

Chapter 1. Content analysis using AmCAT

Introduction

Content analysis is a method of analysis used for measuring relevant concepts in a text in a systematic way. Automatic content analysis refers to content analysis with the help of a computer. The greatest advantage of automatic content analysis is the fact that computers can count fast and effectively. As such, large amounts of text can be analysed quickly. The disadvantage of automatic content analysis, however, is the fact that computers can recognize and count words, but fail to actually read and interpret them. Computers can recognize words alone, not concepts. Therefore, it is the researchers' task to formulate search terms with which they instruct the computer to recognize concepts on the basis of combinations of words. Automatic content analysis can be executed in AmCAT using the *Index Search* function. The current chapter examines the possibilities of this function.

A browser enables login at AmCAT, for example firefox or internet explorer or Chrome. Go to: <http://eur.amcat.vu.nl>. The Login to AmCAT. Your studentnumber is both inlog and password.

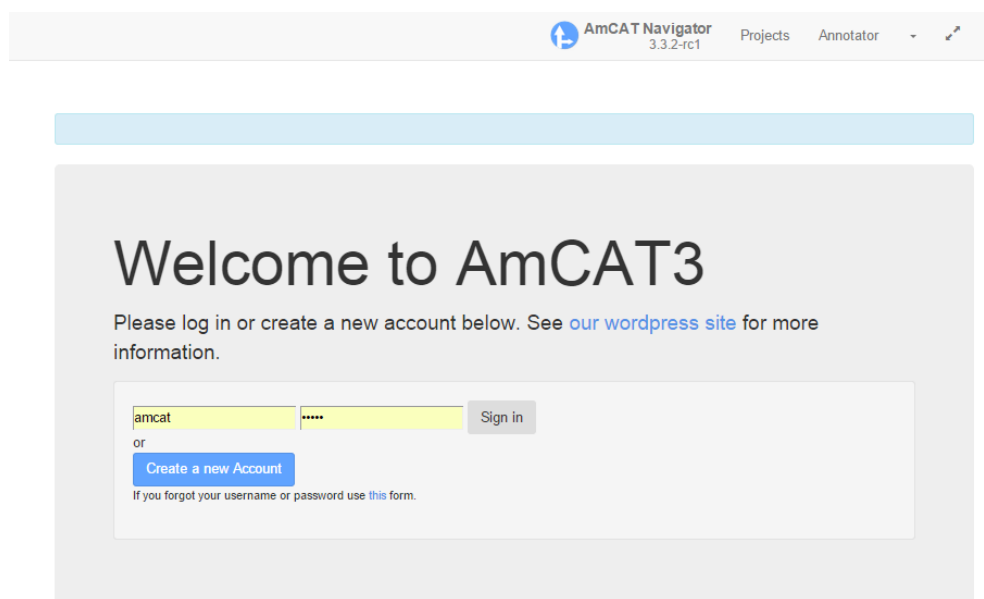


Figure 1. opening screen AmCAT

Here you can choose your project. There are projects that you own and projects of other users (where you have a guest role as reader). All Projects shows you all projects at eur.amcat.nl.

You can click on the project you want to use for your analysis, and you will see the details of that project. Here we use a project with wiki news (we use open source to avoid copy right issues) but this is the same for all projects in which you work with newspaper articles from Lexis Nexis for example.

Project view

Underneath you see an overview of a project. You can see there is one articleset included with 1406 articles.

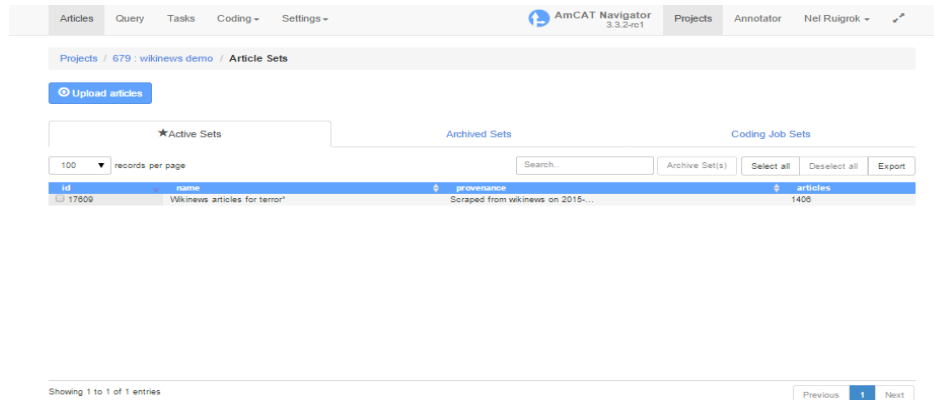


Figure 2. Article sets in a project

Add articles to your project

You can add articles to your own project through this Article section. You see the button Upload Articles. If you click on that you get the following screen:

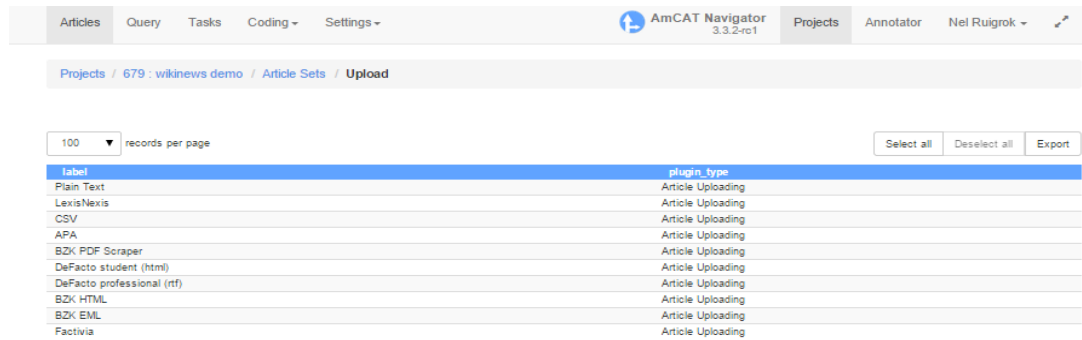


Figure 3. Upload options for articles

Here you can click on several option for different formats of texts that can be uploaded. We work with newspaper articles that are most likely from LexisNexis. Therefore click on the LexisNexis row to get the screen in which you can upload articles that you downloaded from LexisNexis.

You can upload articles from Lexis Nexis also as a Zip-file. This is handy in case of a lot of articles that you need to download (you can only download 200 articles per saving from LN).

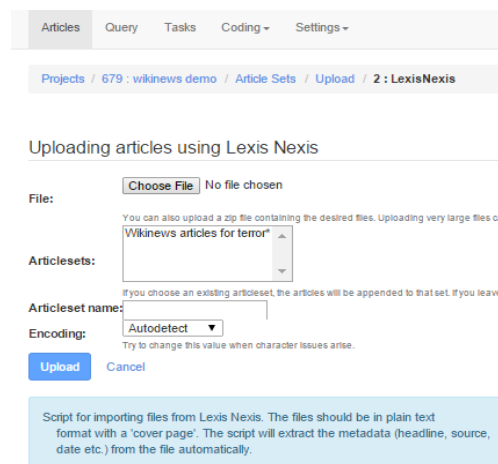


Figure 4. Uploading articles form Lexis Nexis

Lexis Nexis format

In order to upload articles from Lexis Nexis to AmCAT you should collect them from lexis nexis as a txt.file also you should make sure you save it correctly (see box 1 uploading articles through Lexis Nexis).

Box 1. Uploading articles through Lexis Nexis

Selecting articles in Lexis Nexis

- 1) Go to LexisNexis
- 2) Select "Date" and "news sources" and fill in your search term
- 3) Save the articles by clicking on the disc logo in the upper right corner
- 4) Select in the pop-up window by "Format" de option *tekst/text* by ticking at "Page Options" *Voorblad/cover page*.
- 5) Click on download and save under a chosen name on your computer.

Go to AMCAT and choose your project. Click at ADD ARTICLES TO PROJECT.
Browse to your saved file with the articles from lexis nexis.
Upload your file to amcat in your own Project!

Browsing & Analysing the data

When you have a collection of texts it is shown in the articles set. Each articleset can be used separately or all sets can be used together. We go back to the overview with the articlesets and click on the article set (in this case 679: wikinews demo). You will get an overview of all articles in this set:

id	date	headline	mediumid	medium	creator	byline	addressee	section	uri	length	page
114399553	2008-08-06T00:00:00	Somali "Anti-Terrorism Allian...	989907784	Wikinews						600	
114399554	2008-08-11T00:00:00	Police in Britain uncover susp...	989907784	Wikinews						1895	
114399555	2008-09-07T00:00:00	17 terrorism suspects arrested...	989907784	Wikinews						294	
114399556	2004-12-20T00:00:00	Department of Defense report L...	989907784	Wikinews						555	
114399557	2008-06-02T00:00:00	British police shoot man in an...	989907784	Wikinews						405	
114399558	2008-08-28T00:00:00	Pentagon con...	989907784	Wikinews						260	
114399559	2005-07-04T00:00:00	Department of Defense report lambasts communication failure in US War on Terror	989907784	Wikinews						425	
114399560	2007-04-23T00:00:00	Civilians deal...	989907784	Wikinews						249	
114399561	2005-07-07T00:00:00	Archbishop says abortion and e...	989907784	Wikinews						1772	
114399562	2007-06-30T00:00:00	Coordinated terrorist attack h...	989907784	Wikinews						290	
114399563	2011-10-19T00:00:00	M15 terror threat warning leve...	989907784	Wikinews						560	
114399564	2008-09-08T00:00:00	Jailed Swedish journalists tri...	989907784	Wikinews						444	
114399565	2010-01-28T00:00:00	Colombian army officials accus...	989907784	Wikinews						745	
114399566	2005-07-07T00:00:00	International conference agree...	989907784	Wikinews						242	
114399566	2005-07-07T00:00:00	US Homeland Security raises tr...	989907784	Wikinews							

Figure 5. Overview of articles in an article set

You can click on the individual article and in a new tab you get to read the whole article. For now, we go to the Query screen in order to conduct some queries and explore the data! Click on the blue button on the left corner and you get the query screen.

Chapter 2. Searching the content of articles

The query screen gives you a few options. Note that in the Article sets now only the Wikinews articles for terror is selected. This option makes it possible to only search one article set, or more. When you do not select an article set, AmCAT uses all articlesets in the project.

Selection media and period (dates)

In the *Query Search*, it is common practice that search orders are execute on the entire index. This function, however, offers the possibility to execute the analysis on the index's subset. You can define this subset by selecting specific media ('Media') and/or date ('Date'). The standard configuration for Media is *all media*. By clicking the drop down menu you can select one or multiple media in your dataset to which you want to restrict your search order. Use the Ctrl-button to select multiple media. The standard configuration for period is *all dates*. Select the period in the drop down menu by selecting *Is*, *Between*, *From*, or *Before* and enter the chosen dates.

The screenshot shows the AmCAT Navigator interface. At the top, there are navigation tabs: Articles, Query, Tasks, Coding, and Settings. The 'Query' tab is active. Below the tabs, there is a breadcrumb trail: Projects / 679: wikinews demo / Query. The main area is divided into two columns. The left column is titled 'Filter options' and contains several dropdown menus: 'Show sets' (Favourites, Coding Job Sets, All Sets), 'Article sets' (17609 - Wikinews articles for terror), 'Date' (All Dates), and 'Medium' (All Mediums). Below these is the 'Output options' section with radio buttons for 'Summary', 'Article List', and 'Graph/Table'. The right column is titled 'Query options' and features a large text input area for 'Keyword Search string(s)'. Below the input area is a 'Search Syntax Help' link and two buttons: 'Article IDs' and 'Codebook'. At the bottom left of the filter options section is a 'Submit' button. At the bottom center, the word 'Results' is displayed.

Figure 6. Query screen

How to search?

To search in the index it is crucial to make a search string that fits the concept you want to measure. Obama for example is easy since there are few people in the news with the same name. More problematic will be somebody like former President Bush, since not only are there other people with the same name, the name also mentioned something else.

This is one problem with search strings. Another problem/challenge is to get all the words that fit within a concept. A concept like 'Fear' or 'Anger' is extremely difficult to determine with words. Same is true for 'Positive' and 'Negative'. Terrorism for example is doable. However, here you have to make sure you also include words like terror, terrorist and terrorists. This is still quite doable. More complicated is a concept like crime. Think about it...

In AmCAT there are several options to help define good search strings. In a table we provide an overview. Note that we use LOWER cases for all the names. However AND and OR should be in Capitals.

Table 1. Options for creating search terms

Notation in query screen	Description
obama terrorism; obama OR terrorism	Articles that contain obama or terrorism (when there's a space between two words, AmCAT automatically fills OR in in the search)
obama AND terrorism	Articles that contain obama and terrorism
obama NOT terrorism	Articles that contain obama, but not terrorism
"obama fights terrorism"	We are looking for the phrase "word1 word2 wordx". Only when these words are literally in this order directly after each other, they will be found with this search term.
"obama terrorism"~5	Articles that contain Obama and terrorism only within 5 words from each other. 5 is a random number, you can choose any other numbers which are higher than 2.
"(obama OR biden) (iran OR terrorism)"~10	The article contains obama or biden, at a maximum distance of 10 words to <i>either</i> terrorism OR iran.
* or ? (wildcards)	A ? in a search term indicates that the place of ? can be changed in a another character. For example: terroris?, can be terrorism, or terrorist. The place of a * can be changed in a character either in a couple of characters. For example: Economi*, can be changed in economic, economical, economic state, but also in longer words too. Note: you can't put a wildcard at the beginning of a word.

You can combine these techniques to make a more sophisticated search string. Therefore some additional options:

Table 1. Additional options for creating search terms

Notation in query screen	Description
Identifier#	This syntax allows to label your long search term with one single name. The AmCAT-output displays your identifier for the whole search term. For example if the search string is Israel# "(israel* OR idf) (attack* OR invade* OR invas*)"~10, AmCAT will only use Israel in the output for the search string.
() Brackets	Using brackets allows you to make more complicated combinations of modifiers like OR and NOT. Just as in a mathematical formula gets the part of the search term that in brackets is, rank.
Headline: obama	Displays the articles that contain Obama in their headlines. This can also with more complicated search terms. Note: When you fill in: headline: (obama OR bush), displays the articles which contain Obama, Bush or both, but when you fill in: headline: obama or bush, it find the articles which contain obama in the headline, bush in the text, or both.

Functions 'Query Search'

There are various search functions in *the Query Search*. In this chapter, we will illustrate the main functions. The first part contains an overview of four search functions: *Show Summary*, *Show Table*, *Show Graph*, and *Show Cluster Map*. The second part will explain how these different functions can be used to test various hypotheses.

Show Summary

This function provides the possibility to show all articles containing the search terms you entered in the *Query* field. A list is shown with all documents in which one or multiple of your search terms are mentioned. Of each article, the following specifics are displayed: the title, the context in which the search terms are mentioned, the number of times the search terms are mentioned, the date, the medium, and the section of the medium in which the article was published. When clicking the title of one of the articles, the entire text and some additional metadata will show. The search terms you entered are in bold.

This function is skilful to use in order to get a first impression of the content of the debate concerning your concepts in the examined documents. The *Show Summary* function can also be used for verifying the validity and reliability of the formulated search terms. More on this follows in chapter 3. Here an example of a Summary with Obama OR terrorism:

Articles Query Tasks Coding Settings AmCAT Navigator 3.3.2-rc1 Projects Annotator Nel Ruigrok

Projects / 679 : wikinews demo / Query

Filter options

Show sets: Favourites Coding Job Sets All Sets

Article sets: 17609 - Wikinews articles for terror*

Date: All Dates

Medium: All Mediums

Output options

Output: Summary Article List Graph/Table

Submit

Results

0.39 seconds

Found 541 articles

Assign as codingjob ClusterMap Associations Save as Set

Somali "Anti-Terrorism Alliance" gets US funds; so do "Blackhawk Down" warlords
06-06-2006 | words | -
... Alliance for the Restoration of Peace and Counter-Terrorism in war-torn Somalia appeared to involve ... Restoration of Peace and Counter-Terrorism had increased support for Islamists. According to Zorick, the ...

Police in Britain uncover suspected terrorist plot
11-08-2006 | words | -
... attacks in London, the official said. The counter-terrorism operation took several months to complete ... the U.S., and security officials there say aspects of the plan were suggestive of al-Qaeda. Terrorism ...

17 terrorism suspects arrested in Belgium
07-09-2006 | words | -

Department of Defense report lambasts communication failure in US War on Terror
20-12-2004 | words | -
... atmosphere of hostility can manifest in numerous ways. The list of harmful effects includes "terrorism ...

Figure 7. Summary option in query screen

Additional options

The *Show Summary* function has also the possibility of saving this summary as a separate set. This is useful if you want to look at specific parts of the data separately. Through the button on the right “Save as set” this can be done and you can give it a different name. Next time you open the project the set will be in the list of article sets. Note that the articles will not be copied, it is just a categorization of the articles.

Graph/ Table

The *Graph/Table* option enables you to produce a frequency-list or cross-table. There are different options possible. If you click on graph/Table you get the options as shown here.

You can choose what you want to put on the X-as and Y-axis. Default option is the Date that can be either month, year, quarter, week or day. Other options for the X-as are Medium and Total. For example if you have several search terms per medium you want to explore you choose here for medium.

Output options

Output: Summary Article List Graph/Table

X axis:

Date interval:

Y axis:

Output type:

Make values relative to (and exclude) first column:

Figure 8. Graph/Table option in Query screen

The options for Y-axis are also multiple: Total, Search term and medium. So in the case just mentioned you can choose search term here. If you want to look at one search term and different media over time, you choose medium here and Date in the X-axis. The output can either be a table or a graph. For example we now show “Obama terrorism” as a search term over time within only one medium:

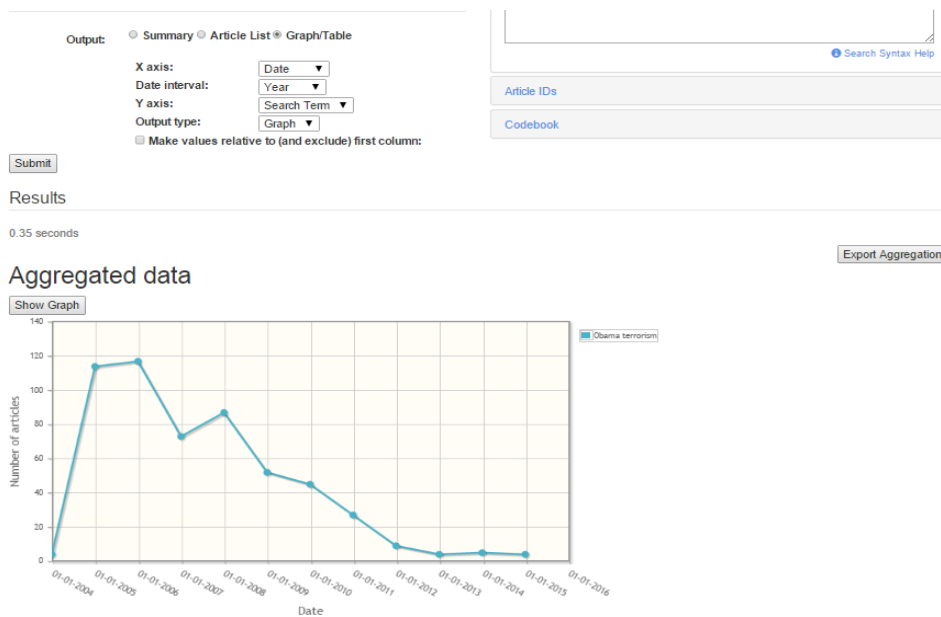


Figure 9. Output from a query using graph option

Note that you can also export the data into an excel file or SPSS file on the right hand side. This is important to remember since for the statistical analysis of the data you should always export the data into SPSS (or other software to conduct analyses).

Verifying search terms

It is important to thoroughly verify your search term: does it actually deliver the correct results? First, you want to know whether the search term does only lead to results related to the concept you defined for your research. In order to verify your search term, use *Show Summary* as we have shown before.

Another way of checking is by using *Show Keyword-in-Context*. This produces a table in which the context wherein your search term appears is presented. This way, you can verify the accuracy of your search term. Here you see an example of the articles in which Obama OR terrorism is searched. We clicked on Article List and clicked on Keyword in Context beside other aspects of the article.

id	date	headline	mediumid	medium	left	keyword	right
4399553	2008-08-08T00:00:00	Somali "Anti-Terrorism" em...	989907784	Wikinews	Somali "Anti-	Terrorism	Alliance" gets US funds; so d...
4399554	2008-08-11T00:00:00	Police in Britain uncover susp...	989907784	Wikinews	attacks in London, the offic...	terrorism	operation took several months...
4399555	2008-09-07T00:00:00	17 "terrorism" suspects...	989907784	Wikinews	17	terrorism	suspects arrested in Belgium
4399556	2004-12-20T00:00:00	Department of Defense report l...	989907784	Wikinews	atmosphere of hostility can m...	terrorism	
4399557	2008-08-02T00:00:00	British police shoot man in an...	989907784	Wikinews	British police shoot man in an...	terrorism	raid
4399560	2007-04-23T00:00:00	Archbishop says abortion and e...	989907784	Wikinews	Archbishop says abortion and e...	terrorism	with a human face'
4399562	2007-06-30T00:00:00	MI6 terror threat warning leve...	989907784	Wikinews	—Home Secretary Jacqui Smith ...	Terrorism	Analysis Centre (JTAC), a
4399563	2011-10-19T00:00:00	Jailed Swedish journalists tri...	989907784	Wikinews	The trial for two Swedish jo...	terrorism	in Ethiopia began yesterday. ...
4399564	2008-09-08T00:00:00	Colombian army officials accus...	989907784	Wikinews	purpose of allowing the offic...	terrorism	
4399565	2010-01-28T00:00:00	International conference agree...	989907784	Wikinews	counter-	terrorism	strategy. In tackling terrori...

Figure 10. Key word in context

Here you see the keyword 'terrorism' and at its left and right the words that appear in the text as well. This way you can determine if the word covers the concept you want to search.

Here again you can export the results to Excel.

Show Cluster Map

The function *Show Cluster Map* enables you to construct a venn diagram. A venn diagram shows the overlap between different sets. Therefore you should insert at least two search terms. A venn diagram is useful in order to visualize the extent to which different concepts are related to one another. In other words, a venn diagram presents associative frames.

For example, if you want to measure how often terrorism is related to either Israel or Palestine, you can fill in the following search terms:

Terrorism#terror*

Israel#israel*

Palestine#palesti*

Note that every search term should start on a separate line!

First you click on summary (in this case you see 1190 articles). Next you click on Clustermap on the right of the screen and in the pop-up you choose the option “show in navigator”. Click on OK and you get the following result.

In total 1190 articles. You see that the majority of the articles do not contain either Palestine or Israel. These are the 1016 articles in the bleu part of the venn diagram. Moreover you see that only two articles are found in which Palestine is mentioned but not Terror or Israel (every yellow dot is 1 article, you can click on them to see the actual text).

Additional options

Using the option *Save as csv* you can save the quantitative data on which the venn diagram is based as a csv-file, after which you can open and process the file in a different program. The csv-file contains a datamatrix which shows for each combination of the search terms you inserted how often this combination occurs.

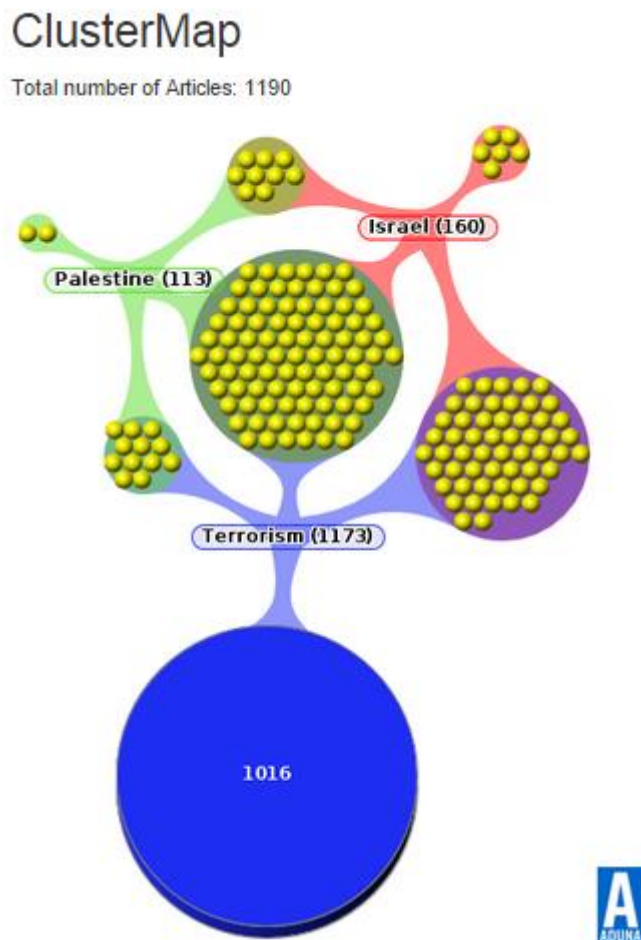


Figure 11. Cluster map

Show Associations

The *Show Associations* function is comparable to the *Show Cluster Map* one. The main difference between the two is that *Show Associations* enables you to quantify the extent to which the overlap between the different concepts occurs. The analysis makes use of a so-called *Conditional Probability (asymmetric)*. The measure is calculated by dividing the number of times concept A and concept B appear within the documents in which both appear by the number of times concept A appears in the entire set of documents.

This way you can predict the chances that concept B appears in a text, given the fact that concept A appears. In order to do so, you fill in the concepts again in the query screen. We use again the concepts as used in the venn diagram:

Terrorism#terror*

Israel#israel*

Palestine#palesti*

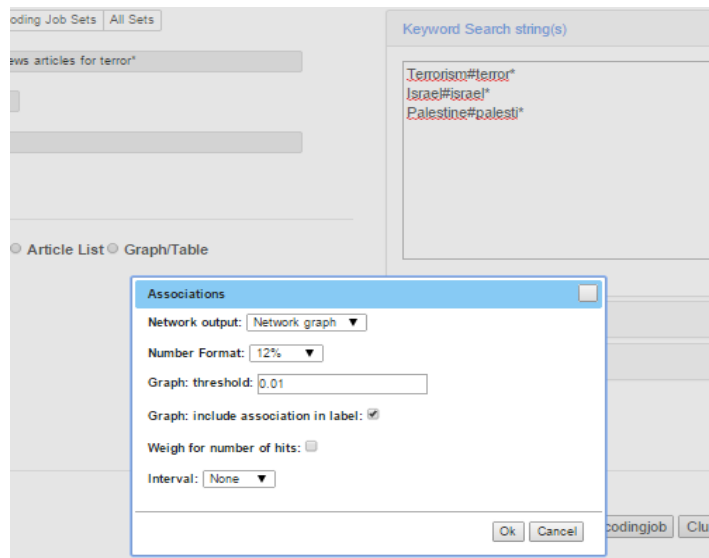


Figure 12. Options for Association graph

You click on summary first and next you click on Associations. In the pop-up screen you choose Network Graph and number format as presented there with 12%. Fill in 0.01 at the threshold (otherwise it will produce arrows as well if there is no relation and we do not want that) and click on include associations in label (this gives you the actual percentages. Click Ok and you get the following picture:

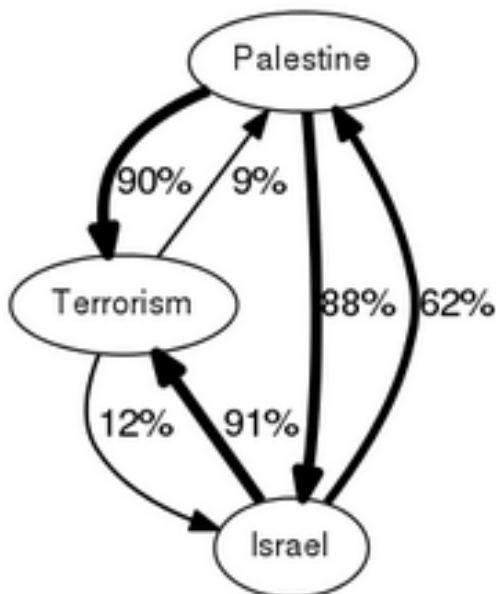


Figure 13. Association graph in AmCAT

This can be read as follows. IF the word Terrorism is mentioned in an article, the chance is 9% that also Palestine is mentioned, while the chance is 12% that Israel is mentioned. On the other hand, if Palestine is mentioned, the chance is 90% that terrorism is mentioned. This percentage is 91% in articles in which Israel is mentioned.

When we look at the both countries in articles we see that when Palestine is mentioned in 88% of the articles also Israel is mentioned. The other way around the percentage is lower, 62%.

These analyses can provide you information about the strength of associations made in articles. See also for example the article: "Shifting frames".