

Text Topics in Visual Analytics

v. Frans Holm, SAS

Text Object

Documentation and demo-data

- Documentation

- https://documentation.sas.com/doc/en/vacdc/v_013/vaobj/n0j471s3qd6kimn1l1do7c6plp5h.htm

- Data – Classification of operations

- https://medinfo.dk/sks/brows.php?s_nod=31170

	Klassifikation af operationer	K
□	Operationer på øje og øjenomgivelser	KC
□	Operationer på øjenhule	KCA
□	Operationer på øjenlåg	KCB
□	Operationer på tåreapparat	KCC
□	[#] Operationer på øjeæble	KCD
□	Operationer på øjenmuskler	KCE
□	Operationer på konjunktiva	KCF
□	Operationer på hornhinde og sklera	KCG
□	[#] Operationer i forreste øjenkammer, kammervinkel, iris og corpus ciliare	KCH
□	[#] Operationer på øjets linse	KCJ
□	[#] Operationer ved sygdomme i choroidea, corpus vitreum og nethinde	KCK
□	[#] Reoperationer efter operation på øje og øjenomgivelser	KCW

Text Topic object

Requirements

- Document Collection
 - Text variable
- Document Details
 - See extra variables
- Data Language
 - Set the Language
- Unique ID
 - Unique Row Identifier
 - Have to be categorical!

The screenshot shows the configuration interface for a Text Topic object. It is divided into four sections, each with an 'Add' button:

- DOCUMENT COLLECTION ***: Contains an '+ Add' button.
- DOCUMENT DETAILS**: Contains an '+ Add' button.
- DATA LANGUAGE ***: Contains a dropdown menu labeled 'Select a language'.
- UNIQUE ID**: Contains a field with a key icon and the text 'Unique_ID'. A 'Close' button is located at the bottom right of this section.

The screenshot shows the SAS Data menu with a context menu open over the 'SKSkode - 10K' item. The context menu options are:

- Add to selected object
- Add to current page
- Add to new page
- Add as a report control
- Add as a page control
- Explain >
- Predict >
- Select >
- Hide
- Duplicate
- New aggregated data
- Set as unique row identifier** (highlighted with a red circle)
- Custom sort...
- New custom category...
- New calculation...
- New geography...
- New parameter...

- Select a language
- Arabic
- Chinese
- Croatian
- Czech
- Danish
- Dutch
- English
- Farsi
- Finnish
- French
- German
- Greek
- Hebrew
- Hindi
- Hungarian
- Indonesian
- Italian
- Japanese
- Kazakh
- Korean
- Norwegian
- Polish
- Portuguese
- Romanian
- Russian
- Slovak
- Slovene
- Spanish
- Swedish
- Tagalog
- Thai
- Turkish
- Vietnamese

Text Object - Options

General

- Term parsing and role identification
 - Custom = Advanced options are shown
- Analyze document sentiment
 - Sentiment analysis determines whether a document has a positive sentiment, negative sentiment, or neutral sentiment based on the content of the document

General

Term parsing and role identification:

Custom

Analyze document sentiment ?

Parsing

Include parts of speech ?

Extract noun groups ?

Extract entities ?

Stem terms ?

Use stop list (if available) ?

Minimum number of documents:

4

Text Object - Options

Parsing

- Include parts of speech
 - specifies that terms are classified by parts of speech (for example, a noun, a verb, or an adjective). The part of speech for each term is displayed in the data tip for the term.
- Extract noun groups
 - specifies whether to identify groups of nouns as terms.
- Extract entities
 - specifies whether to identify text entities such as names, addresses, telephone numbers, and so on. If this option is disabled, then text entities are not treated differently from other text.
- Stem terms
 - specifies whether all forms of a given word are identified as a single term. For example, if you select Stem terms, then the words “sell,” “sells,” “selling,” and “sold” are identified as a single term “sell.”
 - *Anden = And!! (Danish = make no sense)*
- Use stop list (if available)
 - specifies whether to use a stop list to exclude common words such as “the,” “with,” and “is” when identifying terms. If no stop list is available, then a message appears at the bottom of the word cloud.
- Minimum number of documents
 - specifies the minimum number of documents that a term must appear in. Specify a number from 1 to 20. If a term does not appear in the minimum number of documents, then it is not included in the analysis.

General

Term parsing and role identification:

Custom

Analyze document sentiment

Parsing

Include parts of speech

Extract noun groups

Extract entities

Stem terms

Use stop list (if available)

Minimum number of documents:

4

Text Object - Options

Top Discovery

- Maximum topics
 - Specifies the maximum number of topics to create. Specify a number from 2 to 25.
- Cell weight
 - Specifies whether to weight the frequency of each term for every document that it appears in. Selecting Logarithmic de-emphasizes terms that appear many times in relatively few documents.
- Term weight
 - Specifies a weighting algorithm for the terms in the document collection. The Entropy weighting algorithm emphasizes terms that have a low frequency across the document collection.
- Number of terms to use in labels
 - Specifies the number of terms that are included in a topic name. Specify a number from 2 to 8. This option does not affect the number of terms that are used to select topics—only the topic names are changed.

General

Term parsing and role identification:
Automatic

Analyze document sentiment

Topic Discovery

Maximum topics:
6

Cell weight: ?
Logarithmic

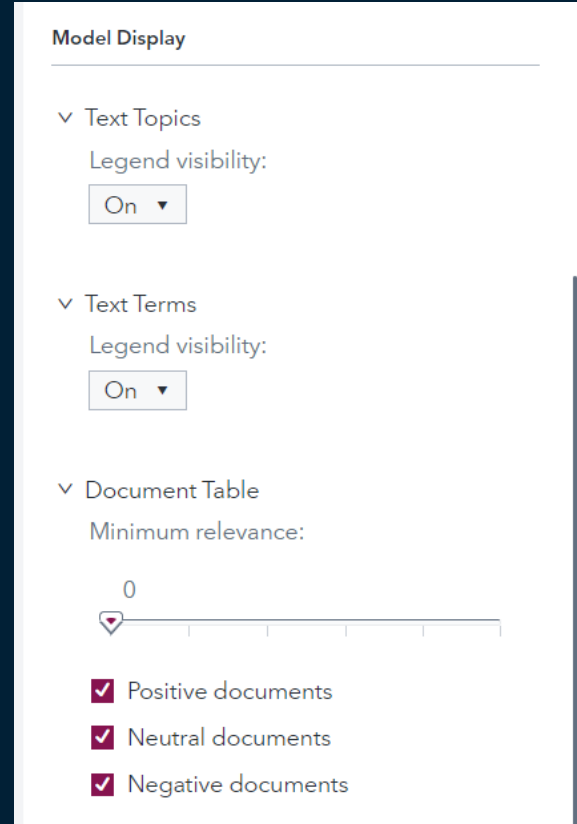
Term weight: ?
Entropy

Number of terms to use in labels:
4

Text Object - Options

Model Display

- Legend visibility
 - Specifies whether the legend is displayed for the topics bar chart and for the terms word cloud.
- Minimum relevance
 - Specifies the minimum relevance value for a document to be displayed when a topic is selected.
- Positive documents
 - Displays documents that have positive sentiment.
- Neutral documents
 - Displays documents that have neutral sentiment.
- Negative documents
 - Displays documents that have negative sentiment.



The screenshot shows the 'Model Display' configuration panel. It is divided into three sections: 'Text Topics', 'Text Terms', and 'Document Table'. Each section has a 'Legend visibility' dropdown set to 'On' and a 'Minimum relevance' slider set to 0. At the bottom, there are three checked checkboxes for 'Positive documents', 'Neutral documents', and 'Negative documents'.

Model Display

Text Topics
Legend visibility: On ▼

Text Terms
Legend visibility: On ▼

Document Table
Minimum relevance: 0

Positive documents
 Neutral documents
 Negative documents

Text Object

Definitions

Document Term matrix

- A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.
- In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

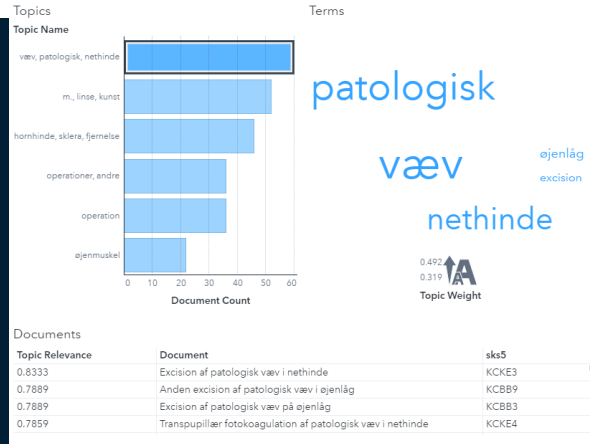
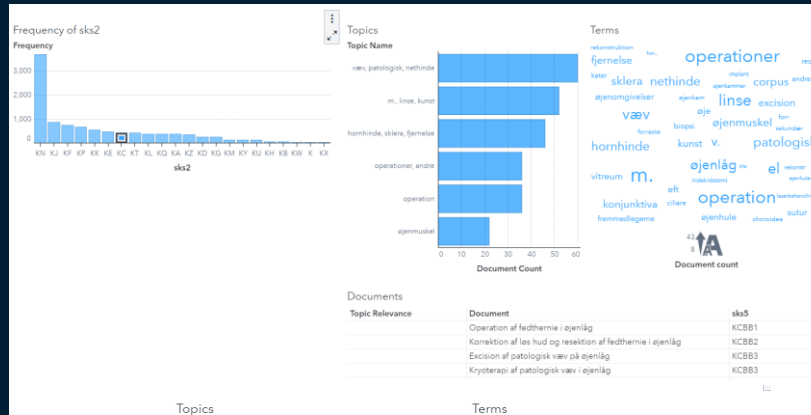
Topic Relevance

- Relevance is the concept of one topic being connected to another topic in a way that makes it useful to consider the second topic when considering the first

Text Object Definitions

Document Count: Terms with the smallest height appears in x documents and the largest X times in the document collection.

Topic Weight: Terms with the smallest height have a weight on x and the largest a weight X in relation to the selected Topic.



Demo

Conclusions

- Not a reporting tool
- But a very good adhoc analytical tool
- You lose your overview when there is many data
 - Filer!
- Easy and quick to use
 - Not very much time is lost by trying
 - But potential much time is gained!