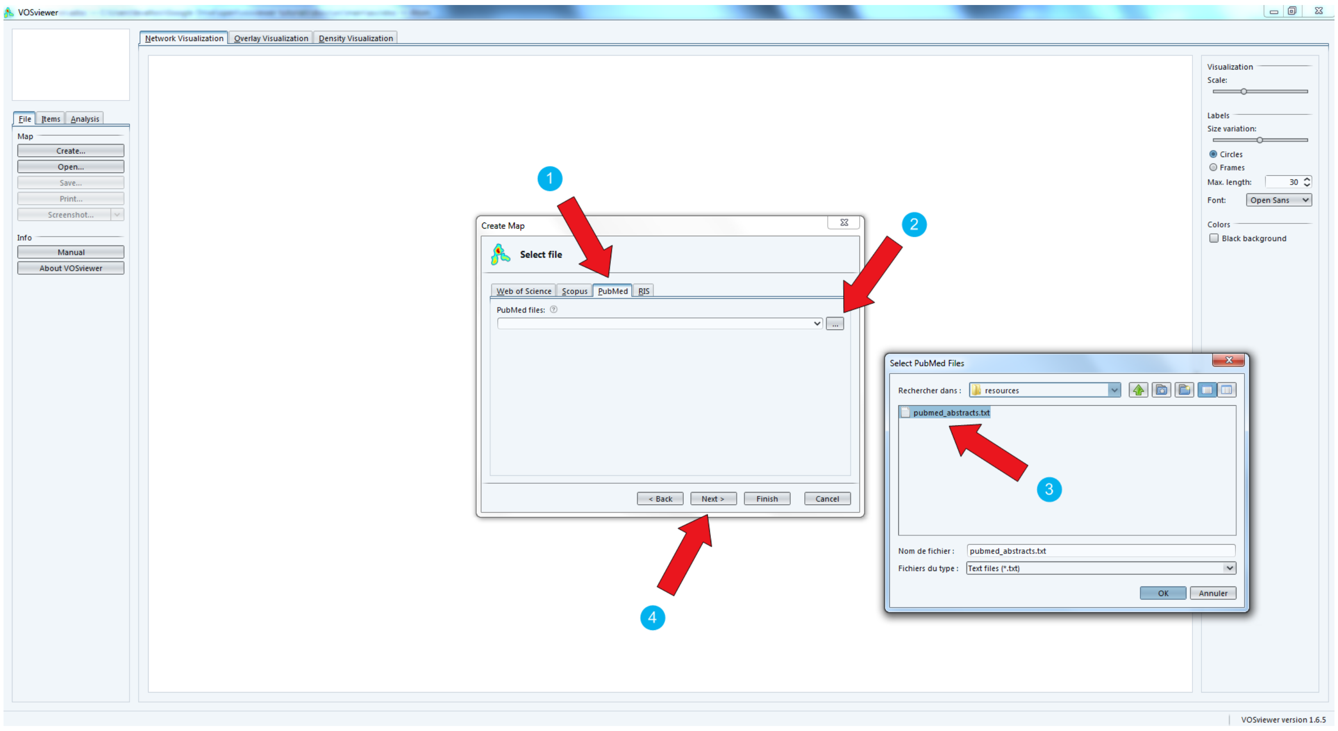


Figure 3. Follow the wizard for biblio data - 2

Now is the step when you decide what kind of network you are interested in.

TIP

> Careful, a network of co-occurrences of keywords does not use the abstracts of the bibliographic entries, but only the keywords supplied by the authors or the editor. These are typically much less informative. There is another way in VosViewer to work with abstracts.

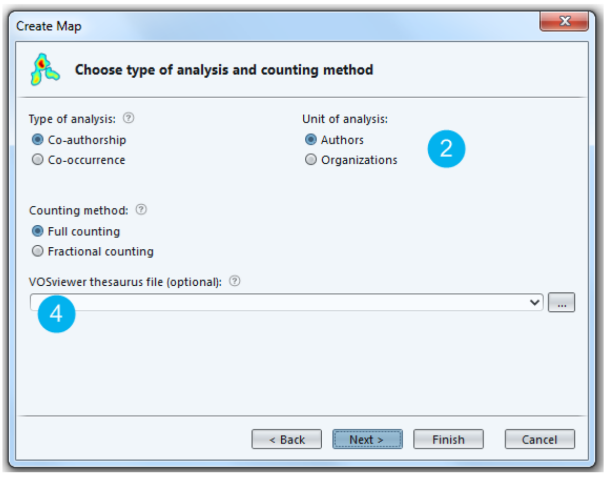
The full web link to the explanation of full vs binary counting: http://seinecle.github.io/gephi-tutiorals/working-with-text-en.html#binary-counting

1. co-authorships -> networks of co-authors

co-occurrences -> networks of the keywords provided by the authors or the editor (not terms in the abstracts!)

3. on counting, see also the explanation here:

http://seinecle.github.io/gephi-tutiorals/working-with-text-en.html#binary-counting



Choose type of analysis and counting method

Type of analysis: [✓ Co-authorship] Co-occurrence

Unit of analysis: [✓ Authors] Organizations

Counting method: [✓ Full counting] Fractional counting

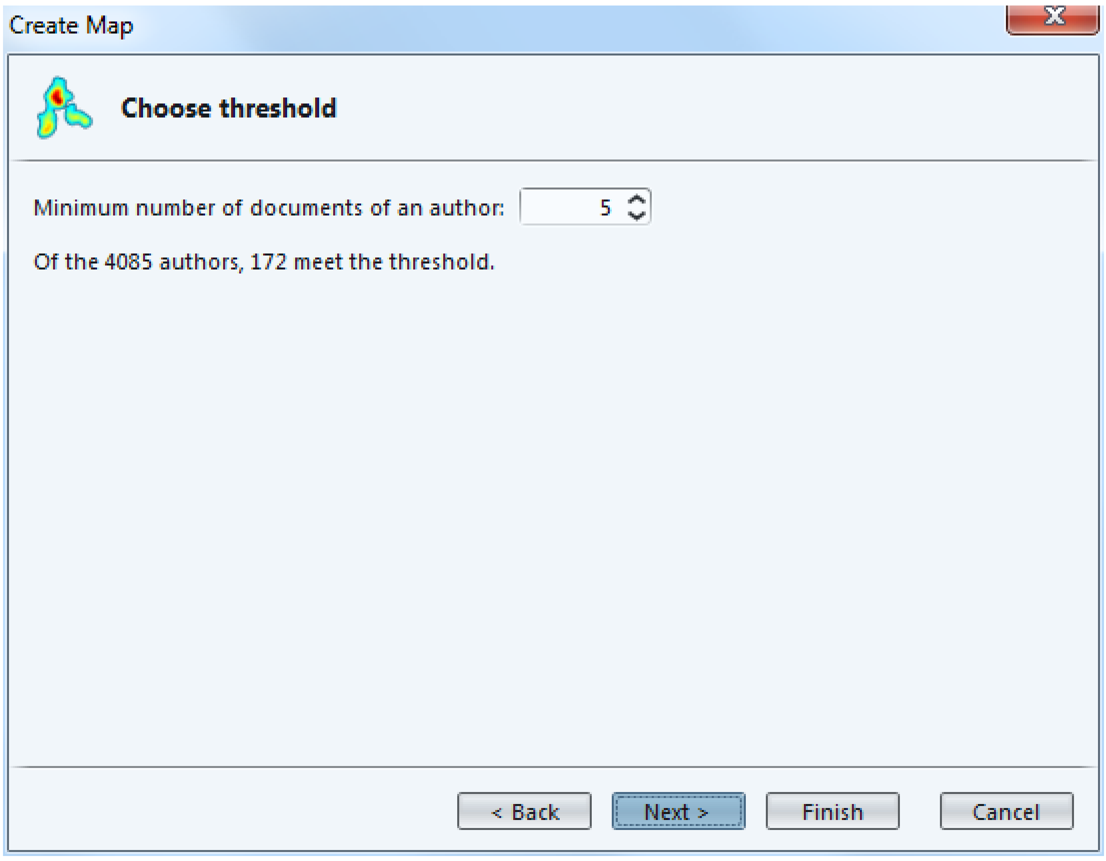
VOSviewer thesaurus file (optional):

2. Are you interested in a network of authors, or a network of the organizations listed in the bibliographic entries?

4. A thesaurus file is often necessary to merge different variants / misspellings of an author name or term.

Figure 4. Follow the wizard for biblio data - 3

You decide how frequent an author should be in your dataset to be included in the visualization. Here, individuals who have authored less than 5 papers in the collection will be left out.



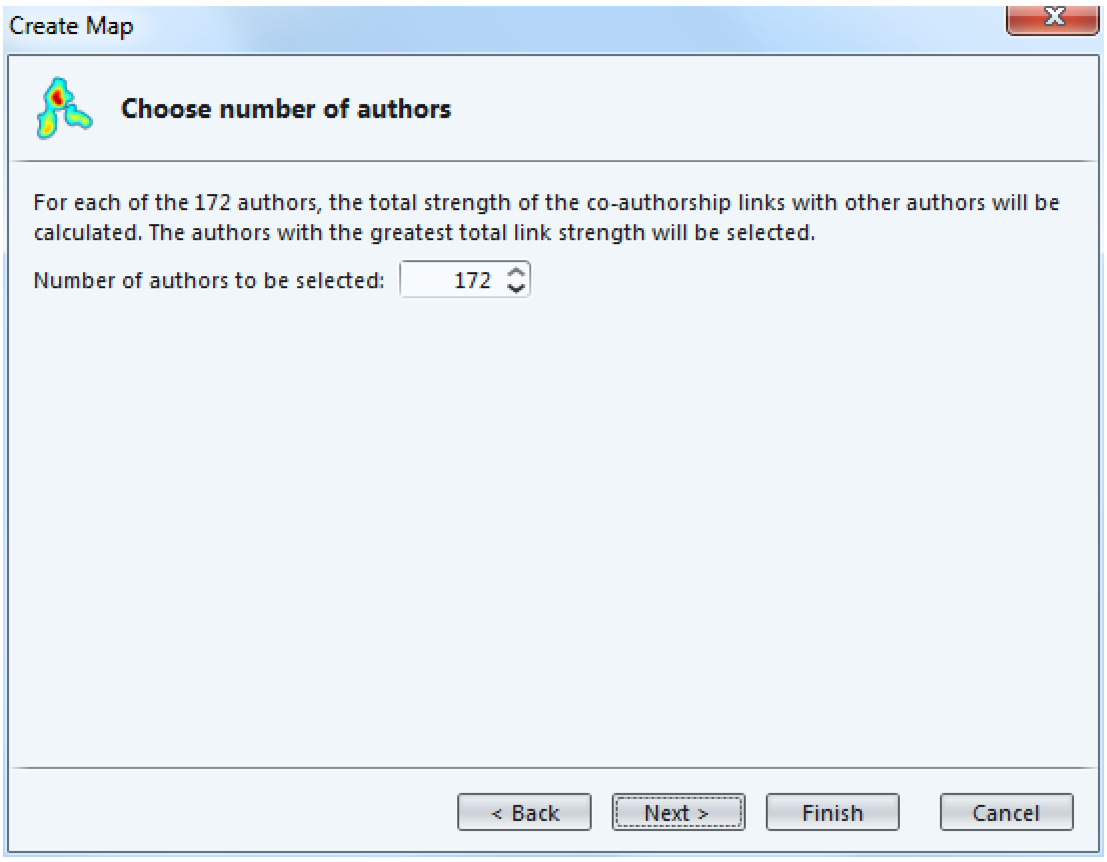
Choose threshold

Minimum number of documents of an author: 5

Of the 4085 authors, 172 meet the threshold.

Figure 5. Follow the wizard for biblio data - 4

Here you can decide that authors who are just weakly connected to others (because they did not coauthor much) can be left out.



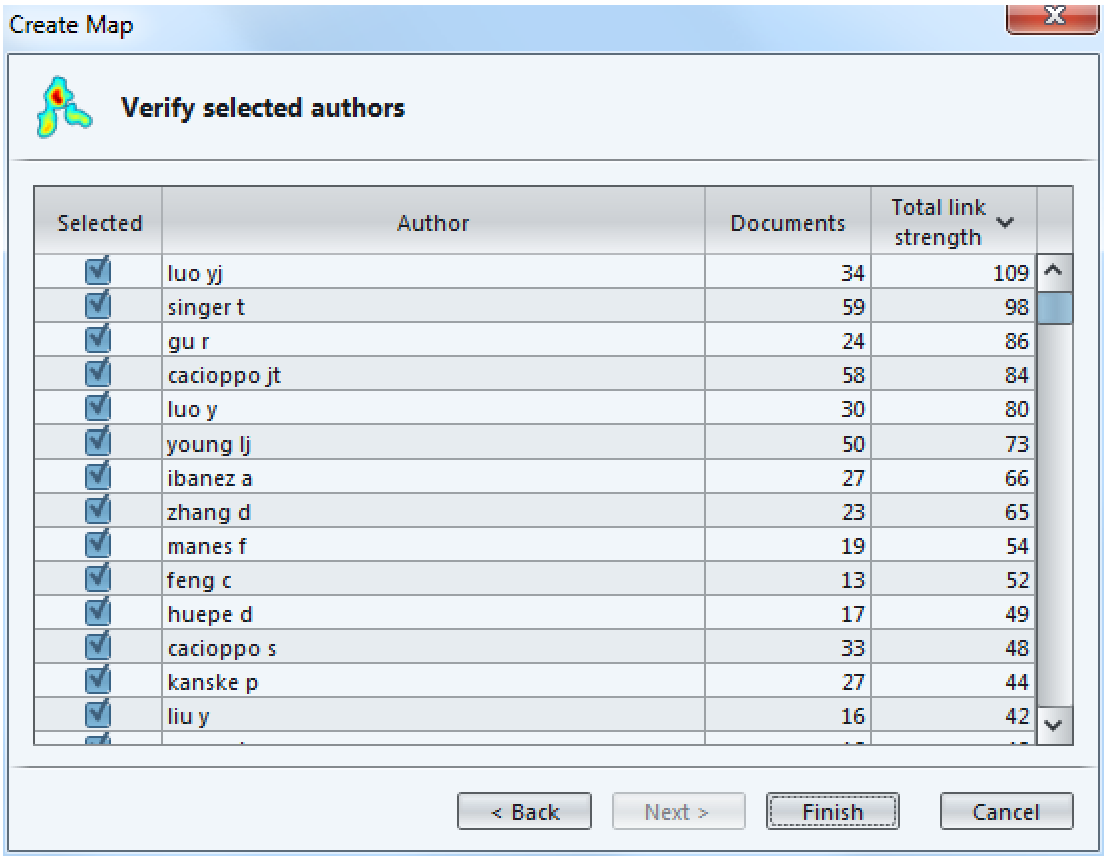
For each of the 172 authors, the total strength of the co-authorship links with other authors will be calculated. The authors with the greatest total link strength will be selected.

Number of authors to be selected: 172

Figure 6. Follow the wizard for biblio data - 5

The next screen is a table to recap which authors are going to be included in the visualization.

We can notice there are near duplicates (`luo yj` and `luo y`), which means we should have used the thesaurus to merge them earlier.

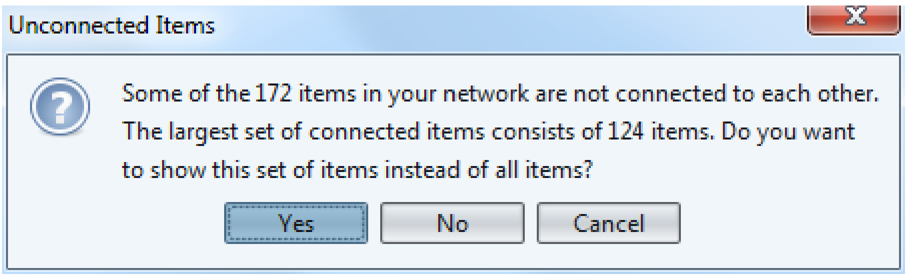


Selected	Author	Documents	Total link strength
✓	luo yj	34	109
✓	singer t	59	98
✓	gu r	24	86
✓	cacioppo jt	58	84
✓	luo y	30	80
✓	young lj	50	73
✓	ibanez a	27	66
✓	zhang d	23	65
✓	manes f	19	54
✓	feng c	13	52
✓	huepe d	17	49
✓	cacioppo s	33	48
✓	kanske p	27	44
✓	liu y	16	42

Figure 7. Follow the wizard for biblio data - 6

The last step of the wizard is an invitation to show only the largest group of connected authors. Smaller groups, isolated from this big group, would be discarded.

Clicking on "No" makes sure we see all groups.



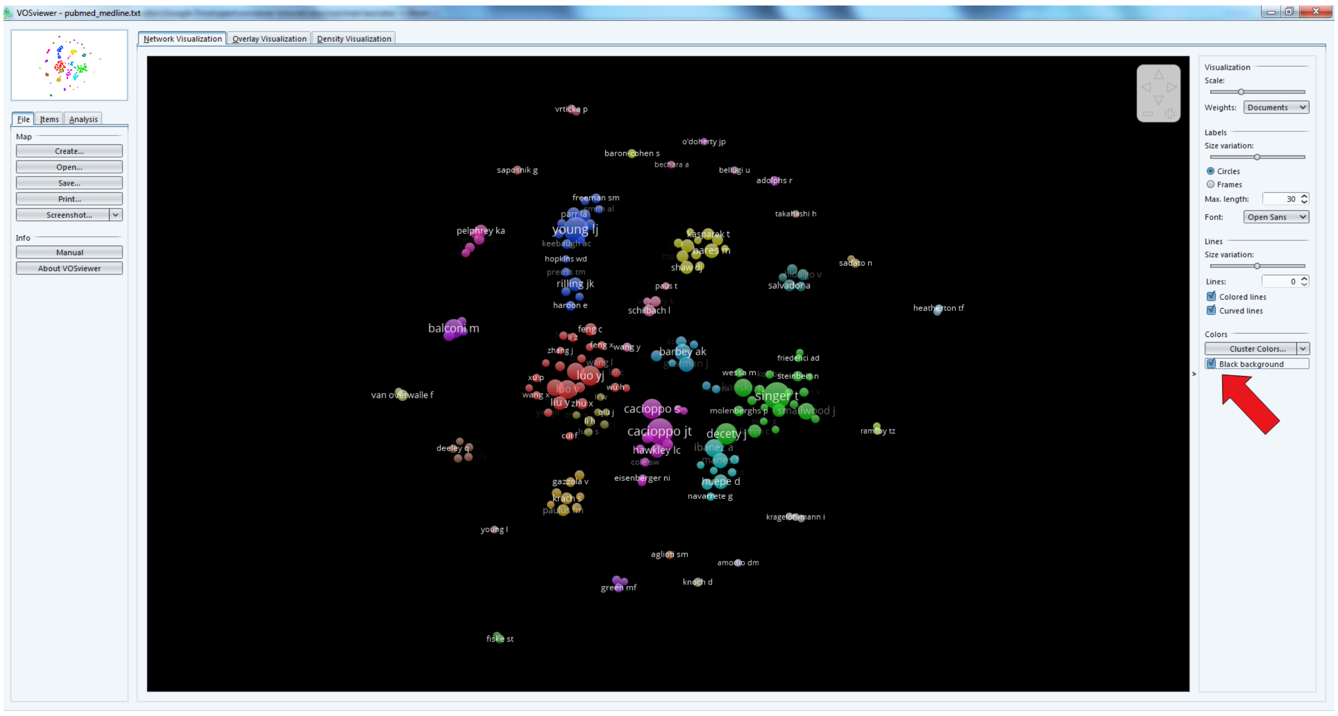
Some of the 172 items in your network are not connected to each other. The largest set of connected items consists of 124 items. Do you want to show this set of items instead of all items?

Figure 8. Follow the wizard for biblio data - 7

The result is a network made of groups of authors and their connections, corresponding to the fact that they co-authored papers.

Bigger nodes represent authors with the most publications.

Notice that the background can be switched to black - it can make it easier to read.



Weights: Documents
Size variation:
Circles
Frames
Max. length: 30
Font: Open Sans
Size variation:
Lines: 0
Colored lines
Curved lines
Cluster Colors...
Black background

Figure 9. Follow the wizard for biblio data - 8

Before going into the details of the visualization, we'll see how to import to other kinds of data source:

1. Importing a text (of a bibliometric nature or not)

We need a dataset to practice.

The dataset must be a single text file, where each unit of text **is a paragraph**. The network will be made of most frequent terms in the network. Terms appearing in the same paragraphs will be connected.

See this page for more precisions on units of text / paragraphs.

download this zip file and unzip it on your computer.

The file `pubmed_abstracts.txt` inside the zip contains 1484 abstracts (short summaries) of research articles from the PubMed database of scientific reports, retrieved by conducting this query:

"social neuroscience" OR "neuroeco*" OR "decision neuroscience"

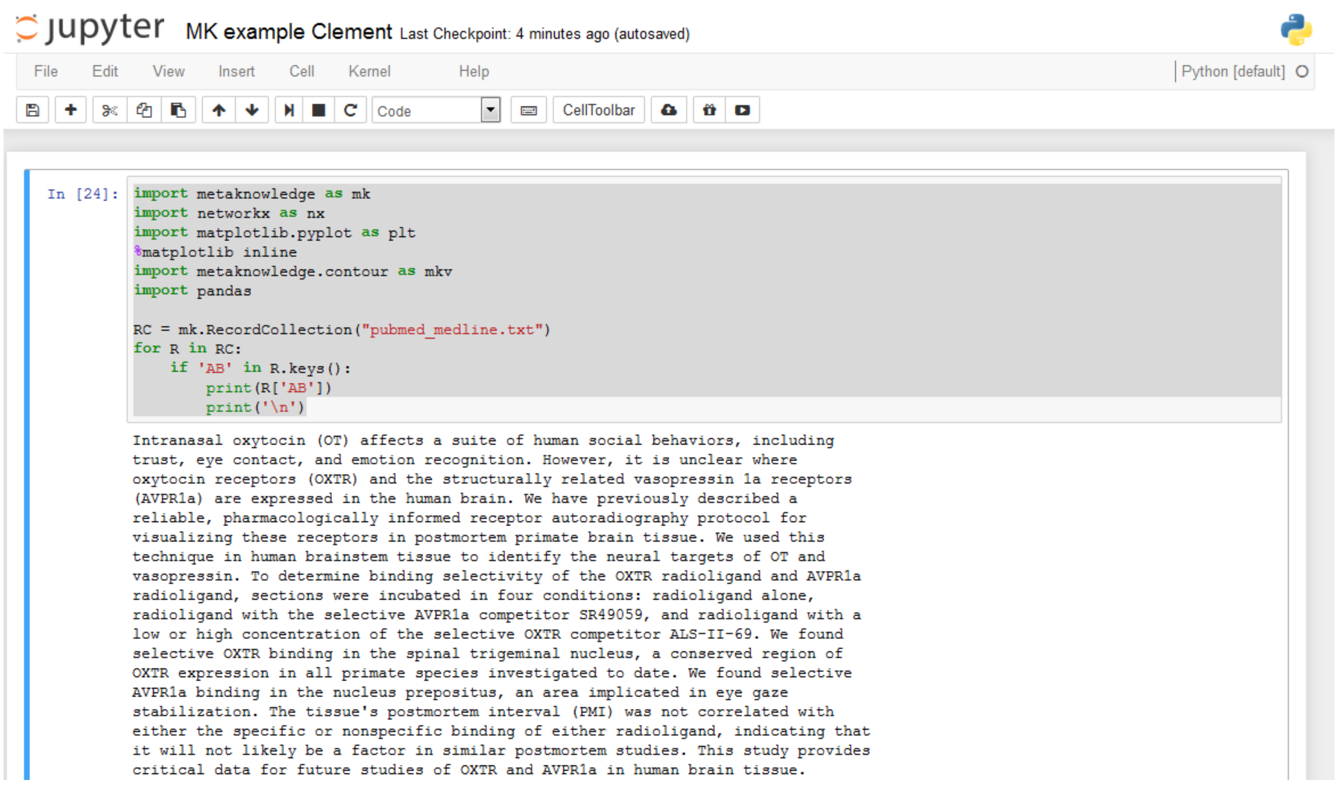
→ The query can be seen at online here. (the query returns more than 1484 results, because some articles have no abstract).

As a sidenote, these abstracts were extracted from the pubmed records in the following way:

- installing Anaconda, a toolkit for Python with many useful things inside
- installing Metaknowledge, a bibliometric toolkit in Python
- running this Jupyter notebook:

```
import metaknowledge as mk
import networkx as nx
import matplotlib.pyplot as plt
%matplotlib inline
import metaknowledge.contour as mkv
import pandas

RC = mk.RecordCollection("pubmed_medline.txt")
for R in RC:
    if 'AB' in R.keys():
        print(R['AB'])
        print('\n')
```



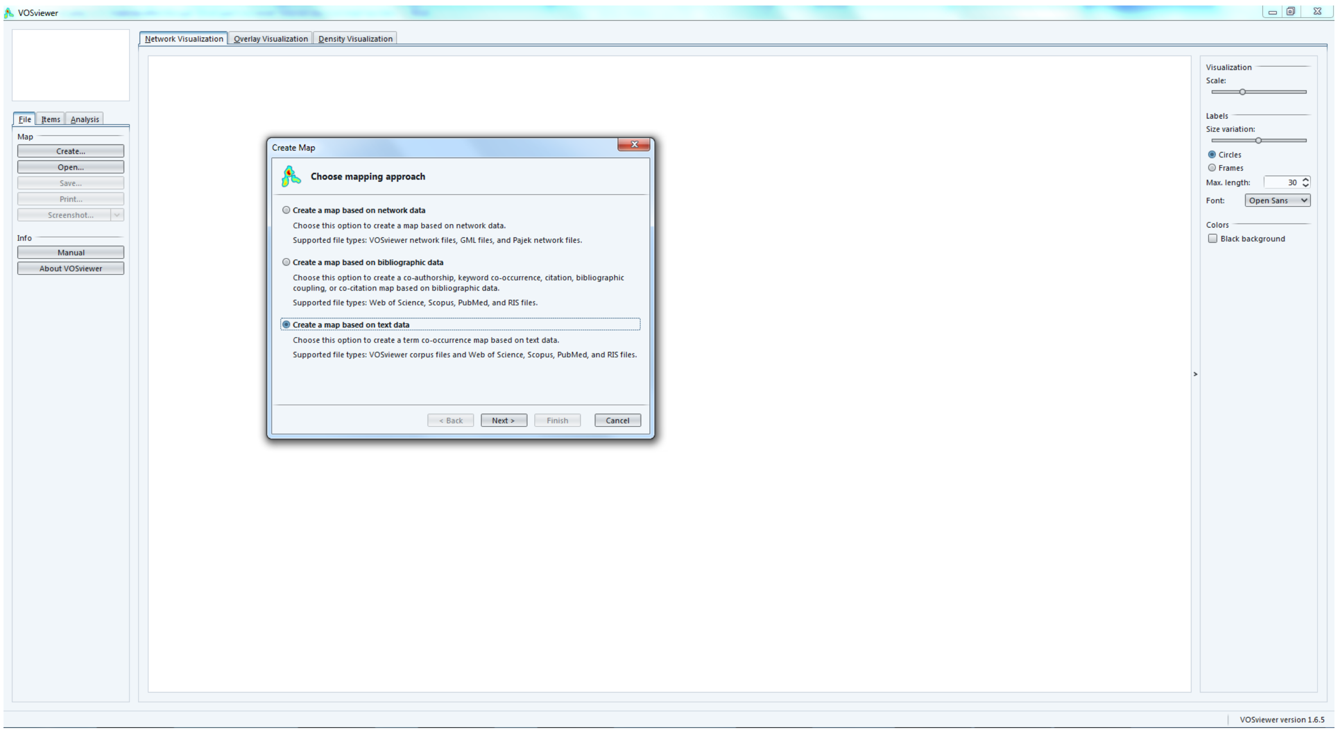
```
In [24]: import metaknowledge as mk
         import networkx as nx
         import matplotlib.pyplot as plt
         %matplotlib inline
         import metaknowledge.contour as mkv
         import pandas

         RC = mk.RecordCollection("pubmed_medline.txt")
         for R in RC:
             if 'AB' in R.keys():
                 print(R['AB'])
                 print('\n')

         Intranasal oxytocin (OT) affects a suite of human social behaviors, including
         trust, eye contact, and emotion recognition. However, it is unclear where
         oxytocin receptors (OXTR) and the structurally related vasopressin 1a receptors
         (AVPR1a) are expressed in the human brain. We have previously described a
         reliable, pharmacologically informed receptor autoradiography protocol for
         visualizing these receptors in postmortem primate brain tissue. We used this
         technique in human brainstem tissue to identify the neural targets of OT and
         vasopressin. To determine binding selectivity of the OXTR radioligand and AVPR1a
         radioligand, sections were incubated in four conditions: radioligand alone,
         radioligand with the selective AVPR1a competitor SR49059, and radioligand with a
         low or high concentration of the selective OXTR competitor ALS-II-69. We found
         selective OXTR binding in the spinal trigeminal nucleus, a conserved region of
         OXTR expression in all primate species investigated to date. We found selective
         AVPR1a binding in the nucleus prepositus, an area implicated in eye gaze
         stabilization. The tissue's postmortem interval (PMI) was not correlated with
         either the specific or nonspecific binding of either radioligand, indicating that
         it will not likely be a factor in similar postmortem studies. This study provides
         critical data for future studies of OXTR and AVPR1a in human brain tissue.
```

Figure 10. Using Metaknowledge to extract abstracts from records

We follow the wizard that will create the text map for us:


Create Map

Choose mapping approach

( ) **Create a map based on network data**

Choose this option to create a map based on network data.

Supported file types: VOSviewer network files, GML files, and Pajek network files.

( ) **Create a map based on bibliographic data**

Choose this option to create a co-authorship, keyword co-occurrence, citation, bibliographic coupling, or co-citation map based on bibliographic data.

Supported file types: Web of Science, Scopus, PubMed, and RIS files.

[✓ **Create a map based on text data**]

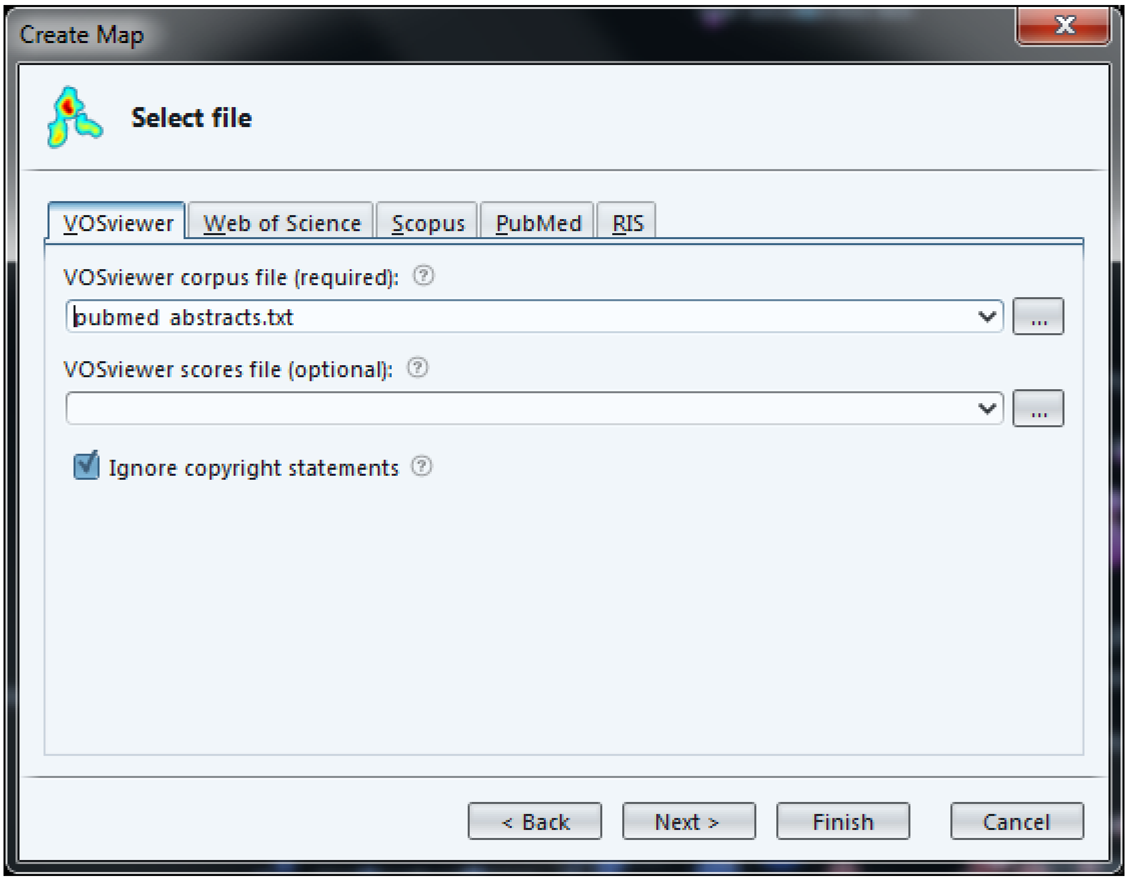
Choose this option to create a term co-occurrence map based on text data.

Supported file types: VOSviewer corpus files and Web of Science, Scopus, PubMed, and RIS files.

Figure 11. Follow the wizard for text maps - 1

Now we can choose the format of our input file.

`Vosviewer` format: this is where we can load a file made of paragraphs:


Create Map

Select file

VOSviewer corpus file (required): pubmed abstracts.txt

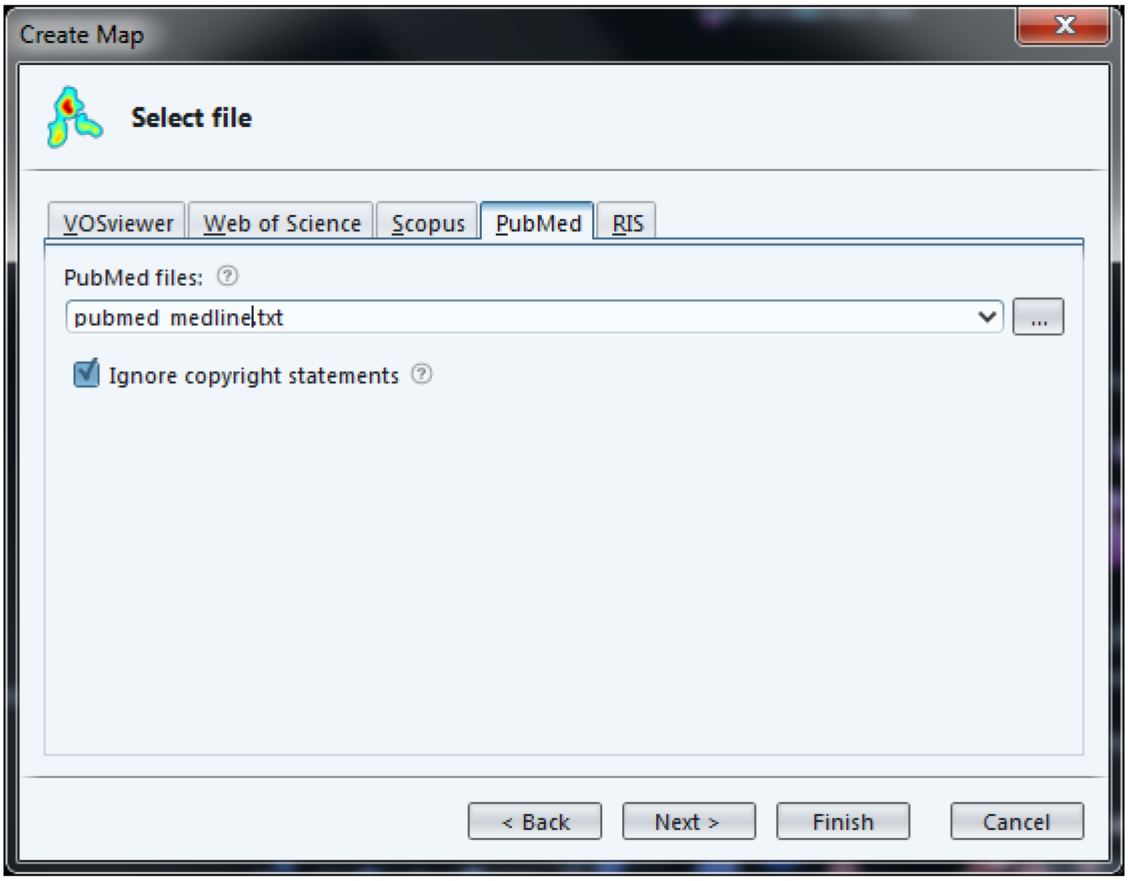
VOSviewer scores file (optional):

[✓] Ignore copyright statements

Figure 12. Choosing the VosViewer format

But VosViewer has scientometrics functions, which actually can extract abstracts directly from MEDLINE entries (or Scopus, Web of Knowledge…)

So we can actually come back to the file we had at the beginninf of this tutorial (`pubmed_medline.txt`) and choose the `PubMed` format:



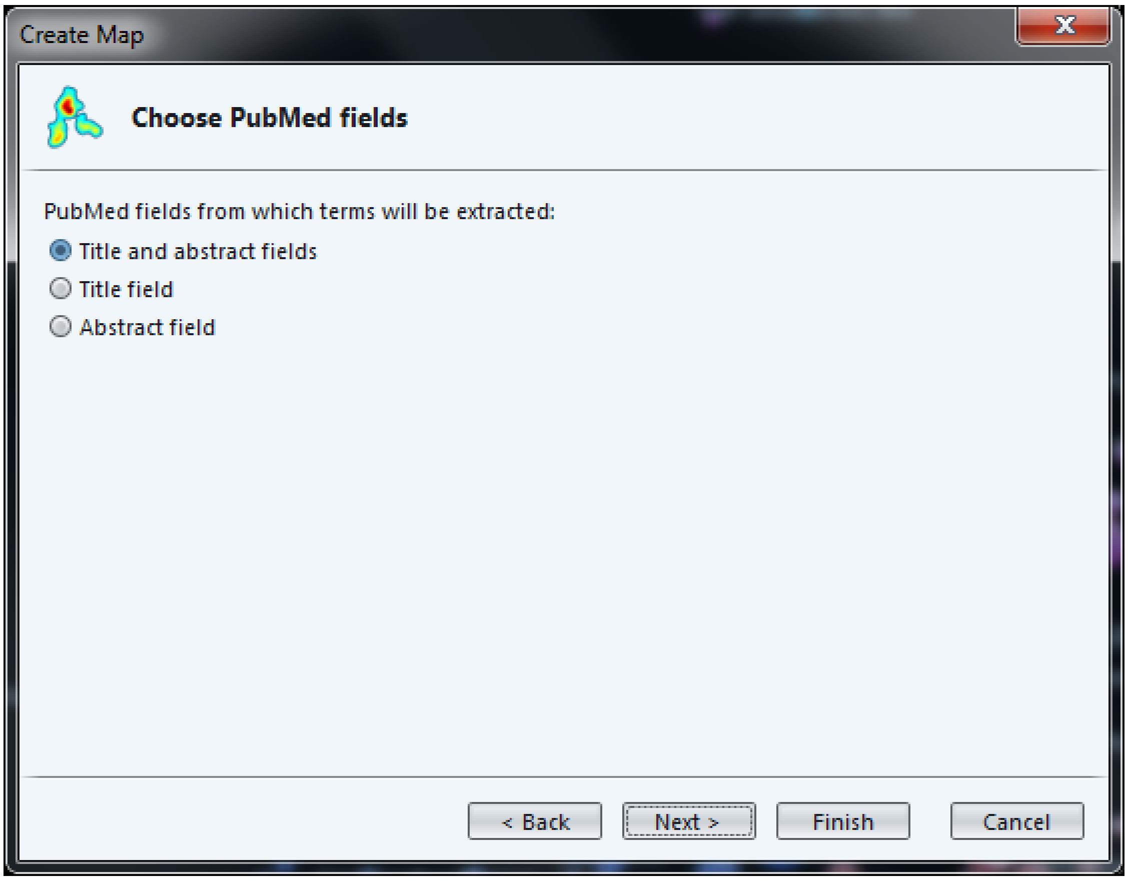
PubMed files: pubmed_medline.txt

[✓] Ignore copyright statements

Figure 13. Choosing the PuMed format

Loading a file in the PubMed / Medline format allows for a finer control:

- we are asked if we want to load abstracts, titles, or both
- it will remove copyright statements for us (such as "Copyright @Elsevier 2016") which often pollute the text.


Create Map

Choose PubMed fields

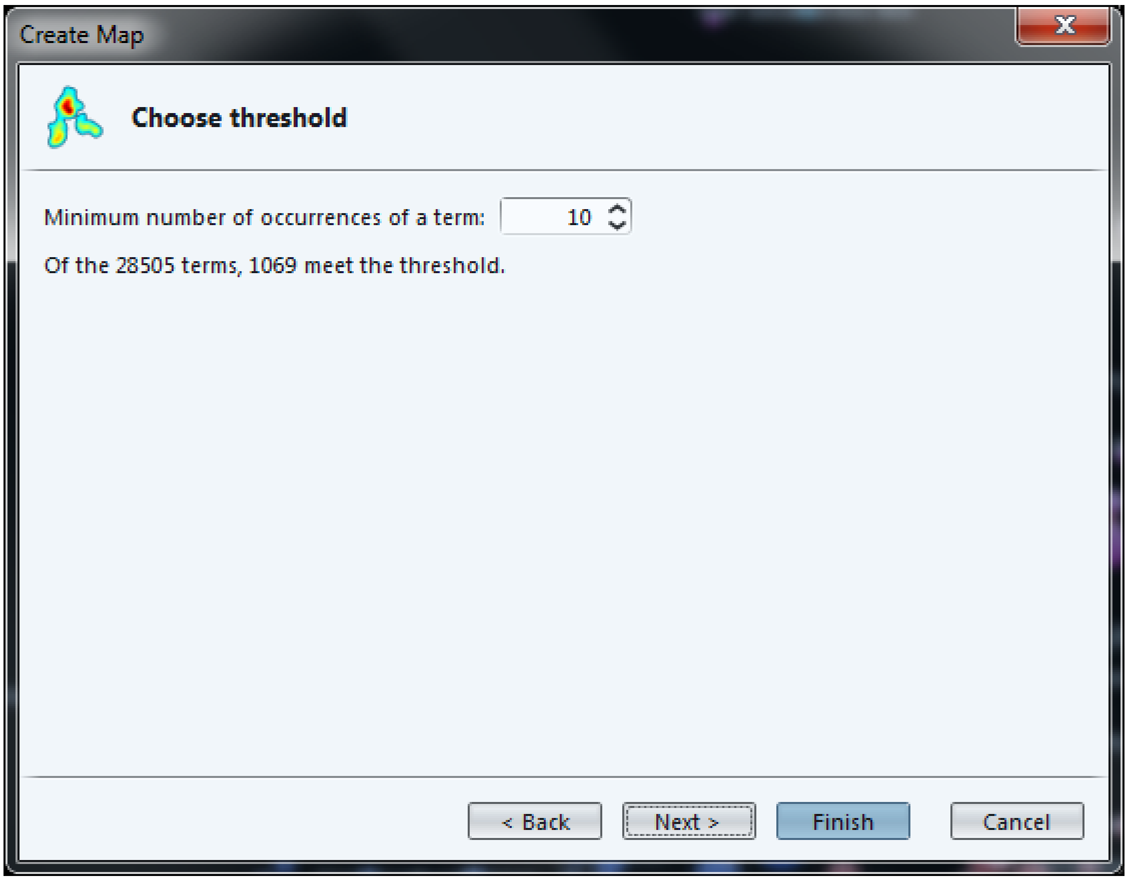
PubMed fields from which terms will be extracted:

[✓ Title and abstract fields]
Title field
Abstract field

Figure 14. Choosing the PuMed format

The next screen, as before, asks if we prefer binary or full counting, and if we have a thesaurus file.

We must then select the minimum frequency of a term (number of times it occurs in total) to be included in the map:


Create Map

Choose threshold

Minimum number of occurrences of a term: 10

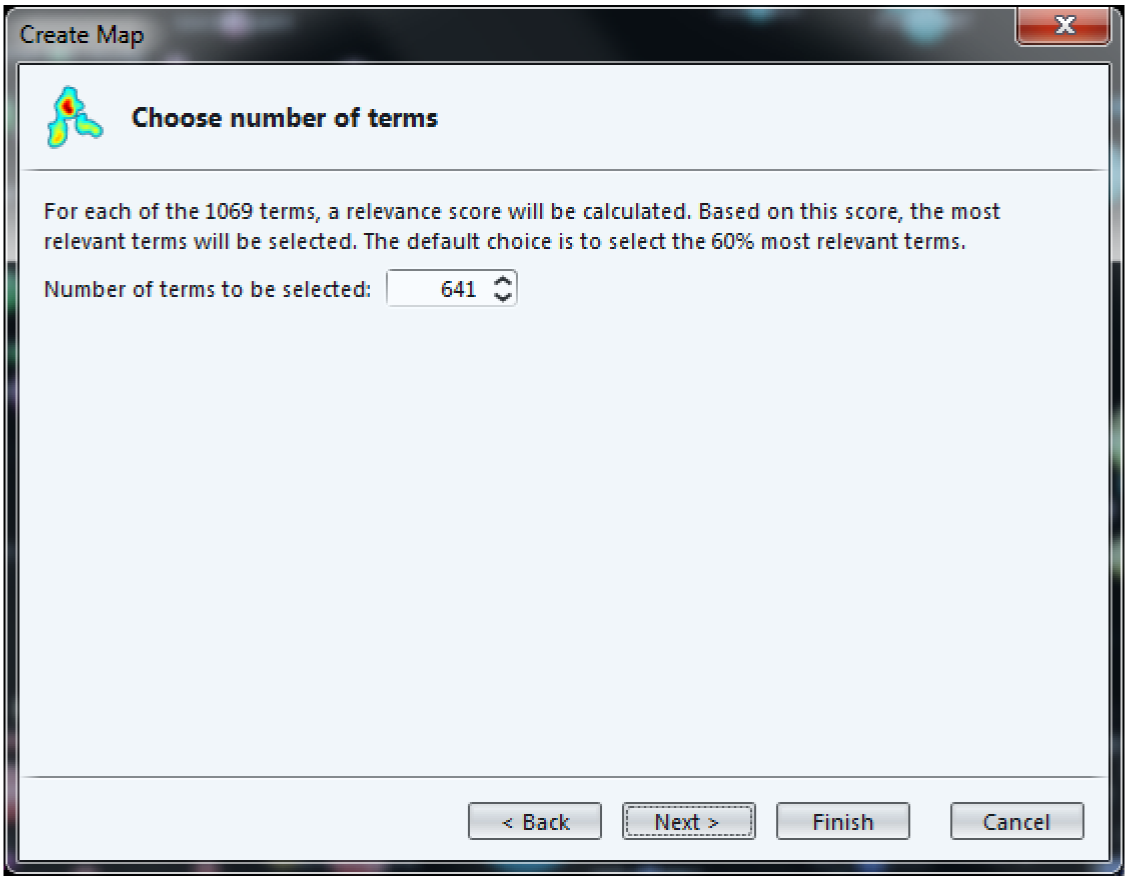
Of the 28505 terms, 1069 meet the threshold.

Figure 15. Minimum frequency for a term to be included

The next screen invites us to select the most "relevant" terms, with a parameter already set to a default value.

While the definition of "relevance" is difficult to explain briefly, its basic idea is that a term is all the more relevant that it tends to be associated with specific words, not with all words equally.

The developers of VosViewer explain their approach here.


Create Map

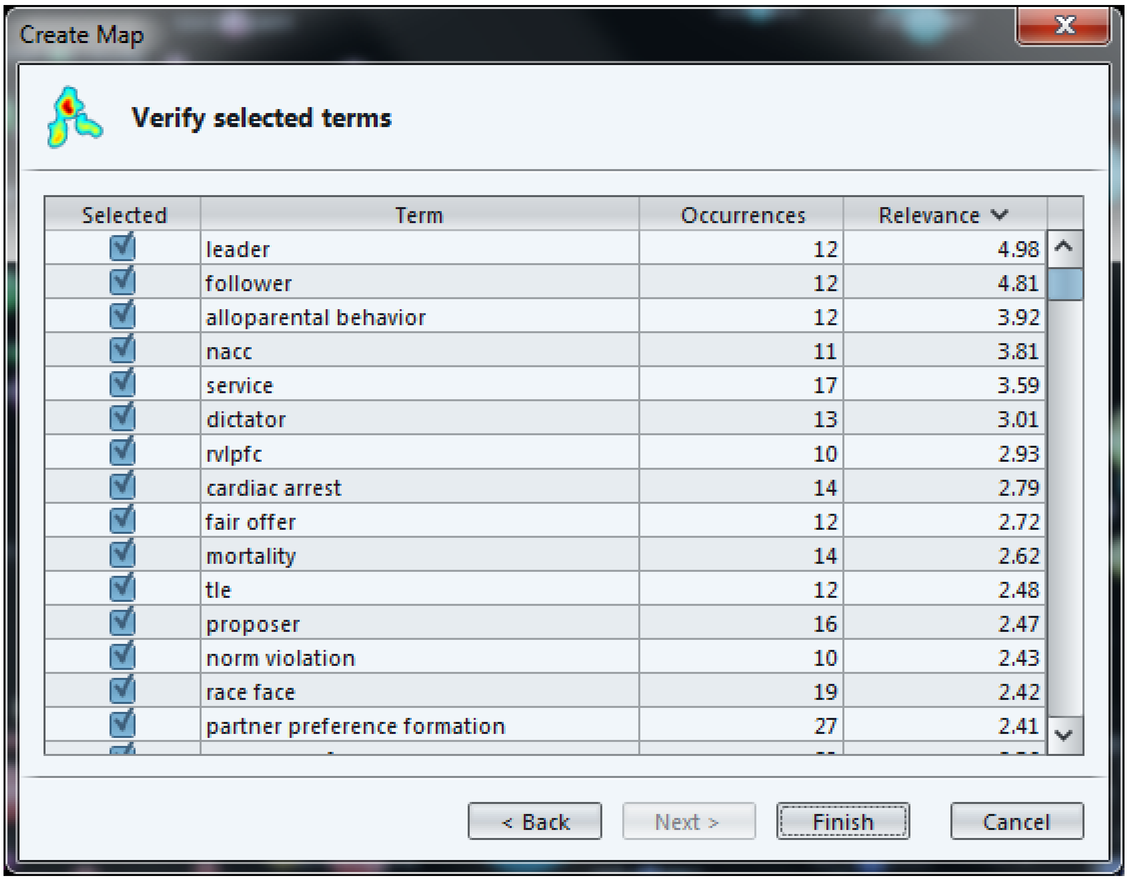
Choose number of terms

For each of the 1069 terms, a relevance score will be calculated. Based on this score, the most relevant terms will be selected. The default choice is to select the 60% most relevant terms.

Number of terms to be selected: 641

Figure 16. Selecting the most relevant words

The last screen of the wizard recaps in a table the terms which will be included in the map, ranked by relevance.



Selected	Term	Occurrences	Relevance
✓	leader	12	4.98
✓	follower	12	4.81
✓	alloparental behavior	12	3.92
✓	nacc	11	3.81
✓	service	17	3.59
✓	dictator	13	3.01
✓	rvlpfc	10	2.93
✓	cardiac arrest	14	2.79
✓	fair offer	12	2.72
✓	mortality	14	2.62
✓	tle	12	2.48
✓	proposer	16	2.47
✓	norm violation	10	2.43
✓	race face	19	2.42
✓	partner preference formation	27	2.41

Figure 17. Recap table

The result:

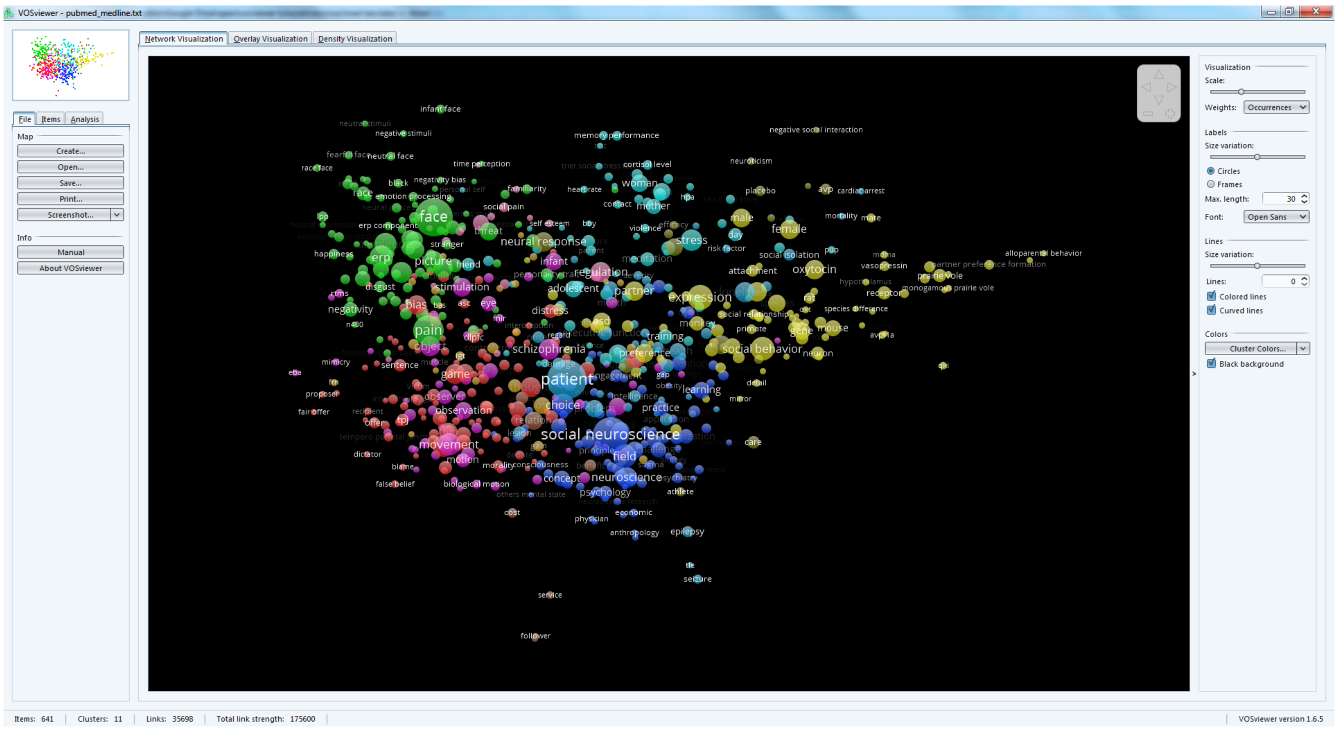
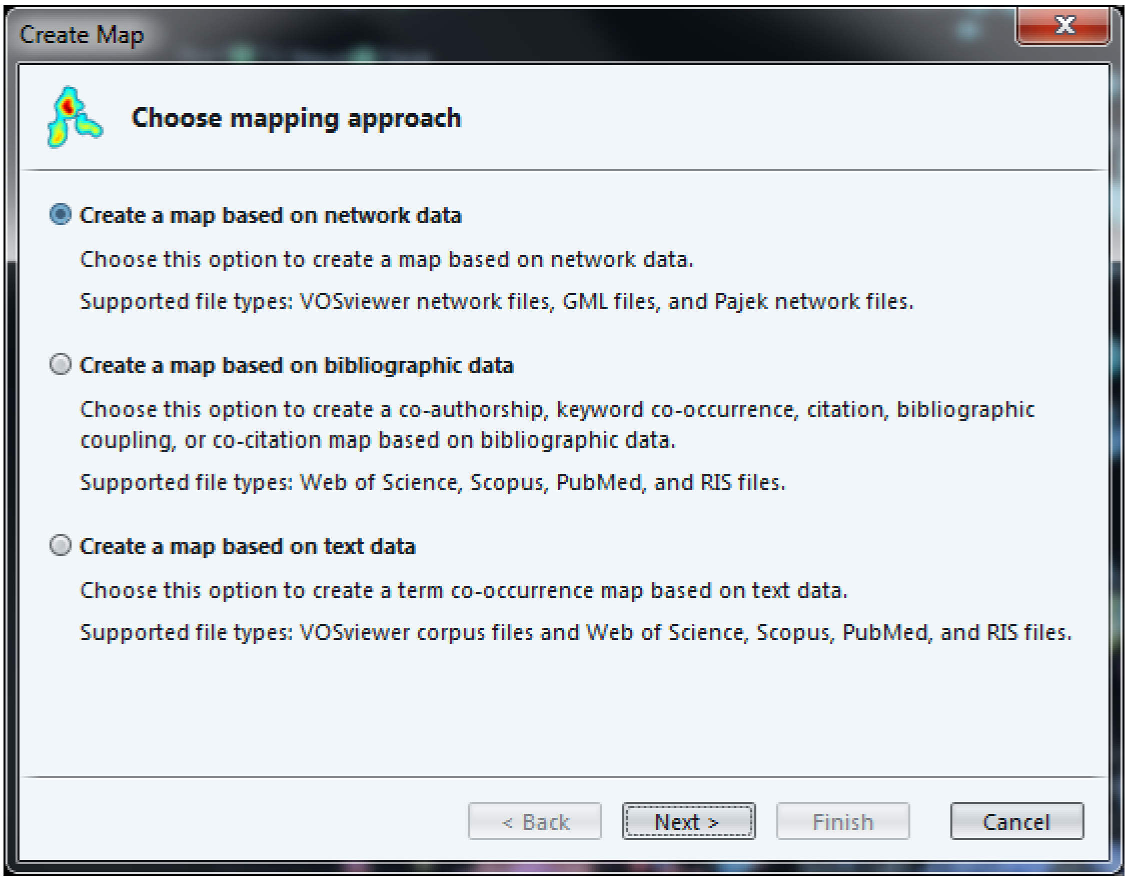


Figure 18. Result - the map of terms

3. Importing a network (gml format or pajek)

This part will not be developed in this tutoriaal as this is straightforward: you have a file in gml or pajek format, you import it:



◉ Create a map based on network data

Choose this option to create a map based on network data.

Supported file types: VOSviewer network files, GML files, and Pajek network files.

○ Create a map based on bibliographic data

Choose this option to create a co-authorship, keyword co-occurrence, citation, bibliographic coupling, or co-citation map based on bibliographic data.

Supported file types: Web of Science, Scopus, PubMed, and RIS files.

○ Create a map based on text data

Choose this option to create a term co-occurrence map based on text data.

Supported file types: VOSviewer corpus files and Web of Science, Scopus, PubMed, and RIS files.

Figure 19. Importing a network

It should however be noted that the import of gml files created with Gephi generates error - possibly a result of the GML format being poorly defined?

(to be continued)

More tutorials on vosviewer

- The manual of vosviewer is included as a pdf file in the zip when downloading the software.
- Vosviewer tutorials website, from which these slides are linked: https://seinecle.github.io/vosviewer-tutorials/

the end