

Texcavator¹ End-user Manual



Fons Laan
Carlos Martinez Ortiz
José de Kruif

Version 1

24 Aug 2016

¹ This manual is the updated version of the manual for the tool that preceded Texcavator: WAHSP.

Contents

1. Introduction
2. Toolbar
3. Screen regions
4. Config options
5. Searching
6. Word Cloud configuration
7. Adding Shico
8. Log-in
9. Acknowledgments

1. Introduction

In this document we will explain how to use the web application *Texcavator*. With your browser you can find the application at <http://texcavator.surfsaralabs.nl/>.

Texcavator allows you to use full-text search on the newspaper archive of the [Dutch Royal Library](#) within the date range 1850-1990. On top of that, it offers visualizations like word clouds, time lines and heat maps. It also provides services to enhance your search experience like filtering, stop word removal, normalization and stemming. Please read this documentation to check all available options.

2. User Interface

The *Texcavator* opening panel is shown in Fig. 1.

Most users use the Guest login. A guest can use all functionalities of Texcavator but will not be able to download subsets of newspaper articles or save queries containing more than 50.000 hits.

You don't need to log in to use Texcavator's basic functionality, but if you want the full set of options available, please log in with your UU account.

Due to copyright considerations, only researches from within the faculty of the Humanities of Utrecht University can apply for a Researcher login.

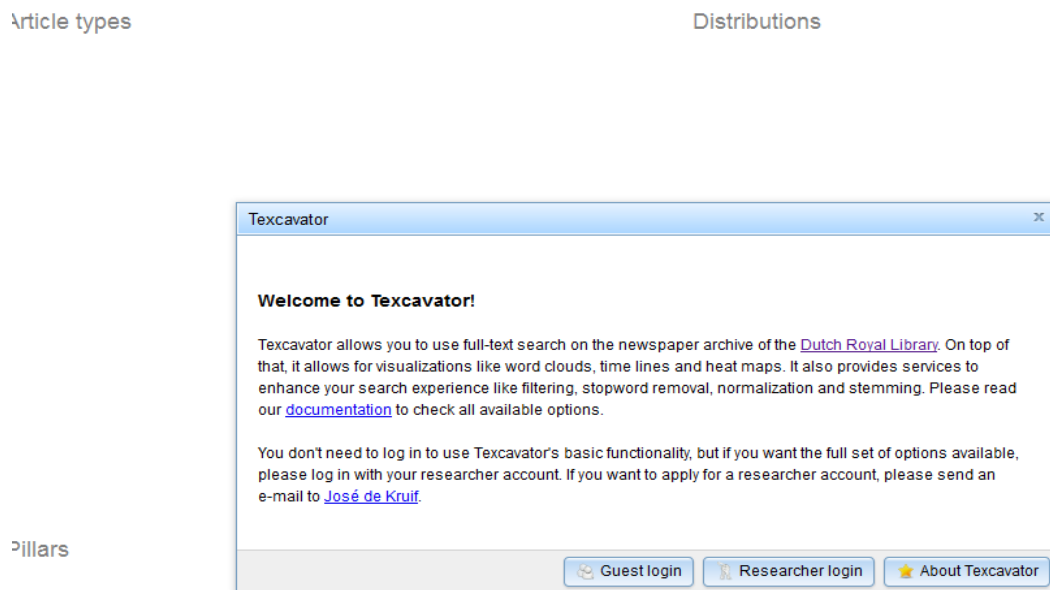


Fig. 1

After login the following screen regions are shown:

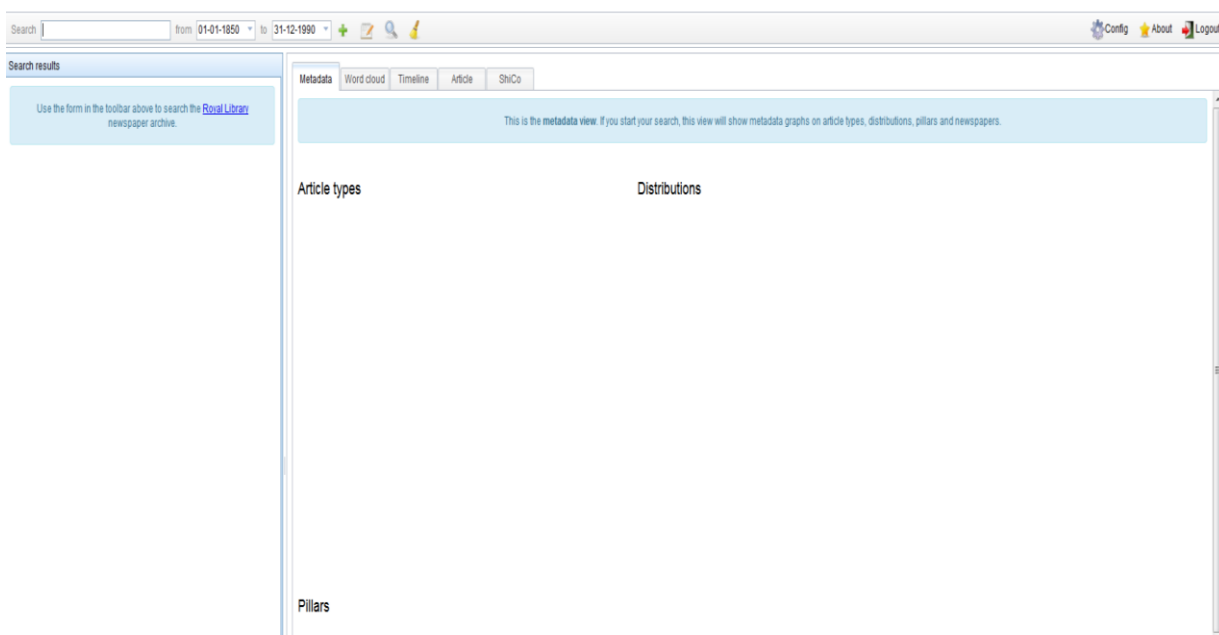









Fig. 2

3. Toolbar

The toolbar at the top (Fig. 3) contains:



Fig. 3

- A search box at the left
- A date widget next to the search box
- A  sign to add a second period to your date widget
- An edit  widget to open a wider screen to adjust your query (this screen expands while you are typing)
- A magnifying  glass that enables you to wipe your query and start another query
- A brush  to clear the entire search form
- The  **Config** widget (config options will be explained later)
Two icons  **About**  **Start** for further documentation.

4. Screens

The Search results panel at the left shows a list of hits consisting of article titles and dates.

The panel at the right contains five tabs:

- A tab for the Metadata view. Scroll to the bottom of this window to inspect pillarization and distribution over newspapers of your search results.
- A tab for the Word Cloud view. Clicking on an article or clicking the cloud button next to a saved query will yield a word cloud.
- A tab for the Timeline view. Clicking on the timeline button next to a saved query creates a timeline graph. By clicking on one of the bars of the graph you can view a word cloud as well as a heat map for the period you selected.

- A tab for the Article view. After searching, select an article to display its text and its image. Clicking on the image will redirect you to the full newspaper page in *Delpher*.
- A tab for ShiCo (ShiCo will be explained later)

5. Searching

A trivial way to search is by using a single query word. Say, we type “iceberg” in the textline area in the search box, have 1850-1990 in the time widget and hit Enter

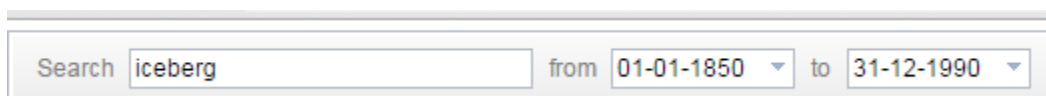


Fig. 4

The left panel shows that 561 articles were found.

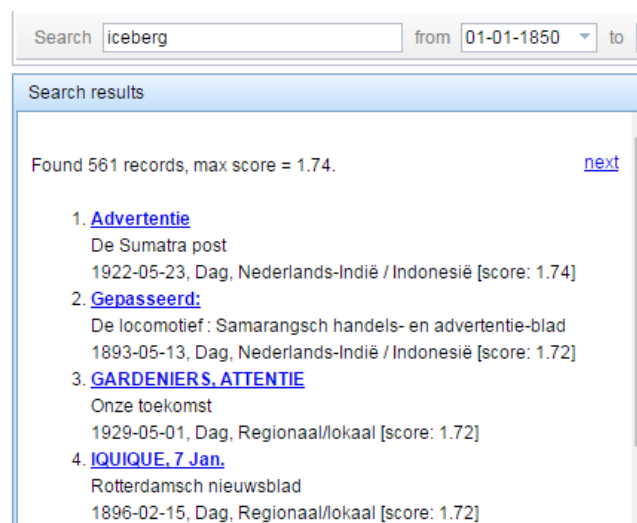


Fig. 5

These 561 hits are ordered according to relevance with a max relevance score of 1.74.² The first chunk is displayed with the titles bold and underlined. Underneath the title is some additional information: the newspaper title, article date, and newspaper 'type' (country-wide, or regional). Clicking [next](#) gives the next chunk of articles. In the right panel you find graphs (Fig. 6Fig. 5) that show the metadata for this query.

² Relevance is computed according to the [TF/IDF](#) formula.

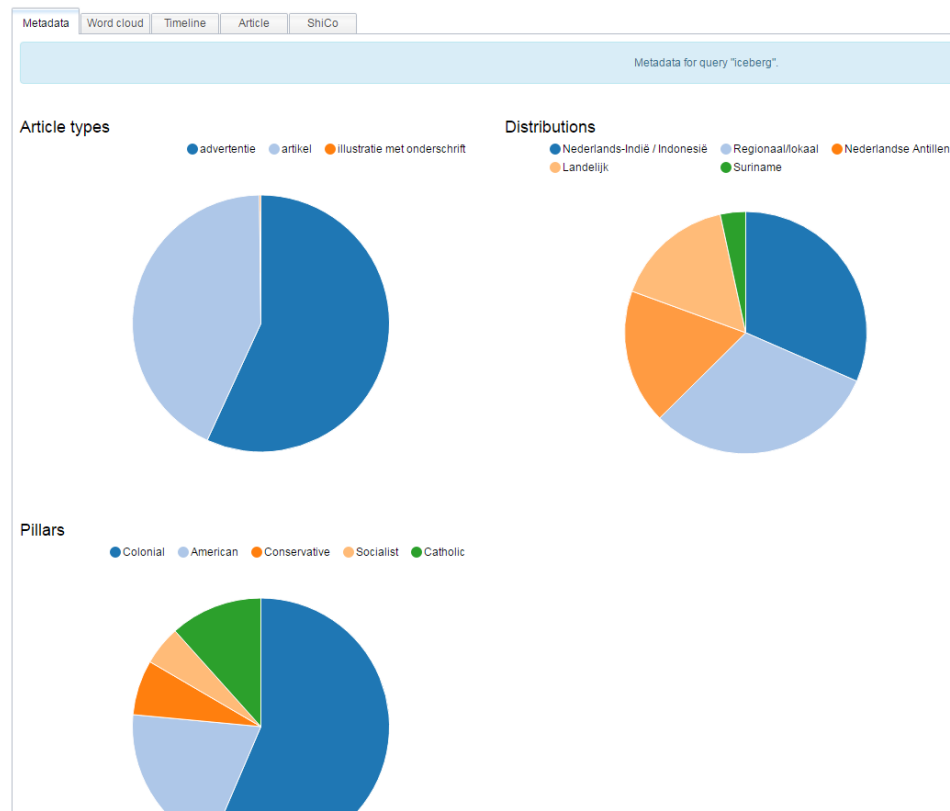


Fig. 6

When you click one of the article titles, the OCR text and the image of the article text are shown in the Article tab (Fig. 7)

Metadata Word cloud Timeline **Article** ShiCo

This is the **article view**. After searching, select an article to display its text and its image. Clicking on the image will redirect you to the full newspaper page in [Delpher](#).

S. O. S. Iceberg.

Zaterdagavond woonden wij, daartoe uitgenoodigd door de Directie van Royal, de voorstelling bij van de Poolfilm S. O. S. Iceberg. Zooals men vermoedelijk weet, is deze film op Groenland opgenomen. Hiertoe werd een speciale expeditie uitgerust, van welke verrichtingen, de Europeesche zoowel als de Amerikaansche pers, geregeld haar lezers op de hoogte hield. Moeiten nog kosten werden gespaard ; vreemdelingen met de noodige ervaring werden speciaal voor deze film aangehuurd. S. O. S. Iceberg is dan ook een product, dat overal goed is ontvangen en een uitstekende pers heeft gehad. Wie Eskimo gezien heeft, mag S.O.S. Iceberg niet verzuimen. Beide filmen spelen in het Poolgebied. Toch bestaat er verschil. In Eskimo ziet men meer het leven, de gewoonten, de zeden van de Eskimo's. In S.O.S. Iceberg wordt inder het grootsche in de Poolnatuur weergegeven. Er is zelden hier te lande een film vertoond met zulke mooie natuurgezichten. Het ineensstorten der ijsmassa's het

S. O. S. Iceberg.

Zaterdagavond woonden wij, daartoe uitgenoodigd door de Directie van Royal, de voorstelling bij van de Poolfilm S. O. S. Iceberg. Zooals men vermoedelijk weet, is deze film op Groenland opgenomen. Hiertoe werd een speciale expeditie uitgerust, van welke verrichtingen, de Europeesche zoowel als de Amerikaansche pers, geregeld haar lezers op de hoogte hield. Moeiten nog kosten werden gespaard ; vreemdelingen met de noodige ervaring werden speciaal voor deze film aangehuurd. S. O. S. Iceberg is dan ook een product, dat overal goed is ontvangen en een uitstekende pers heeft gehad. Wie Eskimo gezien heeft, mag S.O.S. Iceberg niet verzuimen.

Fig. 7

Clicking on the OCR image, will redirect you to the newspaper in [Delpher](#), the online news depot of the Royal Library.

Clicking on the Word cloud tab, will generate the word cloud of the article (Fig. 8)



Fig. 8

The font size of the words is the graphical equivalent of their frequency in the document. Words of too low frequency may not be shown, and in general 'noise' (i.e. stopwords) is also suppressed. When you have chosen the SVG (Scalable Vector Graphics) cloud option in the Config menu, you can see the word frequencies at the top of the cloud when you hover the words with your mouse. And when you click on a word, you can add that word to your personal stopword list, to be removed the next time you generate a cloud. The timeline tab has no relevance for single articles.

When hovering over a term, you get the offer of feeding the term to ShiCo. More on ShiCo on

6. Saving a Query


Important: Queries saved by guests are limited to **50000 hits**. Also, queries saved by guests will be deleted **every day**. Users with research privileges will be able to save queries for a longer period.

When you want to use the results of a query for further investigation, you can save your query (Fig. 9). You can give your query a name and a description (optional).

You can save your current query for additional analysis and later reference in this panel. Give it a proper title and (optionally) a description.

Title:
Iceberg











Description (optional):
Searching for Iceberg without filters


 Save query

Queries saved by guests are limited to 50000 hits.
Also, queries saved by guests will be deleted every day.

Fig. 9

Your saved queries



Mainzer beobachter [74] 2016-09-02T13:29					
Iceberg [561] 2016-09-02T13:28					


 Refresh query list

Queries saved by guests are limited to 50000 hits.
Also, queries saved by guests will be deleted every day.

Fig. 10

Fig. 10 shows the list of your saved queries, they are used to retrieve the OCR data of the articles, create word clouds, timelines, and display newspaper statistics.

The  button will regenerate the article list. The  button will (re)generate the wordcloud.

Clicking the  button generates an interesting graph: the timeline.

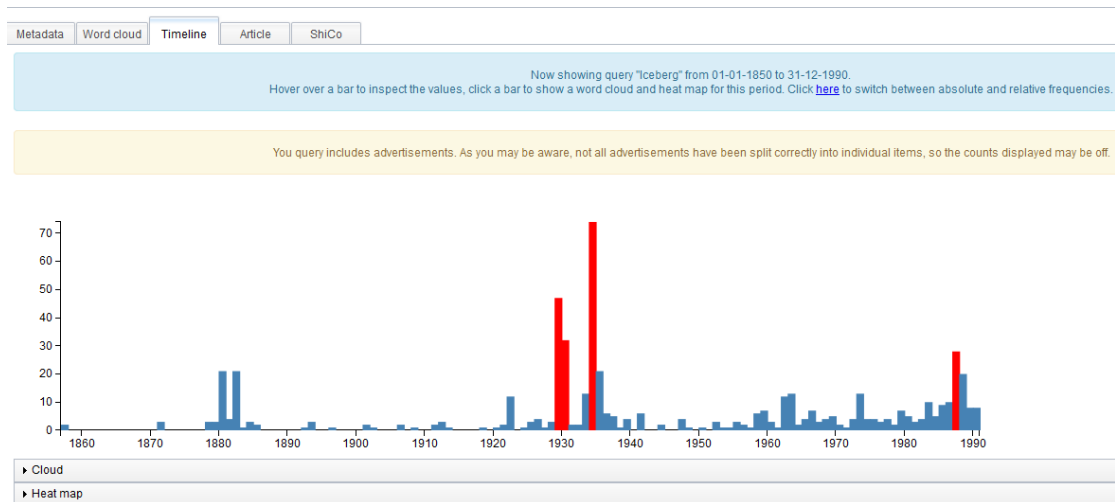
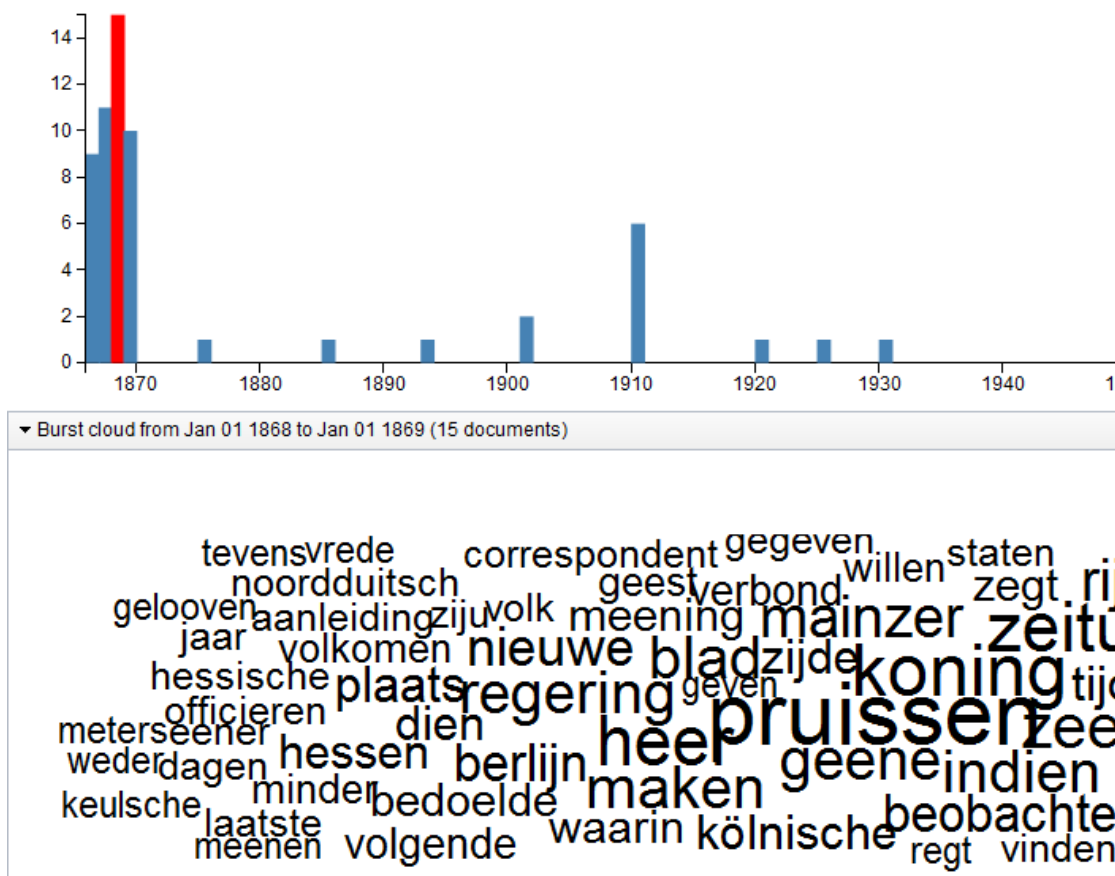


Fig. 11

The timeline graph (Fig. 11) presents a histogram of article frequencies over time. Red bars indicate an unusually high number of hits. Clicking on a bar, generates a word cloud as well as a heat map for the year you chose.



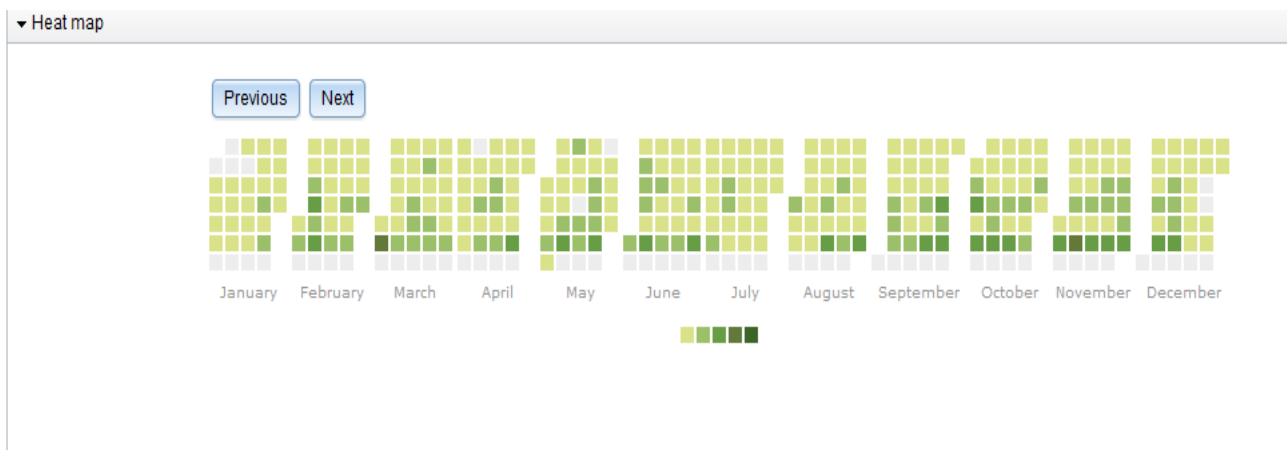


Fig. 12

The heatmap (Fig. 12) shows the hits by date. Click on a square and inspect the articles of that date in the article list in the left window.


For users with research permission there is an extra  button to export query results to json, csv or xml.

Fig. 13

7. Combined and Advanced Queries

The dataset is stored in an Elastic Search search engine. Thus, you can deploy the [elastic search query-string-query features](#). Follow the link to obtain an overview of all the possibilities. Regular expressions are not permitted in *Texcavator*. To give you an idea, the operators you can use are: "(", "\"", "\\", "*", "_exists_", "_missing_", "[?*", "[^+\\-=&|><!(){}\\[\\]^\"~*?:/]" or [a-z_].

The query string is parsed into terms and operators. A term can be a single word or a phrase, e.g. Greta or "Greta Garbo". There is no case-sensitivity.

You can use simple + or – operators and quotes to search for an expression but also use regular expressions.

- + signifies AND operation
- | signifies OR operation
- - negates a single token
- " wraps a number of tokens to signify a phrase for searching
- * at the end of a term signifies a prefix query
- (and) signify precedence
- ~N after a word signifies edit distance (fuzziness)
- ~N after a phrase signifies slop amount

In order to search for any of these special "reserved" characters, they will need to be escaped with \.

A few examples:

Greta Garbo:

Greta Garbo = Greta or Garbo

+Greta +Garbo = Greta AND Garbo

"Greta Garbo" = The exact term Greta Garbo

Wildcards: You can use wildcards to search for (historical) variations in spelling. * and ? at the beginning of a term (*arbo) are too heavy for Texcavator and are not allowed!!

8. Configuration settings

Search Tab

-Article type: You can make the results of your query more precise by choosing a particular type of article, e.g. "iceberg" but only in KB illustration texts.

-Distribution of newspapers: Choose from national, regional or colonial newspapers.

-Pillar: Dutch society was pillarized and each pillar had its own newspaper(s). The pillar distribution can be downloaded.

Configuration

Search Cloud Timeline Export

Search options

Article type

☒ Search KB regular articles

☒ Search KB advertisements

☒ Search KB illustration text

☒ Search KB family messages

Pillar ([download distribution as .csv](#))

☐ American

☐ Catholic

☐ Colonial

☐ Communist

☐ Conservative

☐ Jewish

☐ Miscellaneous

☐ No title

☐ NSB

☐ Pre-1900

Distribution

☒ National NL

☒ Regional NL

☒ Antillen

☒ Surinam

☒ Indonesia

OK

Fig. 14

-Sort order: When you scroll down, you can choose the sort order of your results (Fig. 15).

Configuration

Search Cloud Timeline Export

☒ Regional NL

☒ Antillen

☒ Surinam

☒ Indonesia

☐ Miscellaneous

☐ No title

☐ NSB

☐ Pre-1900

☐ Protestant

☐ Socialist

☐ World War II

50 Number of results to show

Sort order

☒ By score

☐ By date (oldest first)

☐ By date (newest first)

OK

Fig. 15

Cloud tab:

-Reduce font size differences. When the word sizes decline too fast at the cloud edge, this option should improve the result.

-Font scale factor. This scale factor determines the maximum font size.

-Stop words. This removes short words, as specified by a pre-defined list, plus the user-defined stop words from the cloud. The list of stop words can be downloaded.

-Minimum word length. Remove words shorter than x characters. When the stop word list does not block enough noise, this will enhance it.

-Stemming. This applies stemming to the words before computing the cloud.

-Max. # of words in cloud. The number of words returned by the server can be huge. Truncating the list before generating the cloud speeds it up.

-Cloud rendering. Choose between the original HTML canvas cloud, or the newer interactive SVG cloud.

Timeline Tab:

-Normalize document count: to smooth results and/or highlight bursts in red

-Export: export the word count of your cloud using comma's or tabs as separators

9. ShiCo

ShiCo was built by Carlos Martinez Ortiz. Carlos is an engineer at the Escience centre. Carlos also authored this text on ShiCo as well as [the explanation on the algorithms ShiCo uses](#).

ShiCo is a tool for visualizing time shifting concepts. We refer to a concept as the set of words which are related to a given seed word. ShiCo uses a set of semantic models (word2vec) spanning a number of years to explore how concepts change over time -- words related to a given concept at time $t=0$ may differ from the words related to the same concept at time $t=n$.

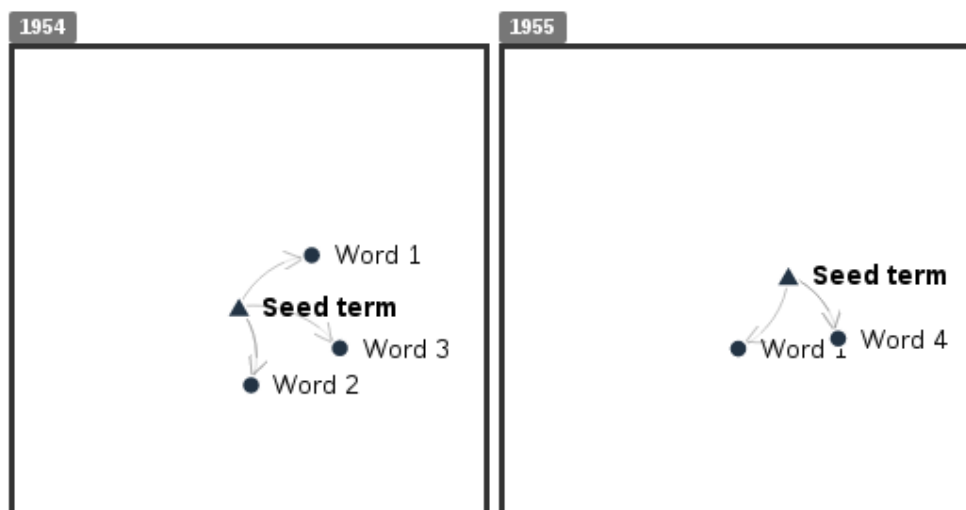


Fig. 16

How to use ShiCo?

This guide will instruct you in the elements for using ShiCo's user interface.

When you first open ShiCo on your browser, you will see a simple search bar:

Figure 17 shows the ShiCo user interface. It features a 'Concept search' section with a search bar labeled 'Search concepts'. To the right of the search bar is a blue button with a '+' symbol. Below the search bar are three green buttons: 'Submit', 'Save parameters', and 'Load parameters'.

Fig. 17

You can enter one or multiple (comma separated) **seed terms**. These seed terms are the entry point for your concept search. Click **Submit** to begin your search. The results from your search will be displayed in the results panel below the search bar.

The search bar has some additional features:

- It allows you to modify the search parameters. Click the **+** button to display additional search parameters.
- It allows you to save the parameters of your current search, or load the parameters of a previous search.

Search parameters

The following is the list of parameters (with a brief explanation) which can be used to control your concept search:

- *Max Terms* The maximum number of words to be included in the vocabulary for each time period.

- *Max related terms* Maximum number of words produced for each search term. These words are then trimmed to produce the final vocabulary (the size of which is controlled by Max Terms)
- *Minimum concept similarity* Minimum similarity concepts must have on semantic space. Words whose similarity (with the seed concept) is smaller than this threshold will be ignored. Also interpreted as Maximum concept distance.
- *Word boost* Additional weight given to seed terms when producing the vocabulary for each time step. Higher weights will cause seed terms to stay in the final vocabulary.
- *Boost method* Method used to determine the weight given to words produced, before the aggregation step. Two methods are available:
 - Sum similarity -- this method uses the similarity (between the seed word and each word) to determine the weight. The similarities are added for each word, as each word can appear in the results of different seed terms.
 - Counts -- count the number of times a word appears as the result of a seed term.
- *Algorithm* The vocabulary monitor contains can use two different algorithms for generating vocabularies. These control which words are used as seed terms for each model:
 - The non-adaptive vocabulary generator uses the same seed terms each time to generate the related terms.
 - The adaptive vocabulary generator uses the related terms generated by one semantic model as seed terms for the next semantic model. This adds an additional possibility: it allows for the semantic models to be used in chronological order, or in reverse chronological order -- searching forwards or backwards in time.
- *Track direction* Direction in which the concept search is conducted:
 - Forward -- starting from the earliest year and moving forward through time.
 - Backward -- starting from the latest year and moving backward through time.
- *Years in interval* Number of years which a single time step will cover.
- *Words per year* Maximum number of words per year time period which are left after the aggregation step.
- *Weighing function* Weighing function of the aggregation step.
- *Function shape* Parameter controlling the shape weighting function of the aggregation step.
- *Do cleaning (uses a cleaning function)*. Apply custom vocabulary cleaning function.
- *Year period* Period of years in which the search will take place. This allows to begin searching at any point in the available time range.

Produced graphics

Once a search is complete, ShiCo displays results in the results panel. Results are displayed using various graphs:

- Stream graph -- this shows each word of the resulting vocabulary as a stream over time. The stream gets wider or narrower according to the weight the word is given in the vocabulary.

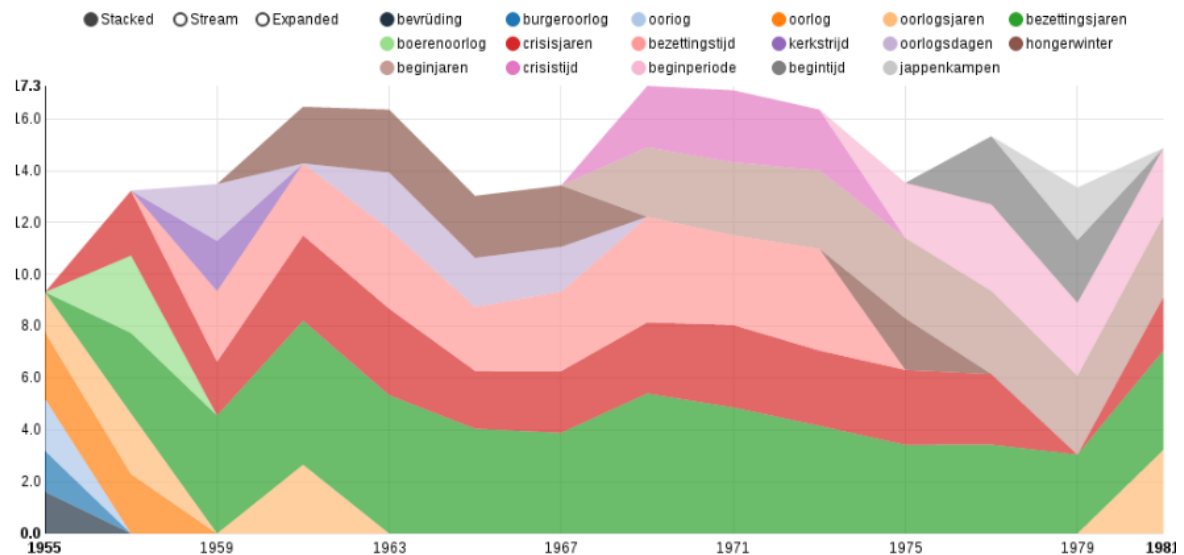


Fig. 18

- Network graphs - this shows a collection of graphs displaying the resulting vocabulary as a network graph. Words which are related to each other are connected with an arrow. The direction of the arrow indicates which word was the product of which seed word.

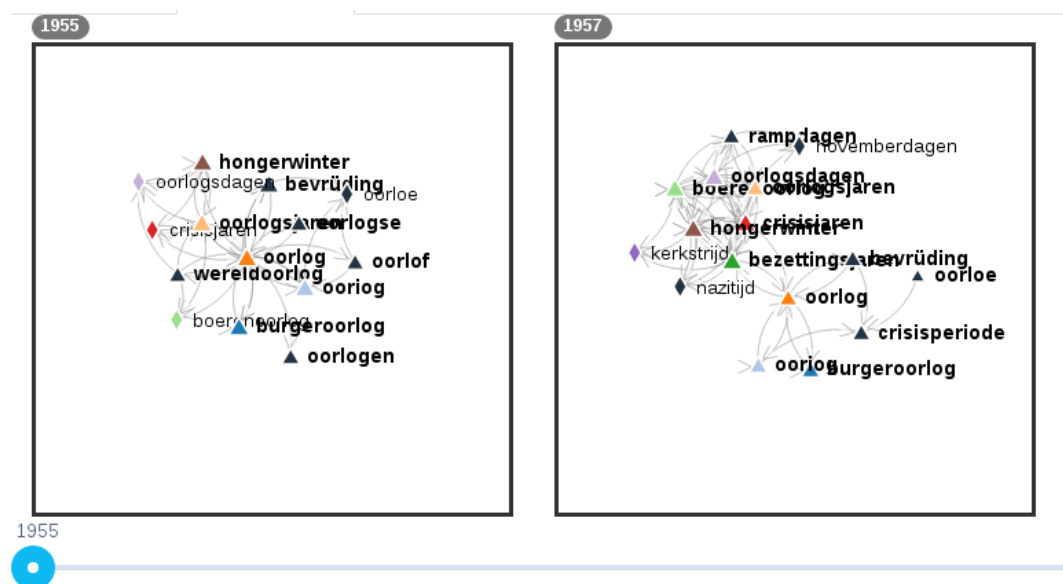


Fig. 19

- Space embedding -- this shows an estimate of the spatial relationship between words in the final vocabulary at every time step. Please keep in mind that these spatial relations are approximate and should be considered with care.

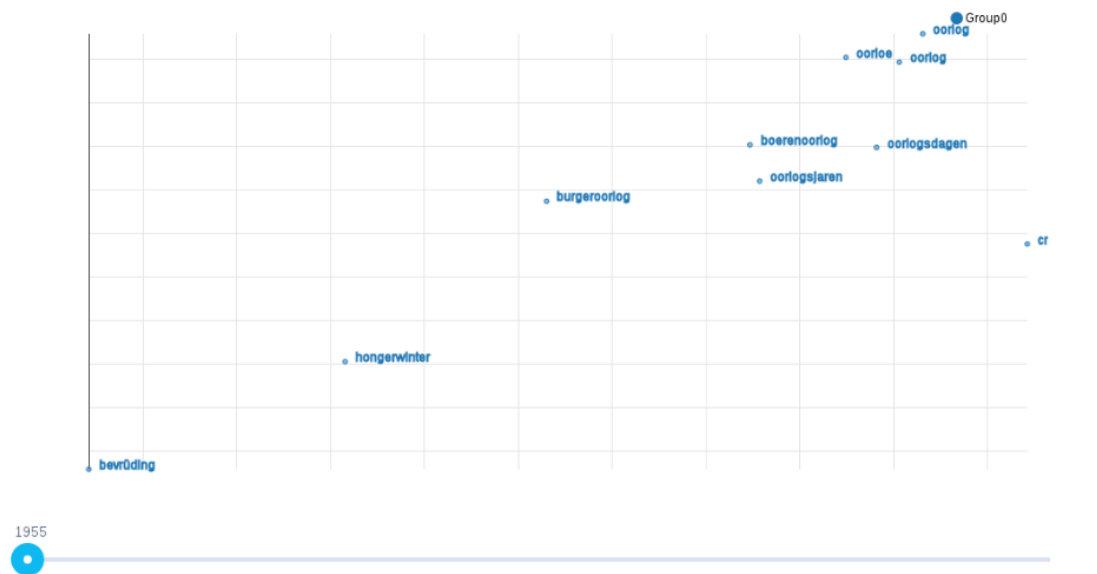


Fig. 20

- Plain text vocabulary -- this shows a text representation of the concept search. This consists, for each time step, of the seed words used and the produced vocabulary.

Saving and loading search parameters

When you click the Save parameters button, a text box with your search parameters will be displayed. Copy these parameters and save them somewhere. Click OK to hide the text box.

When you click the Load parameters button, another text box will be displayed. Enter previously saved search parameters in this box and click Ok to load the parameters.

ShiCo is a tool for visualizing time shifting concepts. We refer to a concept as the set of words which are related to a given seed word. ShiCo uses a set of semantic models (word2vec) spanning a number of years to explore how concepts change over time -- words related to a given concept at time $t=0$ may differ from the words related to the same concept at time $t=n$. For example:

ShiCo allows to explore the relation between terms over time in a way that traces the historicity of concepts and changes and continuities in connotation of terms.

How does ShiCo work?

Given a set of seed terms, ShiCo uses its semantic models to generate a vocabulary of related terms. This process is done for every one of the semantic models available. This is done using a *Vocabulary Monitor*. The vocabulary

monitor generates a vocabulary of related terms and a list of vocabulary links, which explains which of the related terms was generated by which one of the seed terms.

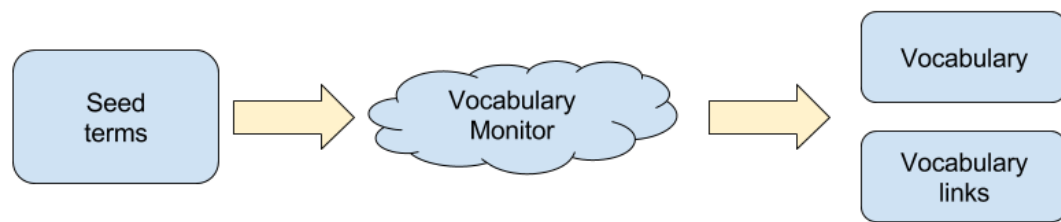


Fig. 21

The produced vocabularies and links are then aggregated using a *Vocabulary Aggregator* -- the vocabulary aggregator groups together results from multiple models into a single final vocabulary. This has a smoothing effect on the produced vocabulary.

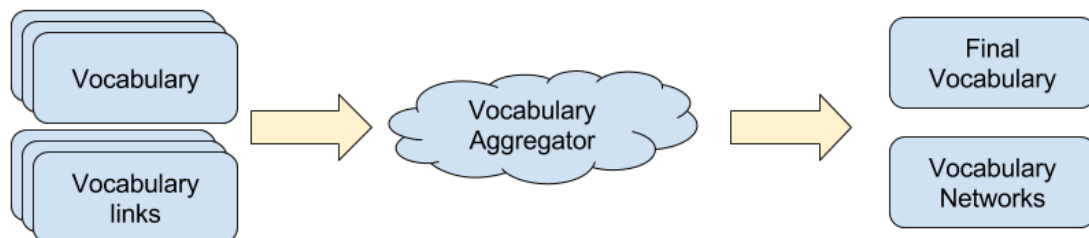


Fig. 22

The vocabulary monitor uses the given terms to query the semantic models. The models provide a list of *related words* and some measure of how closely related they are to the seed term -- this is interpreted as a *distance* between the seed term and the related word:

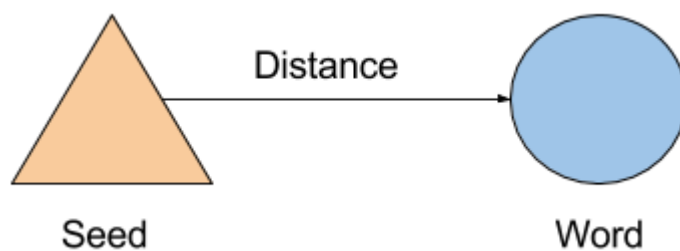


Fig. 23

As mentioned before, from each model the vocabulary monitor generates a vocabulary and list of vocabulary links. The vocabulary is a list of terms related to the seed terms -- each term is assigned a weight. This weight is calculated in one of two ways:

- As a count of the number of times the term appears
- As a sum of the distance between the seed terms and the related term

These weightings start from the assumption that a related term can appear as the result of one or more seed terms, in which case each time it increases the weight assigned to the related term.

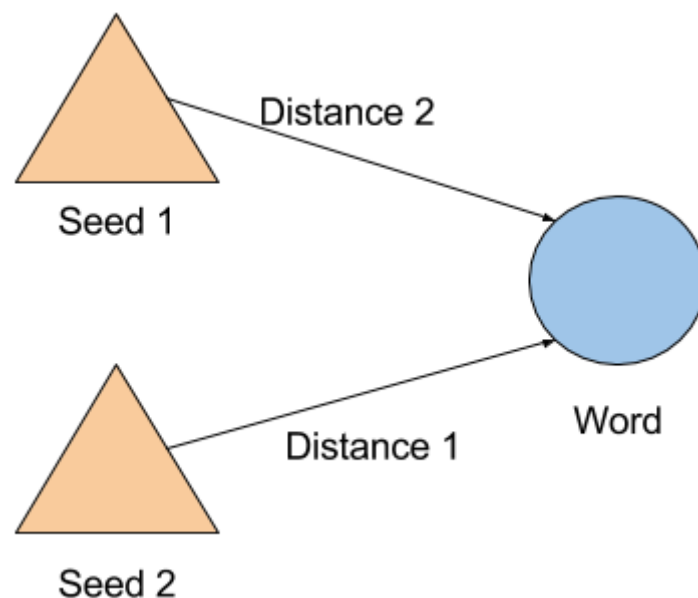


Fig. 24

In this way, the weights of each word are the sum of the weight contribution from each seed.

$\text{weights}(\text{word1}) = \text{weight}(\text{word1}, \text{seed1})$

$\text{weights}(\text{word2}) = \text{weight}(\text{word2}, \text{seed1}) + \text{weight}(\text{word2}, \text{seed2})$

...

The vocabulary monitor limits the number of words in the generated vocabulary, only the N words with the highest weights are included in the vocabulary.

word1: $\text{weights}(\text{word1})$

word2: $\text{weights}(\text{word2})$

...

wordN: $\text{weights}(\text{wordN})$

----- cut point

wordN+1: $\text{weights}(\text{wordN+1})$ # dropped

wordN+2: $\text{weights}(\text{wordN+2})$ # dropped

...

dropped

Algorithms

The vocabulary monitor contains can use two different algorithms for generating vocabularies. These control which words are used as seed terms for each model.

Non-adaptive

The non-adaptive vocabulary generator uses the same seed terms each time to generate the related terms.

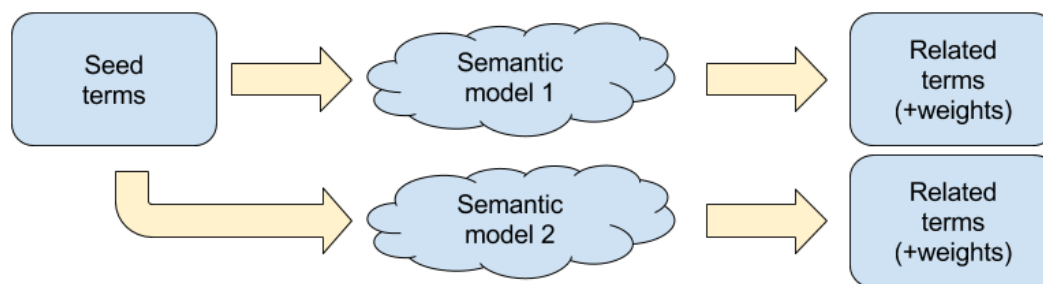


Fig. 25

Adaptive

The adaptive vocabulary generator uses the related terms generated by one semantic model as seed terms for the next semantic model.

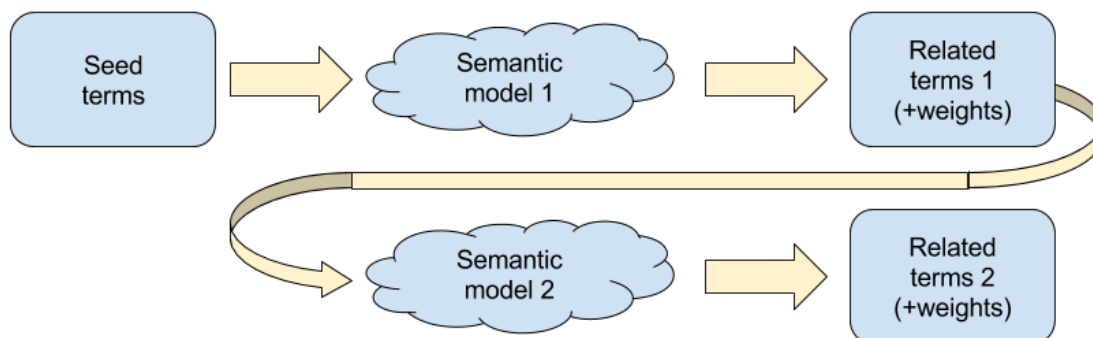


Fig. 26

This adds an additional possibility: it allows for the semantic models to be used in chronological order, or in reverse chronological order -- searching forwards or backwards in time.

Vocabulary Aggregator

A vocabulary aggregator takes a vocabulary produced by a vocabulary monitor and aggregates them over a set time window.

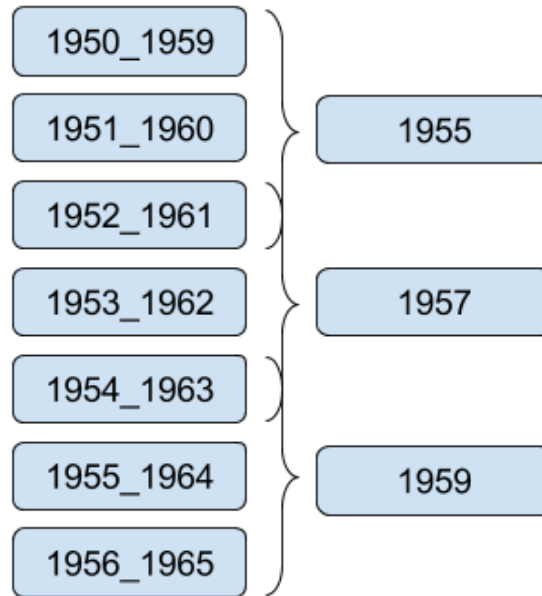
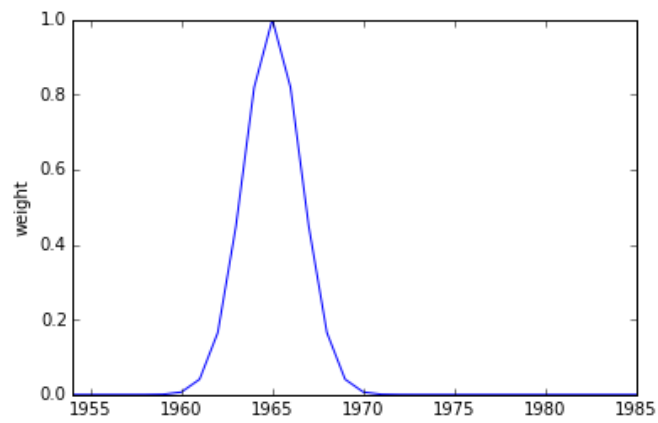


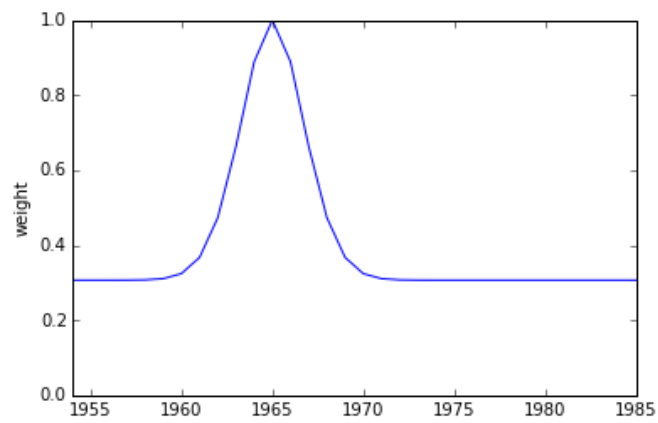
Fig. 27

Weighting functions are used to aggregate topics in a year:
Terms inside the 'window', are weighted by a weighting function.

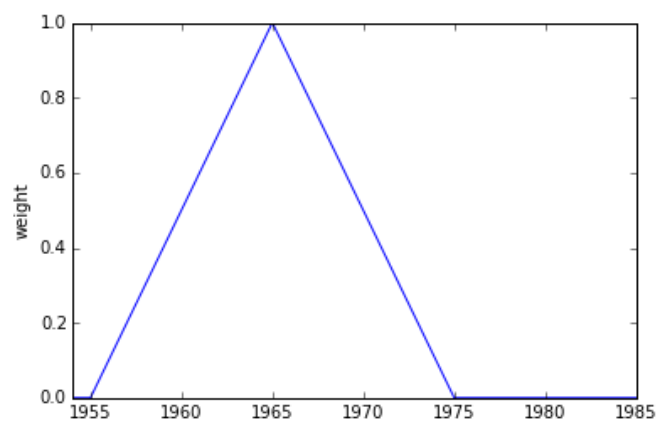
Gaussian



JSD



Linear



Known possible issues:

When Adblock (Plus) is active, the welcome screen of Texcavator does not fully load. This appears to be caused by some assets not being loaded from CDN. The solution is on the user end: make an exception for Texcavator in Adblock.

For further reading:

[elasticsearch Query String Query syntax](#)

[Wikipedia on Dutch orthography](#)