

How to analyze frames using semantic maps of a collection of messages?

## **Pajek Manual**

Esther Vlieger & Loet Leydesdorff

Amsterdam School of Communications Research (ASCoR), University of Amsterdam,  
Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands.

---

## Content

<b>1. Introduction</b>	<b>3</b>
<b>2. How to generate the word/document occurrence matrix?</b>	<b>4</b>
2.1 Frequency List	4
2.2 Full Text	6
<b>3. How to analyze the word/document occurrence matrix?</b>	<b>8</b>
3.1 Variance	8
3.2 Factor Analysis	8
3.3 Cronbach's Alpha	11
<b>4. How to visualize the word/document occurrence matrix?</b>	<b>12</b>
4.1 Drawing the figure	12
4.2 Adjusting the figure to the factor analysis of SPSS	13
4.3 Changing the layout of the figure	14
<b>5. Discussion and further readings</b>	<b>17</b>

## 1. Introduction

In communication studies one is often interested in the content of messages. In addition to content analysis, one can use computer programs to generate semantic maps on the basis of large sets of messages. In this introduction, we explain how a semantic map can be generated from a set of messages. A properly normalized semantic map can be helpful in detecting frames as latent dimensions in sets of texts.

Messages can be contained in a set of documents, a sample of sentences, or any other textual units of analysis. In our design, the textual units of analysis will be considered as the cases, the words contained in these messages as the variables. Thus, we operate with matrices. Matrices which contain words as the variables in the columns and textual units of analysis as cases in the rows (following the convention of SPSS) are called word/document occurrence matrices.

When visualizing a word/document occurrence matrix, a network appears, containing the interrelationships among the words and the textual units. In order to generate this network, one needs to go through various stages using different programs. In this manual we explain how to generate, analyze, and visualize semantic maps from a collection of messages using the programs available at <http://www.leydesdorff.net/indicators> and standard software like SPSS and Pajek.

## 2. How to generate the word/document occurrence matrix?

In this section we explain how to generate a word/document occurrence matrix. In order to generate the word/document occurrences matrix, one first saves the acquired set of messages in such a format that the various programs (at <http://www.leydesdorff.net/indicators>) to be used below can use them as input files. If the messages are short (less than 1000 characters), we can save them as separate lines in a single file using a text editor. (Most secure is saving in WordPad as “TextDocument - MS-DOS Format.”)<sup>1</sup> This file has to be called “text.txt”. In that case we will use the program ti.exe that analyzes title-like phrases. If the messages are longer, the messages need to be saved as separate notepad files, named text1.txt, text2.txt, etc.<sup>2</sup> These files will be read by FullText.exe.

### 2.1 Frequency List

The file text.txt can directly serve as input for the program [FrqList.Exe](#) (shorthand for “frequency list”). This program produces a word frequency list from the file text.txt, needed for assessing which words the analyst wishes to include in the word/document occurrences matrix. As a rule of thumb, more than 75 words are difficult to visualize on a single map, and more than 255 variables are difficult to analyze because of systems limitations in SPSS v. 15 and Excel 2003.

Together with the notepad file, one can use a standard list of stopwords in order to remove the irrelevant words directly from the frequency list. It can be useful to check the frequency list manually, to remove potentially remaining stopwords. If we begin with long texts in different files (text1.txt, text2.txt, ... etc.),<sup>3</sup> these files have first to be combined into a single file

---

<sup>1</sup> In WordPad, one should save as “TextDocument – MS-DOS Format”. In NotePad, use the default (ANSI) for saving. If one uses Word, one should be careful to save the file as a so-called DOS plain text file. When prompted by Word, choose the option to add CR/LF to each line. (CR/LF is an old indication of Carriage returns and Line feeds, like using a typewriter.)

<sup>2</sup> Sometimes, Windows adds the extension .txt automatically. One should take care not to save the files with twice the extension “.txt.txt”. The programs assume only a single “.txt” and will otherwise lead to an error.

<sup>3</sup> Sample files text1.txt, text2.txt, text3.txt, text4.txt can be found at <http://www.leydesdorff.net/software/fulltext/text1.txt>, etc.

text.txt that can be read by FrqList, for the purpose of obtaining a cumulative word frequency list across these files.<sup>4</sup> The use of FrqList is otherwise strictly analogous.

To be able to run FrqList, one needs to install the program in a single folder with the notepad file with all the messages (text.txt) and the list of stopwords (e.g., [stopword.txt](#) for English texts), as can be seen in figure 1.<sup>5</sup>

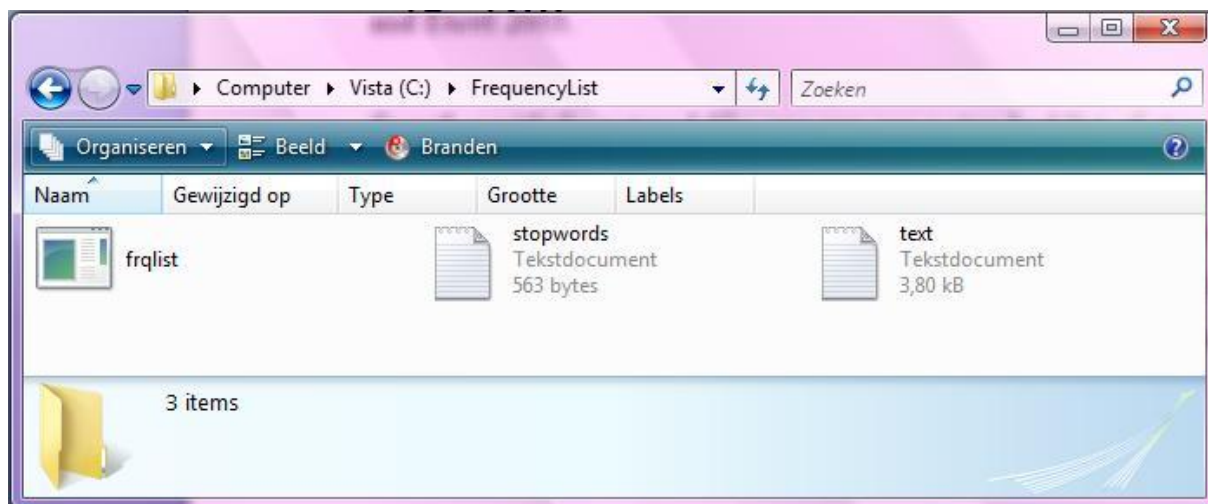


Figure 1 Example FrqList

After running, the program FrqList produces a frequency list: the combined word frequency list is made available as WRDFRQ.txt in the same folder, as can be seen in figure 2. This file can be read into Excel in a next step so that, for example, the single occurrences of words can be discarded from further analysis.

<sup>4</sup> One can combine these files in an editor (e.g., WordPad or NotePad) or alternatively by opening a DOS box. In the DOS box, use “cd” for changing to the folder which contains the files and type: “copy text\*.txt text.txt”. Make sure to erase an older version of text.txt first.

<sup>5</sup> A corresponding list of stopwords for Dutch texts can be found at [http://www.leydesdorff.net/software/ti\\_ne/stopword.txt](http://www.leydesdorff.net/software/ti_ne/stopword.txt).

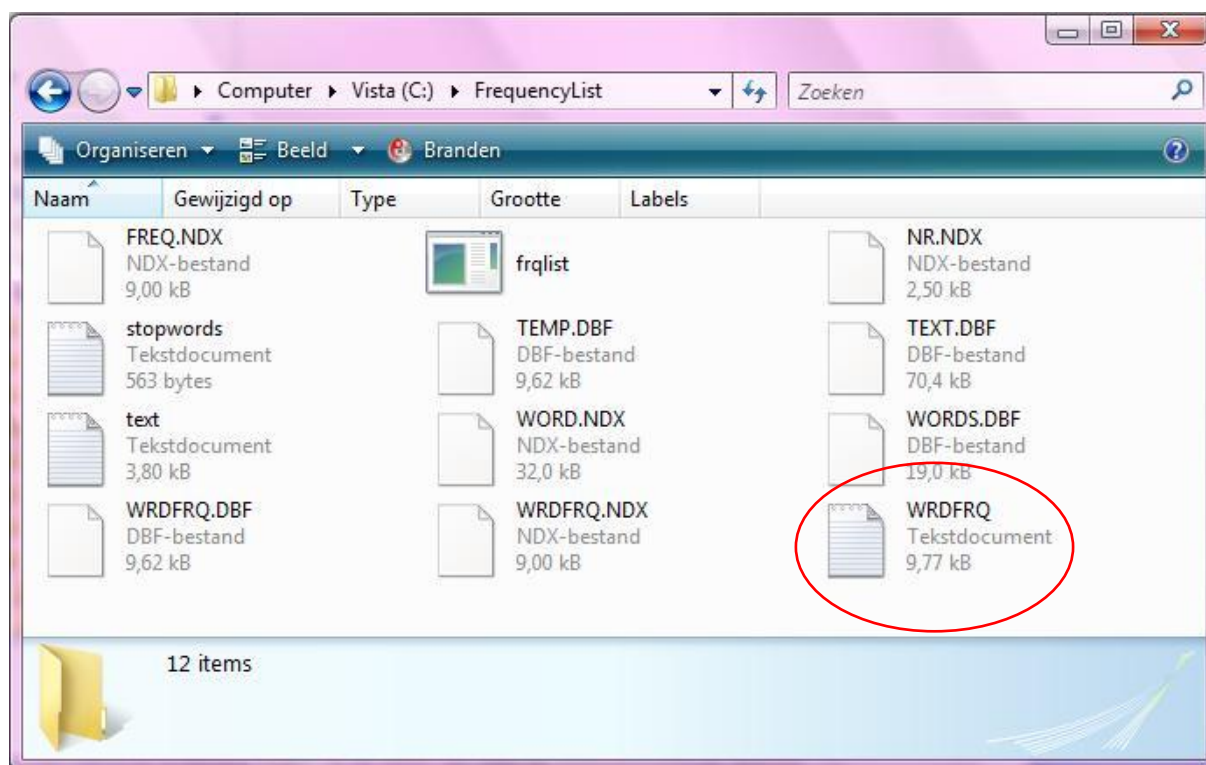


Figure 2 Output FrqList

## 2.2 Full Text

The next step is to import the frequency list into an Excel file in order to separate the words from the frequencies as numerical values. At this stage, the list of words may be too long to use efficiently. To be able to visually interpret the data at a later stage, it can be advised to use a maximum of approximately 75 words. The first 75 words from the frequency list (without the frequencies) need to be saved as a notepad file by the name of words.txt. (Use WordPad or NotePad for saving or obey the conventions for a plain DOS text as above.) This file “words.txt” can serve as input for the programs [Ti.exe](#) or [FullText.exe](#).<sup>6</sup>

One can use ti.exe for the case that the texts are short (< 1000 characters) and organized as separate lines in a single file text.txt, but fulltext.exe is used in the case of a series of longer text files named text1.txt, text2.txt, text3.txt, ..., etc. Both programs need in addition to the information in the textual units, an input file named words.txt (in the same folder) with the information about the words to be included as variables. Prepare this file carefully using the

<sup>6</sup> Dutch versions of these programs are available at [http://www.leydesdorff.net/software/ti\\_ne/ti\\_ne.exe](http://www.leydesdorff.net/software/ti_ne/ti_ne.exe) and [http://www.leydesdorff.net/software/ft\\_ne/ft\\_ne.exe](http://www.leydesdorff.net/software/ft_ne/ft_ne.exe). While the English versions correct for the plural “s”, the Dutch versions also consider “en” as a plural.

instructions about removing stopwords and making selections specified above. You may wish to run FrqList.exe a second time with a manually revised file stopword.txt. (Save this file as a DOS file!)

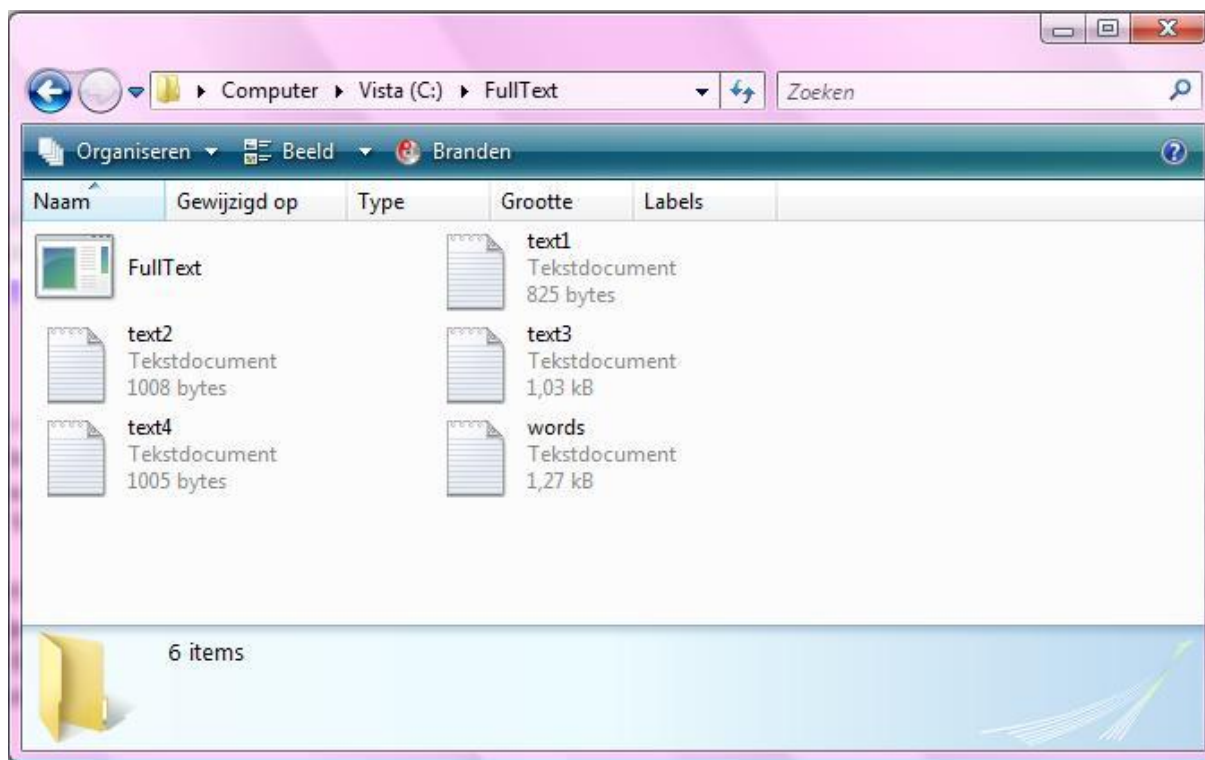


Figure 3 Example FullText

As can be seen in figure 3, the separately saved messages (text1.txt, text2.txt, etc.), together with the file words.txt, form the input for FullText. (Analogously, for Ti.exe one needs the files text.txt and words.txt.) The program produces data files, which can be used as input for the statistical program SPSS and the network visualization program Pajek. By installing the program FullText in the same folder containing the saved messages and words.txt, the program can be run. The output of FullText can also be found in this same folder, as can be seen in figure 4.

Prior to running FullText, the program demands to insert the file name ('words') and the number of texts. After running FullText (or Ti.exe), one can use the files matrix.dbf and labels.sps to statistically analyze the word/document occurrence matrix by using SPSS (The file matrix.dbf contains the data and can be read by SPSS. The file labels.sps is an SPSS

syntax file for labelling the variables with names.). In order to generate a visualization of the semantic map, one can use the file cosine.dat as input to Pajek. How to use these files for Pajek and SPSS will be discussed in chapter 4.

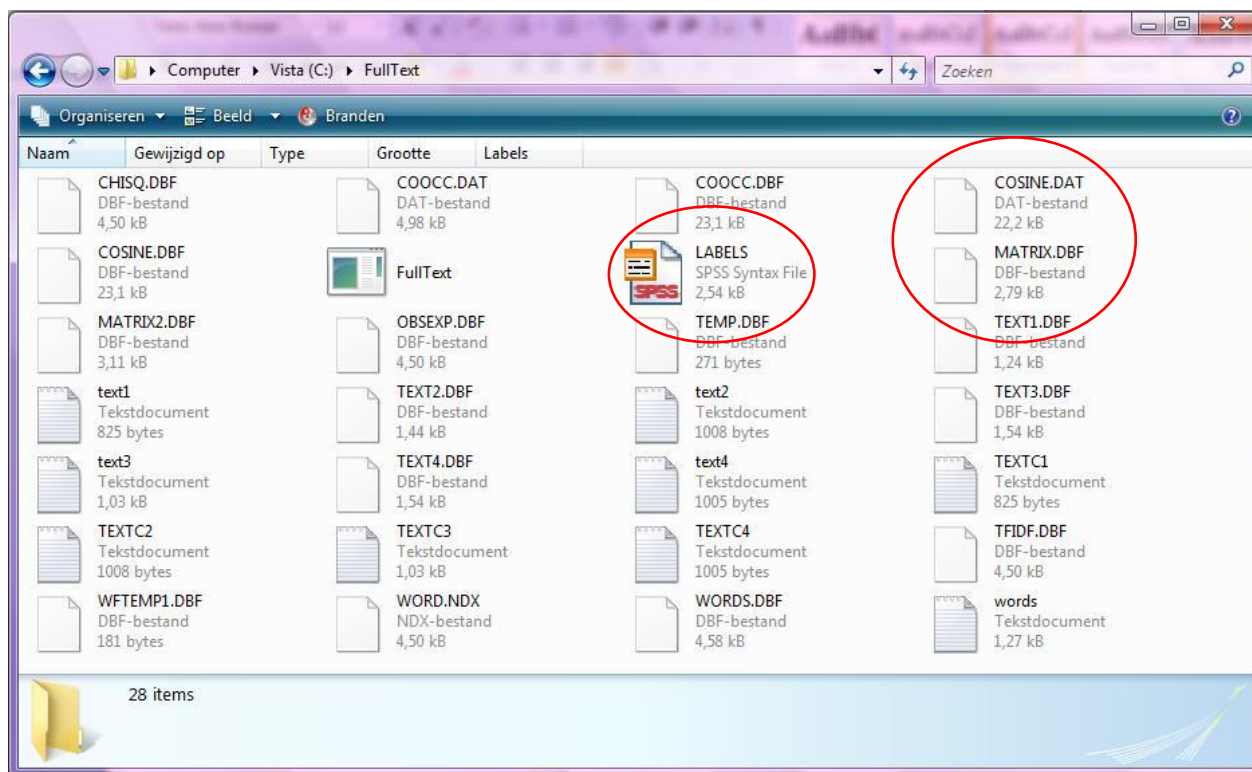


Figure 4 Output FullText



### 3. How to analyze the word/document occurrence matrix?

In order to analyze the data from the matrix, one can use the statistical program SPSS. As discussed in chapter 2, the file matrix.dbf can be read by SPSS ('File – Open – Data – Matrix.dbf'). To label the variables with names, choose 'File – Open – Syntax' in order to read the file labels.sps. Choose 'Run – All'. As can be seen in the syntax file, FullText has deleted the 's' at the end of the words. The aim is to remove the plural forms, but this may have no use when analyzing a word/document occurrence matrix. By comparing to the original words in the file WRDFRQ.txt (which was generated by FrqList) the labels in the variable view of SPSS can be manually adapted. This is only necessary if one wants to use the words as labels; for example, in a table of the SPSS output. When visualizing the word/document occurrence matrix, as we will explain in this manual, the words can be adapted for use in Pajek at a later stage.

#### 3.1 Variance

In order to analyze the word/document occurrence matrix in terms of its latent structure, one may wish to conduct a factor analysis in SPSS. The factor analysis will demonstrate which words belong to which components. Prior to the factor analysis one has to calculate the variance of the variables (the words from the matrix). Words with a variance of zero cannot be used by SPSS in a factor analysis and therefore need to be left out of the process. (The variance is calculated by choosing 'Analyze – Descriptive Statistics – Descriptives', then selecting all the words into the right column and then ticking 'Variance' under 'Options'.)

#### 3.2 Factor analysis

The next step is analyzing the data by means of a factor analysis. Choose 'Analyze – Data Reduction – Factor' in SPSS. This step is visualized in figure 5. Select all the variables in the left column, except the ones with a variance of zero, and select them to the right column. Then, under 'Extraction', tick 'Scree plot' and undo 'Unrotated factor solution'. Then, under 'Rotation', tick 'Varimax' and 'Loading plot(s)' and finally, under 'Options', tick 'Sorted by size' and 'Suppress absolute values lower than', which is set on .10.

Under Extraction it is additionally possible to manually choose the number of factors. When the output of the factor analysis produces too many factors, it may be advised to manually set the number of factors on, for example, six. More than six factors will be difficult to visualize and interpret through Pajek at a later stage.

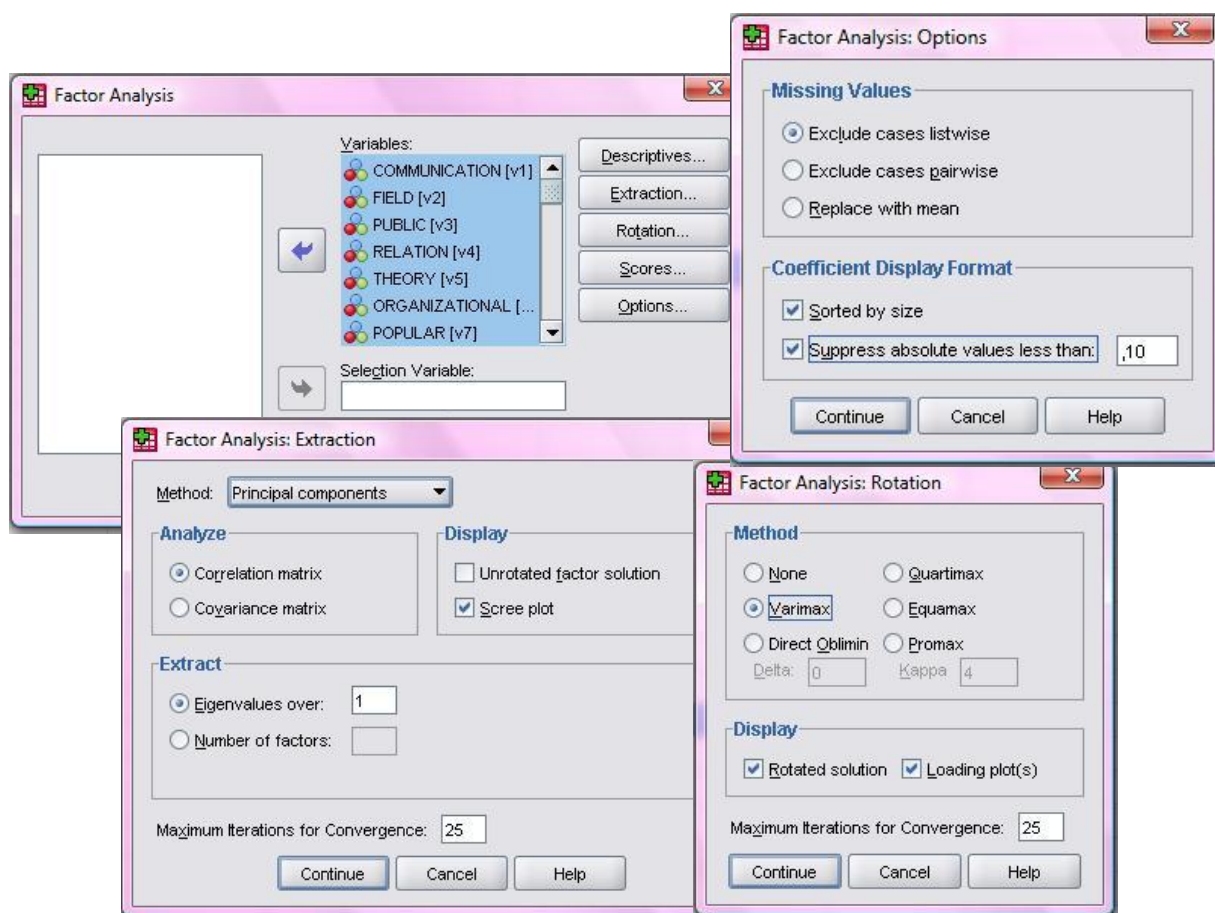


Figure 5 Factor Analysis in SPSS

The options are now set in the right manner to conduct a factor analysis. SPSS produces several tables and figures in the output. The most relevant for our purpose is the Rotated Component Matrix. This matrix shows the number of components (factors) and the loading of the different words on the components. At this stage, one can arrange the words under the different components, which can be used when visualizing the word/document occurrence matrix in the next stage. In figure 6 an example of a few words are visualized with the arrangement under the different components. The different components can be considered as the different frames used in these texts. In the example in figure 6, the texts are built around

three different frames. How this output can be used to visualize the word/document occurrence matrix will be discussed in chapter 4.

**Rotated Component Matrix<sup>a</sup>**

	Component		
	1	2	3
RESEARCH	<b>,875</b>	,436	-,209
FIELD	<b>,780</b>	,256	,572
WITHIN	<b>,568</b>	-,674	-,472
WELL	<b>,568</b>	-,674	-,472
LEVEL	<b>,332</b>		-,940
COMMUNICATION	-,147	<b>,968</b>	,202
APPLIED	,533	<b>,843</b>	
AREA	,483	<b>,841</b>	,243
PUBLIC	,345	<b>,766</b>	,542
THEORETIC	,345	<b>,766</b>	,542
RELATION	,345	<b>,766</b>	,542
TOOLS	,345	<b>,766</b>	,542
CONCEPTUAL	,345	<b>,766</b>	,542
DEVELOPED	,585	<b>,734</b>	-,344
CONCLUDES	-,332		<b>,940</b>
YEARS	,579		<b>,811</b>
ACROSS	,579		<b>,811</b>
APPLY	,579		<b>,811</b>
BASED	,579		<b>,811</b>
TRENDS	,579		<b>,811</b>
VISUAL	,324	-,860	<b>,395</b>
STUDIES	-,343	-,852	<b>,395</b>

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Figure 6 Output Factor Analysis in SPSS (an example with a limited number of words/variables)

In addition to the positive factor loading, one may also wish to take into account that “level” has a negative loading (-.94) on Factor 3.

### ***Cronbach’s Alpha***

Prior to the visualization of the matrix, one may wish to conduct a reliability analysis, by calculating Cronbach’s Alpha ( $\alpha$ ) for each frame (component). This measure controls after the factor analysis whether the frames form a reliable scale. First, one has to determine which words belong to which frames by using the output of the factor analysis in SPSS, like the example in figure 6.

The next step is the calculation of Cronbach's Alpha in SPSS, by choosing 'Analyze – Scale – Reliability Analysis'. Select the words from the first frame (loading on factor 1) into the right column and run the reliability analysis by choosing 'OK'. Figure 7 shows the output of this analysis with Cronbach's Alpha for the example from figure 6, using the second frame (factor) which was composed of nine items (that is, words as variables).

Reliability Statistics	
Cronbach's Alpha	N of Items
.949	9

Figure 7 Output reliability analysis (Cronbach's Alpha) in SPSS

In the example in figure 7, Cronbach's Alpha has a value of .95. In order to guarantee the internal consistency of the scale, Cronbach's Alpha needs to have a minimal value of .65.

This reliability test is to be run for each factor separately. You may wish to adjust or discard the factor so that the reliability is enhanced.

#### 4. How to visualize the word/document occurrence matrix?

In this section we explain how to visualize the word/document occurrence matrix by using Pajek and the output of the factor analysis in SPSS. Pajek can be downloaded at <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>; in this manual we use Pajek 2.05.

In order to visualize the output of FullText, one is advised to use the file cosine.dat, which was generated by FullText (see chapter 2).<sup>7</sup> In the first part of this section the drawing of the figure is discussed. After that we will explain how the figure can be informed by the output of the factor analysis in SPSS. The final part of this section discusses the layout of the figure and how this can be changed.

<sup>7</sup> The cosine-normalized matrix can be compared to the Pearson correlation matrix which is used for the factor analysis, but without the normalization to the mean. Word-frequency distributions are usually not normally distributed and therefore this normalization to the mean is not considered useful for the visualization. The results of the factor analysis inform us about the latent dimensions which are made visible by the visualization as good as possible. Note that a visualization is not an analytical technique.

### *Drawing the figure*

Choose 'File – Network – Read' to open the file cosine.dat in Pajek. To create a partitioned figure, one can choose 'Net – Partitions – Core – All'. To draw the figure, choose 'Draw – Draw partition'. One can change the layout of the figure by choosing 'Layout – Energy – Kamada-Kawai – Free'. In this stage, one has created a figure which shows the different components with different colors, as can be seen in figure 8. However, the algorithm used in Pajek for attributing the colors is different from the results of the factor analysis. We will change this below.

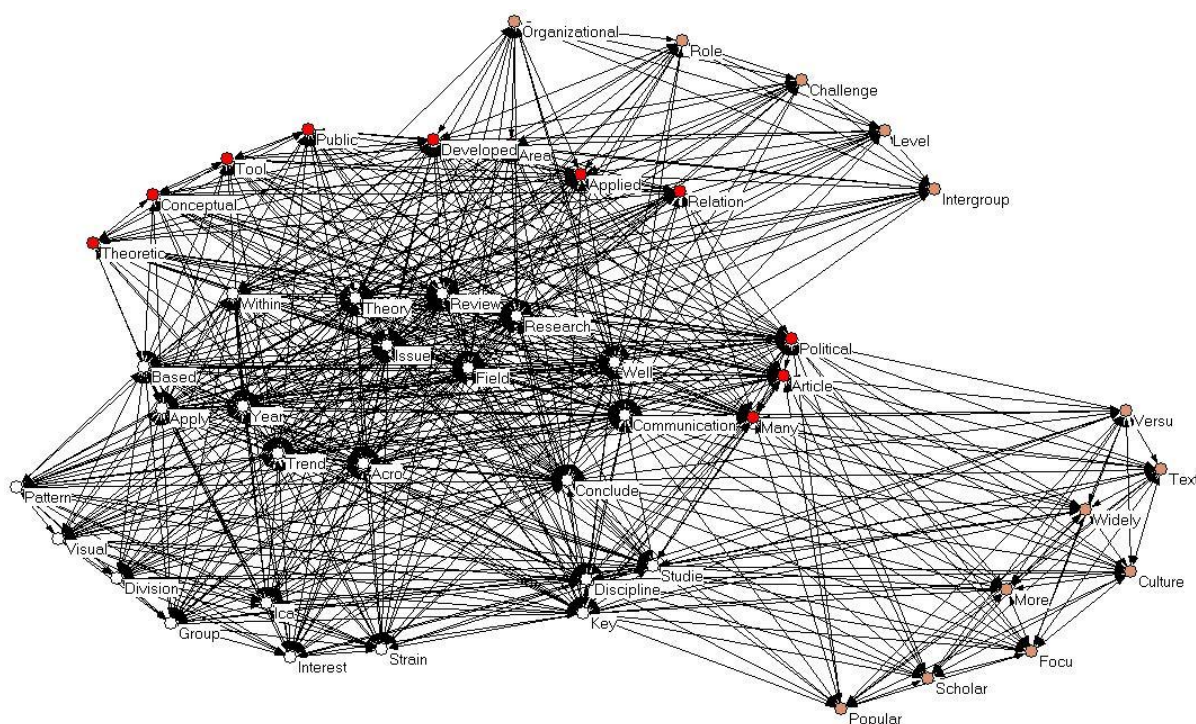


Figure 8 Standard Pajek figure with different components

As discussed in chapter 3 and shown in figure 8, FullText automatically removes an 's' at the end of a word. Also in Pajek it is possible to put back the 's', in case of an incorrect removal. To do so, close the figure and choose 'File – Partition – Edit' in Pajek. In this window one can change the words manually. After closing the window and drawing the partition figure again, the words are changed.<sup>8</sup>

<sup>8</sup> Alternatively, one can change the words in the input file cosine.dat using an ASCII editor such as Notepad.



### *Adjusting the figure to the factor analysis of SPSS*

The next step in visualizing the word/document occurrence matrix is the adjustment of the figure to the output of the factor analysis in SPSS, as discussed in the previous chapter. After the factor analysis in SPSS, each word was assigned to a specific frame. In the example, there were three different frames made visible in the output (figure 6). In spite of the fact that figure 8 also shows three frames in Pajek, there are differences between these frames and the frames from SPSS. These differences are being caused by the fact that Pajek uses the cosine matrix while SPSS uses the correlation matrix and performs an orthogonal rotation.

The figure as shown in figure 8 can be adjusted to the output of the factor analysis in SPSS. This adjustment can be done in the same way as the changing of the words in the former section. By choosing ‘File – Partition – Edit’, the frames can be reclassified by assigning the same numbers to words in the same frame.<sup>9</sup> For example, in Figure 6 the word “public” had its highest factor loading on Factor Two. Thus, we adjust the number of the partition into “2”.

After adjusting thus the figure from figure 8 to the factor analysis in SPSS, a new figure can be drawn, which is shown in figure 9.

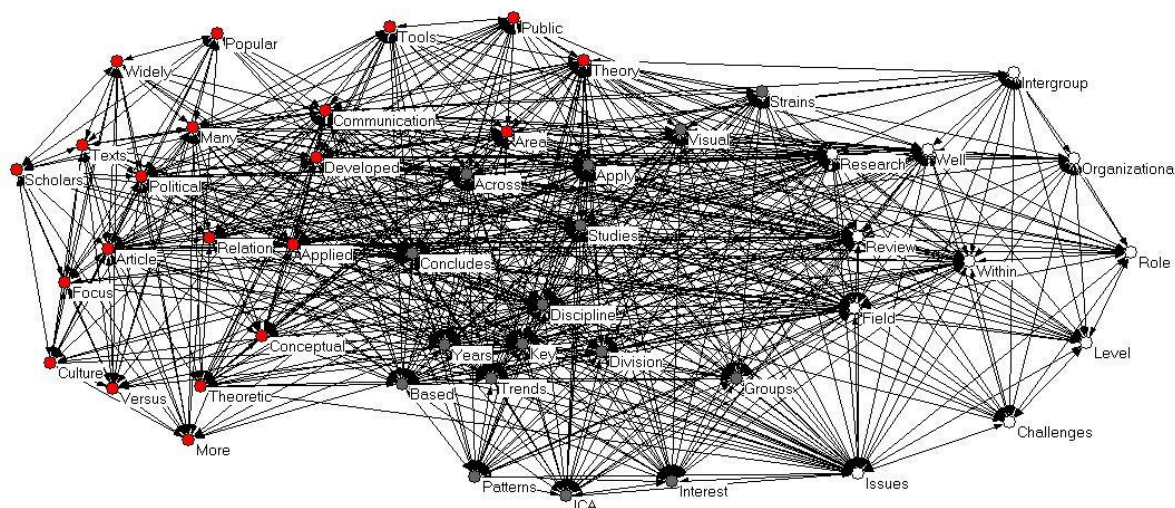


Figure 9 Pajek figure adjusted to factor analysis in SPSS

The initial numbers, corresponding to the different frames, are provided by Pajek using another algorithm than factor analysis (the *k*-core algorithm). Nevertheless, the numbers are

<sup>9</sup> In the Draw screen, Shift-Left click a vertex to increase its partition cluster number by one, Alt-Left click a vertex to decrease it by one.

always one-to-one related to the colours of the vertices in the figure. As can be seen in figures 8 and 9, it is difficult to read the words in the current layout of the figure. The different lines are also difficult to distinguish. In the next section the possibility of changing the layout of the figure in order to clarify it will be discussed.

### ***Changing the layout of the figure***

In this final section, the possibilities of changing the layout of the figure are being discussed. There are several options which can increase the readability of the figure. A few of these options are being introduced here. After following these steps, the figure will be better readable and interpretable.

**Background** The figure can be read most clearly against a white background. To change the background, choose in the Drawing Pane: ‘Options – Colors – Background’ and choose white as the background color.

**Lines** To make sure the different lines can be distinguished, it is possible to remove the lines with a value lower than for instance 0.2 (this depends on the figure, different values can be tried). To do so, close the figure, than choose ‘Net – Transform – Remove – Lines with value – lower than’ and fill in the appropriate value.

It is also possible to adjust the width of the lines to their values. In order to do so, draw the partition figure, then choose the option ‘Options – Lines – Different Widths’. Since the cosine varies between zero and one, a value of 3 or 5 will provide differences.

**Arrows** The arrow heads are not adding anything to the figure, so they can be removed. To do so, choose ‘Options – Size – of Arrows’ and fill out 0.

**Font** The size of the font can be changed through ‘Options – Size – of Font – Select’. Use at least 12 for a PowerPoint presentation in order to make sure the words can be read. To make sure the words do not overlap each other, it is possible to drag the words a little into different directions.

**Vertices** The sizes of the vertices can be made proportional to the (logarithm of) the frequency of the words. In order to do so, choose ‘Options – Size – of Vertices defined in input file’.

To enlarge all the vertices, choose 'Options – Size – of Vertices' and fill in the size. In figure 10, the vertices have been given the size of 10.

### Colors

To change the colors of the vertices, choose 'Options – Colors – Partition Colors – for Vertices'. One can change the colors of the vertices, by clicking on the current color and then filling in the number of the wished color as seen on the color pallet. After that, click on OK and close the color pallet. Then click on one of the vertices you want to change and the entire frame will have the wished color. This can be done for each group of vertices. Make sure the colors are in different shades, in order to visually see the differences between the different frames.

Figure 10 shows the same figure as in figures 8 and 9 after passing through the preceding steps. The words can be read better and the differences in the loadings of the words can be interpreted.

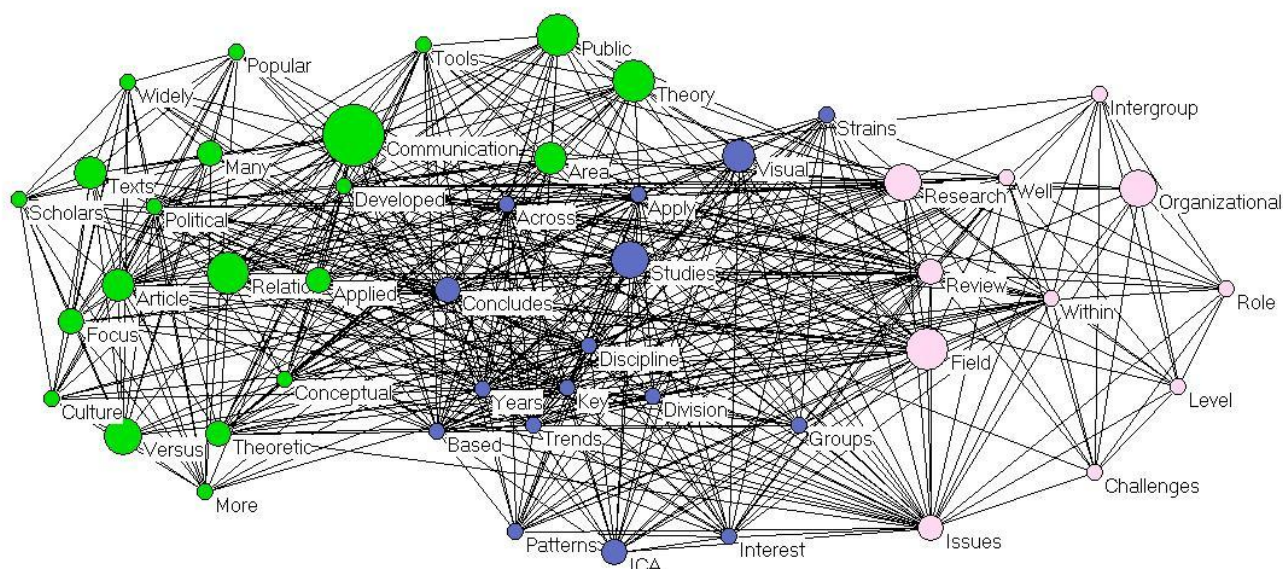


Figure 10 Pajek figure after changing the layout

A final option to complete the above figure is the addition of the frames to the figure (using Word or a program like Paint). Through the different words it is possible to name the different frames. Figure 11 shows an example of this addition to the above figure.



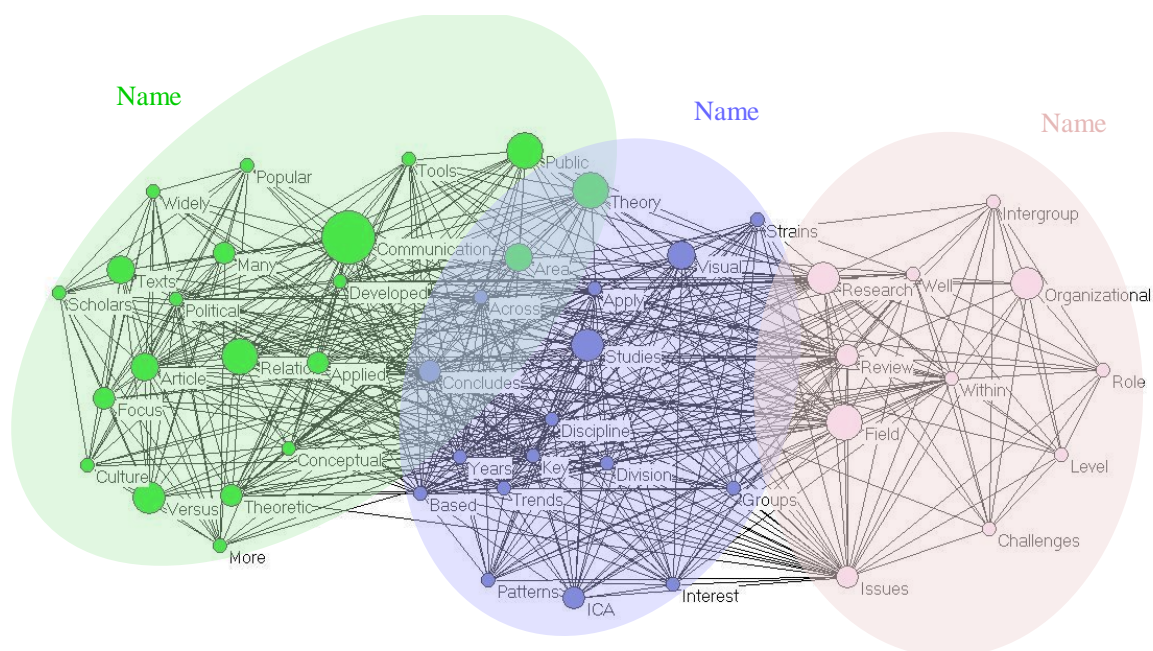


Figure 11 Pajek figure after taking out the different frames

## 5. Discussion and further readings

Our discussion above has been very much oriented towards “getting started” with Pajek for the visualization of latent frames in textual messages. The resulting output can be further embellished for presentation purposes using lesson 6 at <http://www.leydesdorff.net/indicators/>. Other lessons at this same page use the same techniques for other purposes. For example, one can be interested in the cited references in texts and thus wish to make a citation matrix instead of a matrix of co-occurring words. The basic scheme is that of textual units of analysis (messages) to which a set of variables can be attributed. These variables can be words, author names, institutional addresses, cited references, etc. One can then generate the file matrix.dbf and cosine.dat as described above, and use them for analysis in SPSS and/or visualization in Pajek.

In addition to the available statistics in SPSS, Pajek hosts a number of statistics which have been developed over the past few decades in social network analysis. We already mentioned above the  $k$ -core algorithm which groups together nodes (vertices) which are interrelated with at least  $k$  neighbours. An introduction to these statistics is provided by: Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside; at <http://faculty.ucr.edu/~hanneman/nettext/>. An introduction to Pajek is

provided by: De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press.

We appreciate if users of this text feedback suggestions for improvements to us (at [loet@leydesdorff.net](mailto:loet@leydesdorff.net) ).