Starting Guide iLCM

true

14 Juni, 2021

Contents

1 Before the first use

Before starting the first program, the user can decide between loading the iLCM version from a standalone or locally. The standalone version is available at https://hub.docker.com/r/ckahmann/ilcm. To use the version locally, check that docker is installed (https://docs.docker.com/install/). Having done this, the command:

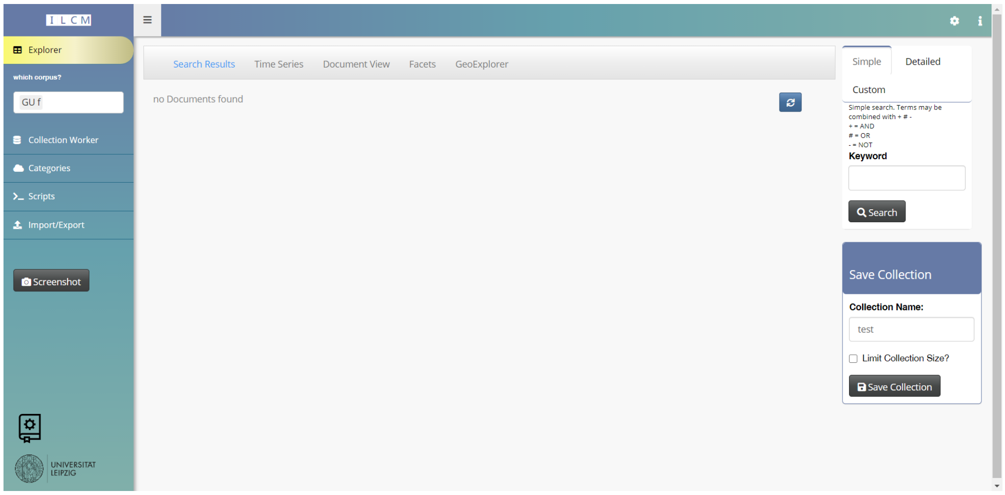
```
docker run -it -d -p 3838:3838 -p 8787:8787 -p 8983:8983 ckahmann/ilcm:latest
```

does pull the image from docker hub and starts a container hosting the needed services. The user can set port mappings to different ports on their local system. Per default, the 3 services are available on this port:

- Shiny-server: http://localhost:3838
- Rstudio-server: http://localhost:8787
- Solr-dashboard: http://localhost:8983

2 Getting started

After starting the iLCM-version the user should see the iLCM starting page:



which corpus?: GU f

no Documents found

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Collection Name: test

□ Limit Collection Size?

Figure 1: Starting page of iLCM

3 Overall structure

The iLCM consists of 5 different basic tabs, which can be switched in the sidebar. These are *Explorer*, *Collection Worker*, *Categories*, *Scripts* and the *Import/Export*.

3.1 Explorer

In the *Explorer* tab, the user can search the database, get extra information about the search results, annotate text, and create a collection of documents, which serves as the starting point for further analysis.

3.2 Collection Worker

The *Collection Worker* is the tab where the user starts various NLP-algorithms. Also, the results of the started analysis can be accessed in the *Collection Worker*.

3.3 Categories

In the *Categories* tab, the user can have a closer look at the made annotations, create new annotation schemes or create further annotations.

3.4 Scripts

Here the user is offered five functionalities. First, if the user is familiar with R, they can change the NLP-algorithms' scripts. The user should be enabled to create different versions of the script. However,keep in mind that they have to produce the same output formats as the original script did. This is the case because the result of visualization is dependent on certain variables and variable names. The following options are given by the possibility of creating or manipulating blacklists, whitelists, dictionaries or vocabulary lists. These can be used in the preprocessing chain when applying different NLP-algorithms.

3.5 Import/Export

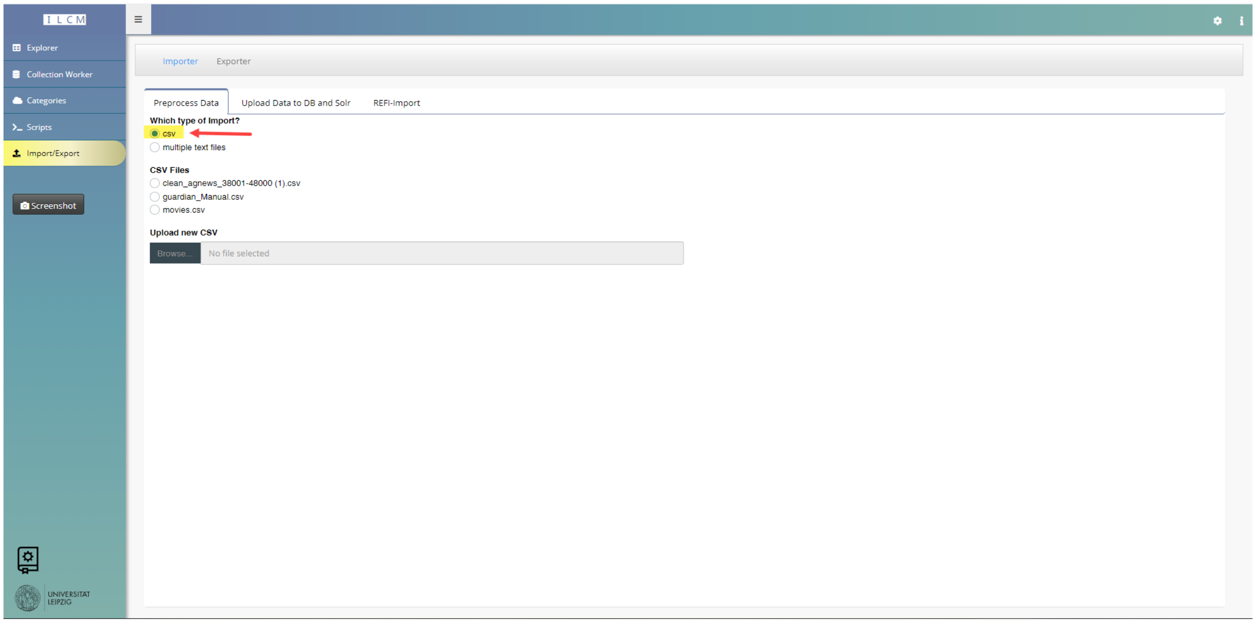
In the *Import/Export* tab, the user can upload their text data. The data will then be preprocessed and uploaded to the database and Solr. After running NLP-task or other operations, they can export collections, results, annotations or other files from the Solr database.

4 Import own Documents

The first thing the user might want to do is importing their data into the tool. For this purpose, there are different possibilities. The first option is importing their data as a csv-file, which can have any format. The second is importing multiple text files, like .pdf, .txt, .doc, .docx, where every file is imported as a single document. As another option they can also import REFI-projects or codebooks from other tools or upload the data directly to the database and Solr.

4.1 CSV-file

To import csv-files the user needs to select the csv-type in the *Importer* tab. After selecting this a *browser* bar should appear bellow. Here the user is able to search within their own directories for the csv-file they wish to upload to the tool.



Which type of Import?

● csv
○ multiple text files

CSV Files

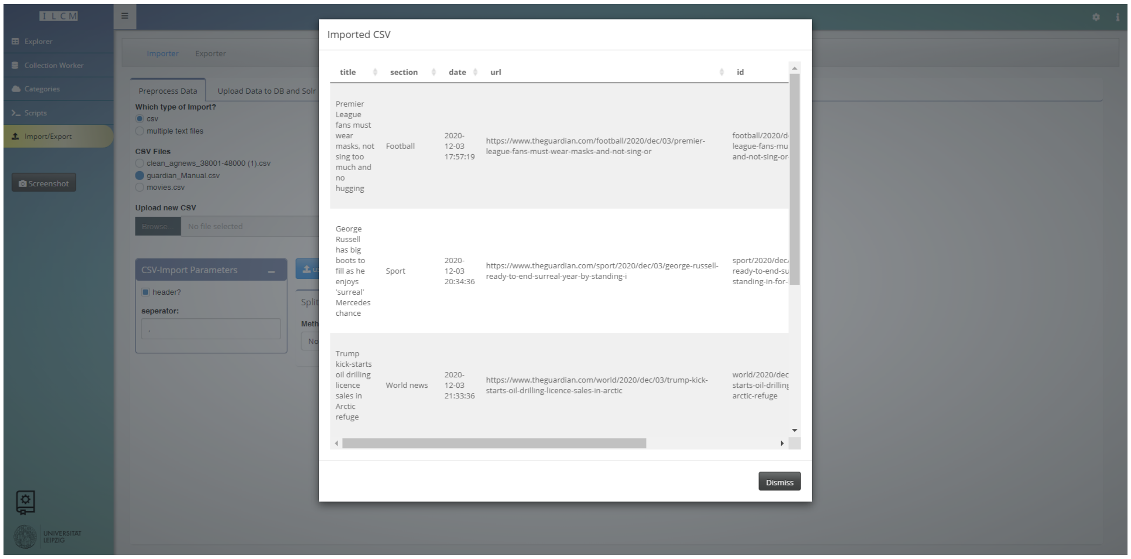
○ clean_agnews_38001-48000 (1).csv
○ guardian_Manual.csv
○ movies.csv

Upload new CSV

No file selected

Figure 2: Import .csv-files

Once the file is loaded into the tool, the user can either check whether the data is imported correctly or directly start the mapping.

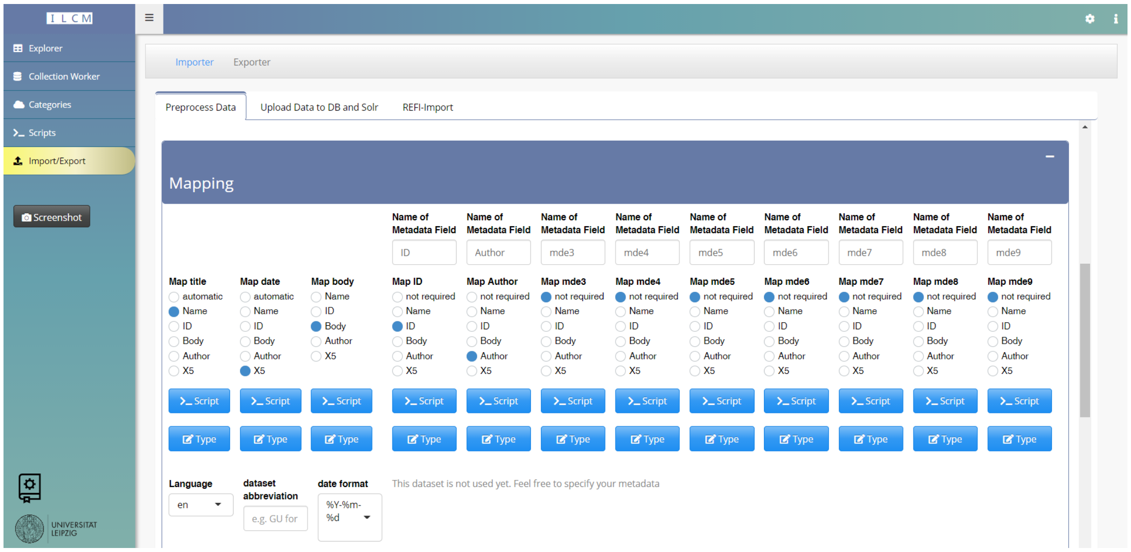


Imported CSV

title	section	date	url	id
Premier League fans must wear masks, not sing too much and no hugging	Football	2020-12-03 17:57:19	https://www.theguardian.com/football/2020/dec/03/premier-league-fans-must-wear-masks-and-not-sing-or	football/2020/d league-fans-mu and-not-sing-or
George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance	Sport	2020-12-03 20:34:36	https://www.theguardian.com/sport/2020/dec/03/george-russell-ready-to-end-surreal-year-by-standing-i	sport/2020/dec ready-to-end-su standing-in-for-
Trump kick-starts oil drilling licence sales in Arctic refuge	World news	2020-12-03 21:33:36	https://www.theguardian.com/world/2020/dec/03/trump-kick-starts-oil-drilling-licence-sales-in-arctic	world/2020/dec starts-oil-drilling arctic-refuge

Figure 3: check if .csv-file is imported correctly

In the database, the texts are saved with certain metadata-fields. Very often, they will not perfectly match the metadata the user has. Therefore the user might want to abuse some given metadata fields for their purposes (e.g. publisher, section...). After clicking *Start Mapping*, the user can see two data tables. First, they need to set the mapping. In the header, the metadata-fields of the database are given. In the rows, the user can see the fields, which were found in the imported csv-file.



Mapping

			Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field
			ID	Author	mde3	mde4	mde5	mde6	mde7	mde8	mde9
Map title	Map date	Map body	Map ID	Map Author	Map mde3	Map mde4	Map mde5	Map mde6	Map mde7	Map mde8	Map mde9
○ automatic	○ automatic	○ Name	○ not required	○ not required	● not required	● not required	● not required	● not required	● not required	● not required	● not required
● Name	○ Name	○ ID	○ Name	○ Name	○ Name	○ Name	○ Name	○ Name	○ Name	○ Name	○ Name
○ ID	○ ID	● Body	● ID	○ ID	○ ID	○ ID	○ ID	○ ID	○ ID	○ ID	○ ID
○ Body	○ Body	○ Author	○ Body	○ Body	○ Body	○ Body	○ Body	○ Body	○ Body	○ Body	○ Body
○ Author	○ Author	○ X5	○ Author	● Author	○ Author	○ Author	○ Author	○ Author	○ Author	○ Author	○ Author
○ X5	● X5		○ X5	○ X5	○ X5	○ X5	○ X5	○ X5	○ X5	○ X5	○ X5

Language: en

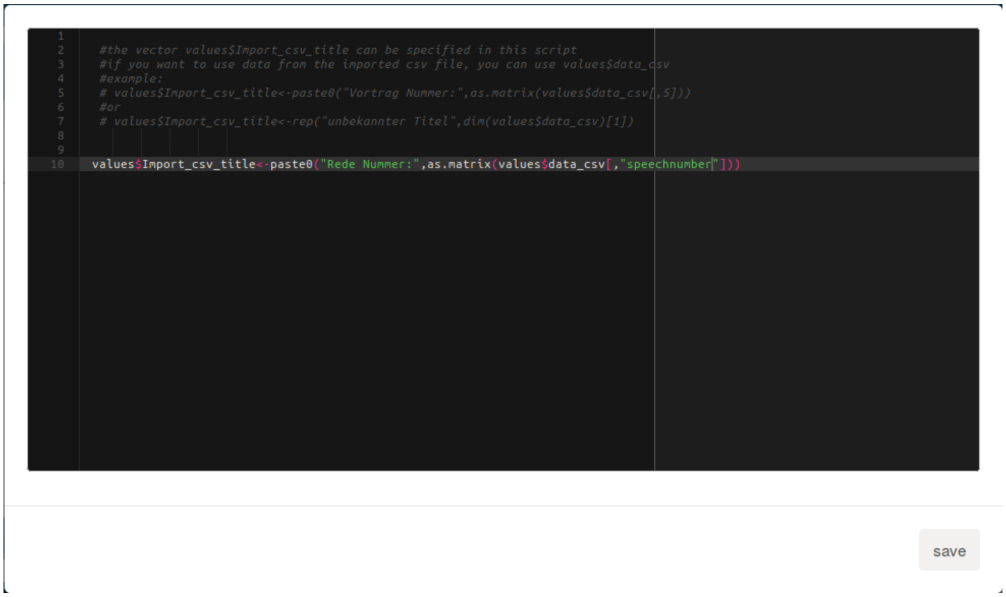
dataset abbreviation: e.g. GU for

date format: %Y-%m-%d

This dataset is not used yet. Feel free to specify your metadata

Figure 4: mapping of csv columns to iLCM database-format

Now it is the users' task to find the proper mapping. In this example, the user can see that the program can automatically guess the corresponding rows of the csv-file, but there is not guaranteed correctness. In map mde1 to mde9, the user can select how to name the columns and which of them should be included. Sometimes the user does not always have the metadata included in the csv-file. Then they have two other possibilities on how to get the data into the iLCM database: The first one is writing some R-commands in a short script for a particular database column.

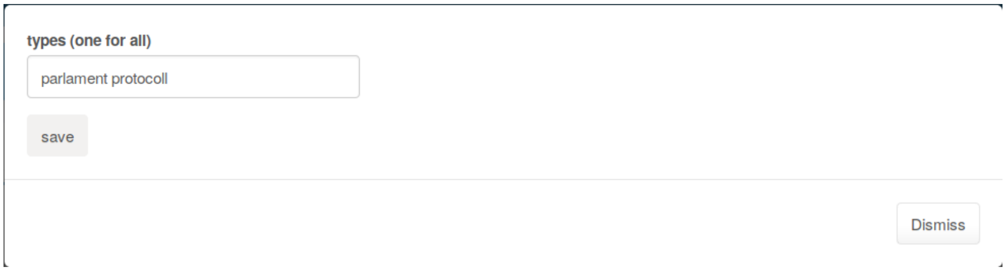


```
 #the vector values$Import_csv_title can be specified in this script
 #if you want to use data from the imported csv file, you can use values$data_csv
 #example:
 # values$Import_csv_title<-paste0("Vortrag Nummer:",as.matrix(values$data_csv[,5]))
 #or
 # values$Import_csv_title<-rep("unbekannter Titel",dim(values$data_csv)[1])


values$Import_csv_title<-paste0("Rede Nummer:",as.matrix(values$data_csv[,"speechnumber"]))
```

Figure 5: use a short R-Script to input a metadata field

Here the column “title” is filled by using the word “speech number” pasted together with the values in the column “speech number” of the csv-file. For other metadata fields - like “publisher” or “type”, it is sometimes the same value for the whole dataset.

If this is the case, the user can just click the “Type-button” to input the value by hand. This value will then be used for all documents.



types (one for all)

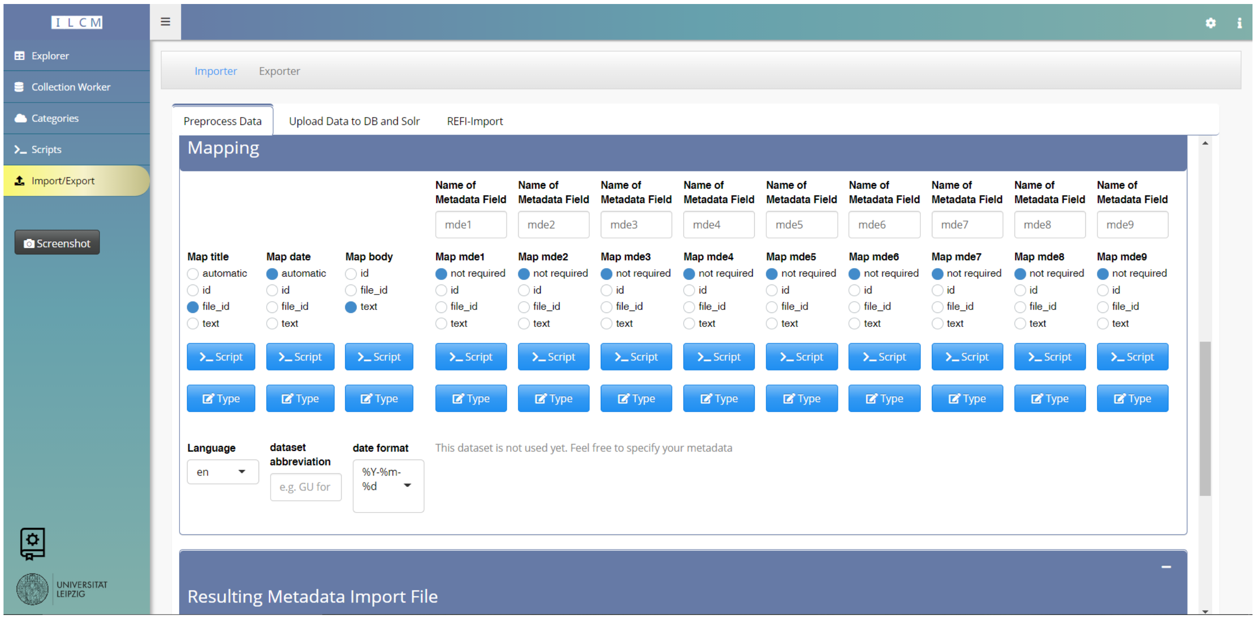
parlament protocoll

Figure 6: Type in a value for the certain database column

On the bottom, the user has to specify a language, the date-format and give an abbreviation for imported data. It is necessary that the abbreviation and the “id_doc” field together form a unique key. So when uploading multiple csv-files, which all belong to the same dataset (same abbreviation), the user needs to use unique id_doc values. In the second table below, the user sees the Metadata input file being created, dependent on the decisions he made in the data table above. Once the user is sure everything is set correctly, he has to click “Start Preprocessing and save csv-files” or “Start preprocessing and directly write to DB”. In both options, the imported data are preprocessed using Spacy and then saved as two csv-files. After clicking the second button, the data gets imported into the database and Solr using the two created csv-files (token and meta).
Once the user restarts the app (open the link in a new tab), the data will be available and ready to work.

4.2 Multiple text files

The input for multiple text files is almost the same as for csv-files. The only big difference is the fact that for the text files only, the fields: “id_doc”, “title”, and “text” exist. The other missing columns have to be filled in by hand if the user wants to include them. Otherwise, he can also leave them empty (“id_doc” and “body” have to be non-empty).



Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field	Name of Metadata Field
mde1	mde2	mde3	mde4	mde5	mde6	mde7	mde8	mde9

Map title	Map date	Map body	Map mde1	Map mde2	Map mde3	Map mde4	Map mde5	Map mde6	Map mde7	Map mde8	Map mde9
○ automatic	● automatic	○ id	● not required	● not required	● not required	● not required	● not required	● not required	● not required	● not required	● not required
○ id	○ id	○ file_id	○ id	○ id	○ id	○ id	○ id	○ id	○ id	○ id	○ id
● file_id	○ file_id	● text	○ file_id	○ file_id	○ file_id	○ file_id	○ file_id	○ file_id	○ file_id	○ file_id	○ file_id
○ text	○ text		○ text	○ text	○ text	○ text	○ text	○ text	○ text	○ text	○ text

Language: en

dataset abbreviation: e.g. GU for

date format: %Y-%m-%d

This dataset is not used yet. Feel free to specify your metadata

Figure 7: example input mapping for multiple text-files

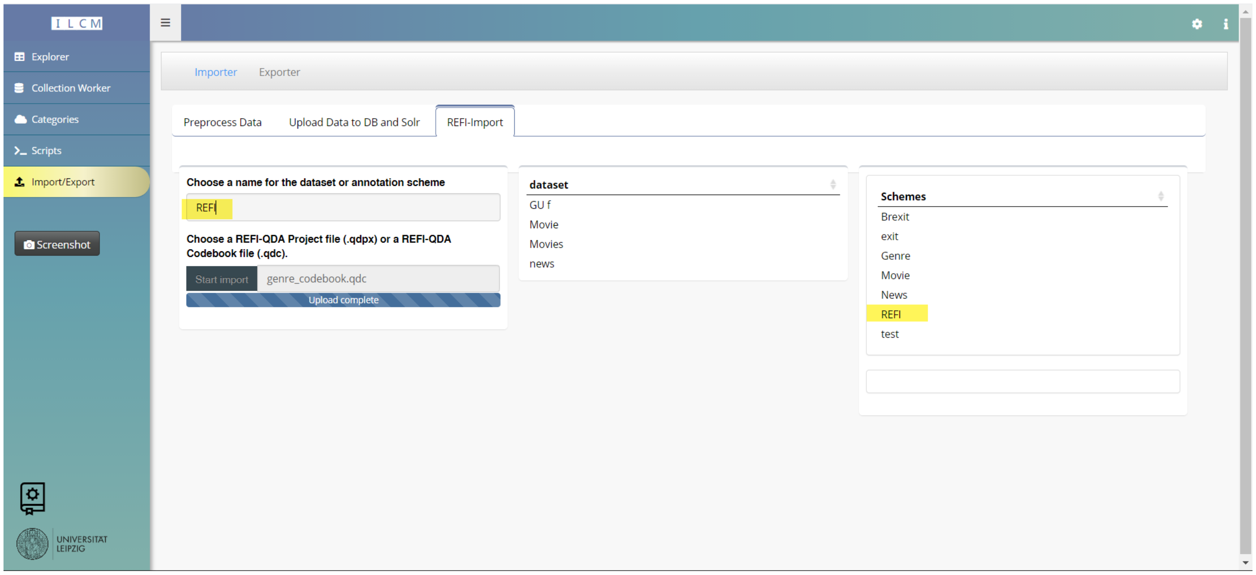
Once the import is successful, the user can choose the dataset in the sidebar-panel below the explorer-button. It is possible to work on multiple datasets at the same time.

4.3 Upload Data to DB and Solr

If the user want to save the current imported data in the database they can go to the next subtab. Here they will see all imported corpora and are able to select one they wish to save in the current Solr-pipeline and the Maria-database. Check the options-menu on the top right of the side (gear-icon) to see if the connection to the MariaDB and SOlr are correct before proceeding to upload the data. For the upload simply click on the *Upload selected data to DB and import to Solr* button. Delete the data by selecting the corpus and clicking the *Delete* button below.

4.4 REFI-import

The user may want to import projects from other software. The REFI-format is a universal format to do so. Therefore the user can click on the corresponding subtab and search for a REFI-QDA Project file (.qdpx) or a REFI-QDA Codebook file (.qdc) in their personal data. On the right side the user also will see the datasets and schemes in the tool that also can be turned into a REFI-project for download. Importing a new project or codebook will update these two lists.



Choose a name for the dataset or annotation scheme

REFI

Choose a REFI-QDA Project file (.qdpx) or a REFI-QDA Codebook file (.qdc).

genre_codebook.qdc

dataset
- GU f
- Movie
- Movies
- news

Schemes
- Brexit
- exit
- Genre
- Movie
- News
- REFI
- test

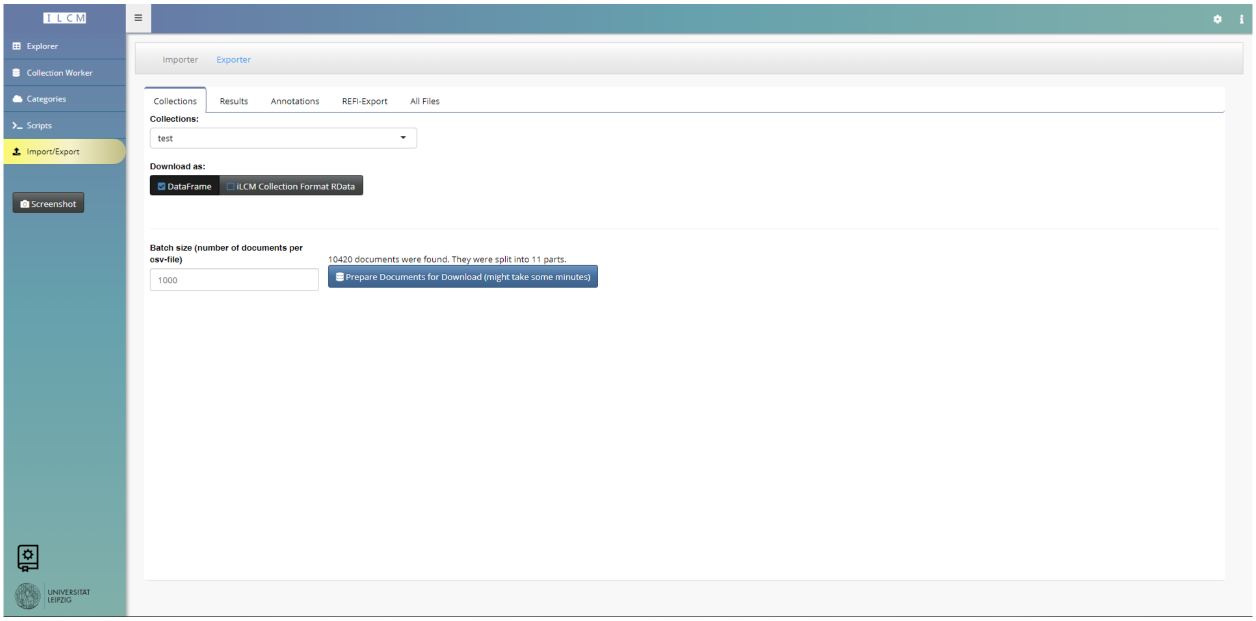
Figure 8: Importing an REFI-codebook under the name 'REFI'

5 Export

After finishing different tasks the user can decide to export the result-data. Different export formats allow the user to reuse the data in other NLP-programs or for visulization, documentation and other.

5.1 Collections

Here the user can download the created collections of documents (see section *Collections*). The user can decide whether the documents should be downloaded as Dataframes in a csv-file or as an RData-file. For the *Dataframes* selection, the user needs to select a batch size defining how many documents a csv-file should contain. Above the download button should appear a small text stating how many documents are in the collection and how many files will be generated for the selected batch size. Keep in mind that this will be a time-consuming process to generate all the files. The RData selection allows the user to download a huge file containing all the documents to the selection. This comes in handy when an experienced user wants to work with this collection in another R-program.



Collections: test

Download as: ☑ DataFrame □ iLCM Collection Format RData

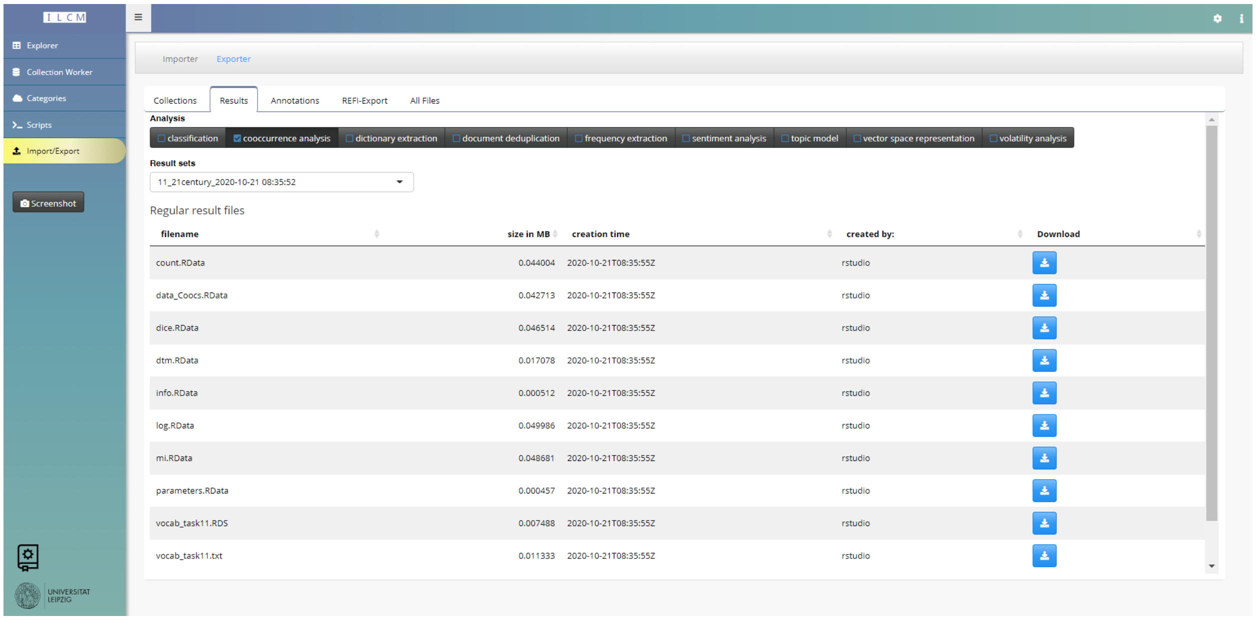
Batch size (number of documents per csv-file): 1000

10420 documents were found. They were split into 11 parts.

Figure 9: Datframe-export for a collection of 10420 documents and batch size of 1000 documents per csv-file

5.2 Results

Here the user has the option to download all the different results of NLP tasks from the tool. First, they can select the analysis task they wish to download by clicking on the corresponding field. Then they need to select a result set, where the name is generated through: *id* of the task, *name* of used collection and *time and date* when the task was started. After selecting the result set, a list with the downloadable data files should appear. Which exact files will appear depends on the selected analysis type. For example, in the *co-occurrence analysis*, there will be a file for the counts of co-occurrences, the dice parameter calculated for the co-occurrence, the log-file of the task, and so on. All these files can be downloaded by clicking on the blue download button in each file row.



Analysis

□ classification ☑ cooccurrence analysis □ dictionary extraction □ document deduplication □ frequency extraction □ sentiment analysis □ topic model □ vector space representation □ volatility analysis

Result sets

11_21century_2020-10-21 08:35:52

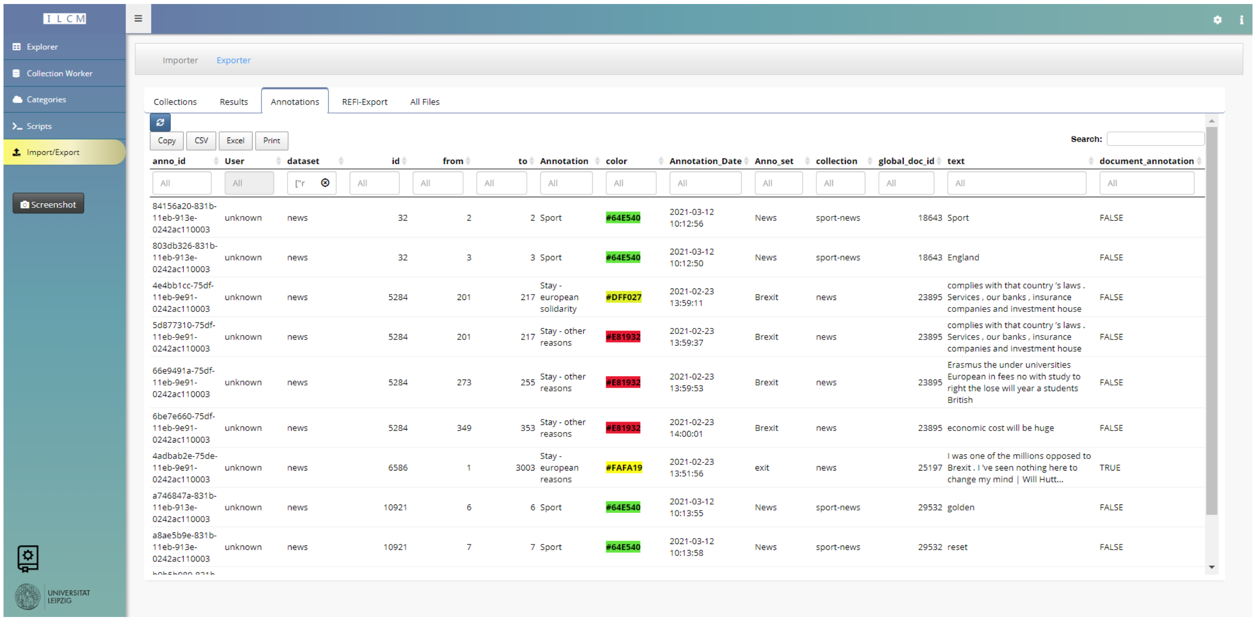
Regular result files

filename	size in MB	creation time	created by:	Download
count.RData	0.044004	2020-10-21T08:35:55Z	rstudio	
data_Coocs.RData	0.042713	2020-10-21T08:35:55Z	rstudio	
dice.RData	0.046514	2020-10-21T08:35:55Z	rstudio	
dtm.RData	0.017078	2020-10-21T08:35:55Z	rstudio	
info.RData	0.000512	2020-10-21T08:35:55Z	rstudio	
log.RData	0.049986	2020-10-21T08:35:55Z	rstudio	
mi.RData	0.048681	2020-10-21T08:35:55Z	rstudio	
parameters.RData	0.000457	2020-10-21T08:35:55Z	rstudio	
vocab_task11.RDS	0.007488	2020-10-21T08:35:55Z	rstudio	
vocab_task11.txt	0.011333	2020-10-21T08:35:55Z	rstudio	

Figure 10: Result-export for a co-occurrence analysis of a movie collection

5.3 Annotations

Made annotations can be downloaded as well. Therefore the user can download all made annotations or filter the annotations within the list by simply searching for the annotations via one row. After the annotations where found, the user can decide whether they want to download it as a csv-file, excel-file, copy the list in the computer clipboard, or even print the annotations. Printing the annotation will allow the user to save the annotations as pdf-file as well.

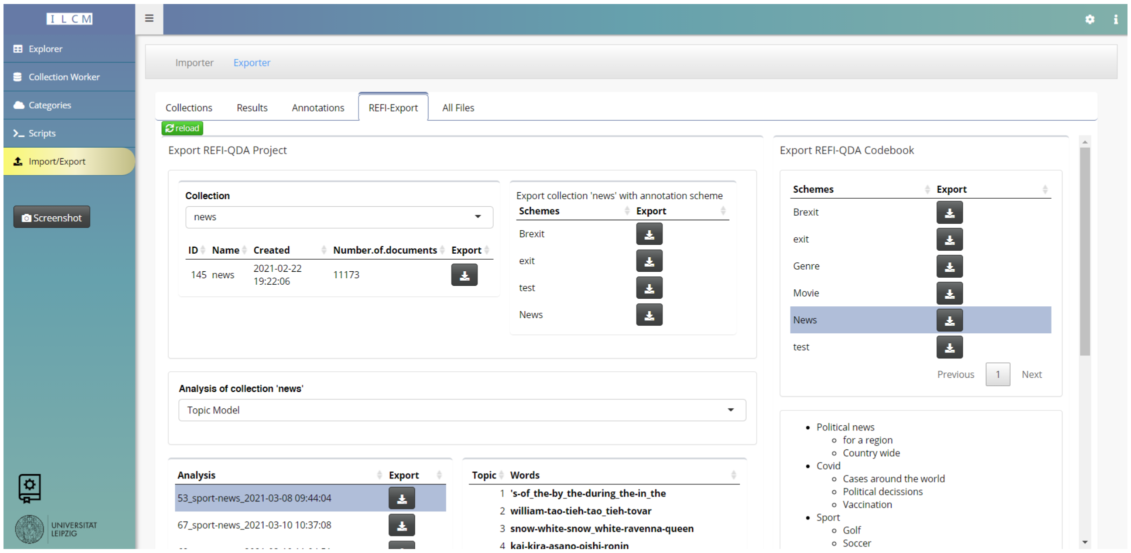


anno_id	User	dataset	id	from	to	Annotation	color	Annotation_Date	Anno_set	collection	global_doc_id	text	document_annotation
All	All	["r	All	All	All	All	All	All	All	All	All	All	All
84156a20-831b-11eb-913e-0242ac110003	unknown	news	32	2	2	Sport	#64E540	2021-03-12 10:12:56	News	sport-news	18643	Sport	FALSE
803db326-831b-11eb-913e-0242ac110003	unknown	news	32	3	3	Sport	#64E540	2021-03-12 10:12:50	News	sport-news	18643	England	FALSE
4e4bb1cc-75df-11eb-9e91-0242ac110003	unknown	news	5284	201	217	Stay - european solidarity	#DFF027	2021-02-23 13:59:11	Brexit	news	23895	complies with that country 's laws . Services , our banks , insurance companies and investment house	FALSE
5d877310-75df-11eb-9e91-0242ac110003	unknown	news	5284	201	217	Stay - other reasons	#E81932	2021-02-23 13:59:37	Brexit	news	23895	complies with that country 's laws . Services , our banks , insurance companies and investment house	FALSE
66e9491a-75df-11eb-9e91-0242ac110003	unknown	news	5284	273	255	Stay - other reasons	#E81932	2021-02-23 13:59:53	Brexit	news	23895	Erasmus the under universities European in fees no with study to right the lose will year a students British	FALSE
6be7e660-75df-11eb-9e91-0242ac110003	unknown	news	5284	349	353	Stay - other reasons	#E81932	2021-02-23 14:00:01	Brexit	news	23895	economic cost will be huge	FALSE
4adbab2e-75de-11eb-9e91-0242ac110003	unknown	news	6586	1	3003	Stay - european reasons	#FAFA19	2021-02-23 13:51:56	exit	news	25197	I was one of the millions opposed to Brexit . I 've seen nothing here to change my mind | Will Hutt...	TRUE
a746847a-831b-11eb-913e-0242ac110003	unknown	news	10921	6	6	Sport	#64E540	2021-03-12 10:13:55	News	sport-news	29532	golden	FALSE
a8ae5b9e-831b-11eb-913e-0242ac110003	unknown	news	10921	7	7	Sport	#64E540	2021-03-12 10:13:58	News	sport-news	29532	reset	FALSE

Figure 11: Annotation export for the news dataset

5.4 REFI-export

The REFI-export comes in handy when the user wish to transfer projects between software packages. Therefore the user can select a specific project and download it's documents or the made annotations. It's also possible to download a *Codebook* for the annotation projects in order to import it to another program. There is also the option to download the analysis results for the chosen collection.



Export REFI-QDA Project

Collection: news

ID	Name	Created	Number.of.documents	Export
145	news	2021-02-22 19:22:06	11173	⬇

Export collection 'news' with annotation scheme

Schemes	Export
Brexit	⬇
exit	⬇
test	⬇
News	⬇

Analysis of collection 'news': Topic Model

Analysis	Export
53_sport-news_2021-03-08 09:44:04	⬇
67_sport-news_2021-03-10 10:37:08	⬇

Topic	Words
1	's-of_the-by_the-during_the-in_the
2	william-tao-tieh-tao_tieh-tovar
3	snow-white-snow_white-ravenna-queen
4	kai-kira-asano-oishi-ronin

Export REFI-QDA Codebook

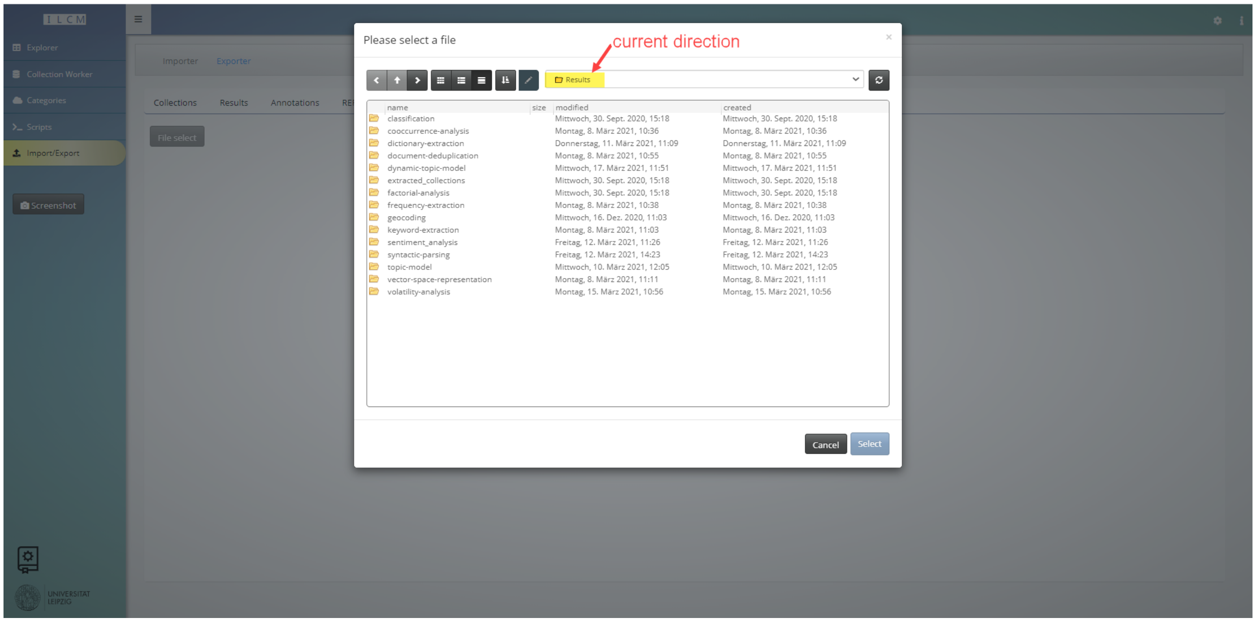
Schemes	Export
Brexit	⬇
exit	⬇
Genre	⬇
Movie	⬇
News	⬇
test	⬇

- Political news
  - for a region
  - Country wide
- Covid
  - Cases around the world
  - Political decissions
  - Vaccination
- Sport
  - Golf
  - Soccer

Figure 12: REFI-export for a news data collection

5.5 All Files

Here the user will see a *File select* button. After clicking on it, a window will open similar to an *Explorer* window on the PC. There the user can jump between different folders. The current folder direction is given on the top right of the window, below other folders will appear. In the example screenshot, the current folder is the *Results* folder containing all NLP-task results. This direction can be changed by opening the drop-down menu and selecting another folder. The folders inside the *Results* folder contain the folders of a single NLP-task of a certain kind (e.g. co-occurrence calculation). The single tasks are named with the id of the task and the name of the collection. Inside these folders, the user will find a list of files generated during the processing of the tasks. Here the user can decide which file they wish to download by clicking on the corresponding file name and the *Select* button on the bottom right of the window. If the user wants to select multiple files, they can do so by holding on to the CTRL button on their keyboard while selecting the file names with their mouse. After the file/files are selected, the small window will disappear, and on the main window, a table of all selected files is shown. Now the user needs to push the *Download* button.



name	size	modified	created
classification		Mittwoch, 30. Sept. 2020, 15:18	Mittwoch, 30. Sept. 2020, 15:18
cooccurrence-analysis		Montag, 8. März 2021, 10:36	Montag, 8. März 2021, 10:36
dictionary-extraction		Donnerstag, 11. März 2021, 11:09	Donnerstag, 11. März 2021, 11:09
document-deduplication		Montag, 8. März 2021, 10:55	Montag, 8. März 2021, 10:55
dynamic-topic-model		Mittwoch, 17. März 2021, 11:51	Mittwoch, 17. März 2021, 11:51
extracted_collections		Mittwoch, 30. Sept. 2020, 15:18	Mittwoch, 30. Sept. 2020, 15:18
factorial-analysis		Mittwoch, 30. Sept. 2020, 15:18	Mittwoch, 30. Sept. 2020, 15:18
frequency-extraction		Montag, 8. März 2021, 10:38	Montag, 8. März 2021, 10:38
geocoding		Mittwoch, 16. Dez. 2020, 11:03	Mittwoch, 16. Dez. 2020, 11:03
keyword-extraction		Montag, 8. März 2021, 11:03	Montag, 8. März 2021, 11:03
sentiment_analysis		Freitag, 12. März 2021, 11:26	Freitag, 12. März 2021, 11:26
syntactic-parsing		Freitag, 12. März 2021, 14:23	Freitag, 12. März 2021, 14:23
topic-model		Mittwoch, 10. März 2021, 12:05	Mittwoch, 10. März 2021, 12:05
vector-space-representation		Montag, 8. März 2021, 11:11	Montag, 8. März 2021, 11:11
volatility-analysis		Montag, 15. März 2021, 10:56	Montag, 15. März 2021, 10:56

Figure 13: Explorer window to download all kinds of files from the iLCM-tool

6 Select imported dataset

After the user has imported a dataset, they can switch to the *Explorer* tab. There they will find under the *Explorer* tab a search field. Here the user can chose between the loaded datasets now called ‘corpus’. After selecting one or more corpora they can start searching, building a collection and finally start the NLP-tasks.

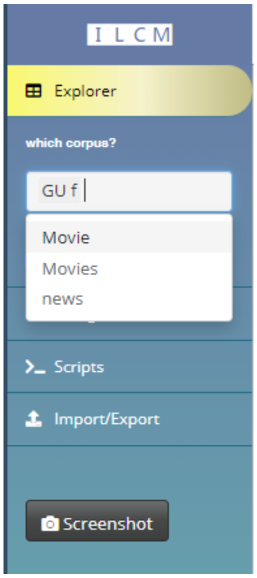
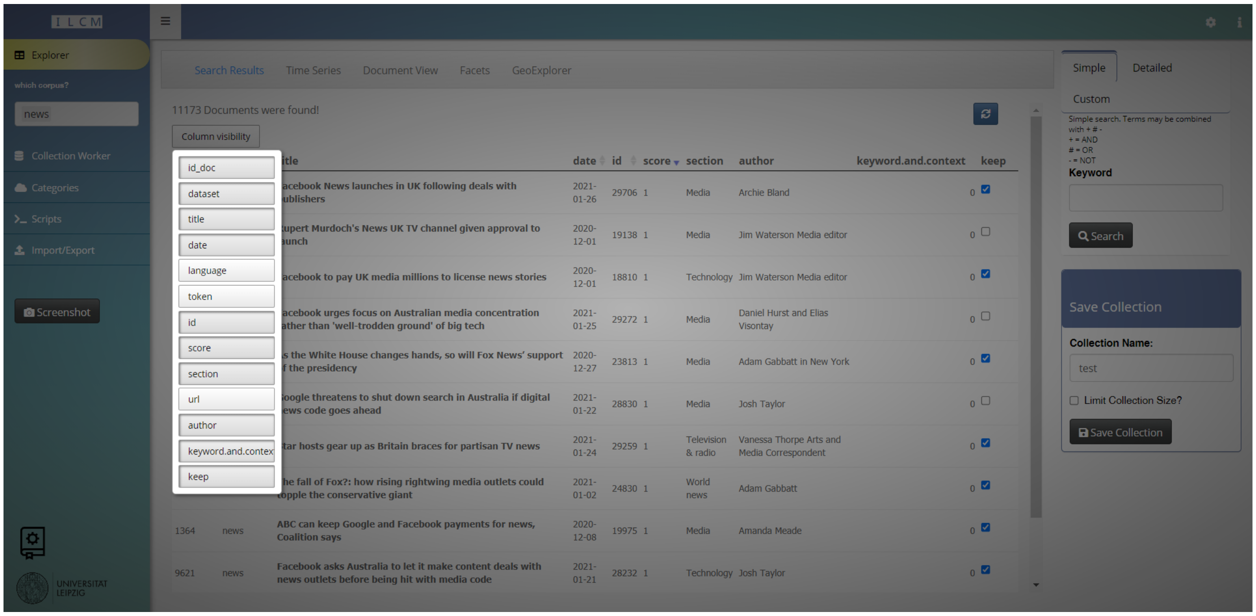


Figure 14: Selecting a dataset/ group of sets

6.1 Search Results

After the user has selected one or more corpora thy can see all documents in the *Explorer* tab under the *Search Results* subtab. They simply need to run an empty search request by clicking on the *search* button on the left site. Then a list will appear with all documents and their metadata information. On top of the page they user will also find the number of found documents. In order to clean the appearance of the list and focus on important information the user can change the column visibility. They therefore need to click on the *Columns visibility* button than a list of all column names will appear and by clicking on the column name the corresponding column will disappear in the documents list. To sort the documents the user can click on the column-names in the list and they will sort in ascending or descending order of the entries (e.g. clicking on the doc id will sort the documents from the smallest to the largest id or reversed).



which corpus? news

11173 Documents were found!

		itle	date	id	score	section	author	keyword.and.context	keep
		acebook News launches in UK following deals with ublishers	2021-01-26	29706	1	Media	Archie Bland	0	☑
		upert Murdoch's News UK TV channel given approval to aunch	2020-12-01	19138	1	Media	Jim Waterson Media editor	0	☐
		acebook to pay UK media millions to license news stories	2020-12-01	18810	1	Technology	Jim Waterson Media editor	0	☑
		acebook urges focus on Australian media concentration ather than 'well-trodden ground' of big tech	2021-01-25	29272	1	Media	Daniel Hurst and Elias Visontay	0	☐
		s the White House changes hands, so will Fox News' support f the presidency	2020-12-27	23813	1	Media	Adam Gabbatt in New York	0	☑
		oogle threatens to shut down search in Australia if digital ews code goes ahead	2021-01-22	28830	1	Media	Josh Taylor	0	☐
		tar hosts gear up as Britain braces for partisan TV news	2021-01-24	29259	1	Television & radio	Vanessa Thorpe Arts and Media Correspondent	0	☑
		he fall of Fox?: how rising rightwing media outlets could topple the conservative giant	2021-01-02	24830	1	World news	Adam Gabbatt	0	☑
1364	news	ABC can keep Google and Facebook payments for news, Coalition says	2020-12-08	19975	1	Media	Amanda Meade	0	☑
9621	news	Facebook asks Australia to let it make content deals with news outlets before being hit with media code	2021-01-21	28232	1	Technology	Josh Taylor	0	☑

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword

Collection Name: test

☐ Limit Collection Size?

Figure 15: Selecting column visibility

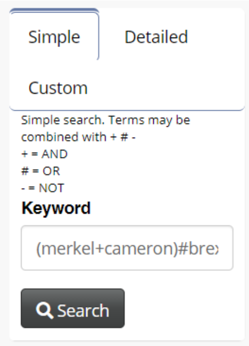
7 Search documents

One fundamental feature of the iLCM is examining vast amounts of text in a very fluent matter. The user can search for specific documents, have a closer look at some metadata for the search results and then save the documents they are interested in as a collection. When searching for documents, the user needs to make sure they are working on a suitable dataset. They can specify the dataset/datasets in the left top select-bar under the sidebar's Explorer-button. The two search options, “simple” and “detailed”, the user gets suggestions while typing, based on the texts in the database.

7.1 Simple search

In the *Simple Search*, the user can enter search terms and combine them with the logical operators: and, or, not. The usage of brackets and quotation marks (for exact matches) for more complex queries is also allowed. The logical operators can be used by typing:

1. + = AND
2. - = NOT
3. # = OR



Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

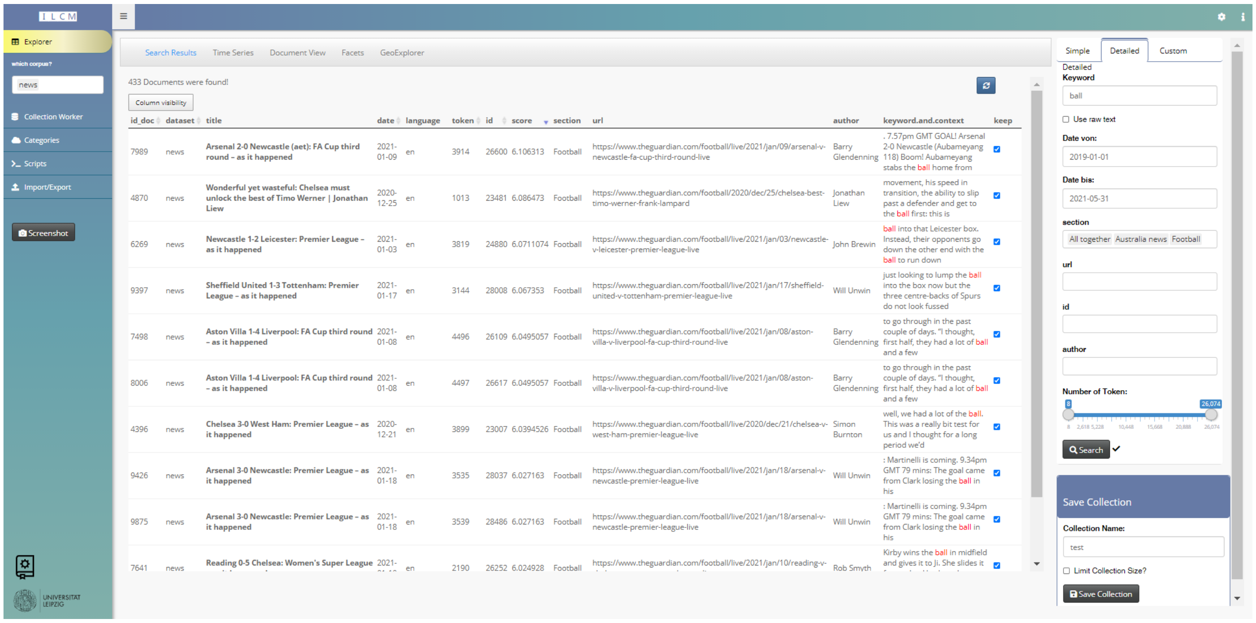
Keyword: (merkel+cameron)#brex

Figure 16: example query for simple search

The example query: *(merkel+cameron)#brexit* will return all documents in which the words *merkel* and *cameron*, or the word *brexit* is existent in the *body*-, the *title*- or in the *author* field.
The resulting documents are then listed in the *Search Results* tab. The user can see the resulting documents with metadata and some keyword and context based on the search terms. Below the search results, a slider enables the user to navigate through the other returned documents. On the complete right column of the search results table, there is a checkbox for every document, where the user can select whether he wants to use this specific document in the collection.

7.2 Detailed search

In the *Detailed Search*, the user has the same options as in the *Simple search.* On top of that, they can filter documents by their metadata. So the user can easily filter all the documents before a particular time or from a specific author or a certain section. The metadata filters are updated whenever the user changes the dataset/datasets they want to work on. The results are again displayed in the *Search Result* table.



which corpus?: news

433 Documents were found!

id_doc	dataset	title	date	language	token	id	score	section	url	author	keyword.and.context	keep
7989	news	Arsenal 2-0 Newcastle (aet): FA Cup third round – as it happened	2021-01-09	en	3914	26600	6.106313	Football	https://www.theguardian.com/football/live/2021/jan/09/arsenal-v-newcastle-fa-cup-third-round-live	Barry Glendenning	. 7.57pm GMT GOAL! Arsenal 2-0 Newcastle (Aubameyang 118) Boom! Aubameyang stabs the ball home from	☑
4870	news	Wonderful yet wasteful: Chelsea must unlock the best of Timo Werner | Jonathan Liew	2020-12-25	en	1013	23481	6.086473	Football	https://www.theguardian.com/football/2020/dec/25/chelsea-best-timo-werner-frank-lampard	Jonathan Liew	movement, his speed in transition, the ability to slip past a defender and get to the ball first: this is	☑
6269	news	Newcastle 1-2 Leicester: Premier League – as it happened	2021-01-03	en	3819	24880	6.0711074	Football	https://www.theguardian.com/football/live/2021/jan/03/newcastle-v-leicester-premier-league-live	John Brewin	ball into that Leicester box. Instead, their opponents go down the other end with the ball to run down	☑
9397	news	Sheffield United 1-3 Tottenham: Premier League – as it happened	2021-01-17	en	3144	28008	6.067353	Football	https://www.theguardian.com/football/live/2021/jan/17/sheffield-united-v-tottenham-premier-league-live	Will Unwin	just looking to lump the ball into the box now but the three centre-backs of Spurs do not look fussed	☑
7498	news	Aston Villa 1-4 Liverpool: FA Cup third round – as it happened	2021-01-08	en	4496	26109	6.0495057	Football	https://www.theguardian.com/football/live/2021/jan/08/aston-villa-v-liverpool-fa-cup-third-round-live	Barry Glendenning	to go through in the past couple of days. "I thought, first half, they had a lot of ball and a few	☑
8006	news	Aston Villa 1-4 Liverpool: FA Cup third round – as it happened	2021-01-08	en	4497	26617	6.0495057	Football	https://www.theguardian.com/football/live/2021/jan/08/aston-villa-v-liverpool-fa-cup-third-round-live	Barry Glendenning	to go through in the past couple of days. "I thought, first half, they had a lot of ball and a few	☑
4396	news	Chelsea 3-0 West Ham: Premier League – as it happened	2020-12-21	en	3899	23007	6.0394526	Football	https://www.theguardian.com/football/live/2020/dec/21/chelsea-v-west-ham-premier-league-live	Simon Burnton	well, we had a lot of the ball. This was a really bit test for us and I thought for a long period we'd	☑
9426	news	Arsenal 3-0 Newcastle: Premier League – as it happened	2021-01-18	en	3535	28037	6.027163	Football	https://www.theguardian.com/football/live/2021/jan/18/arsenal-v-newcastle-premier-league-live	Will Unwin	: Martinelli is coming. 9.34pm GMT 79 mins: The goal came from Clark losing the ball in his	☑
9875	news	Arsenal 3-0 Newcastle: Premier League – as it happened	2021-01-18	en	3539	28486	6.027163	Football	https://www.theguardian.com/football/live/2021/jan/18/arsenal-v-newcastle-premier-league-live	Will Unwin	: Martinelli is coming. 9.34pm GMT 79 mins: The goal came from Clark losing the ball in his	☑
7641	news	Reading 0-5 Chelsea: Women's Super League	2021-	en	2190	26252	6.024928	Football	https://www.theguardian.com/football/live/2021/jan/10/reading-v-	Rob Smyth	Kirby wins the ball in midfield and gives it to Ji. She slides it	☑

Detailed

Keyword: ball

☐ Use raw text

Date von: 2019-01-01

Date bis: 2021-05-31

section: All together Australia news Football

url:

id:

author:

Number of Token: 8 – 26,074 (8, 2,618, 5,228, 10,448, 15,668, 20,888, 26,074)

Search ✓

Save Collection

Collection Name: test

☐ Limit Collection Size?

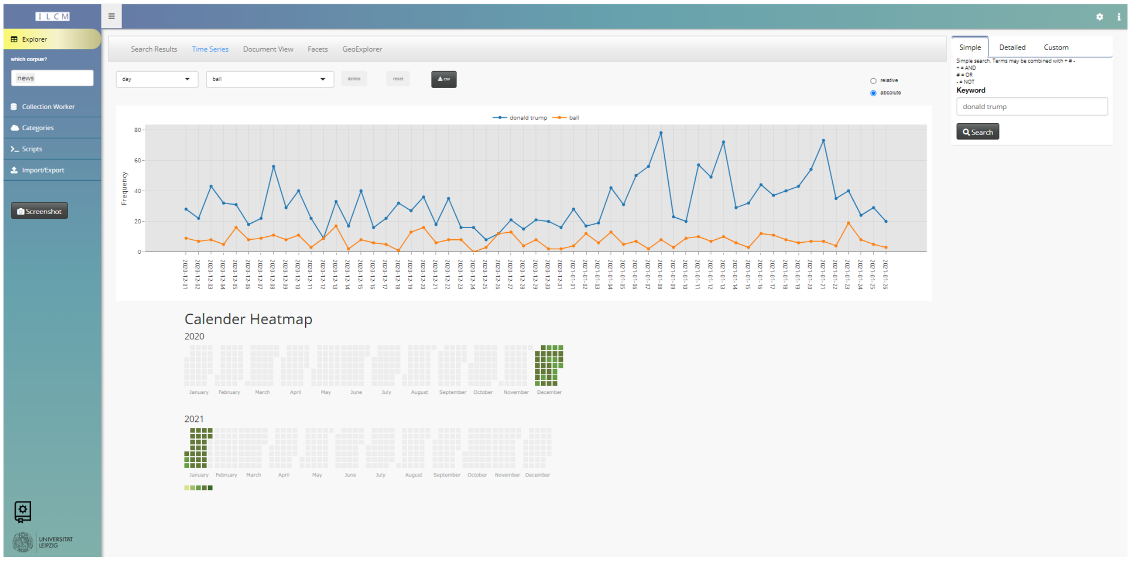
Figure 17: example query for detailed search and found results

7.3 Custom search

Solr is building the Search-backend. As a result, experienced users can use their own queries for Solr(Luscene query). In the *Custom-search* tab also the sub-setting on special datasets has to be done by hand. This is not recommended for inexperienced users, however.

7.4 Time series

Once the users have started a search, they can look at the resulting documents. Furthermore, they can check the time-series metadata for the search query. In the *Time Series* subtab, there are two plots. In the first one, the user can see the number of documents found at a specific point in time. These plots can be compared to the last ten searches; the user has made. The time series can be plotted in days, month, and years. Also, the user can switch between relative and absolute numbers. They can download the time series data by clicking the *csv*-button. When the user clicks in the plot to a specific point in time, it will trigger a new search, whereby just those documents will be returned, which are released on the data the user has clicked on. The second plot is the calendar view. Here the user can also see the number of documents that were found for a specific day. This plot can be handy when there are some patterns in the time series data for a query (e.g. Christmas).



news

day

ball

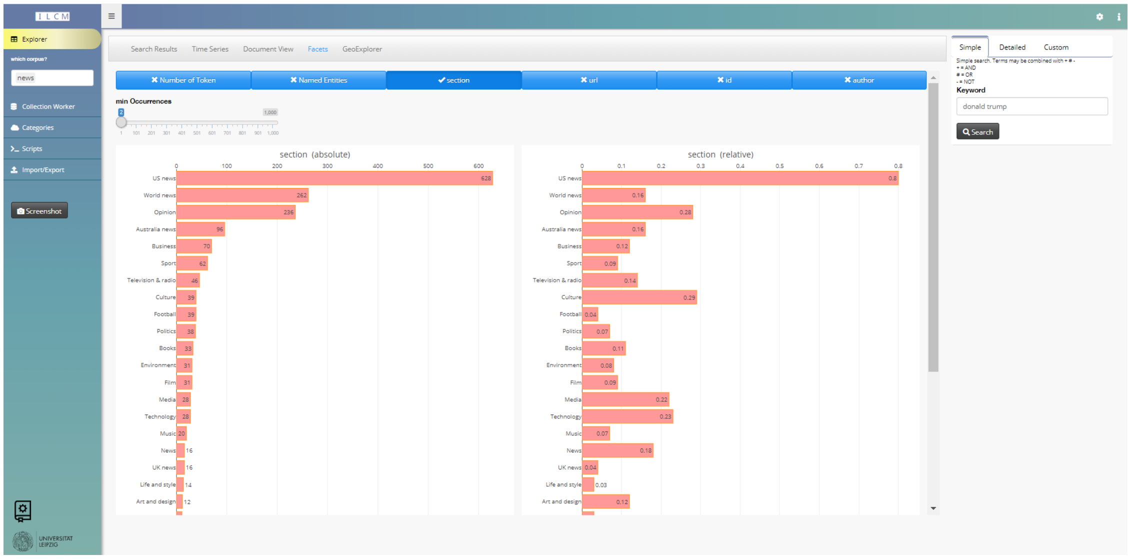
Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword: donald trump

Figure 18: time series data for 2 searches

7.5 Facets

Despite the data of a document, other metadata might be of interest. In the *Facets* subtab, the user can see the top entries for the fields: *author, section, number of tokens,* and *entities*. For the authors, the user can set a threshold for an author's minimum occurrences to be displayed. The graph on the left side shows the absolute numbers ordered beginning from the highest values. The graph on the right side shows the relative values for the entities. In more detail, the user can see the percentage of occurrences for a particular entry, which are covered by the current search.



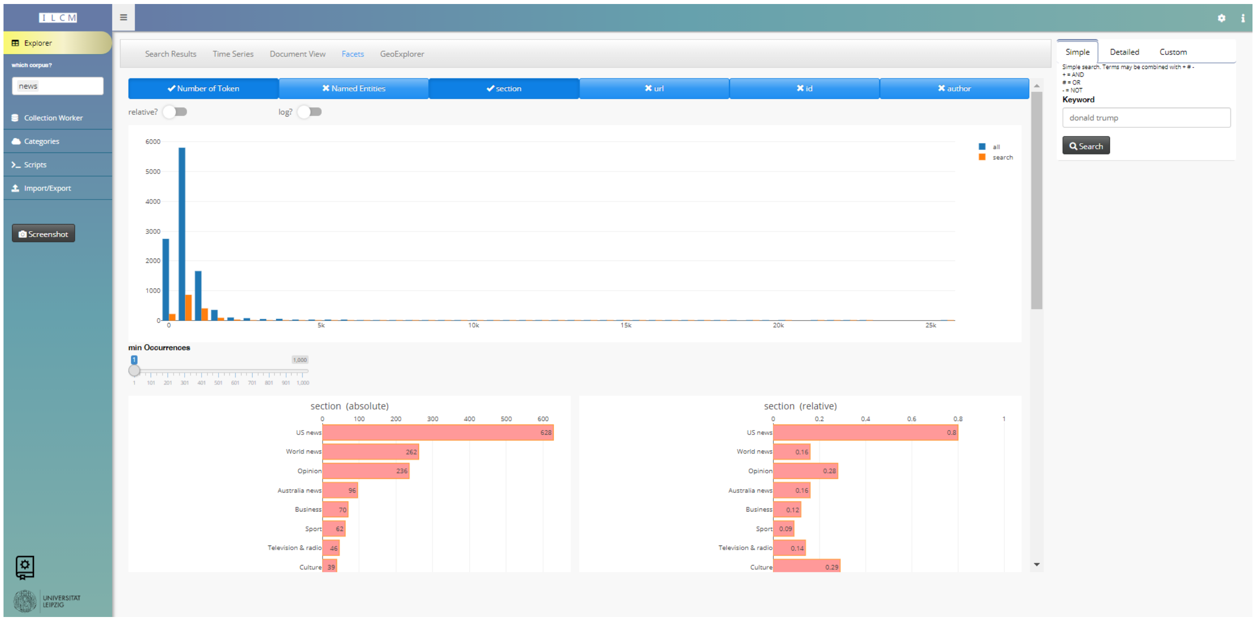
news

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword: donald trump

Figure 19: facet view for attribute 'section'

The number of documents is shown in a bar-chart, where the user can see the documents which exhibit a specific range of words for the current query compared with all documents in the database (for selected datasets). The user can toggle between relative and absolute numbers of documents. Furthermore, he can switch to the logarithmic-mode for a better overview.



which corpus? news

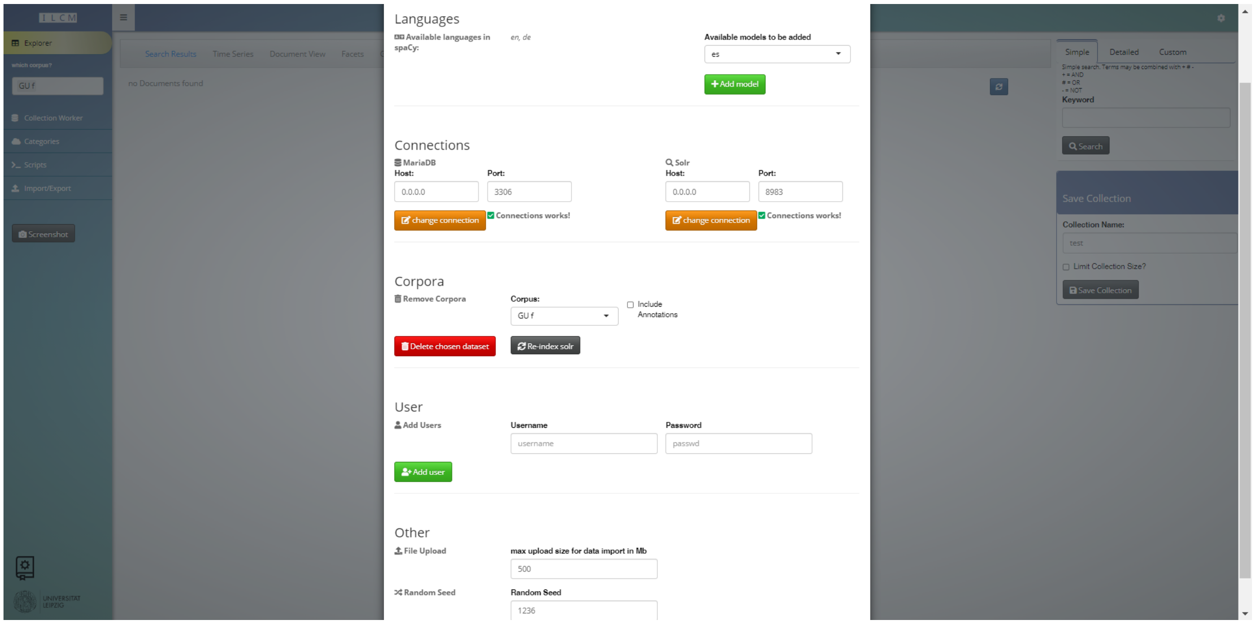
Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword
donald trump

Figure 20: facet view for attribute 'number of tokens'

8 Options

On the top right corner of the iLCM tool there is a small gear-icon. Clicking on it will open the *Option* window. Here the user can see which languages are currently loaded in the tool and which other language models can be loaded to the tool. Below that they will see the connection points for the *MariaDB* and *Solr* database which will define where to save the result files and so on. Then the user has the option to delete some of the corpora which are imported to the system with the option to as well delete all the made annotations for this corpora. On the bottom on the page the user can sign-in to the system. This comes in handy when more than one person is working in one iLCM-tool and they want to see which person started which task or what kind of annotations a specific user has made. The last point on this page is covering the import case. There the user has the option to set boarders to the import function. When the connected database is spare on space the user can set the *max upload size* down in order to prevent the system from overloading the database. And they can also set the random seed for the different NLP-tasks in here, this is useful when it comes to comparability of tasks.



Languages

Available languages in spaCy: en, de

Available models to be added: es

Connections

MariaDB
Host: 0.0.0.0
Port: 3306
☑ Connections works!

Solr
Host: 0.0.0.0
Port: 8983
☑ Connections works!

Corpora

Remove Corpora
Corpus: GU f
□ Include Annotations

User

Add Users
Username: username
Password: passwd

Other

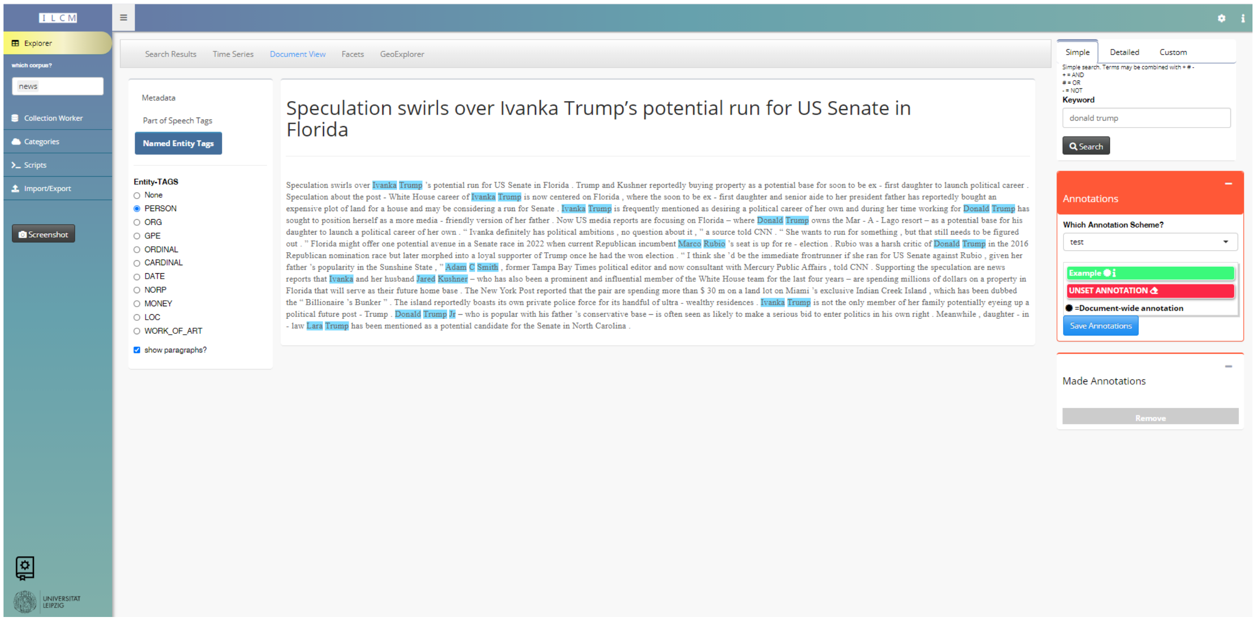
File Upload
max upload size for data import in Mb: 500

Random Seed
Random Seed: 1236

Figure 21: Tool-wide options

9 Document view

Of course, the app also allows working in a close reading approach. For this purpose, the user needs to select a document in the search-results. They can do so by clicking somewhere in the row of the document he is interested in. Next, they will automatically be forwarded to the document view. Here they can see three different boxes. In the middle, they can see the title and the whole text of the document. The metadata of the document is displayed on the top-left. Below, the metadata the user can select different POS- and NER-tags, which will be highlighted in the text. This might be helpful for some annotation tasks. The tags are calculated during the preprocessing using *Spacy*. On the top right side, the user can start a new search. If they want to go back to the results, they need to click the "Search Results" subtab in the navigation bar. Below the search-box, there is the annotation-box.



which corpus? news

Entity-TAGS
○ None
● PERSON
○ ORG
○ GPE
○ ORDINAL
○ CARDINAL
○ DATE
○ NORP
○ MONEY
○ LOC
○ WORK_OF_ART

☑ show paragraphs?

Speculation swirls over Ivanka Trump's potential run for US Senate in Florida

Speculation swirls over Ivanka Trump 's potential run for US Senate in Florida . Trump and Kushner reportedly buying property as a potential base for soon to be ex - first daughter to launch political career . Speculation about the post - White House career of Ivanka Trump is now centered on Florida , where the soon to be ex - first daughter and senior aide to her president father has reportedly bought an expensive plot of land for a house and may be considering a run for Senate . Ivanka Trump is frequently mentioned as desiring a political career of her own and during her time working for Donald Trump has sought to position herself as a more media - friendly version of her father . Now US media reports are focusing on Florida – where Donald Trump owns the Mar - A - Lago resort – as a potential base for his daughter to launch a political career of her own . “ Ivanka definitely has political ambitions , no question about it , ” a source told CNN . “ She wants to run for something , but that still needs to be figured out . ” Florida might offer one potential avenue in a Senate race in 2022 when current Republican incumbent Marco Rubio 's seat is up for re - election . Rubio was a harsh critic of Donald Trump in the 2016 Republican nomination race but later morphed into a loyal supporter of Trump once he had the won election . “ I think she 'd be the immediate frontrunner if she ran for US Senate against Rubio , given her father 's popularity in the Sunshine State , ” Adam C Smith , former Tampa Bay Times political editor and now consultant with Mercury Public Affairs , told CNN . Supporting the speculation are news reports that Ivanka and her husband Jared Kushner – who has also been a prominent and influential member of the White House team for the last four years – are spending millions of dollars on a property in Florida that will serve as their future home base . The New York Post reported that the pair are spending more than $ 30 m on a land lot on Miami 's exclusive Indian Creek Island , which has been dubbed the “ Billionaire 's Bunker ” . The island reportedly boasts its own private police force for its handful of ultra - wealthy residences . Ivanka Trump is not the only member of her family potentially eyeing up a political future post - Trump . Donald Trump Jr – who is popular with his father 's conservative base – is often seen as likely to make a serious bid to enter politics in his own right . Meanwhile , daughter - in - law Lara Trump has been mentioned as a potential candidate for the Senate in North Carolina .

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword: donald trump

Which Annotation Scheme? test

● =Document-wide annotation

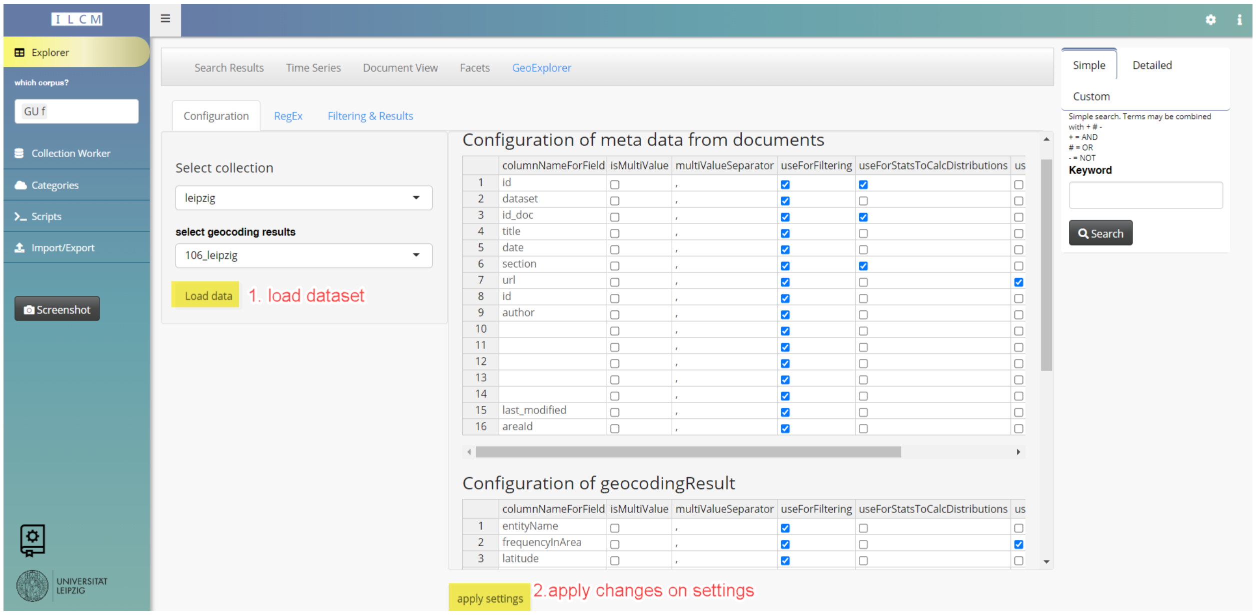
Figure 22: Document View with highlighted NER tag:'Person'

10 Geo-Explorer

To see the geographic information found in a certain collection at least one *Geocoding* task must be preceded for the particular collection. This task will search inside the documents and their meta data for geographic locations (like: 'Leipzig, Germany') and convert them into geographic coordinates. This allows to present the documents as points on a map corresponding which locations were found in the documents. This process is quite time consuming so its recommended to use a small collection for the *Geocoding* task as the process will search for all locations and will present them on a live map.

10.1 Configuration

First the user will set the configuration information right. Therefore they chose a collection and the corresponding *Geocoding* results they wish to view. After hitting the *Load data* button a couple of tables will appear on the right. These are used as parameters for filtering the results, or as stats for the calculation of numeric information/ distributions. The first table will allow the user to chose from the metadata information. Simply tick the corresponding fields in the row of a information from interest. If there are more than one value for the piece of metadata information the user needs to tick the *isMultiValue* column and select the used separator for them. The table on the bottom will allow the user to configure which *Geocoding* information should be used for the final visualization of the results. After setting all parameters up don't forget to hit the *apply settings* button on the bottom of the page.



Select collection: leipzig

select geocoding results: 106_leipzig

1. load dataset

Configuration of meta data from documents

	columnNameForField	isMultiValue	multiValueSeparator	useForFiltering	useForStatsToCalcDistributions	us
1	id	□	,	☑	☑	□
2	dataset	□	,	☑	□	□
3	id_doc	□	,	☑	☑	□
4	title	□	,	☑	□	□
5	date	□	,	☑	□	□
6	section	□	,	☑	☑	□
7	url	□	,	☑	□	☑
8	id	□	,	☑	□	□
9	author	□	,	☑	□	□
10		□	,	☑	□	□
11		□	,	☑	□	□
12		□	,	☑	□	□
13		□	,	☑	□	□
14		□	,	☑	□	□
15	last_modified	□	,	☑	□	□
16	areaId	□	,	☑	□	□

Configuration of geocodingResult

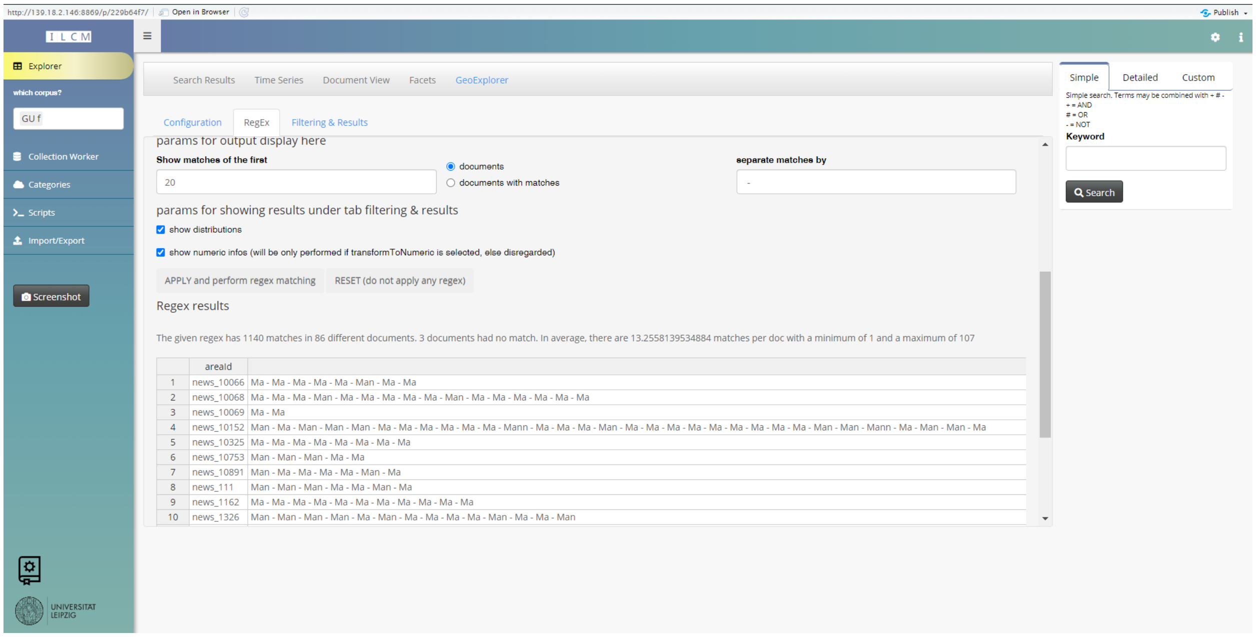
	columnNameForField	isMultiValue	multiValueSeparator	useForFiltering	useForStatsToCalcDistributions	us
1	entityName	□	,	☑	□	□
2	frequencyInArea	□	,	☑	□	☑
3	latitude	□	,	☑	□	□

2.apply changes on settings

Figure 23: Setup the configuration for the Geocoding results

10.2 RegEx

The next subsection will allow the user to set regular expressions that should be matched to the *GeoCoding* results. The regular can help to find specific documents indicating a certain region. For example: If the user is looking for texts about *Great Britain*. Therefore documents containing words like *British, Brits, English, England* maybe indicators that the text is about *Great Britain*. So the regular expression indicating all of these words will look something like _Brit*_ and _Engl*_. To find more information on how to set up regular expression check out information on regular expressions. After choosing the regular expressions for the matching the user also needs to set the separator marks used for indicating decimal places and the separation of thousands digits. At lead the user can select parameters for the result presentation. After all parameters are set the user can click on *APPLY and perform Regex matching*. Then a counter will appear below informing about the number of found matches as well as a list with the results. If the settings should be canceled just press the *RESET* button.



which corpus?: GU f

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT

Keyword

params for output display here

Show matches of the first: 20

● documents
○ documents with matches

separate matches by: -

params for showing results under tab filtering & results

☑ show distributions

☑ show numeric infos (will be only performed if transformToNumeric is selected, else disregarded)

APPLY and perform regex matching | RESET (do not apply any regex)

Regex results

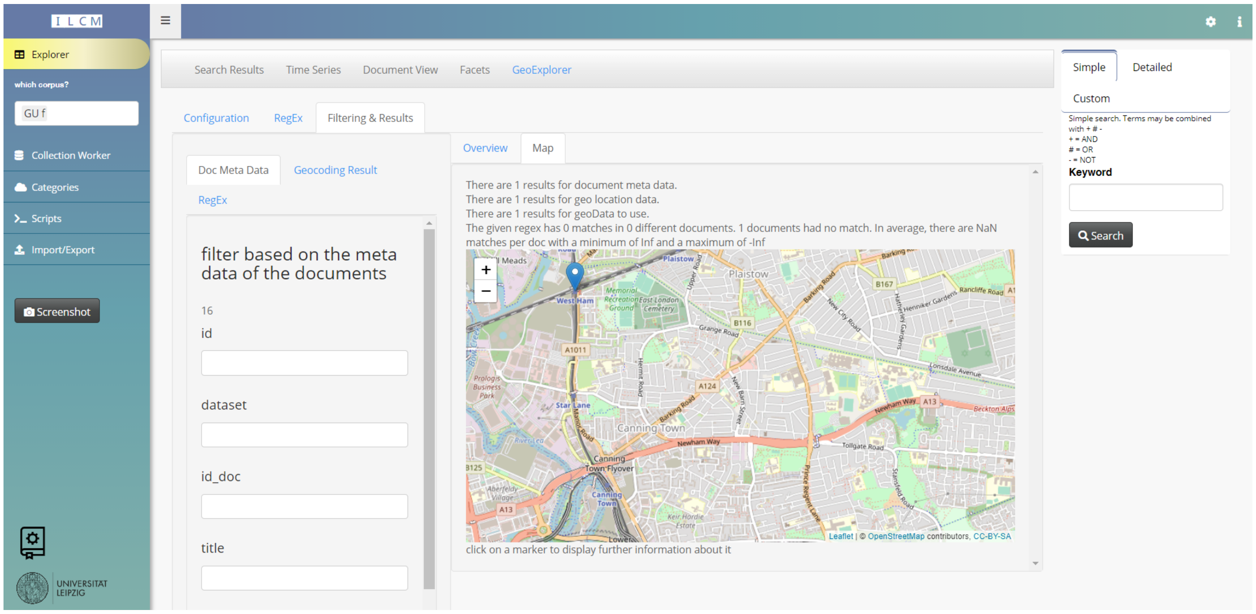
The given regex has 1140 matches in 86 different documents. 3 documents had no match. In average, there are 13.2558139534884 matches per doc with a minimum of 1 and a maximum of 107

	areaId	
1	news_10066	Ma - Ma - Ma - Ma - Ma - Man - Ma - Ma
2	news_10068	Ma - Ma - Ma - Man - Ma - Ma - Ma - Ma - Ma - Man - Ma - Ma - Ma - Ma - Ma - Ma
3	news_10069	Ma - Ma
4	news_10152	Man - Ma - Man - Man - Man - Ma - Ma - Ma - Ma - Ma - Ma - Mann - Ma - Ma - Ma - Man - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Man - Man - Mann - Ma - Man - Man - Ma
5	news_10325	Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma
6	news_10753	Man - Man - Man - Ma - Ma
7	news_10891	Man - Ma - Ma - Ma - Ma - Man - Ma
8	news_111	Man - Man - Man - Ma - Ma - Man - Ma
9	news_1162	Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma - Ma
10	news_1326	Man - Man - Man - Man - Ma - Man - Ma - Ma - Ma - Ma - Man - Ma - Ma - Man

Figure 24: Found matches for the Regular expression 'Man*' in a sport themed collection

10.3 Filtering&Results

In this subtab all the premade settings from the *Configuration* subtabs will be visualized. This page consists of two parts: One to set filtering parameters for either metadata information, regular expression or the direct Geocoding results and one side showing the results of the filtering as an overview or on a map. In the overview part the user will first see a summary text of what was found considering the set filters. First the user will see the statistics for the metadata followed by the statistics of the Geocoding results and the ones for the regular expression task. The stats will consist of a diagramm and a table with further information. The map will show which locations were found by placing marks. Here the user can zoom in and out to see a more detailed view or a more general overview - the navigation works like every other map-tool.



filter based on the meta data of the documents

16

id

dataset

id_doc

title

There are 1 results for document meta data.
There are 1 results for geo location data.
There are 1 results for geoData to use.
The given regex has 0 matches in 0 different documents. 1 documents had no match. In average, there are NaN matches per doc with a minimum of Inf and a maximum of -Inf

click on a marker to display further information about it

Simple search. Terms may be combined with + # -
+ = AND
= OR
- = NOT
Keyword

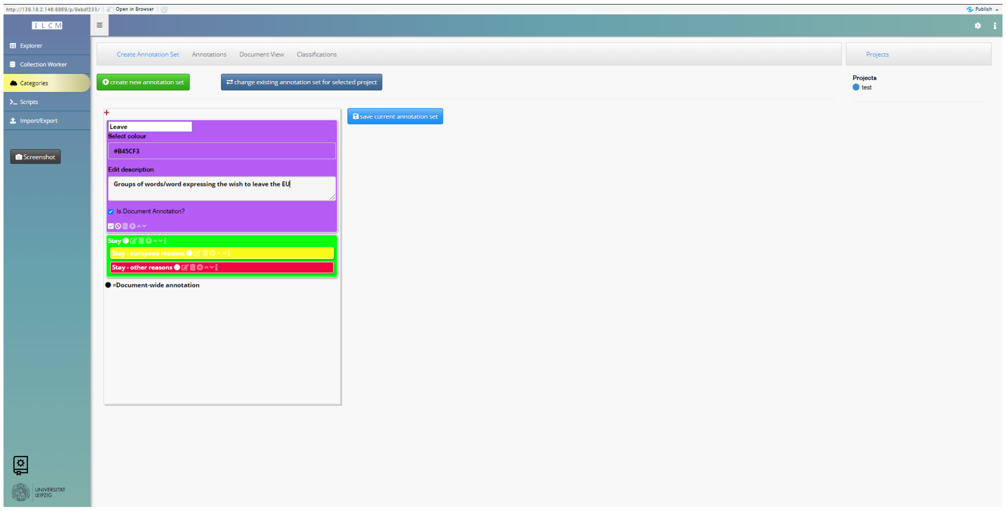
Figure 25: Map with locations of documents (closeup view)

11 Annotate documents

Using the iLCM allows the user to annotate their texts with predefined annotations schemes. These annotations might work as the basis for a classification-task.

11.1 Create annotation schemes

In the *Categories* tabs, the user finds the subtab *Create Annotation Set*. Here they can create new or manipulate existing annotation schemes. The annotation scheme can be hierarchical. An annotation class needs a name, a colour, a description, and it needs to be specified whether the annotation should target the whole document or just a specific part of the text.



Leave

Select colour

#B45CF3

Edit description

Groups of words/word expressing the wish to leave the EU

☑ Is Document Annotation?

Stay

Stay - european reasons

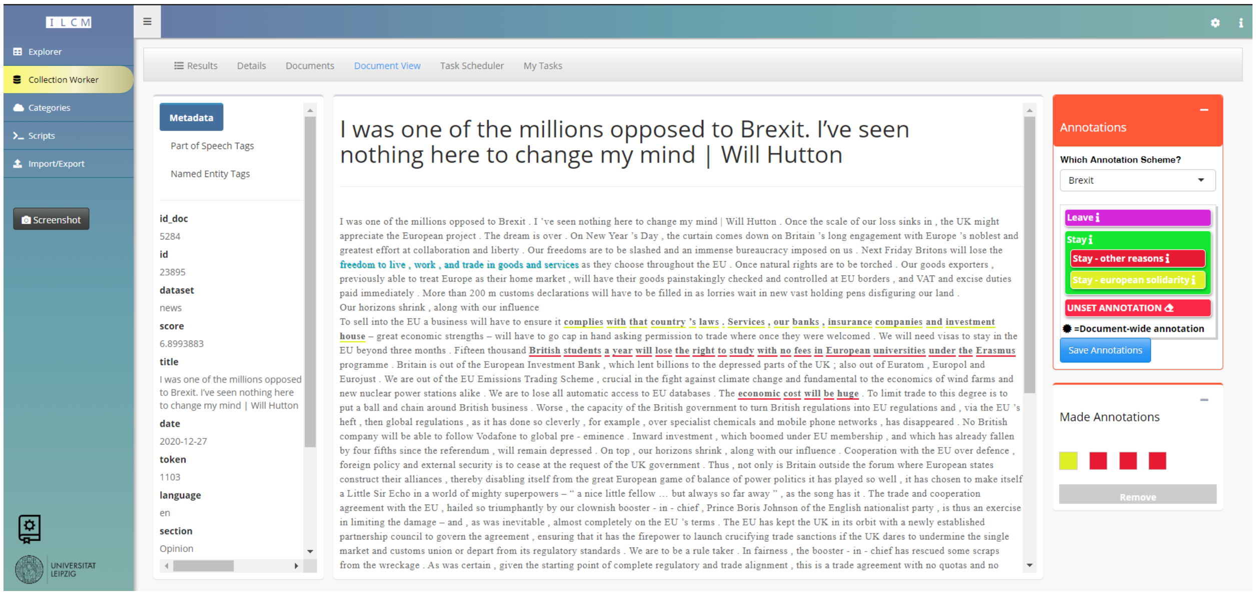
Stay - other reasons

● =Document-wide annotation

Figure 26: Subtab for annotation scheme creation

11.2 Annotate

Once there is an annotation scheme, the user can select it in the annotation box in the *Document View*. Afterwards, the annotation scheme will be displayed, as seen below. For annotation, the user needs to select the part of the text they want to annotate. The text will be highlighted. A click on the annotation class they want to use will then apply to the annotation; once the user is finished with annotating the document, they need to click the save annotation button. This will save the made annotations again. If the user is interested in seeing all made annotations or deleting several annotations, they can switch the *Annotations* subtab inside the tab *Categories*. Clicking on an annotation will then open the corresponding document. The user can also continue to annotate in this area.



id_doc
5284
id
23895
dataset
news
score
6.8993883
title
I was one of the millions opposed to Brexit. I've seen nothing here to change my mind | Will Hutton
date
2020-12-27
token
1103
language
en
section
Opinion

I was one of the millions opposed to Brexit. I've seen nothing here to change my mind | Will Hutton

I was one of the millions opposed to Brexit . I 've seen nothing here to change my mind | Will Hutton . Once the scale of our loss sinks in , the UK might appreciate the European project . The dream is over . On New Year 's Day , the curtain comes down on Britain 's long engagement with Europe 's noblest and greatest effort at collaboration and liberty . Our freedoms are to be slashed and an immense bureaucracy imposed on us . Next Friday Britons will lose the **freedom to live , work , and trade in goods and services** as they choose throughout the EU . Once natural rights are to be torched . Our goods exporters , previously able to treat Europe as their home market , will have their goods painstakingly checked and controlled at EU borders , and VAT and excise duties paid immediately . More than 200 m customs declarations will have to be filled in as lorries wait in new vast holding pens disfiguring our land .
Our horizons shrink , along with our influence
To sell into the EU a business will have to ensure it **complies with that country 's laws . Services , our banks , insurance companies and investment house** – great economic strengths – will have to go cap in hand asking permission to trade where once they were welcomed . We will need visas to stay in the EU beyond three months . Fifteen thousand **British students a year will lose the right to study with no fees in European universities under the Erasmus** programme . Britain is out of the European Investment Bank , which lent billions to the depressed parts of the UK ; also out of Euratom , Europol and Eurojust . We are out of the EU Emissions Trading Scheme , crucial in the fight against climate change and fundamental to the economics of wind farms and new nuclear power stations alike . We are to lose all automatic access to EU databases . The **economic cost will be huge** . To limit trade to this degree is to put a ball and chain around British business . Worse , the capacity of the British government to turn British regulations into EU regulations and , via the EU 's heft , then global regulations , as it has done so cleverly , for example , over specialist chemicals and mobile phone networks , has disappeared . No British company will be able to follow Vodafone to global pre - eminence . Inward investment , which boomed under EU membership , and which has already fallen by four fifths since the referendum , will remain depressed . On top , our horizons shrink , along with our influence . Cooperation with the EU over defence , foreign policy and external security is to cease at the request of the UK government . Thus , not only is Britain outside the forum where European states construct their alliances , thereby disabling itself from the great European game of balance of power politics it has played so well , it has chosen to make itself a Little Sir Echo in a world of mighty superpowers – " a nice little fellow … but always so far away " , as the song has it . The trade and cooperation agreement with the EU , hailed so triumphantly by our clownish booster - in - chief , Prince Boris Johnson of the English nationalist party , is thus an exercise in limiting the damage – and , as was inevitable , almost completely on the EU 's terms . The EU has kept the UK in its orbit with a newly established partnership council to govern the agreement , ensuring that it has the firepower to launch crucifying trade sanctions if the UK dares to undermine the single market and customs union or depart from its regulatory standards . We are to be a rule taker . In fairness , the booster - in - chief has rescued some scraps from the wreckage . As was certain , given the starting point of complete regulatory and trade alignment , this is a trade agreement with no quotas and no

Which Annotation Scheme?

Brexit

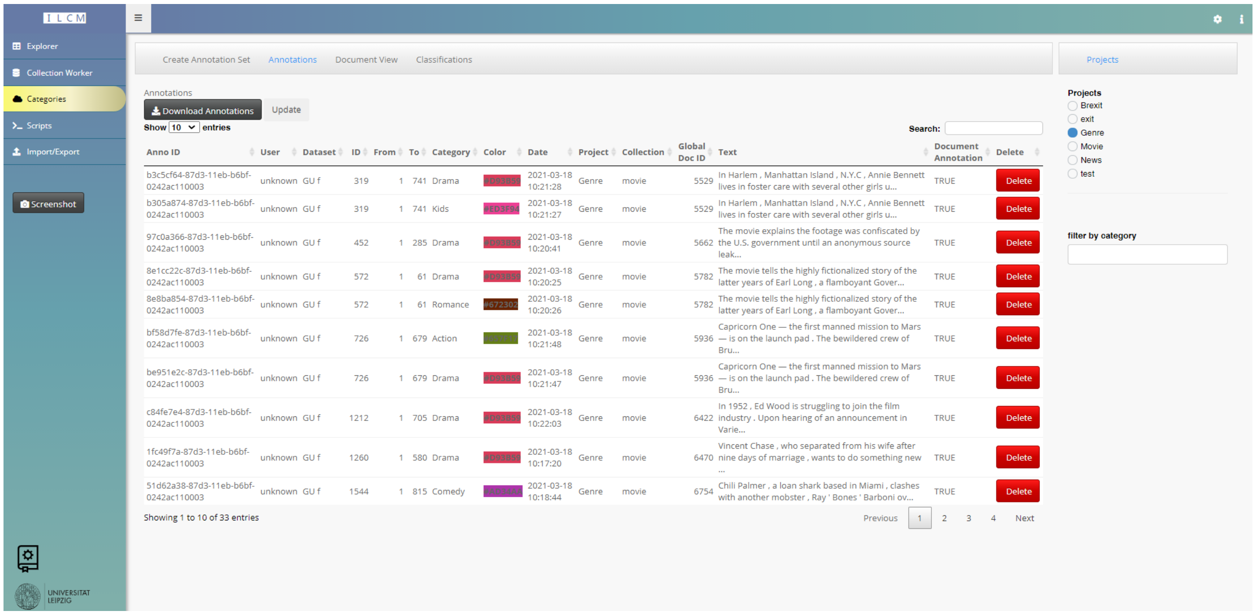
● =Document-wide annotation

🟨 🟥 🟥 🟥

Figure 27: Example for an annotation pro

11.3 Overview of annotations

After the user has made a couple of different annotation schemes and annotated some documents, they may want to get an overview of the annotations. Therefore the user can got to the *Categories* tab and select the *Annotations* subtab. Here they can between the ‘projects’ witch are build on the different annotation schemes. After selecting a project a list will appear depending of all made annotations. On top the user has the option to delete made annotations and search within the annotations or the categories.



Annotations

Show 10 entries

Search:

Anno ID	User	Dataset	ID	From	To	Category	Color	Date	Project	Collection	Global Doc ID	Text	Document Annotation	Delete
b3c5cf64-87d3-11eb-b6bf-0242ac110003	unknown	GU f	319	1	741	Drama	#D93B59	2021-03-18 10:21:28	Genre	movie	5529	In Harlem , Manhattan Island , N.Y.C , Annie Bennett lives in foster care with several other girls u...	TRUE	Delete
b305a874-87d3-11eb-b6bf-0242ac110003	unknown	GU f	319	1	741	Kids	#ED3F94	2021-03-18 10:21:27	Genre	movie	5529	In Harlem , Manhattan Island , N.Y.C , Annie Bennett lives in foster care with several other girls u...	TRUE	Delete
97c0a366-87d3-11eb-b6bf-0242ac110003	unknown	GU f	452	1	285	Drama	#D93B59	2021-03-18 10:20:41	Genre	movie	5662	The movie explains the footage was confiscated by the U.S. government until an anonymous source leak...	TRUE	Delete
8e1cc22c-87d3-11eb-b6bf-0242ac110003	unknown	GU f	572	1	61	Drama	#D93B59	2021-03-18 10:20:25	Genre	movie	5782	The movie tells the highly fictionalized story of the latter years of Earl Long , a flamboyant Gover...	TRUE	Delete
8e8ba854-87d3-11eb-b6bf-0242ac110003	unknown	GU f	572	1	61	Romance	#672302	2021-03-18 10:20:26	Genre	movie	5782	The movie tells the highly fictionalized story of the latter years of Earl Long , a flamboyant Gover...	TRUE	Delete
bf58d7fe-87d3-11eb-b6bf-0242ac110003	unknown	GU f	726	1	679	Action	#657F17	2021-03-18 10:21:48	Genre	movie	5936	Capricorn One — the first manned mission to Mars — is on the launch pad . The bewildered crew of Bru...	TRUE	Delete
be951e2c-87d3-11eb-b6bf-0242ac110003	unknown	GU f	726	1	679	Drama	#D93B59	2021-03-18 10:21:47	Genre	movie	5936	Capricorn One — the first manned mission to Mars — is on the launch pad . The bewildered crew of Bru...	TRUE	Delete
c84fe7e4-87d3-11eb-b6bf-0242ac110003	unknown	GU f	1212	1	705	Drama	#D93B59	2021-03-18 10:22:03	Genre	movie	6422	In 1952 , Ed Wood is struggling to join the film industry . Upon hearing of an announcement in Varie...	TRUE	Delete
1fc49f7a-87d3-11eb-b6bf-0242ac110003	unknown	GU f	1260	1	580	Drama	#D93B59	2021-03-18 10:17:20	Genre	movie	6470	Vincent Chase , who separated from his wife after nine days of marriage , wants to do something new ...	TRUE	Delete
51d62a38-87d3-11eb-b6bf-0242ac110003	unknown	GU f	1544	1	815	Comedy	#AD34AA	2021-03-18 10:18:44	Genre	movie	6754	Chili Palmer , a loan shark based in Miami , clashes with another mobster , Ray ' Bones ' Barboni ov...	TRUE	Delete

Showing 1 to 10 of 33 entries

Projects
○ Brexit
○ exit
● Genre
○ Movie
○ News
○ test

filter by category

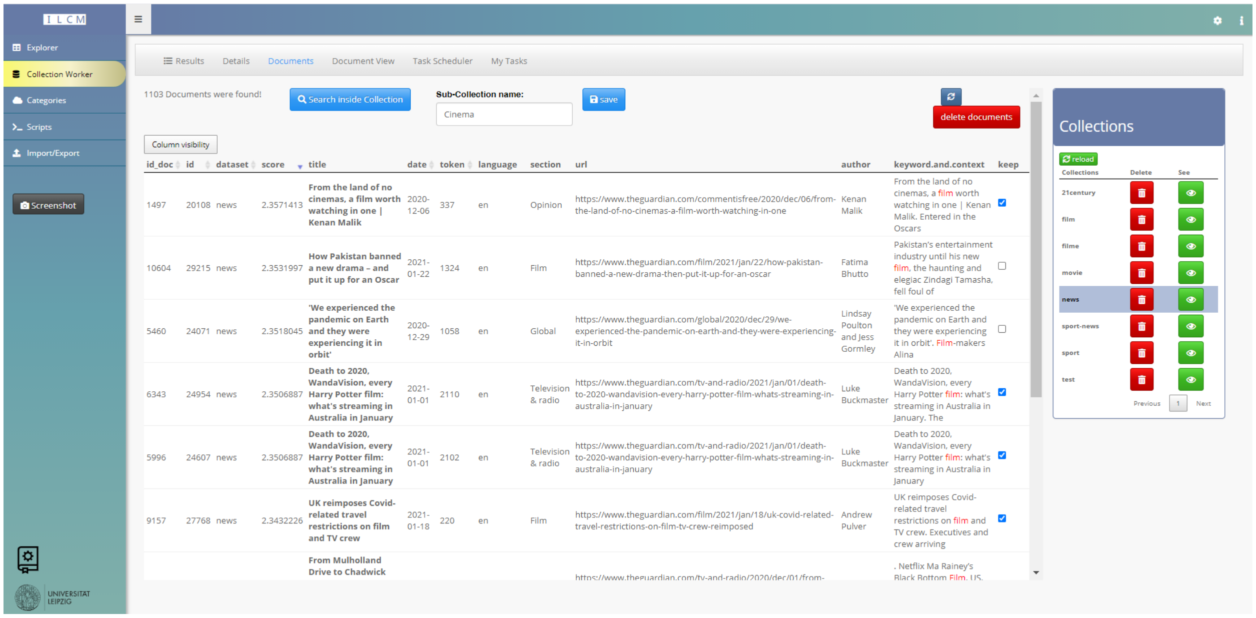
Figure 28: List of all annotations from the project ’Genre’

12 Create collections

Having found the correct documents the user is interested in, they can save these documents in a collection. This is feasible by filling a name in the *save collection* box in the *Search-Results* subtab. This will save all documents in the current search result as a collection. Documents that should be excluded (“keep”-checkbox unchecked) will not be saved. Another option to limit the number of documents in the collection is to tick the *Limit collection size* in the *save collection* window on the bottom right. This will allow the user to select a limitaion method and a limitation number. To reconstruct the process of analysis, the steps of finding the documents and excluding some are saved for every collection in the *Create-collection* statement (hover collection name in subtab *Results* in *Collection Worker* tab). Each collection needs a specific name and should not include the sign: ”_”.

12.1 Create sub-collections

After the first collection was saved, the user can switch to the *Collection worker* tab. Here, they can select a collection by clicking on the collection name in the column on the right side. Further, the user can go to the *Documents* subtab to view the documents of the collection. Now they have the chance to search inside this specific collection and create a sub-collection. After finishing the search task a list of documents will appear, which contain the searched phrases/word. The last column in the list will allow the user to decide whether they want to keep a certain document in the sub-collection or exclude it. Then the user needs to enter a name for the new collection and click on the save-icon. The new sub-collection will then appear in the *Collections* bar on the right side. Note that the made annotation will not be transferred to the new sub-collection.



1103 Documents were found!

Sub-Collection name: Cinema

id_doc	id	dataset	score	title	date	token	language	section	url	author	keyword.and.context	keep
1497	20108	news	2.3571413	From the land of no cinemas, a film worth watching in one | Kenan Malik	2020-12-06	337	en	Opinion	https://www.theguardian.com/commentisfree/2020/dec/06/from-the-land-of-no-cinemas-a-film-worth-watching-in-one	Kenan Malik	From the land of no cinemas, a film worth watching in one | Kenan Malik. Entered in the Oscars	☑
10604	29215	news	2.3531997	How Pakistan banned a new drama – and put it up for an Oscar	2021-01-22	1324	en	Film	https://www.theguardian.com/film/2021/jan/22/how-pakistan-banned-a-new-drama-then-put-it-up-for-an-oscar	Fatima Bhutto	Pakistan's entertainment industry until his new film, the haunting and elegiac Zindagi Tamasha, fell foul of	☐
5460	24071	news	2.3518045	'We experienced the pandemic on Earth and they were experiencing it in orbit'	2020-12-29	1058	en	Global	https://www.theguardian.com/global/2020/dec/29/we-experienced-the-pandemic-on-earth-and-they-were-experiencing-it-in-orbit	Lindsay Poulton and Jess Gormley	'We experienced the pandemic on Earth and they were experiencing it in orbit'. Film-makers Alina	☐
6343	24954	news	2.3506887	Death to 2020, WandaVision, every Harry Potter film: what's streaming in Australia in January	2021-01-01	2110	en	Television & radio	https://www.theguardian.com/tv-and-radio/2021/jan/01/death-to-2020-wandavision-every-harry-potter-film-whats-streaming-in-australia-in-january	Luke Buckmaster	Death to 2020, WandaVision, every Harry Potter film: what's streaming in Australia in January. The	☑
5996	24607	news	2.3506887	Death to 2020, WandaVision, every Harry Potter film: what's streaming in Australia in January	2021-01-01	2102	en	Television & radio	https://www.theguardian.com/tv-and-radio/2021/jan/01/death-to-2020-wandavision-every-harry-potter-film-whats-streaming-in-australia-in-january	Luke Buckmaster	Death to 2020, WandaVision, every Harry Potter film: what's streaming in Australia in January	☑
9157	27768	news	2.3432226	UK reimposes Covid-related travel restrictions on film and TV crew	2021-01-18	220	en	Film	https://www.theguardian.com/film/2021/jan/18/uk-covid-related-travel-restrictions-on-film-tv-crew-reimposed	Andrew Pulver	UK reimposes Covid-related travel restrictions on film and TV crew. Executives and crew arriving	☑
				From Mulholland Drive to Chadwick					https://www.theguardian.com/tv-and-radio/2020/dec/01/from-		. Netflix Ma Rainey's Black Bottom Film. US.	

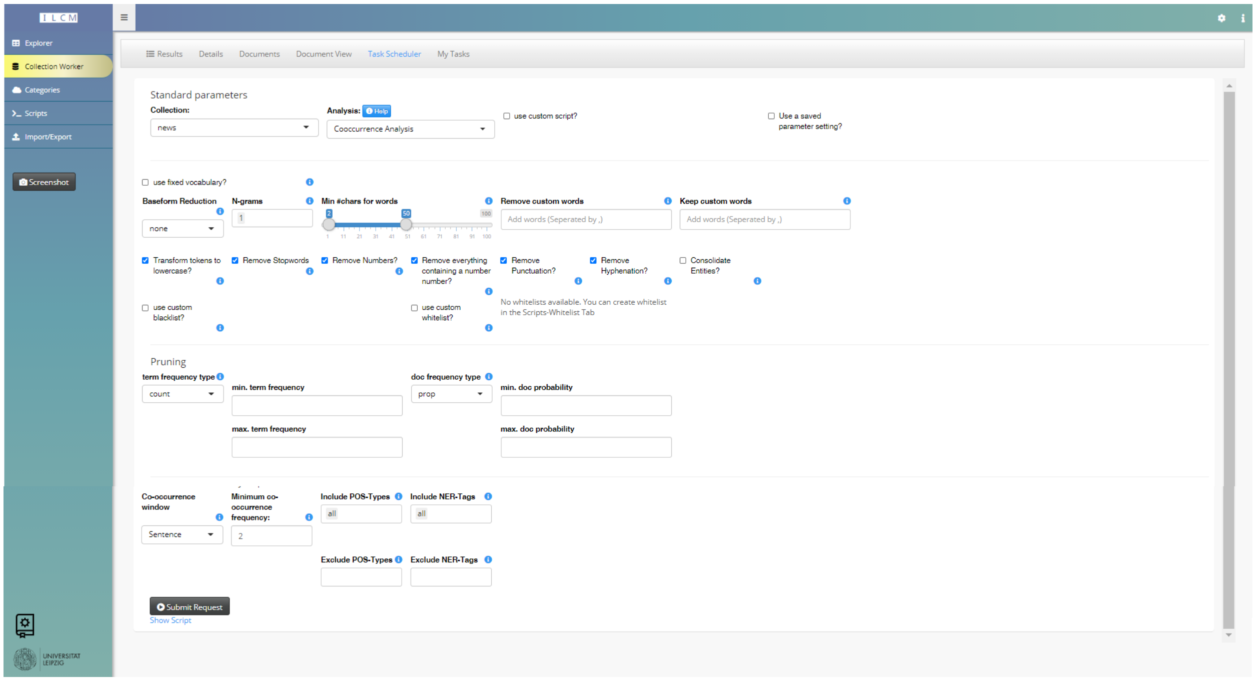
Collections	Delete	See
21century	🗑	👁
film	🗑	👁
filme	🗑	👁
movie	🗑	👁
news	🗑	👁
sport-news	🗑	👁
sport	🗑	👁
test	🗑	👁

Previous 1 Next

Figure 29: Subcollection of news articles after searching for the term ’film’

13 Start an NLP-task

Besides the search-tasks, the iLCM can also apply some of the latest NLP-algorithms to the saved collections. In the *Collection Worker* tab, there is a subtab called *Task Scheduler*. Initially, a collection and an NLP-task needs to be specified. The following tasks are implemented:



Standard parameters

Collection: news

Analysis: Cooccurrence Analysis

☐ use custom script?

☐ Use a saved parameter setting?

☐ use fixed vocabulary?

Baseform Reduction: none

N-grams: 1

Min #chars for words: 2 – 50 (1 … 100)

Remove custom words: Add words (Seperated by ,)

Keep custom words: Add words (Seperated by ,)

☑ Transform tokens to lowercase?

☑ Remove Stopwords

☑ Remove Numbers?

☑ Remove everything containing a number number?

☑ Remove Punctuation?

☑ Remove Hyphenation?

☐ Consolidate Entities?

☐ use custom blacklist?

☐ use custom whitelist?

No whitelists available. You can create whitelist in the Scripts-Whitelist Tab

Pruning

term frequency type: count

min. term frequency:

max. term frequency:

doc frequency type: prop

min. doc probability:

max. doc probability:

Co-occurrence window: Sentence

Minimum co-occurrence frequency: 2

Include POS-Types: all

Include NER-Tags: all

Exclude POS-Types:

Exclude NER-Tags:

Figure 30: Task Scheduler subtab

- 1. **Co-occurrence Analysis** aims to find terms that appear significantly often together. These words can then be displayed in a network. E.g. a network of people appearing together in texts can be created.

- 2. **Frequency Extraction** counts the number of several dictionaries occurrences at specific points in time. A dictionary is composed of several words.

- 3. **Dictionary Extraction** counts the occurrence of several dictionaries at specific points in time. A dictionary is composed of several words.

- 4. **Volatility Analysis** (Context Volatility) measures the amount of context change for a specific word for a certain period.

- 5. **Topic Modeling** - A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.

- 6. **Classification** of the documents of the chosen collection depending on certain classification parameters, such as a given annotation set.

- 7. **Geocoding** find a potential location in the documents identifying where a particular document was written.

- 8. **Sentiment Analysis** finds word/phrases that lead to the reader to a certain emotion and calculates which emotion dominates a document.

- 9. **Document Deduplication** check for duplicates between the documents of a selected collection.

- 10. **Dynamic Topic Model** - dynamic modelling for "topics" that occur in a collection of documents.

- 11. **Keyword Extraction** - extract words that appear significant often in the documents of a specific collection.
- 12. **Syntactic Parsing** finds relationships between words in a sentence.
- 13. **Vector Space Representation** of the words/phrases of a collection for further work with word vectors.
- 14. Download Collection as CSV.
- 15. Save collection as a token-object.
- 16. Save collection as meta-object.

13.1 Standard parameter

The NLP-algorithms may need preprocessing before working adequately. This might be due to reasons of calculation time or achieving better results. There are some standard parameters the user can set for most algorithm. These parameters are:

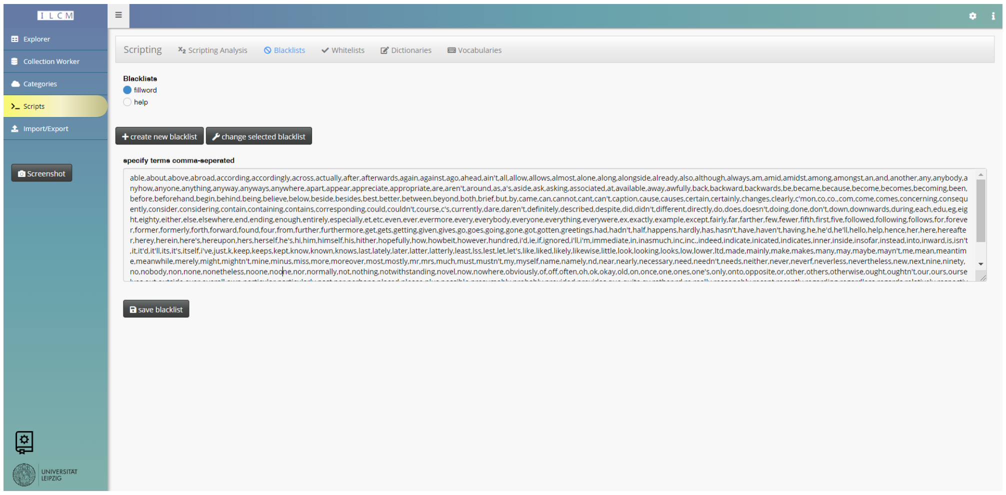
- **Use fixed vocabulary**: Some tasks allow the users to decide if they want to use a fixed vocabulary in the analysis. Here the user has to enable the use of the fixed vocabulary and then chose one of the existing vocabulary lists. (Note that when the fixed vocabulary list is excluded, other parameters disappear as they are no longer relevant for the analysis).
- **Baseform reduction**: The users can specify whether they want to use lemmatization, stemming or no baseform reduction mechanism
- **Min number** of chars per word: The user can specify the minimal and maximal amount of characters a word needs to have. This helps immensely to get rid of errors in the tokenization.
- **N-grams**: The user can specify whether they want to work with uni-/bi- or trigrams. Keep in mind that using bi-/trigrams will increase the vocabulary size, leading to a more extended and more resource-intensive calculation.
- **Remove custom words**: The user can enter some words they want explicitly to be removed from the calculation. If there are more than a few words, it might be more useful to use the blacklist parameter instead.
- **Remove stopwords**: The user can enable/disable the stopword removal. The standard *quanteda* stopwords are used.
- **Transform to lower case**: The user can specify whether they want to transform all words to lowercase.
- **Remove numbers**: The user can enable/disable the removal of numbers
- **Remove punctuation**: The user can enable/disable the removal of punctuation (e.g.: “,”,“.”,“-”,”*“,”#”,…).
- **Remove hyphenation**: The user can enable/disable the removal of hyphenation.
- **Consolidate entities**: During the preparation of new data, named entities are extracted using Spacy. These might be composed of multiple words. E.g. the two words. *Angela* and *Merkel* from the entity “person”. Consolidating the entities describes the process of binding words that form an entity together. These related words will then be used as a single word (“Angela_Merkel”).
- **Use custom blacklist**: If the user wants to exclude multiple words from the analysis (e.g. bigger stopword lists, known words fragments in the data, …), they can create a blacklist and

use it during the preprocessing. Here the user has to enable using a blacklist and then choose one of the existing blacklists.

- **Use custom whitelist**: If the user wants to include multiple words from the analysis (e.g. bigger lists of important words, known fragments in the data, ...), they can create a whitelist and use it during the preprocessing. Here the user has to enable the use of a whitelist and then choose one of the existing whitelists.
- **Pruning**: The pruning aims to remove words from the analysis whose frequency/document frequency is under or over a certain threshold. We use the pruning parameter setting of quanteda (quanteda pruning).
- **Include POS- or NER-tags**: In most of the NLP-task a POS (Part of Speech)- and NER(Named Entity Recognition)-tagging is necessary in order to complete the task. Therefore the user can choose which Tags should be considered to be important within the analysis. As an example it would be useful when the user wants to create *Topic Models* where the topics just consist out of nouns (POS-tag) with a specific NER-classification (e.g just mentioned locations).
- **Exclude POS- or NER-tags**: In contrast to including specified POS- or NER-tags the user can also exclude them. For example when the user wants to start a *Co-occurrence Analysis* the co-occurrence between words and punctuation marks (POS-tags) are not interesting therefore they can exclude them. Or they maybe are not interested in occurrence of firm-names (NER-tag) and therefore they can exclude them.

13.2 Create a blacklist

The blacklist can be created in the *Scripts* tab using the *Blacklists* subtab. Here the user can create new blacklists or manipulate existing ones. The words have to be entered comma separated.



Blacklists
● fillword
○ help

specify terms comma-seperated

able,about,above,abroad,according,accordingly,across,actually,after,afterwards,again,against,ago,ahead,ain't,all,allow,allows,almost,alone,along,alongside,already,also,although,always,am,amid,amidst,among,amongst,an,and,another,any,anybody,a
nyhow,anyone,anything,anyway,anyways,anywhere,apart,appear,appreciate,appropriate,are,aren't,around,as,a's,aside,ask,asking,associated,at,available,away,awfully,back,backward,backwards,be,became,because,become,becomes,becoming,been,
before,beforehand,begin,behind,being,believe,below,beside,besides,best,better,between,beyond,both,brief,but,by,came,can,cannot,cant,can't,caption,cause,causes,certain,certainly,changes,clearly,c'mon,co,co.,com,come,comes,concerning,consequ
ently,consider,considering,contain,containing,contains,corresponding,could,couldn't,course,c's,currently,dare,daren't,definitely,described,despite,did,didn't,different,directly,do,does,doesn't,doing,done,don't,down,downwards,during,each,edu,eg,eig
ht,eighty,either,else,elsewhere,end,ending,enough,entirely,especially,et,etc,even,ever,evermore,every,everybody,everyone,everything,everywere,ex,exactly,example,except,fairly,far,farther,few,fewer,fifth,first,five,followed,following,follows,for,foreve
r,former,formerly,forth,forward,found,four,from,further,furthermore,get,gets,getting,given,gives,go,goes,going,gone,got,gotten,greetings,had,hadn't,half,happens,hardly,has,hasn't,have,haven't,having,he,he'd,he'll,hello,help,hence,her,here,hereafte
r,herey,herein,here's,hereupon,hers,herself,he's,hi,him,himself,his,hither,hopefully,how,howbeit,however,hundred,i'd,ie,if,ignored,i'll,i'm,immediate,in,inasmuch,inc,inc.,indeed,indicate,inicated,indicates,inner,inside,insofar,instead,into,inward,is,isn't
,it,it'd,it'll,its,it's,itself,i've,just,k,keep,keeps,kept,know,known,knows,last,lately,later,latter,latterly,least,lss,lest,let,let's,like,liked,likely,likewise,little,look,looking,looks,low,lower,ltd,made,mainly,make,makes,many,may,maybe,mayn't,me,mean,meantim
e,meanwhile,merely,might,mightn't,mine,minus,miss,more,moreover,most,mostly,mr,mrs,much,must,mustn't,my,myself,name,namely,nd,near,nearly,necessary,need,needn't,needs,neither,never,neverf,neverless,nevertheless,new,next,nine,ninety,
no,nobody,non,none,nonetheless,noone,noone,nor,normally,not,nothing,notwithstanding,novel,now,nowhere,obviously,of,off,often,oh,ok,okay,old,on,once,one,ones,one's,only,onto,opposite,or,other,others,otherwise,ought,oughtn't,our,ours,ourse

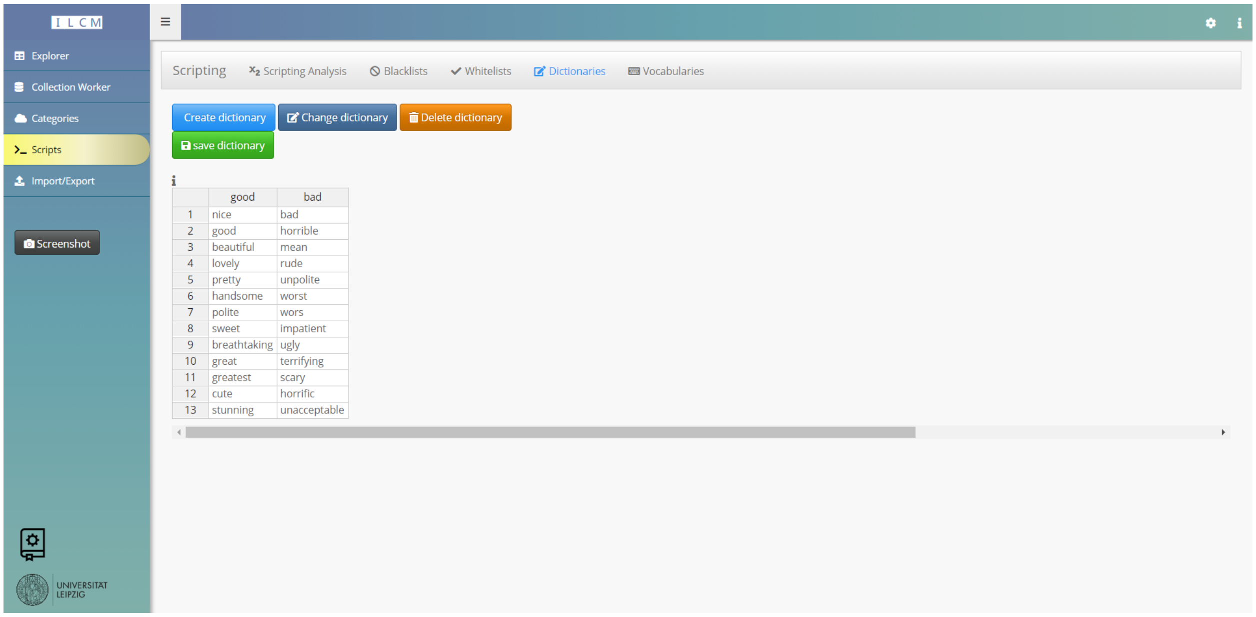
Figure 31: Blacklist Creation

13.3 Create whitelists

The whitelist can be created in the *Scripts* tab using the *Whitelists* subtab. Here the user can create new whitelists or manipulate existing ones. The words have to be entered comma separated.

13.4 Create dictionaries

The dictionaries can be created or changed and even deleted through the *Scripts* tab under the *Dictionaries* subtab directory. A dictionary consists of different columns. In the following example, a dictionary describing good or bad connoted words is used. To create a new column simply right-click on the table, a menu will open with different kinds of options to manipulate the current table. The concrete number of columns and rows is variable. After creating or changing dictionaries, the user needs to press the *Save dictionary* button. This will lead to a small window where the user needs to select a name for the dictionary and the columns' names.



	good	bad
1	nice	bad
2	good	horrible
3	beautiful	mean
4	lovely	rude
5	pretty	unpolite
6	handsome	worst
7	polite	wors
8	sweet	impatient
9	breathtaking	ugly
10	great	terrifying
11	greatest	scary
12	cute	horrific
13	stunning	unacceptable

Figure 32: Dictionary Creation

13.5 Create vocabularies

Vocabularies can be created similar to blacklists or whitelists. Therefore the user can enter words for the vocabulary (comma-separated) or add through importing from finished analysis models. When selecting the last option, a window will open where the user can select the type of analysis and the id of the task (see *Collection Worker* under *Results*) and select the corresponding vocabulary list.

13.6 Special parameters

Besides the standard parameters for the preprocessing chain, each algorithm can be specified using parameters. The whole parameter setting for each started task is saved in the database. This secures the reproducibility of all calculated results.

1. **Parameters co-occurrence analysis**

- Doc/Sent co-occurrence: The calculation of co-occurrences can be done on different units of text. The iLCM offers the possibilities: document and sentence. This means a pair of words form a co-occurrence if they appear together in a sentence or a whole document.

- Minimum co-occurrence frequency: When working with vast amounts of data, the vocabulary can grow very fast. The co-occurrence matrix is a sparse term-term matrix. When the user works with a huge vocabulary, the intern object-size of the term-term matrix threshold for a minimum co-occurrence frequency, this means a combination of a term is assigned a non-0 value only if the words appear more often together than the set threshold. In most cases, this can be done without hesitation because the significance value for the word combination should be shallow anyway.
- POS-tags: For some co-occurrence analysis applications, it might be interesting to work on certain POS-tags words. This can be achieved with this parameter. Here the user can choose multiple POS-tags they want to keep. All other words that do not belong to one of the chosen POS-tags will be excluded from the analysis.
- NER-tags: Also, for NER-tags, the user can specify certain tags they want to use. The words that do not belong to the chosen tags will also be excluded. With this option, it can be pretty easy to create a network of persons. The user needs to select the NER-tag "Person", and the result will be a network of co-occurring persons.

2. **Parameter frequency extraction**

- POS-tags: Calculate the frequency counts just for words that are marked with one of the chosen POS-tags
- NER-tags: Calculate the frequency counts just for words that are marked with one of the chosen NER-tags

3. **Parameter dictionary extraction**

- Context filter list: In the *Dictionary extraction*, the dictionaries' frequency in the documents is counted. However, sometimes, the users are only interested in the counts if they appear in a particular context. In the context filter list, the user can enter some words that have to be found to be counted in order of the dictionary occurrence.
- Regular expression mode: Sometimes, the context is more complex than just a certain word requirement. Here the user can enable the regular expression mode, and as a result, the context filter list *Input* will be interpreted as a regular expression.
- Context unit: The text unit users want to control the presence of their context filter and, in the end, count the occurrence of the dictionaries. This can be done either on a sentence or document level.
- Dictionary: In the dictionary input, the user has to define his dictionaries. This has to be done in a certain format:
  Conceptname1:term1;term2;term3;...
  Conceptname2:term4;term5;term6;...
  These dictionaries can give a fundamental example for the starting point of a sentiment analysis:
  good:happy;satisfied;awesome...
  bad:sad;unsatisfied;terrible...

4. **Parameters volatility analysis**

- Minimum co-occurrence frequency: When working with vast amounts of data, the vocabulary can grow very fast. The co-occurrence matrix is a sparse term-term matrix. When the user works with a huge vocabulary, the term-term matrix's intern object-size can exceed the available hardware resources. This problem can be overcome by setting a threshold for a minimum co-occurrence frequency. This means a combination of terms is assigned a non-0 value only if the words appear more often together than the set threshold. In most cases, this can be done without hesitation because the significance value for the word combination should be meagre anyway. The co-occurrence matrices for several times slices form the basis of the context volatility calculation.

- Time interval: The aim of measuring context volatility is to calculate the context change amount in a specified time. With this parameter, we set the time unit. This can be days, month or even years.
- History: The *History* parameter determines the number of points in time compared to the specific point of time the context volatility is calculated. The number the user sets is corresponding to the set time interval.
- Number of Core: This parameter specifies the number of cores that are used for the calculation. Though at the fact, that the context volatility calculation is quite complex, it might be helpful to use more than one core.
- Method: The method parameter specifies how the difference of the expected value for the history of the significance of co-occurrence and the actual measured values is compared. This can be done by averaging the distance for every co-occurrence (standard) or calculating the cosine distance of the 2 vectors (expected values, actual measured values).
- Doc/Sent Co-Occurrence: The calculation of co-occurrence can be done on different units of text. The iLCM offers the possibilities: *document* and *sentence*. This means a pair of words form a co-occurrence if they appear together in a sentence/text.
- POS-tags: Calculate the context volatility using words marked with one of the chosen POS-tags.
- NER-tags: Calculate the context volatility just using words that are marked with one of the chosen NER-tags.

5. **Parameters topic modelling**

- Number of topics: Here, the users can specify the number of topics they want to be calculated. The number has to be at least 3 (2 does not work with the principal component for visualization with LDA-vis). The user can get an idea of the correct number of topics for a dataset by looking at the topic coherence page in the _ Result_ tab.
- Alpha: Alpha is the parameter of the Dirichlet prior to the per-document topic distributions.
- Method: There are 2 different topic model backends integrated so far. Here the user can switch between them. For more information have a closer look at the libraries: topicmodels::LDA and LDA::lda.collapsed.Gibbs.sampler
- POS-tags: Calculate the topic model using words marked with one of the chosen POS-tags.
- NER-tags: Calculate the topic model using words marked with one of the chosen NER-tags.

6. **Parameter classification**

- Context Unit: The classification of the data can be done on different units of text. The iLCM offers the possibilities: *document* and *sentence*. This means words get a specific classification depending on parameters corresponding to sentences or the whole text.
- Project: The user needs to decide to which project the current classification should belong. They can decide within a list similar to the annotation schemes.
- Mode: There are 4 different options of how the classification can be done: producing 50 new active learning examples, active learning on whole documents evaluating training set, classify the entire collection. Depending on the selected mode, there are different other parameters to specify the active learning method.

7. **Parameter sentiment analysis**

- Sentiment Dictionary: Essential for the sentiment analysis is the selected dictionary of predefined sentiments connected with the language of the data set. The user can currently decide between an English or German sentiment dictionary.
- Build document score by: The user can choose between different methods to calculate the document's sentiment (e.g. the mean of all found sentiments in the text).

8. **Parameter document deduplication**

- Similarity measurement: Here, the user can choose how the deduplication of documents will be calculated (Jaccard similarity or Jaccard bag similarity).

9. **Parameter dynamic topic modelling**

- Number of topics: Here, the users can specify the number of topics they want to be calculated. The number has to be at least 3 (2 does not work with the principal component for visualization with LDA-vis). The user can get an idea of the correct number of topics for a dataset by looking at the topic coherence page in the *Result* tab.
- Alpha: Alpha is the parameter of the Dirichlet prior to the per-document topic distributions.
- Chain variance: The user decides by which factor the dynamic topic modelling chains should variate.
- How to split dates: The user can decide how the different documents should be separated - they can choose between a separation by time or via automatic chunking.

10. **Parameter keyword extraction**

- Method: There are 4 different methods on how the keywords in the data can be extracted. Here the user can switch between them. For more information: ...
- Minimal number of co-occurrence: The user can select a minimal number of occurrences per keyword (e.g. each word that appears at least 4 times in a document should be considered a keyword of this document).
- Maximal number of co-occurrence: The user can also select a maximal number of occurrences per keyword. Words that frequently appear in a text are more likely to be stopwords or no specific keywords for this document.
- Separator: Here, the user can decide what separation sign should separate the found keywords.

11. **Parameters syntactic parsing**

- Number of corse: Here, the user can decide how many cores should be used to calculate syntactic parsing. This depends on how fast the user wants results and how many cores are available in the current system.

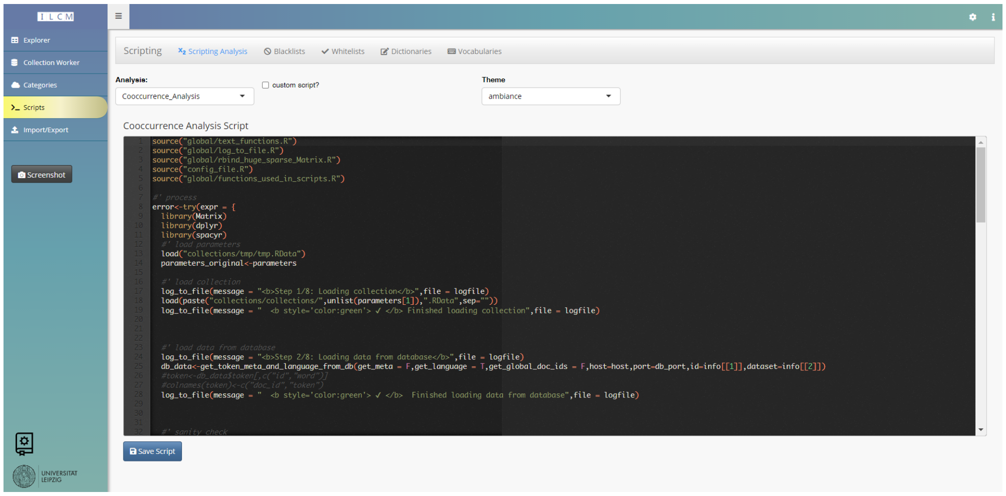
12. **Parameters vector space representation**

- Minimum occurrence: The user needs to select a minimum number of occurrences for the words to become a vector.
- Number of vectors: Here, the user needs to specify how many vectors should be created during the task.
- Number of threads: The user needs to specify the number of threads used during the vector calculation.
- Windows: Here, the user can select a window size. This size will determine how many words will be considered to calculate the vector from a target word.

- Number of iteration: Here, the user can select several iterations to create the word vectors. This needs to be a trade-off between the accuracy of the word representation and overfitting the vectors for a specific collection of documents. Also, more iteration means the task will need longer to finish.
- Number of negative samples: The user can select several negative samples to prevent the vectors from overfitting. However, the number should be chosen in order to the size of the currently used collection.

13.7 Change script

Having specified all parameters, the user can start the script. R-experienced users may be unsatisfied with how the algorithms are implemented. Therefore they can also change existing scripts by clicking the *Show Script* button. This button will lead the user to the *Scripts* tab, and the users can see and edit the existing script for the chosen algorithm. The essential thing while editing a script is to keep the same output format as the original script to assure a correct visualization. The manipulated script can be saved and used for later calculation by ticking the *use custom script* field in the *Task Scheduler* subtab.



Analysis: Cooccurrence_Analysis

□ custom script?

Theme: ambiance

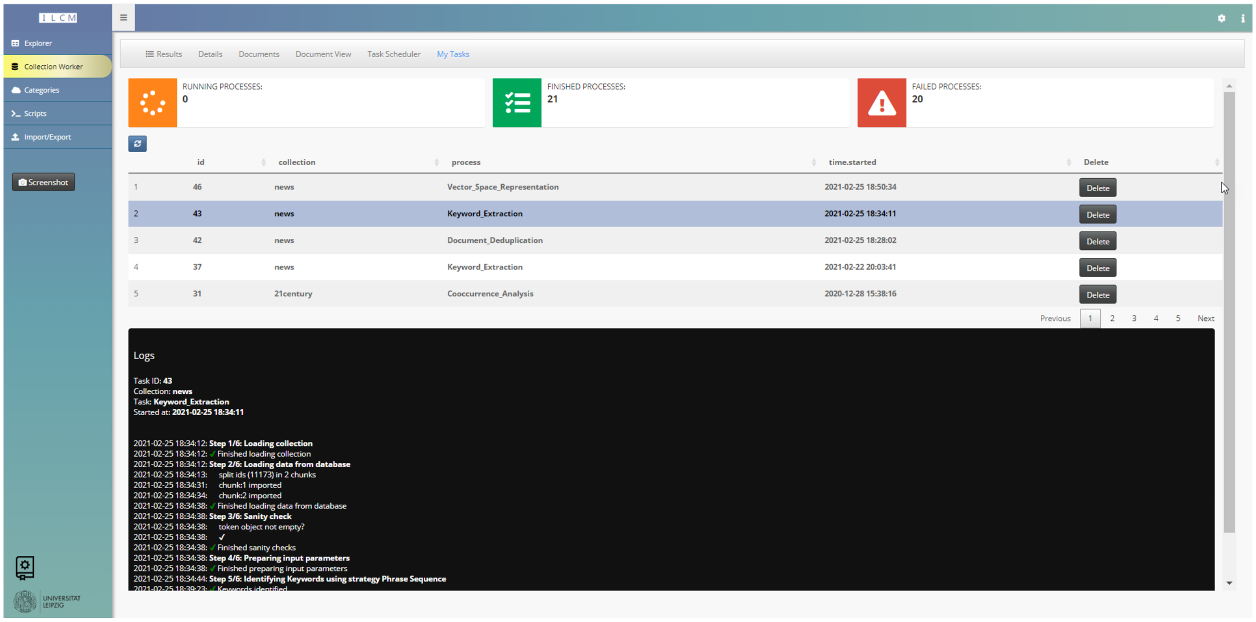
Cooccurrence Analysis Script

```
source("global/text_functions.R")
source("global/log_to_file.R")
source("global/rbind_huge_sparse_Matrix.R")
source("config_file.R")
source("global/functions_used_in_scripts.R")

#' process
error<-try(expr = {
  library(Matrix)
  library(dplyr)
  library(spacyr)
  #' load parameters
  load("collections/tmp/tmp.RData")
  parameters_original<-parameters

  #' load collection
  log_to_file(message = "<b>Step 1/8: Loading collection</b>",file = logfile)
  load(paste("collections/collections/",unlist(parameters[1]),".RData",sep=""))
  log_to_file(message = "  <b style='color:green'> ✓ </b> Finished loading collection",file = logfile)



  #' load data from database
  log_to_file(message = "<b>Step 2/8: Loading data from database</b>",file = logfile)
  db_data<-get_token_meta_and_language_from_db(get_meta = F,get_language = T,get_global_doc_ids = F,host=host,port=db_port,id=info[[1]],dataset=info[[2]])
  #token<-db_data$token[,c("id","word")]
  #colnames(token)<-c("doc_id","token")
  log_to_file(message = "  <b style='color:green'> ✓ </b>  Finished loading data from database",file = logfile)



  #' sanity check
```

Figure 33: Manipulate Script for Co-occurrence Analysis

13.8 Check logs

In the *My Tasks* subtab, the users can also look at the amount of running, finished and failed tasks. They can also check the logs for each started task. Therefore they have to click on the box (running, finished or failed) and then select the task in the shown table below by clicking in the corresponding row. The log-file will then be loaded and shown:



RUNNING PROCESSES: **0**

FINISHED PROCESSES: **21**

FAILED PROCESSES: **20**

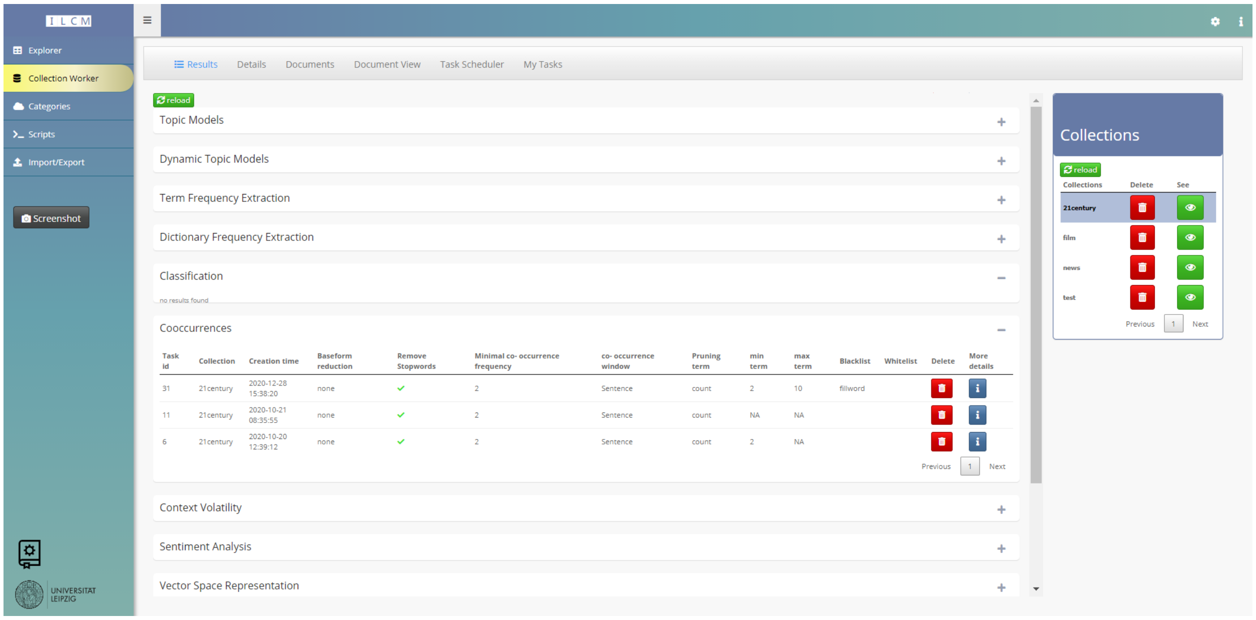
	id	collection	process	time.started	Delete
1	46	news	Vector_Space_Representation	2021-02-25 18:50:34	Delete
2	43	news	Keyword_Extraction	2021-02-25 18:34:11	Delete
3	42	news	Document_Deduplication	2021-02-25 18:28:02	Delete
4	37	news	Keyword_Extraction	2021-02-22 20:03:41	Delete
5	31	21century	Cooccurrence_Analysis	2020-12-28 15:38:16	Delete

```
Logs

Task ID: 43
Collection: news
Task: Keyword_Extraction
Started at: 2021-02-25 18:34:11

2021-02-25 18:34:12: Step 1/6: Loading collection
2021-02-25 18:34:12: ✓ Finished loading collection
2021-02-25 18:34:12: Step 2/6: Loading data from database
2021-02-25 18:34:13:    split ids (11173) in 2 chunks
2021-02-25 18:34:31:    chunk:1 imported
2021-02-25 18:34:34:    chunk:2 imported
2021-02-25 18:34:38: ✓ Finished loading data from database
2021-02-25 18:34:38: Step 3/6: Sanity check
2021-02-25 18:34:38:    token object not empty?
2021-02-25 18:34:38:    ✓
2021-02-25 18:34:38: ✓ Finished sanity checks
2021-02-25 18:34:38: Step 4/6: Preparing input parameters
2021-02-25 18:34:38: ✓ Finished preparing input parameters
2021-02-25 18:34:44: Step 5/6: Identifying Keywords using strategy Phrase Sequence
2021-02-25 18:39:23: ✓ Keywords identified
```

Figure 34: My Tasks subtab

14 Results

The *Results* subtab is part of the *Collection Worker* tab. Here the user can find 2 columns. On the left, the user can see a box for each analysis algorithm. When expanding the boxes, the corresponding finished tasks will be displayed with their parameter setting (if possible). Clicking on a specific task will redirect the user to the *Details* subtab showing the tasks' results.



no results found

Task id	Collection	Creation time	Baseform reduction	Remove Stopwords	Minimal co- occurrence frequency	co- occurrence window	Pruning term	min term	max term	Blacklist	Whitelist	Delete	More details
31	21century	2020-12-28 15:38:20	none	✓	2	Sentence	count	2	10	fillword		🗑	i
11	21century	2020-10-21 08:35:55	none	✓	2	Sentence	count	NA	NA			🗑	i
6	21century	2020-10-20 12:39:12	none	✓	2	Sentence	count	2	NA			🗑	i

Collections	Delete	See
21century	🗑	👁
film	🗑	👁
news	🗑	👁
test	🗑	👁

Figure 35: Results subtab with collection '21century' marked

14.1 Collections

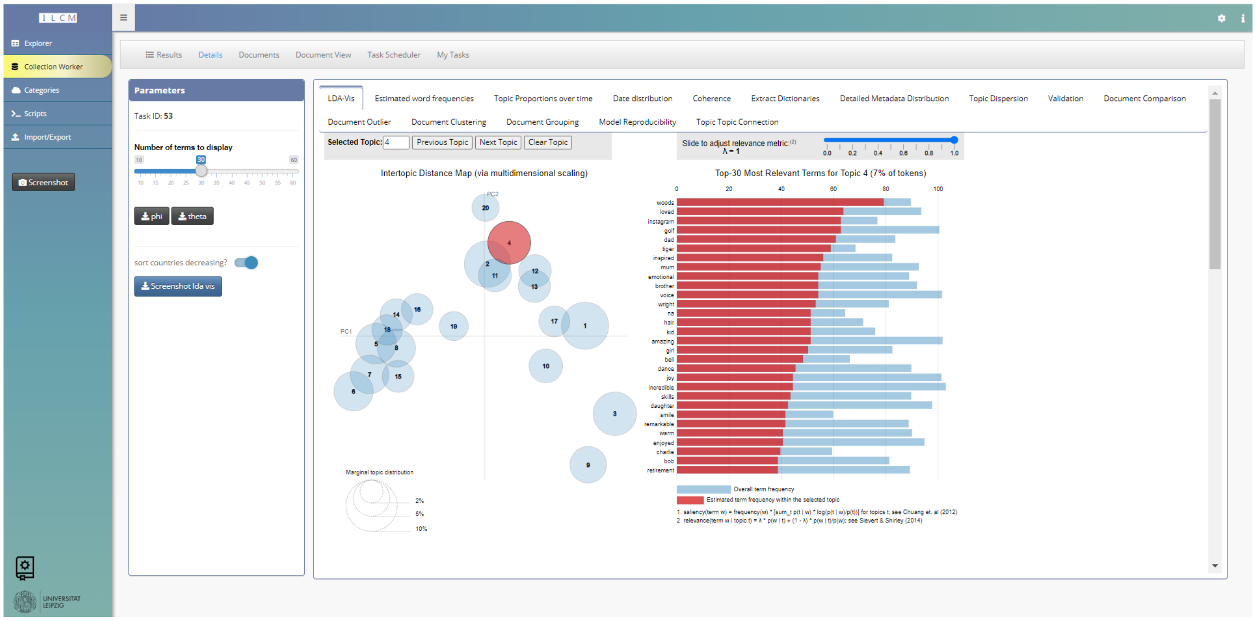
In the right column, there is a box for the collections. Here the users have 3 options. They can have a closer look at the time series plot for every collection (click on *See*) or delete collections. By selecting a row in the *Collections* table, the current collection's intern variable is set. Like that, only the current collection results will be shown in the *Result* column. Furthermore, when switching to the subtab *Document* in the *Collection Worker*, the chosen collection documents are shown and can be read and annotated. When hovering over the name of a collection, its creation-statement will be displayed.

15 Details

In the *Details* subtab, the results of the chosen task are visualized. The *Details* tab depends on the chosen task and its underlying NLP-algorithm.

15.1 Details Topic models

- **LDA-Vis**: In the *LDA-Vis* panel the user sees the LDA-vis visualization for the calculated topic model. In the parameters-box they can change the number of words shown per selected topic. Besides there is the possibility to download the matrices for phi and theta in a csv-format.



Task ID: 53

Number of terms to display: 30

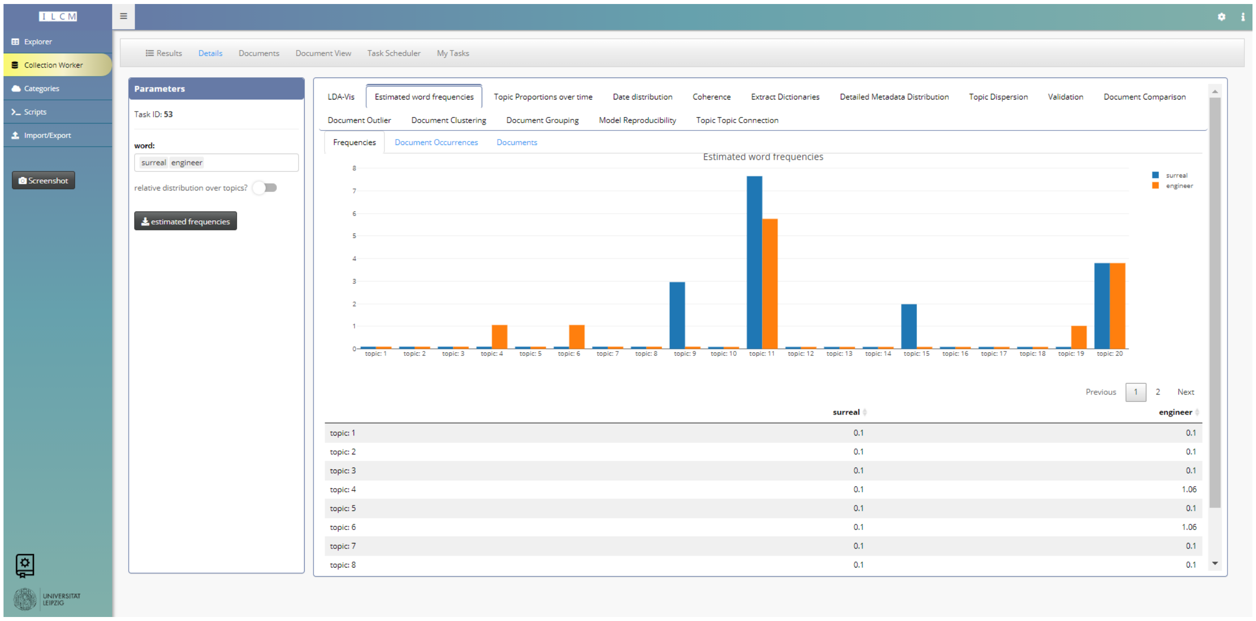
Selected Topic: 4

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 36: LDA-Vis example for topic model calculated on guardian data for section 'sport'

- **Estimated Word Frequencies**: In the parameter box of *Estimate word frequencies*, the user needs to select at least one word to calculate its frequency. They can select between all the words used in the topic modelling documents. With clicking in the *words* text field, a list of possible words should show up. The user can choose between them. There are 3 subtabs implemented. The “normal” *Frequency* of appearance for a word in the calculated topic will show a plot of all the topics and the number of occurrence of the word/words within the topics. Below this plot, the user will find a data table with the topic and the different frequency values of the selected words. In the *Document Occurrence* subtab, specific documents are mentioned that state the selected word/words and a counter for the occurrence for these words inside the document. In the *Document* tab, the user can look closely at the documents mentioned in the last subtab and take a closer look at the text. The selected words will be highlighted. The user also can download the estimated frequencies in the parameter box on the left side.



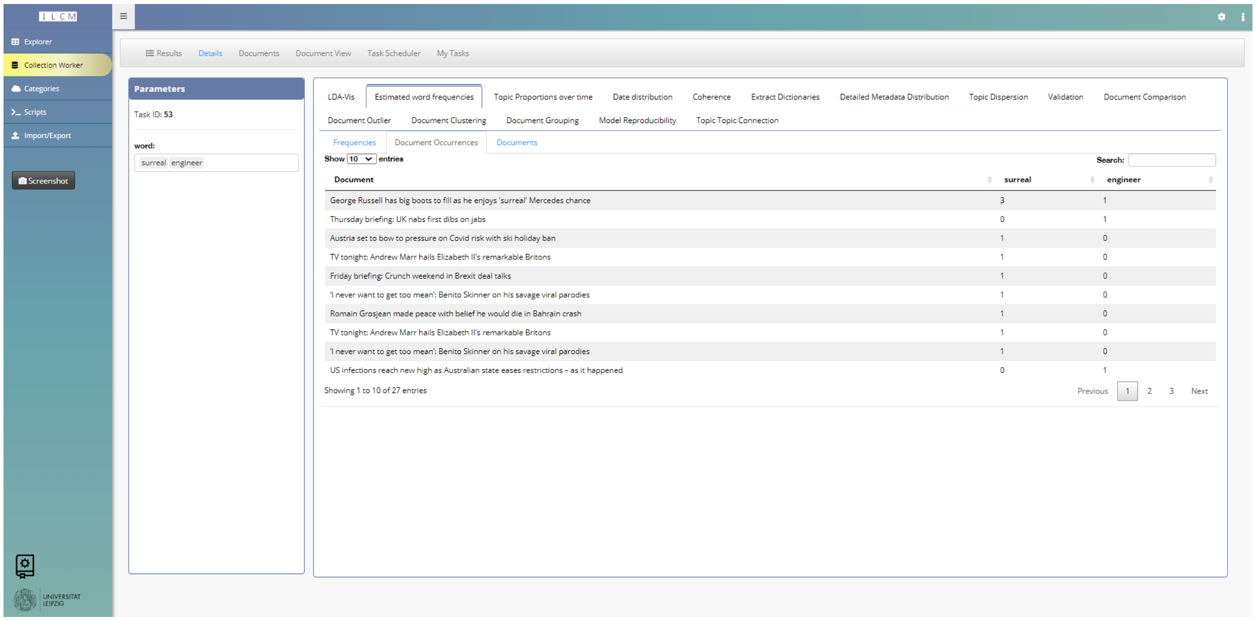
Task ID: 53

word: surreal engineer

relative distribution over topics?

	surreal	engineer
topic: 1	0.1	0.1
topic: 2	0.1	0.1
topic: 3	0.1	0.1
topic: 4	0.1	1.06
topic: 5	0.1	0.1
topic: 6	0.1	1.06
topic: 7	0.1	0.1
topic: 8	0.1	0.1

Figure 37: Frequency estimation over the topics



Task ID: 53

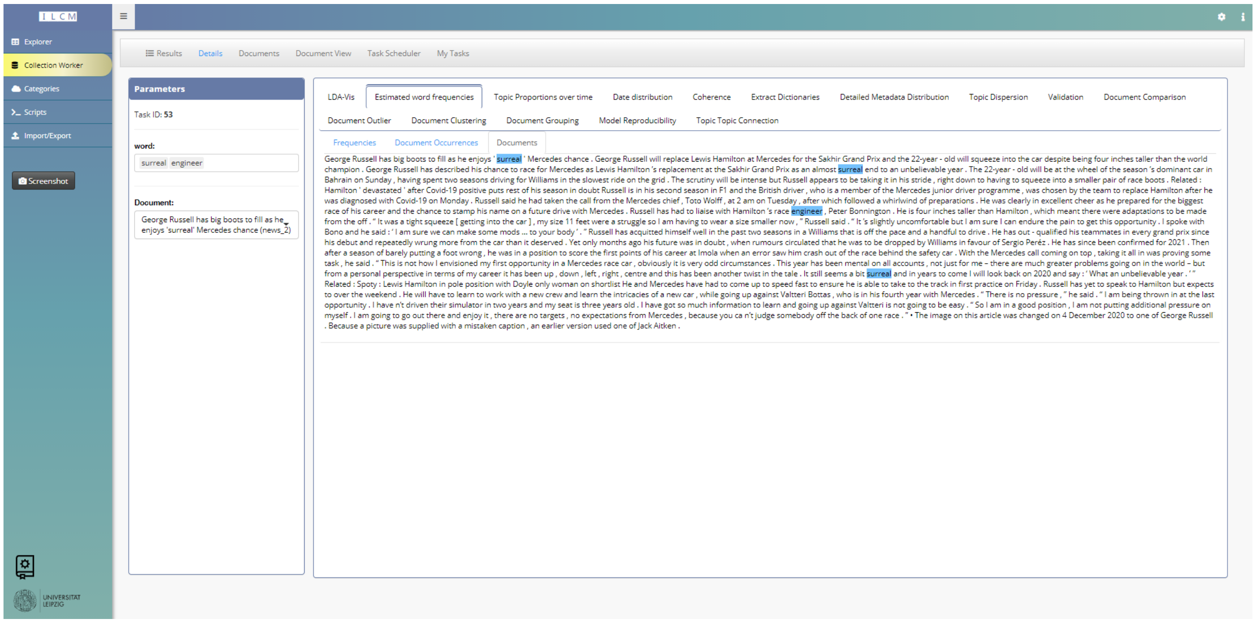
word: surreal engineer

Show 10 entries

Document	surreal	engineer
George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance	3	1
Thursday briefing: UK nabs first dibs on jabs	0	1
Austria set to bow to pressure on Covid risk with ski holiday ban	1	0
TV tonight: Andrew Marr hails Elizabeth II's remarkable Britons	1	0
Friday briefing: Crunch weekend in Brexit deal talks	1	0
‘I never want to get too mean’: Benito Skinner on his savage viral parodies	1	0
Romain Grosjean made peace with belief he would die in Bahrain crash	1	0
TV tonight: Andrew Marr hails Elizabeth II's remarkable Britons	1	0
‘I never want to get too mean’: Benito Skinner on his savage viral parodies	1	0
US infections reach new high as Australian state eases restrictions – as it happened	0	1

Showing 1 to 10 of 27 entries

Figure 38: Occurence of the words within the documents



Task ID: **53**

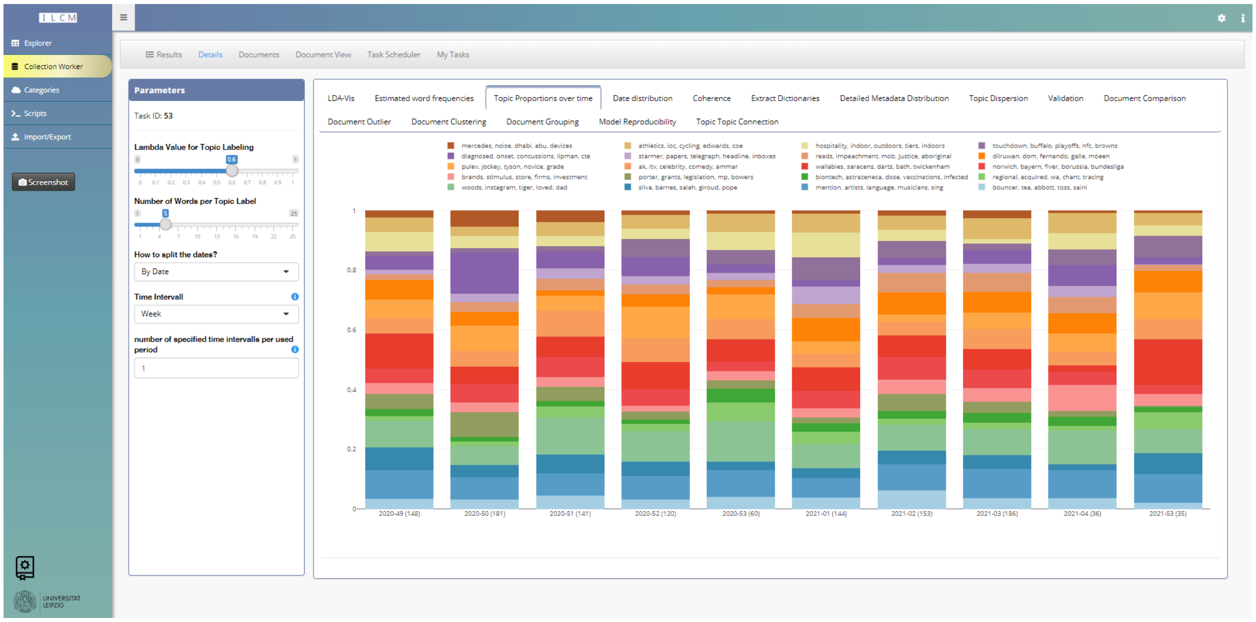
word: surreal engineer

Document: George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)

George Russell has big boots to fill as he enjoys ' surreal ' Mercedes chance . George Russell will replace Lewis Hamilton at Mercedes for the Sakhir Grand Prix and the 22-year - old will squeeze into the car despite being four inches taller than the world champion . George Russell has described his chance to race for Mercedes as Lewis Hamilton 's replacement at the Sakhir Grand Prix as an almost surreal end to an unbelievable year . The 22-year - old will be at the wheel of the season 's dominant car in Bahrain on Sunday , having spent two seasons driving for Williams in the slowest ride on the grid . The scrutiny will be intense but Russell appears to be taking it in his stride , right down to having to squeeze into a smaller pair of race boots . Related : Hamilton ' devastated ' after Covid-19 positive puts rest of his season in doubt Russell is in his second season in F1 and the British driver , who is a member of the Mercedes junior driver programme , was chosen by the team to replace Hamilton after he was diagnosed with Covid-19 on Monday . Russell said he had taken the call from the Mercedes chief , Toto Wolff , at 2 am on Tuesday , after which followed a whirlwind of preparations . He was clearly in excellent cheer as he prepared for the biggest race of his career and the chance to stamp his name on a future drive with Mercedes . Russell has had to liaise with Hamilton 's race engineer , Peter Bonnington . He is four inches taller than Hamilton , which meant there were adaptations to be made from the off . " It was a tight squeeze [getting into the car] , my size 11 feet were a struggle so I am having to wear a size smaller now , " Russell said . " It 's slightly uncomfortable but I am sure I can endure the pain to get this opportunity . I spoke with Bono and he said : ' I am sure we can make some mods ... to your body ' . " Russell has acquitted himself well in the past two seasons in a Williams that is off the pace and a handful to drive . He has out - qualified his teammates in every grand prix since his debut and repeatedly wrung more from the car than it deserved . Yet only months ago his future was in doubt , when rumours circulated that he was to be dropped by Williams in favour of Sergio Peréz . He has since been confirmed for 2021 . Then after a season of barely putting a foot wrong , he was in a position to score the first points of his career at Imola when an error saw him crash out of the race behind the safety car . With the Mercedes call coming on top , taking it all in was proving some task , he said . " This is not how I envisioned my first opportunity in a Mercedes race car , obviously it is very odd circumstances . This year has been mental on all accounts , not just for me – there are much greater problems going on in the world – but from a personal perspective in terms of my career it has been up , down , left , right , centre and this has been another twist in the tale . It still seems a bit surreal and in years to come I will look back on 2020 and say : ' What an unbelievable year . ' " Related : Spoty : Lewis Hamilton in pole position with Doyle only woman on shortlist He and Mercedes have had to come up to speed fast to ensure he is able to take to the track in first practice on Friday . Russell has yet to speak to Hamilton but expects to over the weekend . He will have to learn to work with a new crew and learn the intricacies of a new car , while going up against Valtteri Bottas , who is in his fourth year with Mercedes . " There is no pressure , " he said . " I am being thrown in at the last opportunity . I have n't driven their simulator in two years and my seat is three years old . I have got so much information to learn and going up against Valtteri is not going to be easy . " So I am in a good position , I am not putting additional pressure on myself . I am going to go out there and enjoy it , there are no targets , no expectations from Mercedes , because you ca n't judge somebody off the back of one race . " • The image on this article was changed on 4 December 2020 to one of George Russell . Because a picture was supplied with a mistaken caption , an earlier version used one of Jack Aitken .

Figure 39: A closer look at the documents that state the selected words

- **Topic Proportion Over Time**: Here, the user can see the topic proportion of the current topic modelling over time. Therefore the user can select a lambda value for the topic labelling, the number of words per topic label and how to split the dates (by date or via chunking). They also can select a specific time interval (year, month, week, days) and the number of specified intervals perused period. This will result in a visualization that shows the proportions of the topic within a specific period.


ILCM
Explorer
Collection Worker
Categories
Scripts
Import/Export
Screenshot
UNIVERSITAT LEIPZIG
Results
Details
Documents
Document View
Task Scheduler
My Tasks
Parameters
LDA-Vis
Estimated word frequencies
Topic Proportions over time
Date distribution
Coherence
Extract Dictionaries
Detailed Metadata Distribution
Topic Dispersion
Validation
Document Comparison
Document Outlier
Document Clustering
Document Grouping
Model Reproducibility
Topic Topic Connection
mercedes, noise, dhabi, abu, devices
diagnosed, onset, concussions, lipman, cte
pulev, jockey, tyson, novice, grade
brands, stimulus, store, firms, investment
woods, instagram, tiger, loved, dad
athletics, ioc, cycling, edwards, coe
starmer, papers, telegraph, headline, inboxes
ak, itv, celebrity, comedy, ammar
porter, grants, legislation, mp, bowers
silva, barnes, salah, giroud, pope
hospitality, indoor, outdoors, tiers, indoors
reads, impeachment, mob, justice, aboriginal
wallabies, saracens, darts, bath, twickenham
biontech, astrazeneca, dose, vaccinations, infected
mention, artists, language, musicians, sing
touchdown, buffalo, playoffs, nfc, browns
dilruwan, dom, fernando, galle, moeen
norwich, bayern, fiver, borussia, bundesliga
regional, acquired, wa, chant, tracing
bouncer, tea, abbott, toss, saini
1
0.8
0.6
0.4
0.2
0
2020-49 (148)
2020-50 (181)
2020-51 (141)
2020-52 (120)
2020-53 (60)
2021-01 (144)
2021-02 (153)
2021-03 (186)
2021-04 (36)
2021-53 (35)

Task ID: **53**

Lambda Value for Topic Labeling: 0.6 (0 – 1)

Number of Words per Topic Label: 5 (1 – 25)

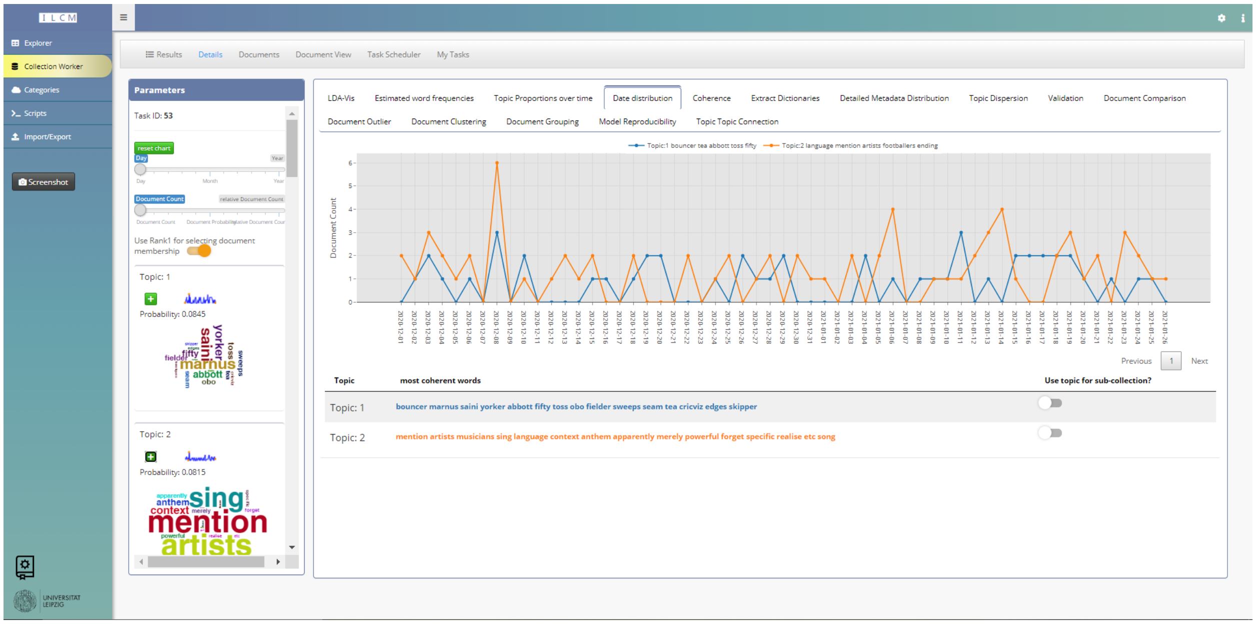
How to split the dates? By Date

Time Intervall Week

number of specified time intervalls per used period 1

Figure 40: Topic proportions over weekly intervals

- **Date Distribution**: The parameter box of *Document Distribution* panel shows a word cloud for each topic. Inside the word clouds plots, the most likely words for every topic are displayed. At the top of the parameters section, the user can change the time series for the time series plot. Besides, they can switch between document count (how many documents are found), document probability (sum of all probabilities for the document belonging to the point in time for the selected topic) and the relative document count (percentage of the number of documents belonging to a specific topic, compared to all documents at the time). The belonging of documents is defined in the input below. The default setting uses the first rank mechanism (a document belongs to the topic for which it has the highest probability). If the user disables this option, they can set a probability threshold. A document then belongs to all topics; it has a higher probability than the threshold. In the top left corner of each word cloud box, the user can add the related topic (green plus-button) to the time series plot. Furthermore, they can see the overall probability of the topic and get an idea of the blue sparkline's times series plot. The time series plot will be calculated concerning the set options in the parameter box. Once the user has added at least one topic to the time series data, a data table is created with the topic number and the most likely words per topic. The user can select one or multiple topics to create a new sub-collection that contains those documents belonging to the chosen topics. The membership of the documents is calculated based on the settings at the top of the parameter box. To achieve this, the user needs to specify a new collection name. With the new sub-collection, the user can start a new NLP-task or annotate documents like the collection that forms a document search.



Task ID: 53

Use Rank1 for selecting document membership

Topic: 1

Probability: 0.0845

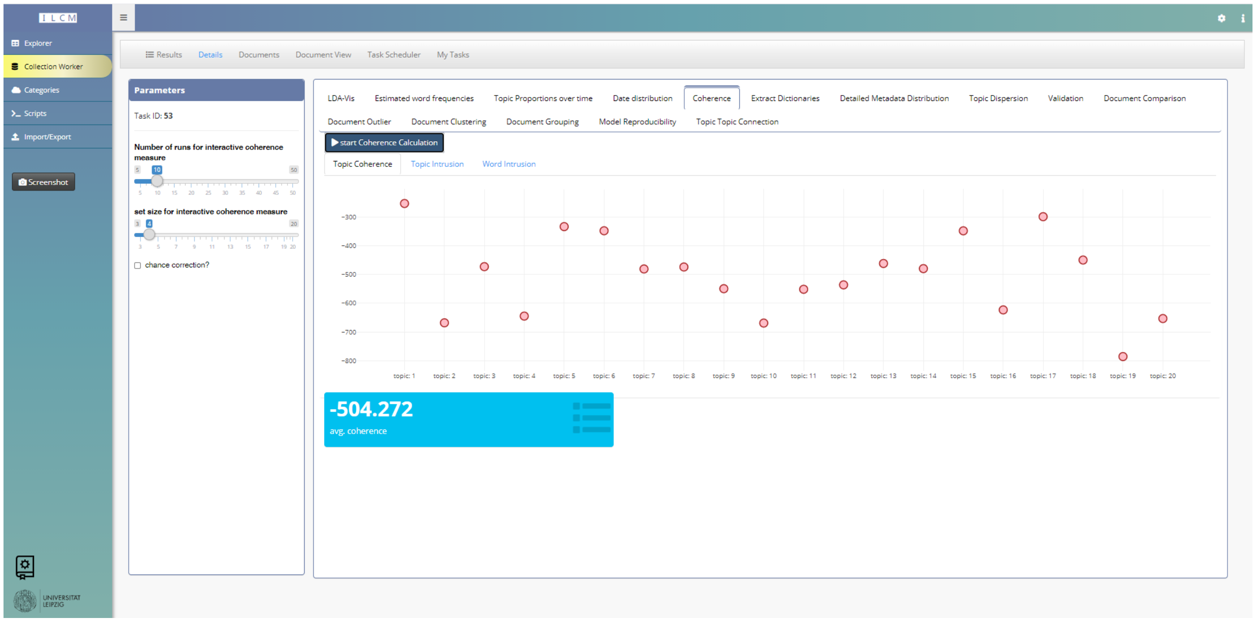
Topic: 2

Probability: 0.0815

Topic	most coherent words	Use topic for sub-collection?
Topic: 1	bouncer marnus saini yorker abbott fifty toss obo fielder sweeps seam tea cricviz edges skipper	○
Topic: 2	mention artists musicians sing language context anthem apparently merely powerful forget specific realise etc song	○

Figure 41: Time series graph for topic model

- **Coherence**: The coherence test enables the evaluation of the calculated topic model. There are 3 different measures implemented: Topic Coherence, Topic Intrusion and Word Intrusion. The topic coherence is in contrast to the 2 intrusion measures calculated without the users' interaction. For the topic intrusion, the user has to read a document and then tell which of the topics does not describe the document. In the word intrusion test, the user sees a group of words, whereby all except one are very likely on the same topic. The task then is identifying the word that does not fit together with the other words. The 2 parameters *number of runs* and *set size* determine the number of examples the user has to work through before the results are saved and the number of topics/words the user has to find the intruder. The chance correction will subtract the estimated random hits from the performance measure.



Task ID: 53

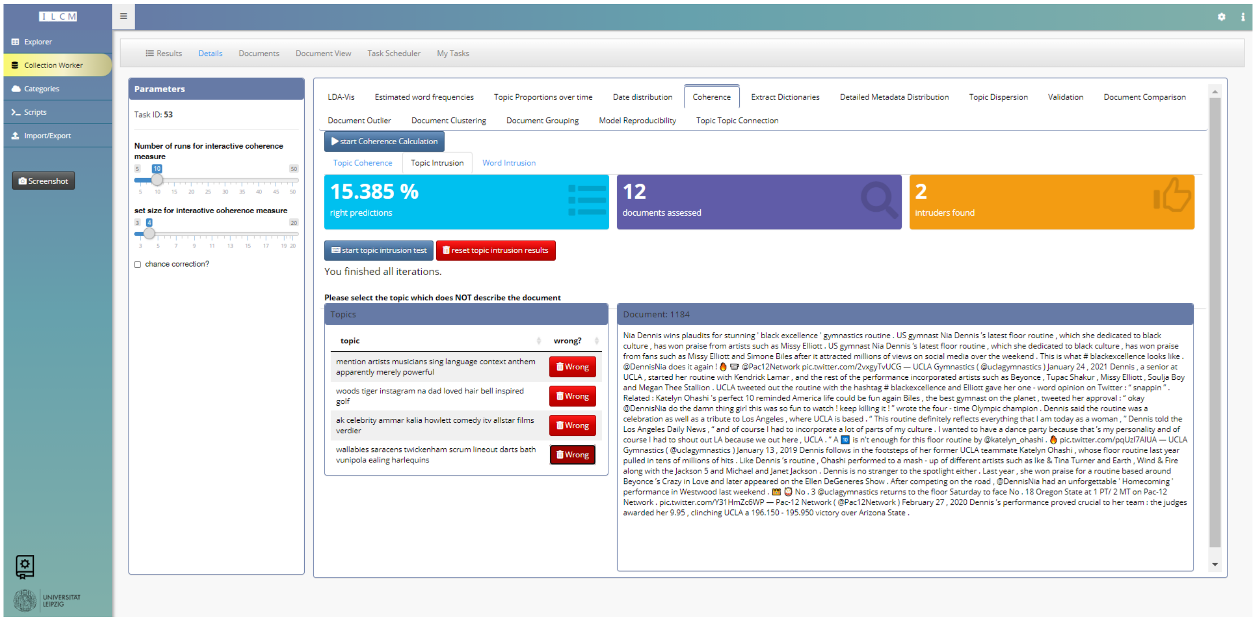
Number of runs for interactive coherence measure: 10

set size for interactive coherence measure: 4

□ chance correction?

-504.272
avg. coherence

Figure 42: Example plot for a topic coherence result



Task ID: 53

Number of runs for interactive coherence measure: 10

set size for interactive coherence measure: 4

□ chance correction?

15.385 %
right predictions

12
documents assessed

2
intruders found

You finished all iterations.

Please select the topic which does NOT describe the document

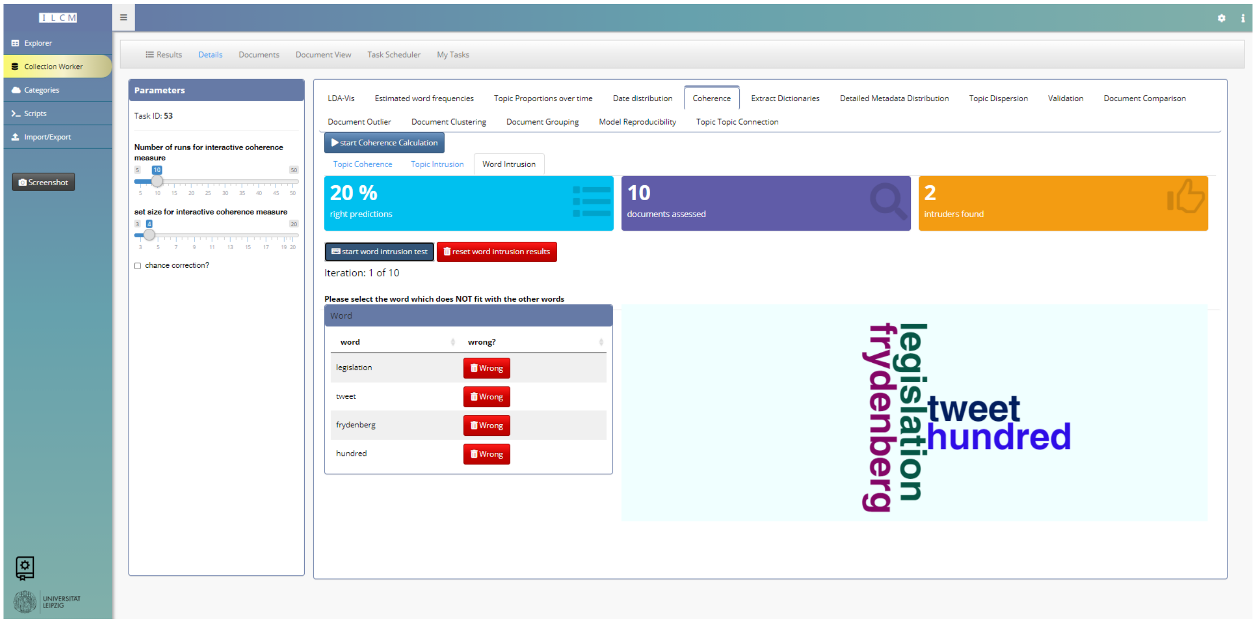
Topics

topic	wrong?
mention artists musicians sing language context anthem apparently merely powerful	Wrong
woods tiger instagram na dad loved hair bell inspired golf	Wrong
ak celebrity ammar kalia howlett comedy itv allstar films verdier	Wrong
wallabies saracens twickenham scrum lineout darts bath vunipola ealing harlequins	Wrong

Document: 1184

Nia Dennis wins plaudits for stunning ' black excellence ' gymnastics routine . US gymnast Nia Dennis 's latest floor routine , which she dedicated to black culture , has won praise from artists such as Missy Elliott . US gymnast Nia Dennis 's latest floor routine , which she dedicated to black culture , has won praise from fans such as Missy Elliott and Simone Biles after it attracted millions of views on social media over the weekend . This is what # blackexcellence looks like . @DennisNia does it again ! 🔥 📼 @Pac12Network pic.twitter.com/2vxgyTvUCG — UCLA Gymnastics (@uclagymnastics) January 24 , 2021 Dennis , a senior at UCLA , started her routine with Kendrick Lamar , and the rest of the performance incorporated artists such as Beyonce , Tupac Shakur , Missy Elliott , Soulja Boy and Megan Thee Stallion . UCLA tweeted out the routine with the hashtag # blackexcellence and Elliott gave her one - word opinion on Twitter : “ snappin ” . Related : Katelyn Ohashi 's perfect 10 reminded America life could be fun again Biles , the best gymnast on the planet , tweeted her approval : “ okay @DennisNia do the damn thing girl this was so fun to watch ! keep killing it ! ” wrote the four - time Olympic champion . Dennis said the routine was a celebration as well as a tribute to Los Angeles , where UCLA is based . “ This routine definitely reflects everything that I am today as a woman , ” Dennis told the Los Angeles Daily News , “ and of course I had to incorporate a lot of parts of my culture . I wanted to have a dance party because that 's my personality and of course I had to shout out LA because we out here , UCLA . ” A 🔟 is n't enough for this floor routine by @katelyn_ohashi . 🔥 pic.twitter.com/pqUzI7AIUA — UCLA Gymnastics (@uclagymnastics) January 13 , 2019 Dennis follows in the footsteps of her former UCLA teammate Katelyn Ohashi , whose floor routine last year pulled in tens of millions of hits . Like Dennis 's routine , Ohashi performed to a mash - up of different artists such as Ike & Tina Turner and Earth , Wind & Fire along with the Jackson 5 and Michael and Janet Jackson . Dennis is no stranger to the spotlight either . Last year , she won praise for a routine based around Beyonce 's Crazy in Love and later appeared on the Ellen DeGeneres Show . After competing on the road , @DennisNia had an unforgettable ' Homecoming ' performance in Westwood last weekend . 👑 🥁 No . 3 @uclagymnastics returns to the floor Saturday to face No . 18 Oregon State at 1 PT/ 2 MT on Pac-12 Network . pic.twitter.com/Y31HmZc6WP — Pac-12 Network (@Pac12Network) February 27 , 2020 Dennis 's performance proved crucial to her team : the judges awarded her 9.95 , clinching UCLA a 196.150 - 195.950 victory over Arizona State .

Figure 43: Example for a topic intrusion test



Parameters

Task ID: 53

Number of runs for interactive coherence measure: 10

set size for interactive coherence measure: 4

□ chance correction?

20 % right predictions

10 documents assessed

2 intruders found

Iteration: 1 of 10

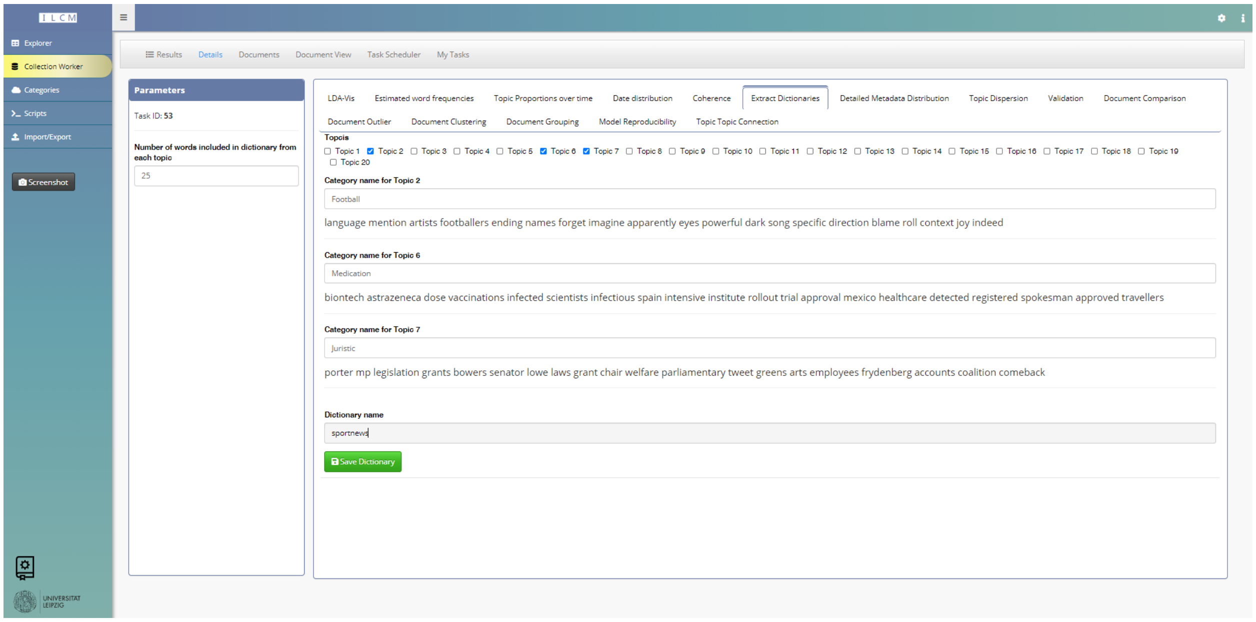
Please select the word which does NOT fit with the other words

Word

word	wrong?
legislation	Wrong
tweet	Wrong
frydenberg	Wrong
hundred	Wrong

Figure 44: Example for a word intrusion test

- **Extract Dictionaries**: Here, the user can select different topics to extract the corresponding words into a dictionary. Each selected Topic becomes a category the user needs to make up a name for. They also need to name the dictionary in order to be able to save the dictionary.



Parameters

Task ID: 53

Number of words included in dictionary from each topic: 25

Topcis

□ Topic 1 ☑ Topic 2 □ Topic 3 □ Topic 4 □ Topic 5 ☑ Topic 6 ☑ Topic 7 □ Topic 8 □ Topic 9 □ Topic 10 □ Topic 11 □ Topic 12 □ Topic 13 □ Topic 14 □ Topic 15 □ Topic 16 □ Topic 17 □ Topic 18 □ Topic 19 □ Topic 20

Category name for Topic 2: Football

language mention artists footballers ending names forget imagine apparently eyes powerful dark song specific direction blame roll context joy indeed

Category name for Topic 6: Medication

biontech astrazeneca dose vaccinations infected scientists infectious spain intensive institute rollout trial approval mexico healthcare detected registered spokesman approved travellers

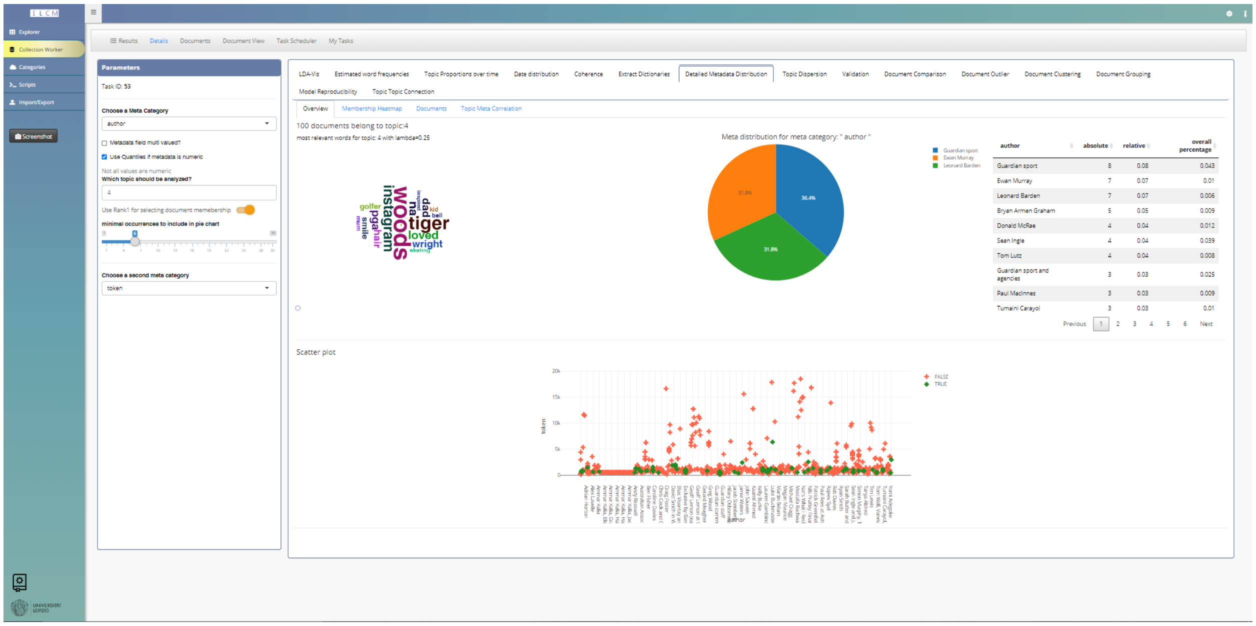
Category name for Topic 7: Juristic

porter mp legislation grants bowers senator lowe laws grant chair welfare parliamentary tweet greens arts employees frydenberg accounts coalition comeback

Dictionary name: sportnews

Figure 45: An example of a potential dictionary consisting of words from three different topics

- **Detailed Metadata Distribution**: This subtab allows the user to combine the documents' meta-information with the topic model information. The user has 4 different visualization options. In the first one, the user gets an overview of selected metadata information corresponding to a selected topic. In the parameter box on the left, they can choose between the shown meta-information (depending on the columns of the csv-file the collection was built on) and the number of the topic they wish to take a look at as well as a minimal occurrence of a certain value to appear in the plot. They also get the chance to chose a second metadata category. These parameters will then calculate the word cloud of the topic and a pie chart of the chosen meta category, and a data table below that the user finds the proportions of the first meta category to the second one. The next panel shows a *Membership Heatmap* of the first selected meta category above all topics. Next, the user will find a data table listing all the selected topic documents along with their metadata information. The last section is the *Topic Meta correlation*, where the user can see the relationship between all of the topics and the documents' found metadata categories.



Parameters

Task ID: 53

Choose a Meta Category: author

□ Metadata field multi valued?

☑ Use Quantiles if metadata is numeric

Not all values are numeric

Which topic should be analyzed? 4

Use Rank1 for selecting document memebership

minimal occurrences to include in pie chart: 1 – 6 – 30

Choose a second meta category: token

100 documents belong to topic:4

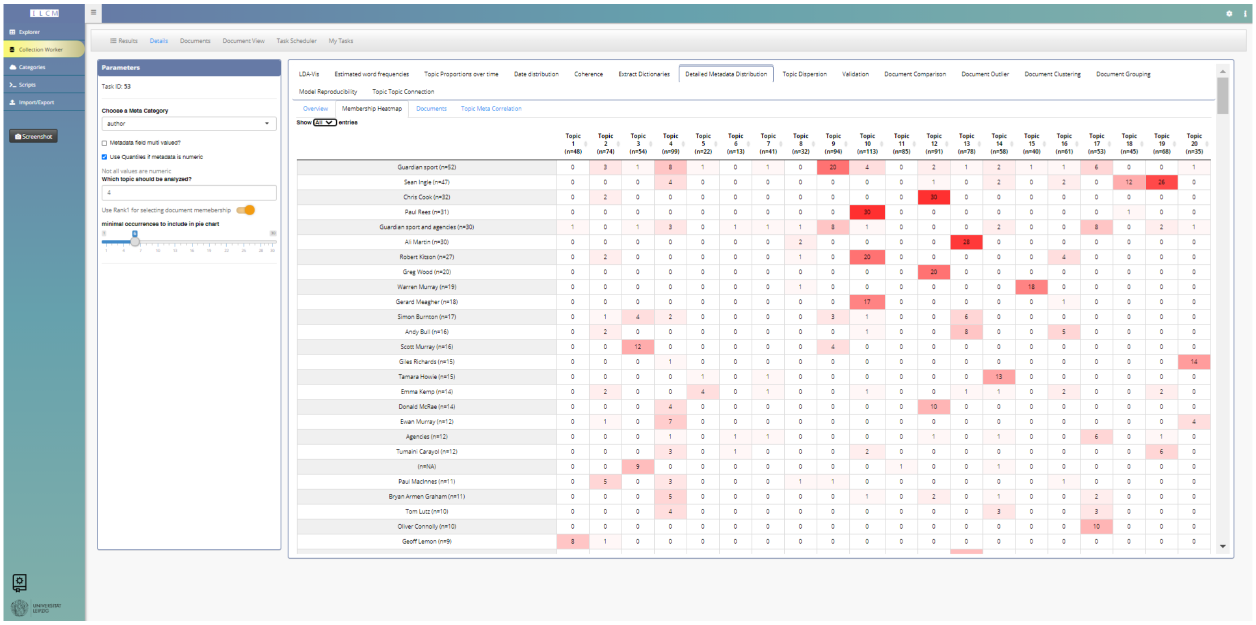
most relevant words for topic: 4 with lambda=0.25

Meta distribution for meta category: " author "

author	absolute	relative	overall percentage
Guardian sport	8	0.08	0.043
Ewan Murray	7	0.07	0.01
Leonard Barden	7	0.07	0.006
Bryan Armen Graham	5	0.05	0.009
Donald McRae	4	0.04	0.012
Sean Ingle	4	0.04	0.039
Tom Lutz	4	0.04	0.008
Guardian sport and agencies	3	0.03	0.025
Paul MacInnes	3	0.03	0.009
Tumaini Carayol	3	0.03	0.01

Previous 1 2 3 4 5 6 Next

Figure 46: Overview of metadata distributions for a selected Topic and metadata category



Parameters

Task ID: 53

Choose a Meta Category: author

☐ Metadata field multi valued?

☑ Use Quantiles if metadata is numeric

Not all values are numeric

Which topic should be analyzed? 4

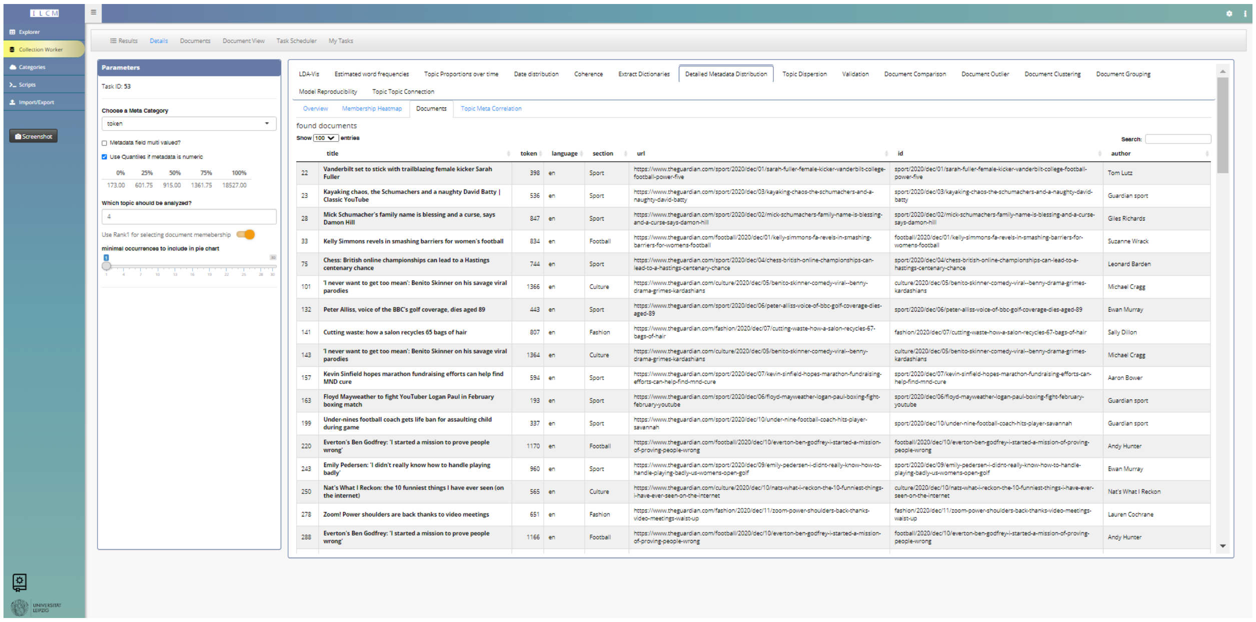
Use Rank1 for selecting document memebership ●

minimal occurrences to include in pie chart: 6

Show All entries

	Topic 1 (n=48)	Topic 2 (n=74)	Topic 3 (n=54)	Topic 4 (n=99)	Topic 5 (n=22)	Topic 6 (n=13)	Topic 7 (n=41)	Topic 8 (n=32)	Topic 9 (n=94)	Topic 10 (n=113)	Topic 11 (n=85)	Topic 12 (n=91)	Topic 13 (n=78)	Topic 14 (n=58)	Topic 15 (n=40)	Topic 16 (n=61)	Topic 17 (n=53)	Topic 18 (n=45)	Topic 19 (n=68)	Topic 20 (n=35)
Guardian sport (n=52)	0	3	1	8	1	0	1	0	20	4	0	2	1	2	1	1	6	0	0	1
Sean Ingle (n=47)	0	0	0	4	0	0	0	0	0	0	0	1	0	2	0	2	0	12	26	0
Chris Cook (n=32)	0	2	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0
Paul Rees (n=31)	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	1	0	0
Guardian sport and agencies (n=30)	1	0	1	3	0	1	1	1	8	1	0	0	0	2	0	0	8	0	2	1
Ali Martin (n=30)	0	0	0	0	0	0	0	2	0	0	0	0	28	0	0	0	0	0	0	0
Robert Kitson (n=27)	0	2	0	0	0	0	0	1	0	20	0	0	0	0	0	4	0	0	0	0
Greg Wood (n=20)	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
Warren Murray (n=19)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	18	0	0	0	0	0
Gerard Meagher (n=18)	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	1	0	0	0	0
Simon Burnton (n=17)	0	1	4	2	0	0	0	0	3	1	0	0	6	0	0	0	0	0	0	0
Andy Bull (n=16)	0	2	0	0	0	0	0	0	0	1	0	0	8	0	0	5	0	0	0	0
Scott Murray (n=16)	0	0	12	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
Giles Richards (n=15)	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
Tamara Howie (n=15)	0	0	0	0	1	0	1	0	0	0	0	0	0	13	0	0	0	0	0	0
Emma Kemp (n=14)	0	2	0	0	4	0	1	0	0	1	0	0	1	1	0	2	0	0	2	0
Donald McRae (n=14)	0	0	0	4	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0
Ewan Murray (n=12)	0	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Agencies (n=12)	0	0	0	1	0	1	1	0	0	0	0	1	0	1	0	0	6	0	1	0
Tumaini Carayol (n=12)	0	0	0	3	0	1	0	0	0	2	0	0	0	0	0	0	0	0	6	0
(n=NA)	0	0	9	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
Paul MacInnes (n=11)	0	5	0	3	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0
Bryan Armen Graham (n=11)	0	0	0	5	0	0	0	0	0	1	0	2	0	1	0	0	2	0	0	0
Tom Lutz (n=10)	0	0	0	4	0	0	0	0	0	0	0	0	0	3	0	0	3	0	0	0
Oliver Connolly (n=10)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
Geoff Lemon (n=9)	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 47: Heatmap of proportion of a selected metadata category over the topics



Parameters

Task ID: 53

Choose a Meta Category: token

☐ Metadata field multi valued?

☑ Use Quantiles if metadata is numeric

0%	25%	50%	75%	100%
173.00	601.75	915.00	1361.75	18527.00

Which topic should be analyzed? 4

Use Rank1 for selecting document memebership ●

minimal occurrences to include in pie chart: 1

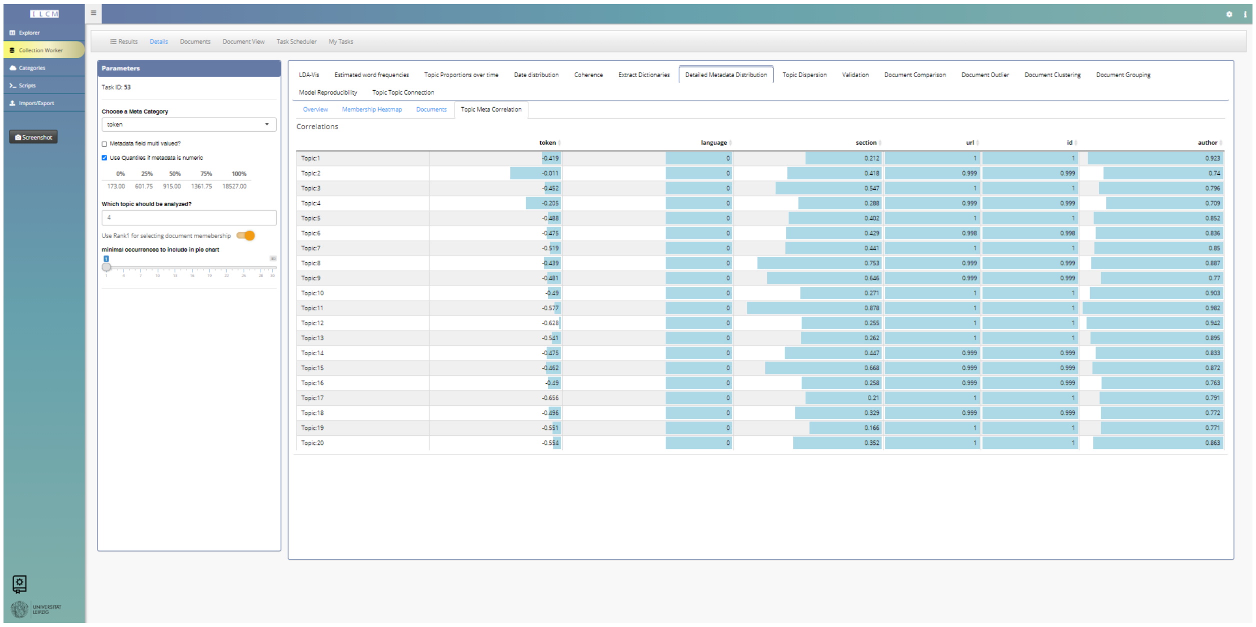
found documents

Show 100 entries

Search:

	title	token	language	section	url	id	author
22	Vanderbilt set to stick with trailblazing female kicker Sarah Fuller	398	en	Sport	https://www.theguardian.com/sport/2020/dec/01/sarah-fuller-female-kicker-vanderbilt-college-football-power-five	sport/2020/dec/01/sarah-fuller-female-kicker-vanderbilt-college-football-power-five	Tom Lutz
23	Kayaking chaos, the Schumachers and a naughty David Batty | Classic YouTube	536	en	Sport	https://www.theguardian.com/sport/2020/dec/03/kayaking-chaos-the-schumachers-and-a-naughty-david-batty	sport/2020/dec/03/kayaking-chaos-the-schumachers-and-a-naughty-david-batty	Guardian sport
28	Mick Schumacher's family name is blessing and a curse, says Damon Hill	847	en	Sport	https://www.theguardian.com/sport/2020/dec/02/mick-schumachers-family-name-is-blessing-and-a-curse-says-damon-hill	sport/2020/dec/02/mick-schumachers-family-name-is-blessing-and-a-curse-says-damon-hill	Giles Richards
33	Kelly Simmons revels in smashing barriers for women's football	834	en	Football	https://www.theguardian.com/football/2020/dec/01/kelly-simmons-fa-revels-in-smashing-barriers-for-womens-football	football/2020/dec/01/kelly-simmons-fa-revels-in-smashing-barriers-for-womens-football	Suzanne Wrack
75	Chess: British online championships can lead to a Hastings centenary chance	744	en	Sport	https://www.theguardian.com/sport/2020/dec/04/chess-british-online-championships-can-lead-to-a-hastings-centenary-chance	sport/2020/dec/04/chess-british-online-championships-can-lead-to-a-hastings-centenary-chance	Leonard Barden
101	'I never want to get too mean': Benito Skinner on his savage viral parodies	1366	en	Culture	https://www.theguardian.com/culture/2020/dec/05/benito-skinner-comedy-viral--benny-drama-grimes-kardashians	culture/2020/dec/05/benito-skinner-comedy-viral--benny-drama-grimes-kardashians	Michael Cragg
132	Peter Alliss, voice of the BBC's golf coverage, dies aged 89	443	en	Sport	https://www.theguardian.com/sport/2020/dec/06/peter-alliss-voice-of-bbc-golf-coverage-dies-aged-89	sport/2020/dec/06/peter-alliss-voice-of-bbc-golf-coverage-dies-aged-89	Ewan Murray
141	Cutting waste: how a salon recycles 65 bags of hair	807	en	Fashion	https://www.theguardian.com/fashion/2020/dec/07/cutting-waste-how-a-salon-recycles-67-bags-of-hair	fashion/2020/dec/07/cutting-waste-how-a-salon-recycles-67-bags-of-hair	Sally Dillon
143	'I never want to get too mean': Benito Skinner on his savage viral parodies	1364	en	Culture	https://www.theguardian.com/culture/2020/dec/05/benito-skinner-comedy-viral--benny-drama-grimes-kardashians	culture/2020/dec/05/benito-skinner-comedy-viral--benny-drama-grimes-kardashians	Michael Cragg
157	Kevin Sinfield hopes marathon fundraising efforts can help find MND cure	594	en	Sport	https://www.theguardian.com/sport/2020/dec/07/kevin-sinfield-hopes-marathon-fundraising-efforts-can-help-find-mnd-cure	sport/2020/dec/07/kevin-sinfield-hopes-marathon-fundraising-efforts-can-help-find-mnd-cure	Aaron Bower
163	Floyd Mayweather to fight YouTuber Logan Paul in February boxing match	193	en	Sport	https://www.theguardian.com/sport/2020/dec/06/floyd-mayweather-logan-paul-boxing-fight-february-youtube	sport/2020/dec/06/floyd-mayweather-logan-paul-boxing-fight-february-youtube	Guardian sport
199	Under-nines football coach gets life ban for assaulting child during game	337	en	Sport	https://www.theguardian.com/sport/2020/dec/10/under-nine-football-coach-hits-player-savannah	sport/2020/dec/10/under-nine-football-coach-hits-player-savannah	Guardian sport
220	Everton's Ben Godfrey: 'I started a mission to prove people wrong'	1170	en	Football	https://www.theguardian.com/football/2020/dec/10/everton-ben-godfrey-i-started-a-mission-of-proving-people-wrong	football/2020/dec/10/everton-ben-godfrey-i-started-a-mission-of-proving-people-wrong	Andy Hunter
243	Emily Pedersen: 'I didn't really know how to handle playing badly'	960	en	Sport	https://www.theguardian.com/sport/2020/dec/09/emily-pedersen-i-didnt-really-know-how-to-handle-playing-badly-us-womens-open-golf	sport/2020/dec/09/emily-pedersen-i-didnt-really-know-how-to-handle-playing-badly-us-womens-open-golf	Ewan Murray
250	Nat's What I Reckon: the 10 funniest things I have ever seen (on the internet)	565	en	Culture	https://www.theguardian.com/culture/2020/dec/10/nats-what-i-reckon-the-10-funniest-things-i-have-ever-seen-on-the-internet	culture/2020/dec/10/nats-what-i-reckon-the-10-funniest-things-i-have-ever-seen-on-the-internet	Nat's What I Reckon
278	Zoom! Power shoulders are back thanks to video meetings	651	en	Fashion	https://www.theguardian.com/fashion/2020/dec/11/zoom-power-shoulders-back-thanks-video-meetings-waist-up	fashion/2020/dec/11/zoom-power-shoulders-back-thanks-video-meetings-waist-up	Lauren Cochrane
288	Everton's Ben Godfrey: 'I started a mission to prove people wrong'	1166	en	Football	https://www.theguardian.com/football/2020/dec/10/everton-ben-godfrey-i-started-a-mission-of-proving-people-wrong	football/2020/dec/10/everton-ben-godfrey-i-started-a-mission-of-proving-people-wrong	Andy Hunter

Figure 48: Insides in the documents of one topic and their metadata



Parameters

Task ID: 53

Choose a Meta Category: token

□ Metadata field multi valued?

☑ Use Quantiles if metadata is numeric

0%	25%	50%	75%	100%
173.00	601.75	915.00	1361.75	18527.00

Which topic should be analyzed? 4

Use Rank1 for selecting document memebership ●

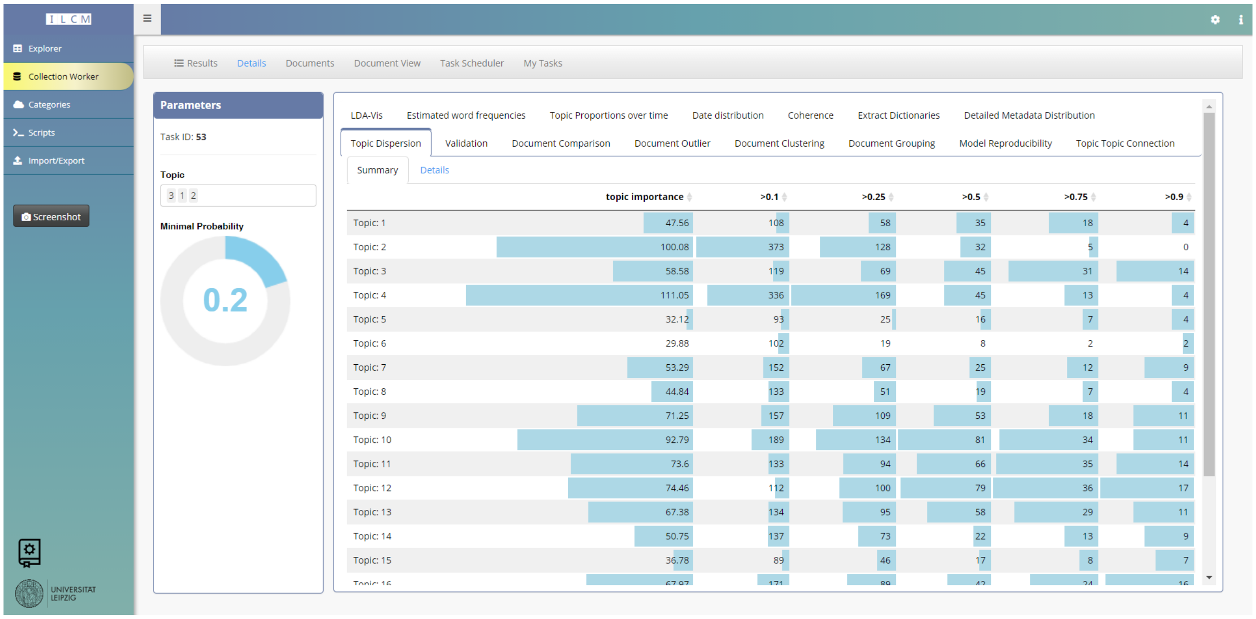
minimal occurrences to include in pie chart: 1

Correlations

	token	language	section	url	id	author
Topic:1	-0.419	0	0.212	1	1	0.923
Topic:2	-0.011	0	0.418	0.999	0.999	0.74
Topic:3	-0.452	0	0.547	1	1	0.796
Topic:4	-0.205	0	0.288	0.999	0.999	0.709
Topic:5	-0.488	0	0.402	1	1	0.852
Topic:6	-0.475	0	0.429	0.998	0.998	0.836
Topic:7	-0.519	0	0.441	1	1	0.85
Topic:8	-0.439	0	0.753	0.999	0.999	0.887
Topic:9	-0.481	0	0.646	0.999	0.999	0.77
Topic:10	-0.49	0	0.271	1	1	0.903
Topic:11	-0.577	0	0.878	1	1	0.982
Topic:12	-0.628	0	0.255	1	1	0.942
Topic:13	-0.541	0	0.262	1	1	0.895
Topic:14	-0.475	0	0.447	0.999	0.999	0.833
Topic:15	-0.462	0	0.668	0.999	0.999	0.872
Topic:16	-0.49	0	0.258	0.999	0.999	0.763
Topic:17	-0.656	0	0.21	1	1	0.791
Topic:18	-0.496	0	0.329	0.999	0.999	0.772
Topic:19	-0.551	0	0.166	1	1	0.771
Topic:20	-0.554	0	0.352	1	1	0.863

Figure 49: Proportion of different metadata categories over different topica

- **Topic Dispersion**: This section shows the importance of different topics for all existing documents. First, the user sees a summary of all the topics and their importance for the documents. The columns show the overall importance and specific thresholds for the importance of a topic and the exact number of documents that evaluate the importance of the topic above this threshold. In the parameter box on the left, the user finds the minimal probability a topic must have to be part of a document's topic set (set when starting the topic model task). Above that, the user finds a text-box to select the topics they want a detailed analysis for. These will show in the *Details* window. Here the user finds different plots that show the topic probability distribution for the selected topics.



Parameters

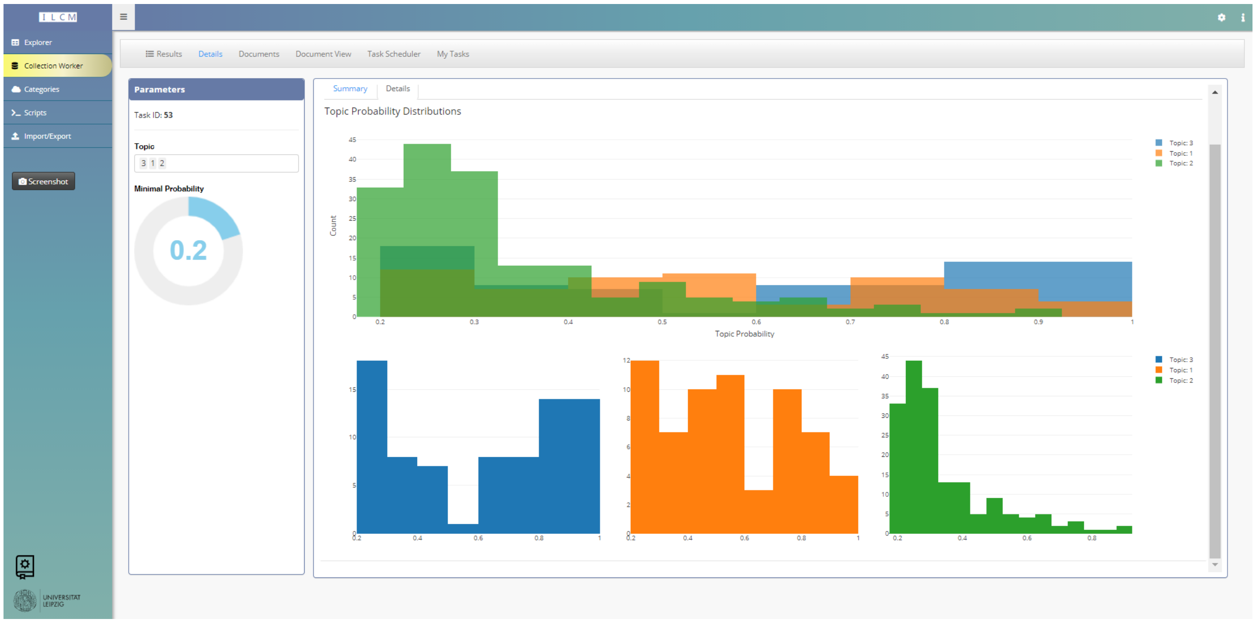
Task ID: 53

Topic: 3 1 2

Minimal Probability: 0.2

	topic importance	>0.1	>0.25	>0.5	>0.75	>0.9
Topic: 1	47.56	108	58	35	18	4
Topic: 2	100.08	373	128	32	5	0
Topic: 3	58.58	119	69	45	31	14
Topic: 4	111.05	336	169	45	13	4
Topic: 5	32.12	93	25	16	7	4
Topic: 6	29.88	102	19	8	2	2
Topic: 7	53.29	152	67	25	12	9
Topic: 8	44.84	133	51	19	7	4
Topic: 9	71.25	157	109	53	18	11
Topic: 10	92.79	189	134	81	34	11
Topic: 11	73.6	133	94	66	35	14
Topic: 12	74.46	112	100	79	36	17
Topic: 13	67.38	134	95	58	29	11
Topic: 14	50.75	137	73	22	13	9
Topic: 15	36.78	89	46	17	8	7
Topic: 16	67.97	171	89	42	24	16

Figure 50: Overall view of the topic dispersion



Parameters

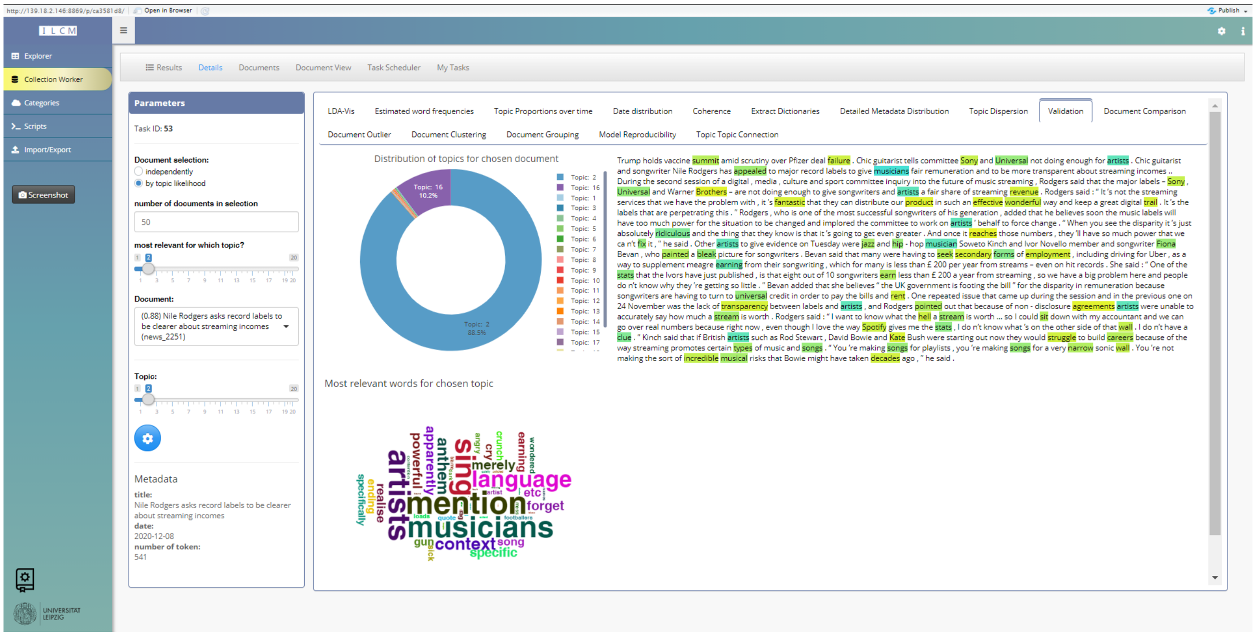
Task ID: 53

Topic: 3 1 2

Minimal Probability: 0.2

Figure 51: Detailes on topic dispersion for selected topics

- **Validation**: In this section, the user gets an overview of essential documents for a specific topic. Therefore, they can choose whether the documents should be selected independently from the currently selected topic or by topic likelihood, leading to documents that represent the chosen topic. When selecting the independent variant, the user can switch between all possible documents. In the topic likelihood part, the user first needs to choose which topic the most relevant topics should be calculated and how many of them should be calculated. In both cases, the system will plot for each document distribution of topics, a word cloud for the most relevant words of a selected topic (see parameter box), and it will show the text of the document. In this text representation, the most important and least important words for a selected topic are highlighted. The user can change the colour of the highlights as well as the method for calculating these words. These features will show when the user clicks on the gear icon in the left parameter box. In the parameter box, the user will also find metadata information about the current document. This allows the user to validate whether the current topic model has calculated useful topics or not.



Parameters

Task ID: 53

Document selection:
○ independently
● by topic likelihood

number of documents in selection
50

most relevant for which topic?
1 2 20

Document:
(0.88) Nile Rodgers asks record labels to be clearer about streaming incomes (news_2251)

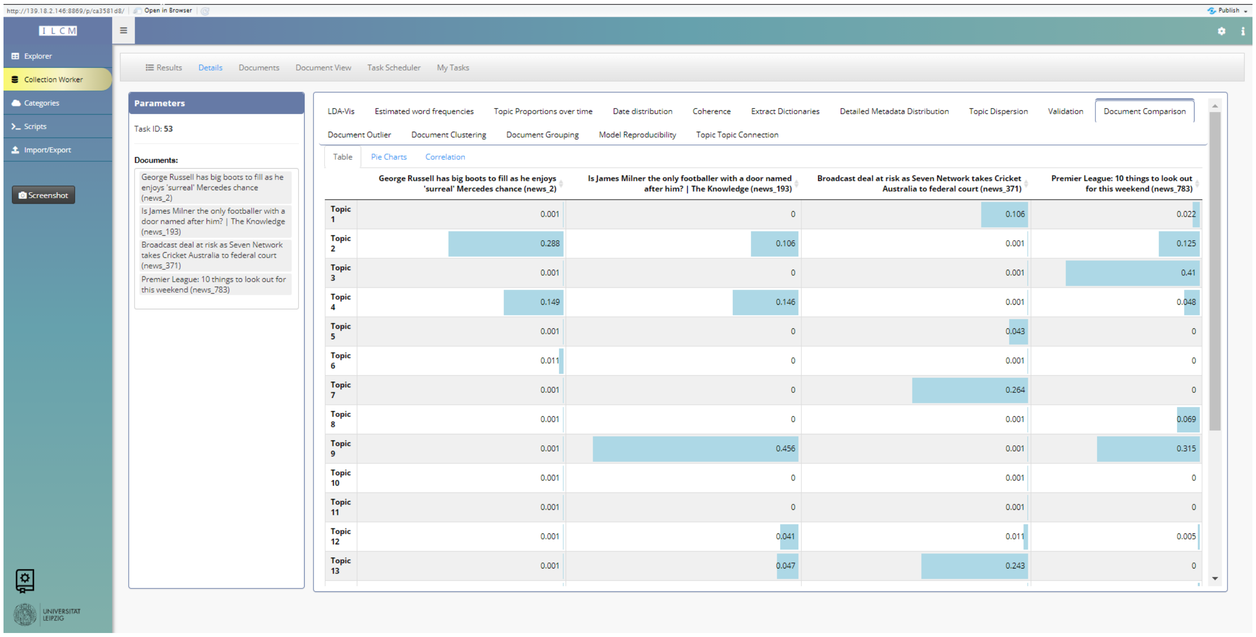
Topic:
1 2 20

Metadata
title:
Nile Rodgers asks record labels to be clearer about streaming incomes
date:
2020-12-08
number of token:
541

Trump holds vaccine summit amid scrutiny over Pfizer deal failure . Chic guitarist tells committee Sony and Universal not doing enough for artists . Chic guitarist and songwriter Nile Rodgers has appealed to major record labels to give musicians fair remuneration and to be more transparent about streaming incomes .. During the second session of a digital , media , culture and sport committee inquiry into the future of music streaming , Rodgers said that the major labels – Sony , Universal and Warner Brothers – are not doing enough to give songwriters and artists a fair share of streaming revenue . Rodgers said : “ It ’s not the streaming services that we have the problem with , it ’s fantastic that they can distribute our product in such an effective wonderful way and keep a great digital trail . It ’s the labels that are perpetrating this . ” Rodgers , who is one of the most successful songwriters of his generation , added that he believes soon the music labels will have too much power for the situation to be changed and implored the committee to work on artists ’ behalf to force change . “ When you see the disparity it ’s just absolutely ridiculous and the thing that they know is that it ’s going to get even greater . And once it reaches those numbers , they ’ll have so much power that we ca n’t fix it , ” he said . Other artists to give evidence on Tuesday were jazz and hip - hop musician Soweto Kinch and Ivor Novello member and songwriter Fiona Bevan , who painted a bleak picture for songwriters . Bevan said that many were having to seek secondary forms of employment , including driving for Uber , as a way to supplement meagre earning from their songwriting , which for many is less than £ 200 per year from streams – even on hit records . She said : “ One of the stats that the Ivors have just published , is that eight out of 10 songwriters earn less than £ 200 a year from streaming , so we have a big problem here and people do n’t know why they ’re getting so little . ” Bevan added that she believes “ the UK government is footing the bill ” for the disparity in remuneration because songwriters are having to turn to universal credit in order to pay the bills and rent . One repeated issue that came up during the session and in the previous one on 24 November was the lack of transparency between labels and artists , and Rodgers pointed out that because of non - disclosure agreements artists were unable to accurately say how much a stream is worth . Rodgers said : “ I want to know what the hell a stream is worth … so I could sit down with my accountant and we can go over real numbers because right now , even though I love the way Spotify gives me the stats , I do n’t know what ’s on the other side of that wall . I do n’t have a clue . ” Kinch said that if British artists such as Rod Stewart , David Bowie and Kate Bush were starting out now they would struggle to build careers because of the way streaming promotes certain types of music and songs . “ You ’re making songs for playlists , you ’re making songs for a very narrow sonic wall . You ’re not making the sort of incredible musical risks that Bowie might have taken decades ago , ” he said .

Figure 52: Overview of a selected document in the validation section

- **Document Comparison**: Here, the user can select specific documents for comparison. They can choose between 3 different representations: A table view or a pie chart of the selected documents and the proportion of the topics within these documents, and the correlation of the documents. In the last part, the user can select a correlation method and a colour scheme. The euclidean and cosine similarity for the documents will also be shown in this section.



Parameters

Task ID: **53**

Documents:

George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)

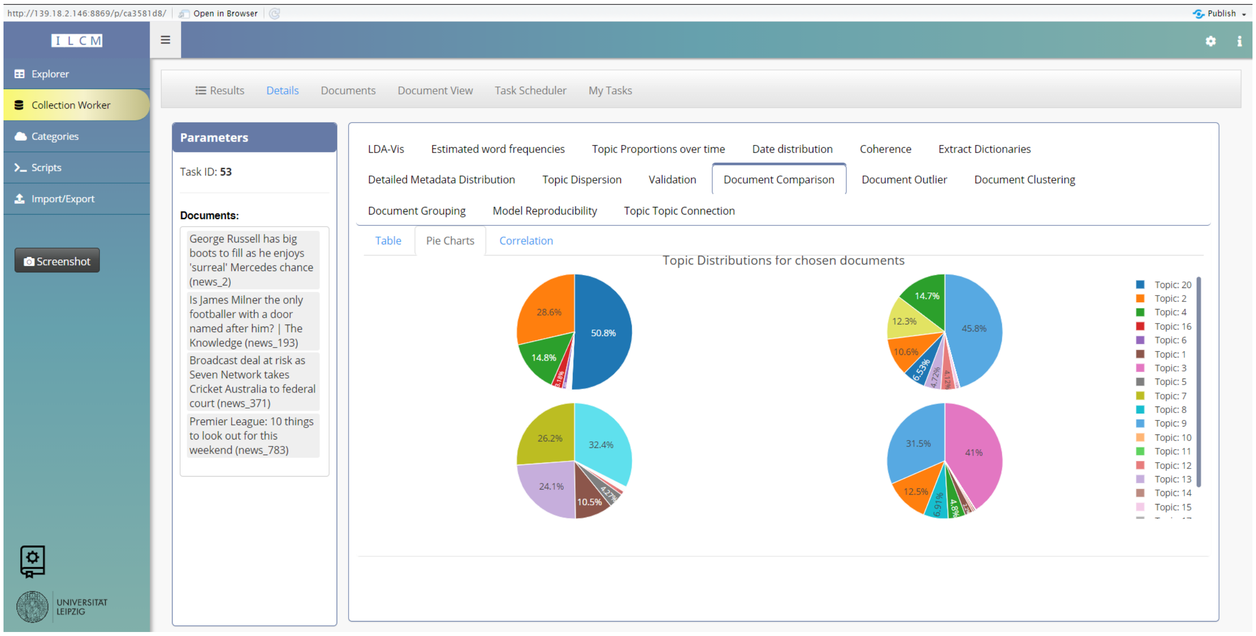
Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)

Broadcast deal at risk as Seven Network takes Cricket Australia to federal court (news_371)

Premier League: 10 things to look out for this weekend (news_783)

	George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)	Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)	Broadcast deal at risk as Seven Network takes Cricket Australia to federal court (news_371)	Premier League: 10 things to look out for this weekend (news_783)
Topic 1	0.001	0	0.106	0.022
Topic 2	0.288	0.106	0.001	0.125
Topic 3	0.001	0	0.001	0.41
Topic 4	0.149	0.146	0.001	0.048
Topic 5	0.001	0	0.043	0
Topic 6	0.011	0	0.001	0
Topic 7	0.001	0	0.264	0
Topic 8	0.001	0	0.001	0.069
Topic 9	0.001	0.456	0.001	0.315
Topic 10	0.001	0	0.001	0
Topic 11	0.001	0	0.001	0
Topic 12	0.001	0.041	0.011	0.005
Topic 13	0.001	0.047	0.243	0

Figure 53: Table representation of the proportion of topics whithin selected documents



Parameters

Task ID: **53**

Documents:

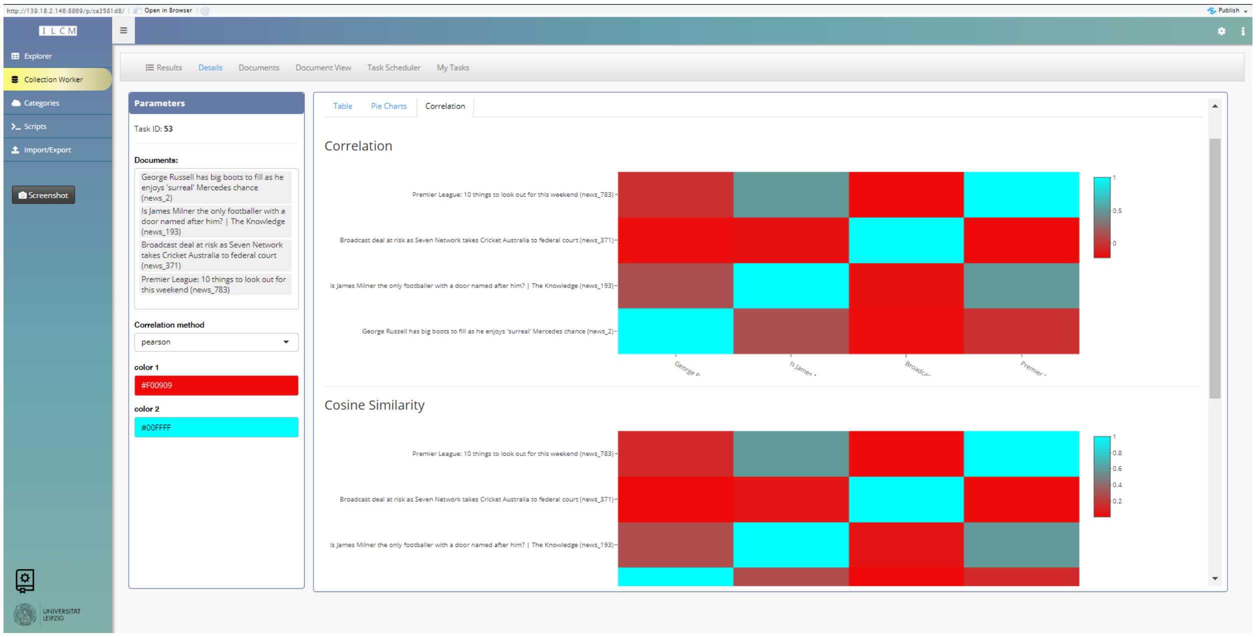
George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)

Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)

Broadcast deal at risk as Seven Network takes Cricket Australia to federal court (news_371)

Premier League: 10 things to look out for this weekend (news_783)

Figure 54: pie chart of the proportion of topics whithin selected documents


http://139.18.2.146:8869/p/ca3581d8/
Open in Browser
Publish
ILCM
Explorer
Collection Worker
Categories
Scripts
Import/Export
Screenshot
Results
Details
Documents
Document View
Task Scheduler
My Tasks
Table
Pie Charts
Correlation
Premier League: 10 things to look out for this weekend (news_783)
Broadcast deal at risk as Seven Network takes Cricket Australia to federal court (news_371)
Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)
George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)
George R
Is James
Broadcas
Premier
1
0.5
0
0.8
0.6
0.4
0.2

Parameters

Task ID: 53

Documents:

George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)
Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)
Broadcast deal at risk as Seven Network takes Cricket Australia to federal court (news_371)
Premier League: 10 things to look out for this weekend (news_783)

Correlation method: pearson

color 1: #F00909

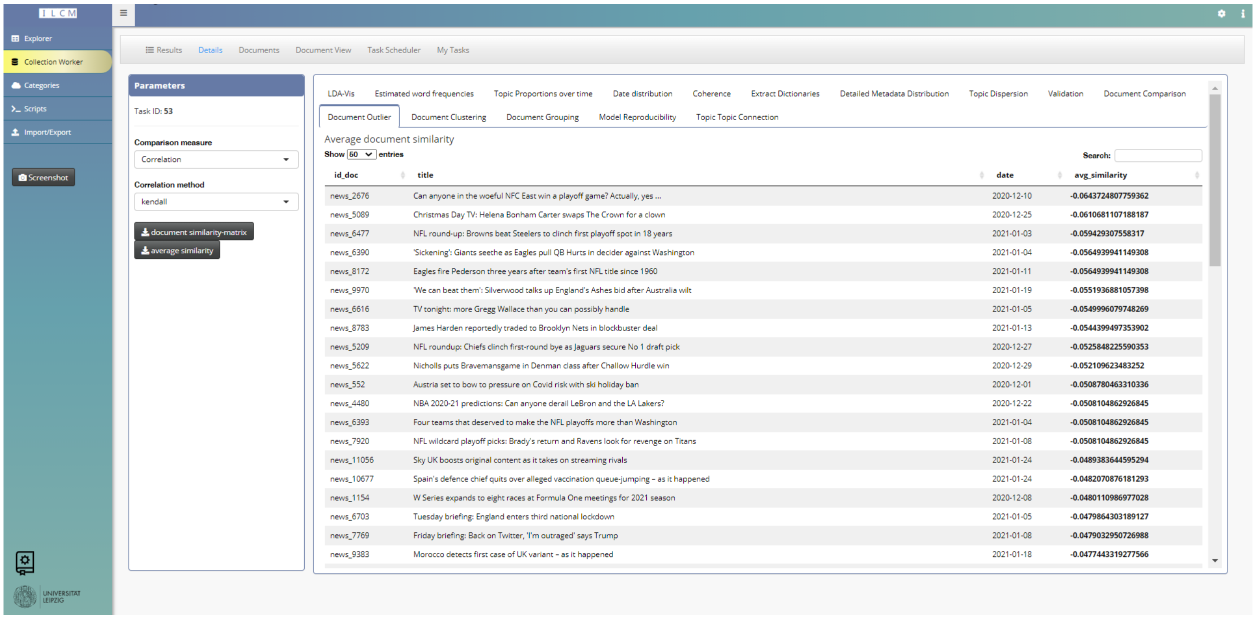
color 2: #00FFFF

Correlation

Cosine Similarity

Figure 55: Correlation between the selected documents

- **Document Outlier**: This part allows the user to download and view comparison measures between the documents. The user can select between 3 comparison measurements: Correlation, euclidean distance and cosine similarity in the left parameter box. After selecting a measurement, the plot will show a data table consisting of the documents and their metadata and the average similarity for the documents. The user can then download the document similarity matrix for the current settings or the average similarities.



Parameters

Task ID: 53

Comparison measure: Correlation

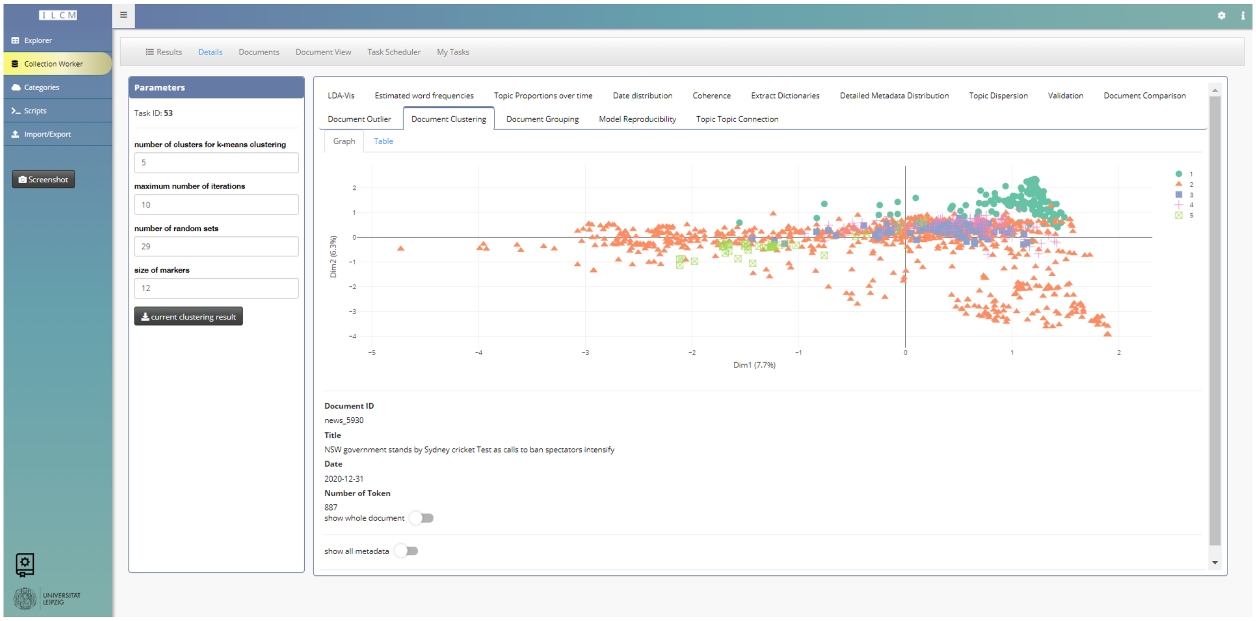
Correlation method: kendall

Average document similarity

id_doc	title	date	avg_similarity
news_2676	Can anyone in the woeful NFC East win a playoff game? Actually, yes ...	2020-12-10	-0.0643724807759362
news_5089	Christmas Day TV: Helena Bonham Carter swaps The Crown for a clown	2020-12-25	-0.0610681107188187
news_6477	NFL round-up: Browns beat Steelers to clinch first playoff spot in 18 years	2021-01-03	-0.059429307558317
news_6390	'Sickening': Giants seethe as Eagles pull QB Hurts in decider against Washington	2021-01-04	-0.0564939941149308
news_8172	Eagles fire Pederson three years after team's first NFL title since 1960	2021-01-11	-0.0564939941149308
news_9970	'We can beat them': Silverwood talks up England's Ashes bid after Australia wilt	2021-01-19	-0.0551936881057398
news_6616	TV tonight: more Gregg Wallace than you can possibly handle	2021-01-05	-0.0549996079748269
news_8783	James Harden reportedly traded to Brooklyn Nets in blockbuster deal	2021-01-13	-0.0544399497353902
news_5209	NFL roundup: Chiefs clinch first-round bye as Jaguars secure No 1 draft pick	2020-12-27	-0.0525848225590353
news_5622	Nicholls puts Bravemansgame in Denman class after Challow Hurdle win	2020-12-29	-0.052109623483252
news_552	Austria set to bow to pressure on Covid risk with ski holiday ban	2020-12-01	-0.0508780463310336
news_4480	NBA 2020-21 predictions: Can anyone derail LeBron and the LA Lakers?	2020-12-22	-0.0508104862926845
news_6393	Four teams that deserved to make the NFL playoffs more than Washington	2021-01-04	-0.0508104862926845
news_7920	NFL wildcard playoff picks: Brady's return and Ravens look for revenge on Titans	2021-01-08	-0.0508104862926845
news_11056	Sky UK boosts original content as it takes on streaming rivals	2021-01-24	-0.0489383644595294
news_10677	Spain's defence chief quits over alleged vaccination queue-jumping – as it happened	2021-01-24	-0.0482070876181293
news_1154	W Series expands to eight races at Formula One meetings for 2021 season	2020-12-08	-0.0480110986977028
news_6703	Tuesday briefing: England enters third national lockdown	2021-01-05	-0.0479864303189127
news_7769	Friday briefing: Back on Twitter, 'I'm outraged' says Trump	2021-01-08	-0.0479032950726988
news_9383	Morocco detects first case of UK variant – as it happened	2021-01-18	-0.0477443319277566

Figure 56: Table representation of document comparision

- **Document Clustering**: This section allows the user to cluster their data. In the parameter box, they need to choose the number of clusters, the number of iterations, and the number of random seeds. After selecting these, a Graph will show up with different markers correlating to the number of clusters. This gives the user the possibility to see the spreading of the different clusters. The user can also change the size of the icons in the parameter box. If they want to get more information about a specific point in the graph, they can hover over the marker and get the corresponding document's title. For more information, they can also click on the marker to see the document's metadata or the whole text. In the other subtab, the user can see the clusters as data tables where each column represents one cluster, and the rows represent the documents. If the user is satisfied with the clustering, they also can download the current clustering result in the parameter box.



Task ID: 53

number of clusters for k-means clustering

5

maximum number of iterations

10

number of random sets

29

size of markers

12

Document ID

news_5930

Title

NSW government stands by Sydney cricket Test as calls to ban spectators intensify

Date

2020-12-31

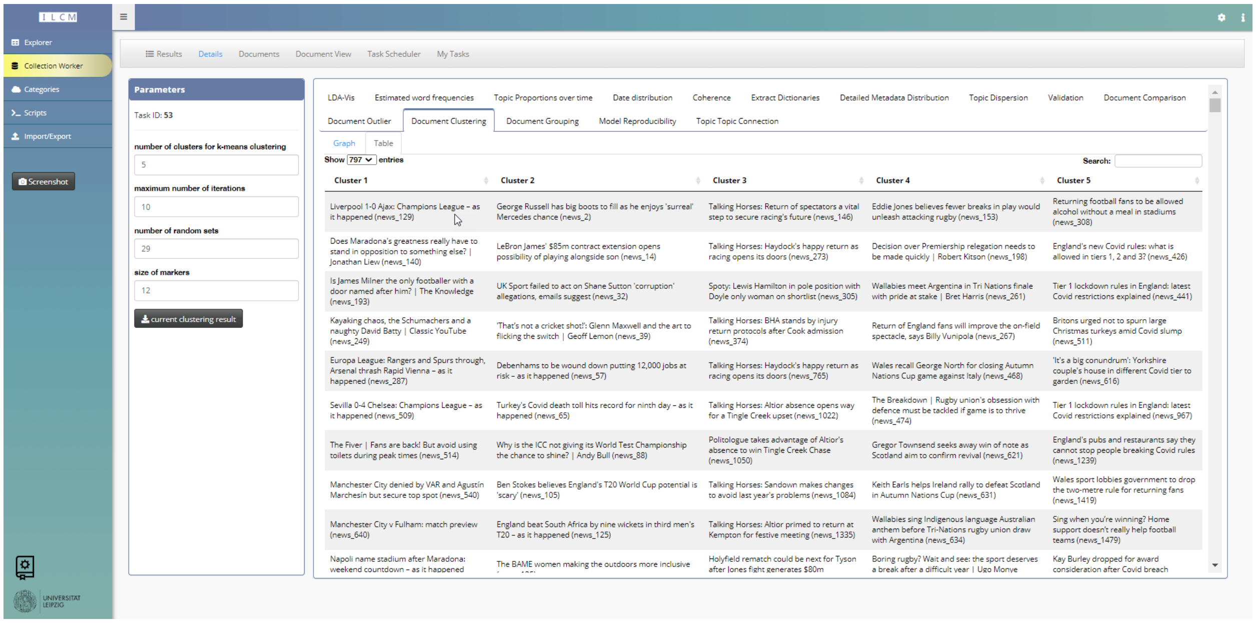
Number of Token

887

show whole document

show all metadata

Figure 57: Graph visulization of 5 clusters



Parameters

Task ID: 53

number of clusters for k-means clustering: 5

maximum number of iterations: 10

number of random sets: 29

size of markers: 12

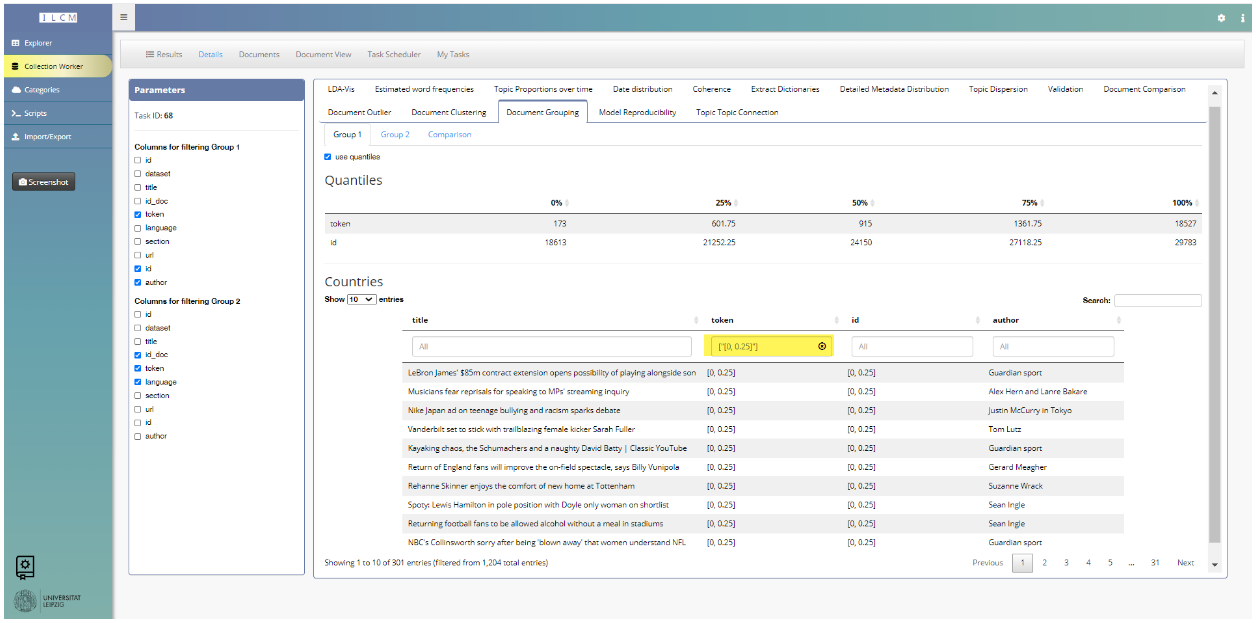
Show 797 entries

Search:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Liverpool 1-0 Ajax: Champions League – as it happened (news_129)	George Russell has big boots to fill as he enjoys 'surreal' Mercedes chance (news_2)	Talking Horses: Return of spectators a vital step to secure racing's future (news_146)	Eddie Jones believes fewer breaks in play would unleash attacking rugby (news_153)	Returning football fans to be allowed alcohol without a meal in stadiums (news_308)
Does Maradona's greatness really have to stand in opposition to something else? | Jonathan Liew (news_140)	LeBron James' $85m contract extension opens possibility of playing alongside son (news_14)	Talking Horses: Haydock's happy return as racing opens its doors (news_273)	Decision over Premiership relegation needs to be made quickly | Robert Kitson (news_198)	England's new Covid rules: what is allowed in tiers 1, 2 and 3? (news_426)
Is James Milner the only footballer with a door named after him? | The Knowledge (news_193)	UK Sport failed to act on Shane Sutton 'corruption' allegations, emails suggest (news_32)	Spoty: Lewis Hamilton in pole position with Doyle only woman on shortlist (news_305)	Wallabies meet Argentina in Tri Nations finale with pride at stake | Bret Harris (news_261)	Tier 1 lockdown rules in England: latest Covid restrictions explained (news_441)
Kayaking chaos, the Schumachers and a naughty David Batty | Classic YouTube (news_249)	'That's not a cricket shot!': Glenn Maxwell and the art to flicking the switch | Geoff Lemon (news_39)	Talking Horses: BHA stands by injury return protocols after Cook admission (news_374)	Return of England fans will improve the on-field spectacle, says Billy Vunipola (news_267)	Britons urged not to spurn large Christmas turkeys amid Covid slump (news_511)
Europa League: Rangers and Spurs through, Arsenal thrash Rapid Vienna – as it happened (news_287)	Debenhams to be wound down putting 12,000 jobs at risk – as it happened (news_57)	Talking Horses: Haydock's happy return as racing opens its doors (news_765)	Wales recall George North for closing Autumn Nations Cup game against Italy (news_468)	'It's a big conundrum': Yorkshire couple's house in different Covid tier to garden (news_616)
Sevilla 0-4 Chelsea: Champions League – as it happened (news_509)	Turkey's Covid death toll hits record for ninth day – as it happened (news_65)	Talking Horses: Altior absence opens way for a Tingle Creek upset (news_1022)	The Breakdown | Rugby union's obsession with defence must be tackled if game is to thrive (news_474)	Tier 1 lockdown rules in England: latest Covid restrictions explained (news_967)
The Fiver | Fans are back! But avoid using toilets during peak times (news_514)	Why is the ICC not giving its World Test Championship the chance to shine? | Andy Bull (news_88)	Politologue takes advantage of Altior's absence to win Tingle Creek Chase (news_1050)	Gregor Townsend seeks away win of note as Scotland aim to confirm revival (news_621)	England's pubs and restaurants say they cannot stop people breaking Covid rules (news_1239)
Manchester City denied by VAR and Agustín Marchesín but secure top spot (news_540)	Ben Stokes believes England's T20 World Cup potential is 'scary' (news_105)	Talking Horses: Sandown makes changes to avoid last year's problems (news_1084)	Keith Earls helps Ireland rally to defeat Scotland in Autumn Nations Cup (news_631)	Wales sport lobbies government to drop the two-metre rule for returning fans (news_1419)
Manchester City v Fulham: match preview (news_640)	England beat South Africa by nine wickets in third men's T20 – as it happened (news_125)	Talking Horses: Altior primed to return at Kempton for festive meeting (news_1335)	Wallabies sing Indigenous language Australian anthem before Tri-Nations rugby union draw with Argentina (news_634)	Sing when you're winning? Home support doesn't really help football teams (news_1479)
Napoli name stadium after Maradona: weekend countdown – as it happened	The BAME women making the outdoors more inclusive	Holyfield rematch could be next for Tyson after Jones fight generates $80m	Boring rugby? Wait and see: the sport deserves a break after a difficult year | Ugo Monye	Kay Burley dropped for award consideration after Covid breach

Figure 58: Table visulization of 5 clusters

- **Document Grouping**: In this section, the user can split the documents into groups depending on their metadata parameters. This comes in handy when the user wants to compare different groups of documents to the calculated topics. Let us say the user wishes to see the different impact of documents with different token sizes. Therefore the user will select in group one and two at least the token parameter. After this, a data table will be shown, and on top of it, the user will find a search bar. By clicking on the bar, the possible token lengths will appear, and the user can select a size for group one and in the next subtab for group two. In the *Comparison* subtab, the user will see the final comparison between the documents. They will find two-word clouds on top of the page summarizing the two groups' documents titles below that the user will see a data table or a heatmap of the impact the two groups will have on the calculated topics and overall correlation between the two groups.



Parameters

Task ID: **68**

Columns for filtering Group 1

- □ id
- □ dataset
- □ title
- □ id_doc
- ☑ token
- □ language
- □ section
- □ url
- ☑ id
- ☑ author

Columns for filtering Group 2

- □ id
- □ dataset
- □ title
- ☑ id_doc
- ☑ token
- ☑ language
- □ section
- □ url
- □ id
- □ author

☑ use quantiles

Quantiles

	0%	25%	50%	75%	100%
token	173	601.75	915	1361.75	18527
id	18613	21252.25	24150	27118.25	29783

Countries

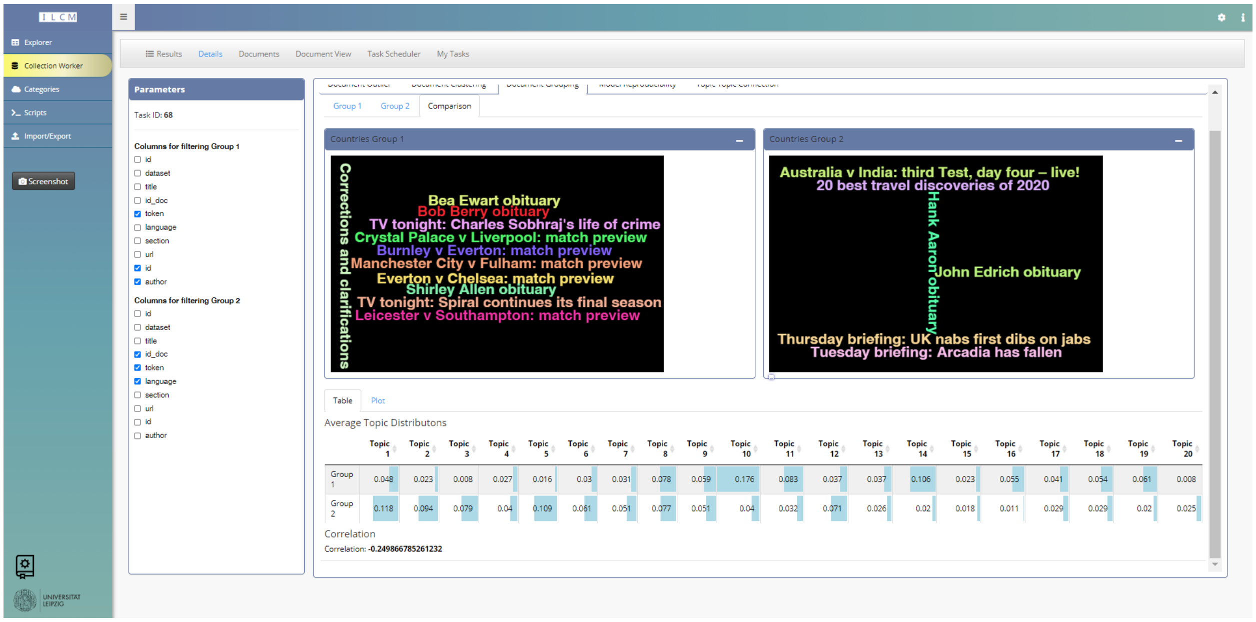
Show 10 entries Search:

title	token	id	author
All	["[0, 0.25]"]	All	All
LeBron James' $85m contract extension opens possibility of playing alongside son	[0, 0.25]	[0, 0.25]	Guardian sport
Musicians fear reprisals for speaking to MPs' streaming inquiry	[0, 0.25]	[0, 0.25]	Alex Hern and Lanre Bakare
Nike Japan ad on teenage bullying and racism sparks debate	[0, 0.25]	[0, 0.25]	Justin McCurry in Tokyo
Vanderbilt set to stick with trailblazing female kicker Sarah Fuller	[0, 0.25]	[0, 0.25]	Tom Lutz
Kayaking chaos, the Schumachers and a naughty David Batty | Classic YouTube	[0, 0.25]	[0, 0.25]	Guardian sport
Return of England fans will improve the on-field spectacle, says Billy Vunipola	[0, 0.25]	[0, 0.25]	Gerard Meagher
Rehanne Skinner enjoys the comfort of new home at Tottenham	[0, 0.25]	[0, 0.25]	Suzanne Wrack
Spoty: Lewis Hamilton in pole position with Doyle only woman on shortlist	[0, 0.25]	[0, 0.25]	Sean Ingle
Returning football fans to be allowed alcohol without a meal in stadiums	[0, 0.25]	[0, 0.25]	Sean Ingle
NBC's Collinsworth sorry after being 'blown away' that women understand NFL	[0, 0.25]	[0, 0.25]	Guardian sport

Showing 1 to 10 of 301 entries (filtered from 1,204 total entries)

Previous 1 2 3 4 5 ... 31 Next

Figure 59: Data table of group one after specifying the token parameter


ILCM
Explorer
Collection Worker
Categories
Scripts
Import/Export
Screenshot
Results
Details
Documents
Document View
Task Scheduler
My Tasks
Document Outlier
Document Clustering
Document Grouping
Model Reproducibility
Topic Topic Connection
Group 1
Group 2
Comparison
Countries Group 1
Corrections and clarifications
Bea Ewart obituary
Bob Berry obituary
TV tonight: Charles Sobhraj's life of crime
Crystal Palace v Liverpool: match preview
Burnley v Everton: match preview
Manchester City v Fulham: match preview
Everton v Chelsea: match preview
Shirley Allen obituary
TV tonight: Spiral continues its final season
Leicester v Southampton: match preview
Countries Group 2
Australia v India: third Test, day four – live!
20 best travel discoveries of 2020
Hank Aaron obituary
John Edrich obituary
Thursday briefing: UK nabs first dibs on jabs
Tuesday briefing: Arcadia has fallen
Table
Plot

Parameters

Task ID: **68**

Columns for filtering Group 1

- □ id
- □ dataset
- □ title
- □ id_doc
- ☑ token
- □ language
- □ section
- □ url
- ☑ id
- ☑ author

Columns for filtering Group 2

- □ id
- □ dataset
- □ title
- ☑ id_doc
- ☑ token
- ☑ language
- □ section
- □ url
- □ id
- □ author

Average Topic Distributons

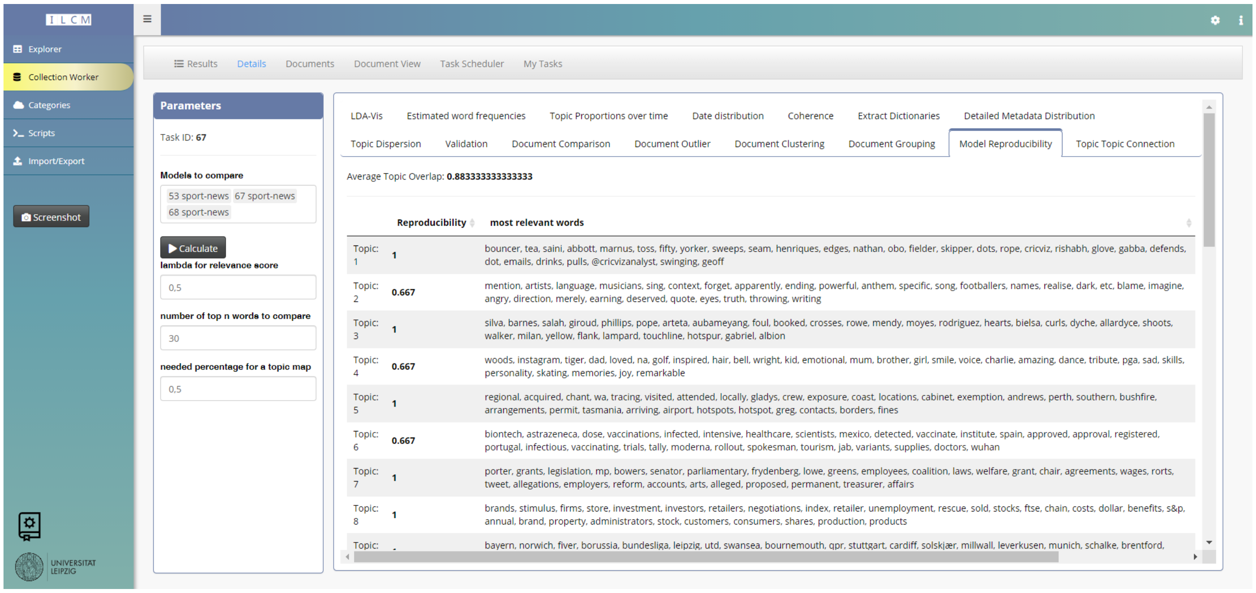
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
Group 1	0.048	0.023	0.008	0.027	0.016	0.03	0.031	0.078	0.059	0.176	0.083	0.037	0.037	0.106	0.023	0.055	0.041	0.054	0.061	0.008
Group 2	0.118	0.094	0.079	0.04	0.109	0.061	0.051	0.077	0.051	0.04	0.032	0.071	0.026	0.02	0.018	0.011	0.029	0.029	0.02	0.025

Correlation

Correlation: **-0.249866785261232**

Figure 60: Comparision of the two groups after specifying different tokens

- **Model Reproducibility**: This section will check the reproducibility of selected topic model analyses. The user needs to select the topic models he wants to compare and some other parameters, and then they need to click on *Calculate* to start the calculation process. Then a data table will show up that sums up the reproducibility of the different topics of the selected topic models. The user will also see the *Average Topic Overlap* above this table. This can help in term of the reliability of the found results.



Parameters

Task ID: **67**

Models to compare

53 sport-news 67 sport-news 68 sport-news

lambda for relevance score

0.5

number of top n words to compare

30

needed percentage for a topic map

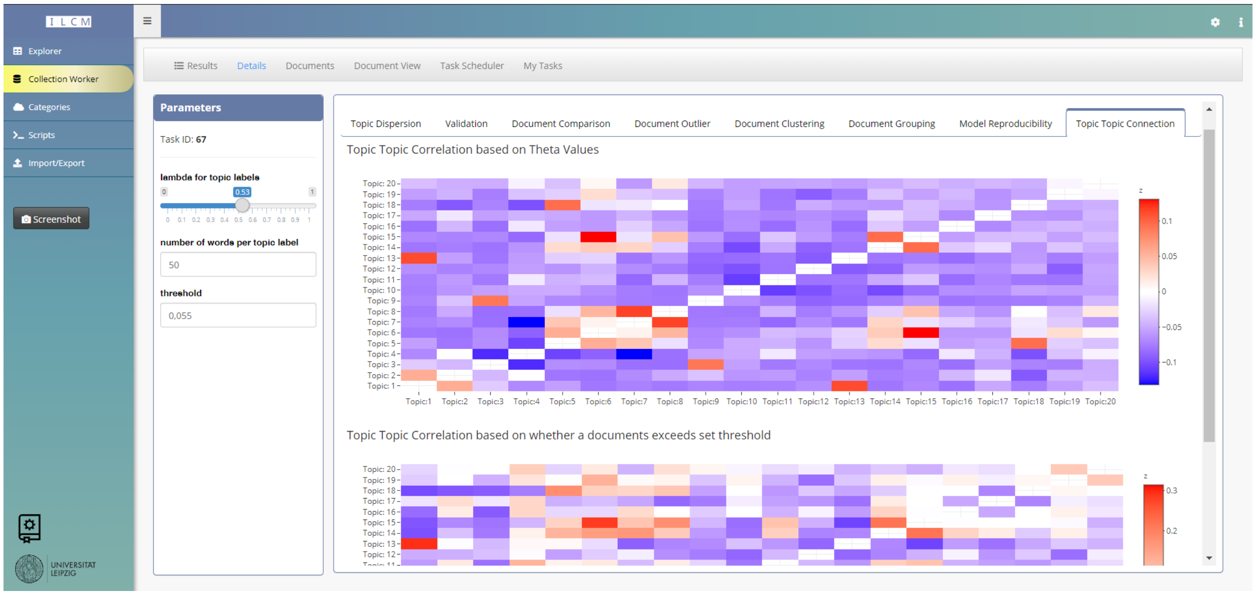
0.5

Average Topic Overlap: **0.883333333333333**

	Reproducibility	most relevant words
Topic: 1	**1**	bouncer, tea, saini, abbott, marnus, toss, fifty, yorker, sweeps, seam, henriques, edges, nathan, obo, fielder, skipper, dots, rope, cricviz, rishabh, glove, gabba, defends, dot, emails, drinks, pulls, @cricvizanalyst, swinging, geoff
Topic: 2	**0.667**	mention, artists, language, musicians, sing, context, forget, apparently, ending, powerful, anthem, specific, song, footballers, names, realise, dark, etc, blame, imagine, angry, direction, merely, earning, deserved, quote, eyes, truth, throwing, writing
Topic: 3	**1**	silva, barnes, salah, giroud, phillips, pope, arteta, aubameyang, foul, booked, crosses, rowe, mendy, moyes, rodriguez, hearts, bielsa, curls, dyche, allardyce, shoots, walker, milan, yellow, flank, lampard, touchline, hotspur, gabriel, albion
Topic: 4	**0.667**	woods, instagram, tiger, dad, loved, na, golf, inspired, hair, bell, wright, kid, emotional, mum, brother, girl, smile, voice, charlie, amazing, dance, tribute, pga, sad, skills, personality, skating, memories, joy, remarkable
Topic: 5	**1**	regional, acquired, chant, wa, tracing, visited, attended, locally, gladys, crew, exposure, coast, locations, cabinet, exemption, andrews, perth, southern, bushfire, arrangements, permit, tasmania, arriving, airport, hotspots, hotspot, greg, contacts, borders, fines
Topic: 6	**0.667**	biontech, astrazeneca, dose, vaccinations, infected, intensive, healthcare, scientists, mexico, detected, vaccinate, institute, spain, approved, approval, registered, portugal, infectious, vaccinating, trials, tally, moderna, rollout, spokesman, tourism, jab, variants, supplies, doctors, wuhan
Topic: 7	**1**	porter, grants, legislation, mp, bowers, senator, parliamentary, frydenberg, lowe, greens, employees, coalition, laws, welfare, grant, chair, agreements, wages, rorts, tweet, allegations, employers, reform, accounts, arts, alleged, proposed, permanent, treasurer, affairs
Topic: 8	**1**	brands, stimulus, firms, store, investment, investors, retailers, negotiations, index, retailer, unemployment, rescue, sold, stocks, ftse, chain, costs, dollar, benefits, s&p, annual, brand, property, administrators, stock, customers, consumers, shares, production, products
Topic:		bayern, norwich, fiver, borussia, bundesliga, leipzig, utd, swansea, bournemouth, qpr, stuttgart, cardiff, solskjær, millwall, leverkusen, munich, schalke, brentford,

Figure 61: Reproducibility of 3 topic models over the same sport-news dataset

- **Topic-Topic Connection**: This section will show the topic-topic correlation between the current topic model analysis topics. The user can set a lambda value for the topic labels, the number of words per topic label (that get reviewed by the system to calculate the correlation) and a specific threshold. On the right, the user will see two plots: One of them is showing the topic correlation based on the topic's theta values, and the other is showing the correlation depending on whether the documents exceeds the set threshold.



Parameters

Task ID: **67**

lambda for topic labels

0.53

number of words per topic label

50

threshold

0.055

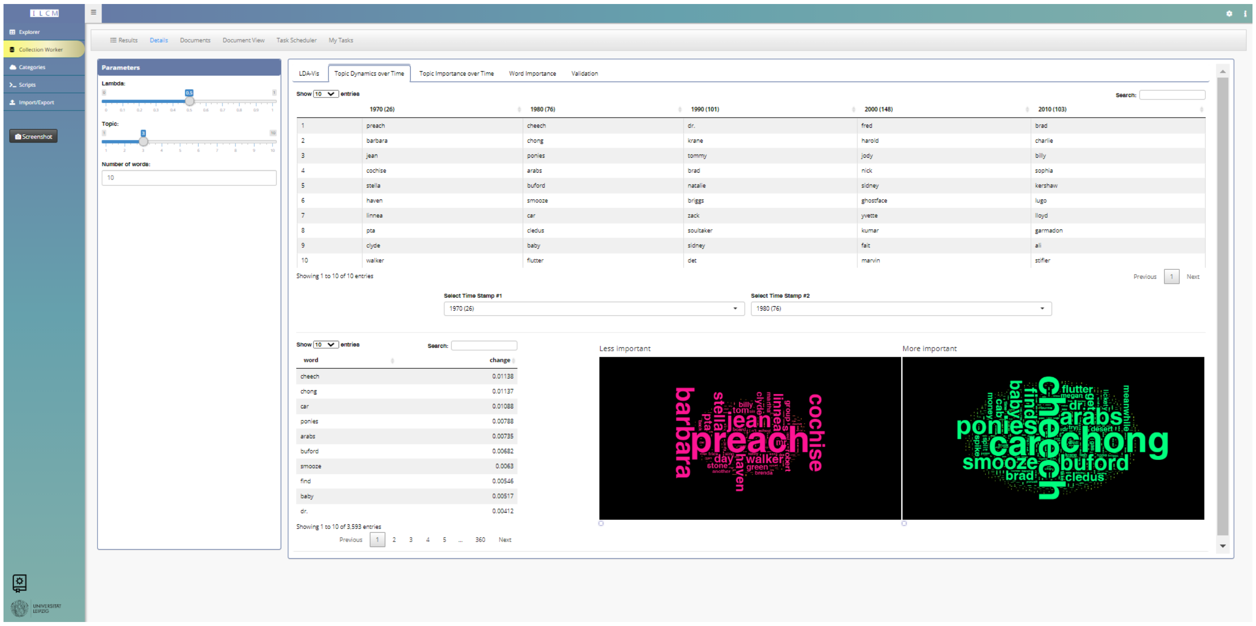
Topic Topic Correlation based on Theta Values

Topic Topic Correlation based on whether a documents exceeds set threshold

Figure 62: Heat map of the topic topic correlation

15.2 Details Dynamic Topic Models

- **LDA-Vis**: In the *LDA-Vis* panel, the user sees the LDA-vis visualization for the calculated dynamic topic model. The parameters box can change the time parameter to see the different topic dispersion within different time intervals.
- **Topic Dynamic over Time**: Here, the user will find a data table of all periods and the top words of a selected topic over these times. On the bottom, the user will find the option to select two time periods for comparison. After selecting the two-word clouds, one with the most and one with the least essential words for a topic within the two selected time stamps will appear. Besides this, a list of words is shown with a *change* parameter that will show the changed importance of this words over time. In the parameter box, the user can set the specific topic a number of words that should be shown and a lambda parameter for calculating which words should be shown.



Parameters

Lambda: 0.5

Topic: 3

Number of words: 10

Show 10 entries

Search:

	1970 (26)	1980 (76)	1990 (101)	2000 (148)	2010 (103)
1	preach	cheech	dr.	fred	brad
2	barbara	chong	krane	harold	charlie
3	jean	ponies	tommy	jody	billy
4	cochise	arabs	brad	nick	sophia
5	stella	buford	natalie	sidney	kershaw
6	haven	smooze	briggs	ghostface	lugo
7	linnea	car	zack	yvette	lloyd
8	pta	cledus	soultaker	kumar	garmadon
9	clyde	baby	sidney	fait	ali
10	walker	flutter	det	marvin	stifler

Showing 1 to 10 of 10 entries

Previous 1 Next

Select Time Stamp #1: 1970 (26)

Select Time Stamp #2: 1980 (76)

Show 10 entries

Search:

word	change
cheech	0.01138
chong	0.01137
car	0.01088
ponies	0.00788
arabs	0.00735
buford	0.00682
smooze	0.0063
find	0.00546
baby	0.00517
dr.	0.00412

Showing 1 to 10 of 3,593 entries

Previous 1 2 3 4 5 ... 360 Next

Figure 63: Dynamic of topic 3 for movie data over time

- **Topic Importance over Time**: This subtab will show the correlation of the importance of the topics over the given time-stamps. At the top, the user will get an overview of the correlation of the topics within the timestamps, and on the bottom, there will be a visualization of how the importance of the topics has changed between the timestamps.

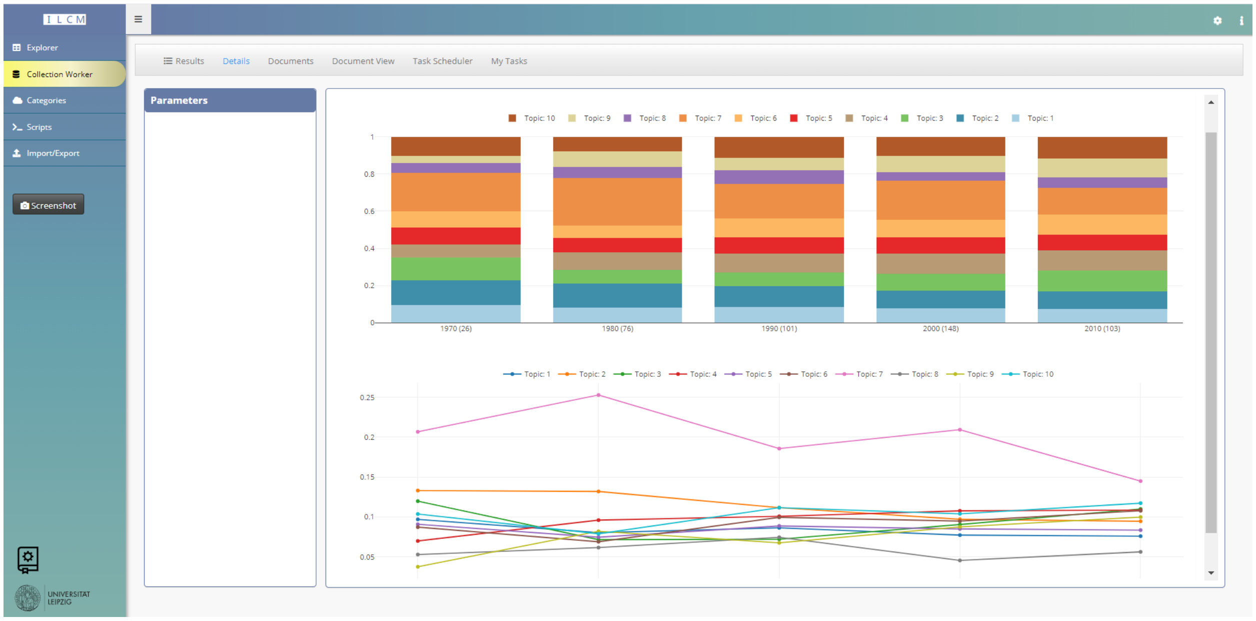
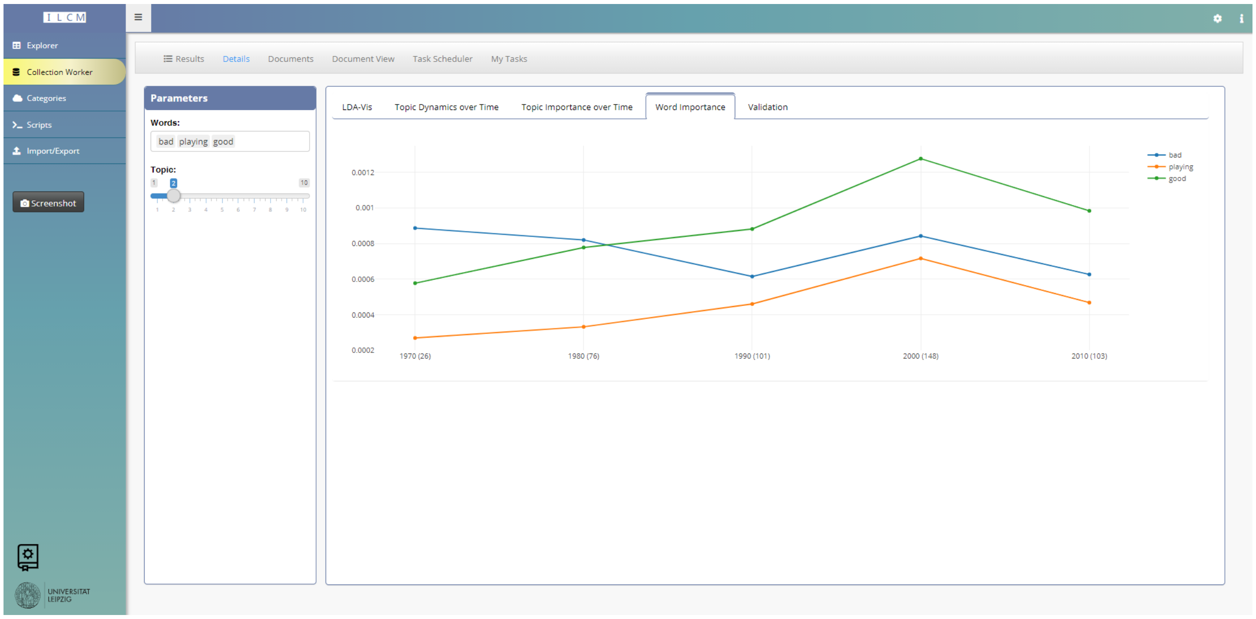


Figure 64: Importance of 10 different topics over the given time intervals

- **Word importance**: Here, the user can see the importance of selected words within a certain topic over time. Therefore the parameter box offers to select the words the user wishes to compare and a slider to select the topic the importance of the word should be calculated for.

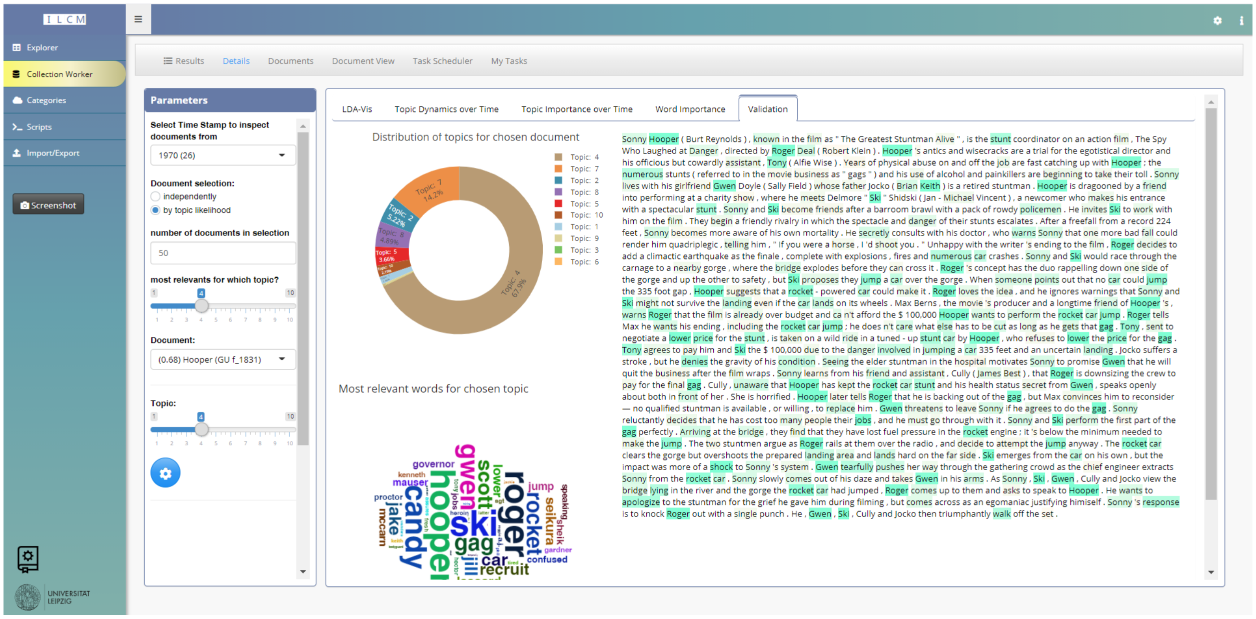


Words: bad playing good

Topic: 2

Figure 65: Importance of the words 'bad', 'good' and 'playing' within topic 2 of a movie dataset over time

- **Validation**: This allows the user to review certain documents from the collection and check to spread topics in the document. Therefore the user needs to specify a timestamp, the number of documents in the selection, a document to view and a topic to present in the parameter box. The text of the selected document will be shown along with highlighted words that indicate the selected topic. Additionally, the user will see a pie chart for the topic distribution for the chosen document and the most relevant words for the chosen topic.



Select Time Stamp to inspect documents from: 1970 (26)

Document selection:
○ independently
● by topic likelihood

number of documents in selection: 50

most relevants for which topic?: 4

Document: (0.68) Hooper (GU f_1831)

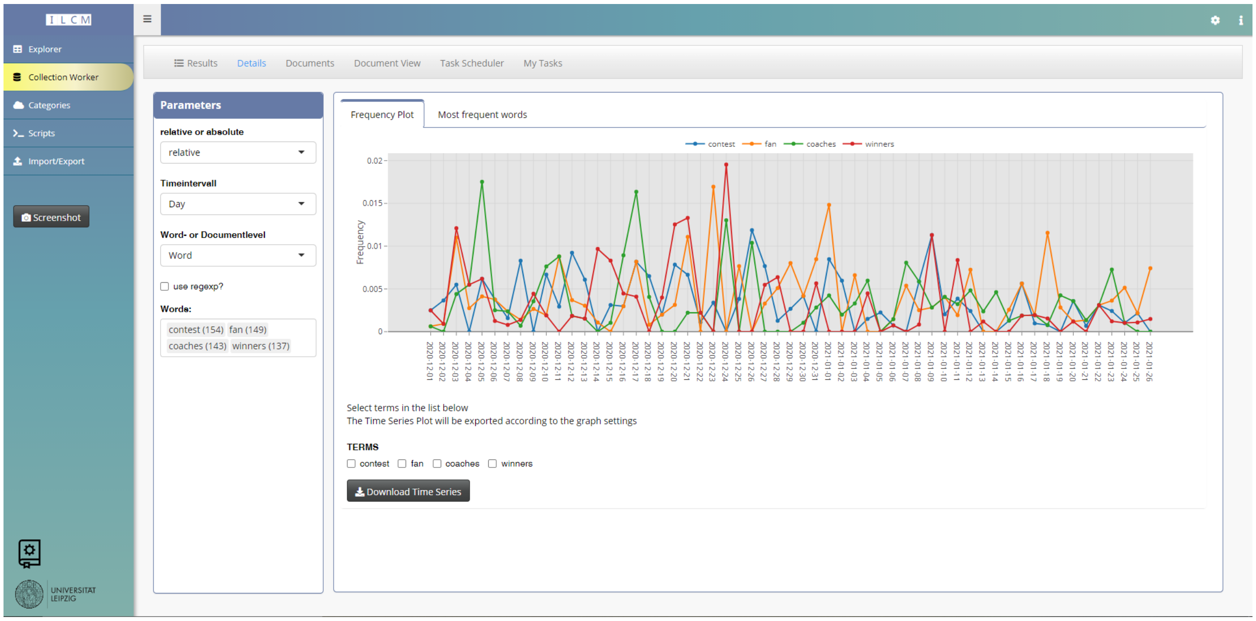
Topic: 4

Sonny Hooper (Burt Reynolds) , known in the film as " The Greatest Stuntman Alive " , is the stunt coordinator on an action film , The Spy Who Laughed at Danger , directed by Roger Deal (Robert Klein) . Hooper 's antics and wisecracks are a trial for the egotistical director and his officious but cowardly assistant , Tony (Alfie Wise) . Years of physical abuse on and off the job are fast catching up with Hooper ; the numerous stunts (referred to in the movie business as " gags ") and his use of alcohol and painkillers are beginning to take their toll . Sonny lives with his girlfriend Gwen Doyle (Sally Field) whose father Jocko (Brian Keith) is a retired stuntman . Hooper is dragooned by a friend into performing at a charity show , where he meets Delmore " Ski " Shidski (Jan - Michael Vincent) , a newcomer who makes his entrance with a spectacular stunt . Sonny and Ski become friends after a barroom brawl with a pack of rowdy policemen . He invites Ski to work with him on the film . They begin a friendly rivalry in which the spectacle and danger of their stunts escalates . After a freefall from a record 224 feet , Sonny becomes more aware of his own mortality . He secretly consults with his doctor , who warns Sonny that one more bad fall could render him quadriplegic , telling him , " If you were a horse , I 'd shoot you . " Unhappy with the writer 's ending to the film , Roger decides to add a climactic earthquake as the finale , complete with explosions , fires and numerous car crashes . Sonny and Ski would race through the carnage to a nearby gorge , where the bridge explodes before they can cross it . Roger 's concept has the duo rappelling down one side of the gorge and up the other to safety , but Ski proposes they jump a car over the gorge . When someone points out that no car could jump the 335 foot gap , Hooper suggests that a rocket - powered car could make it . Roger loves the idea , and he ignores warnings that Sonny and Ski might not survive the landing even if the car lands on its wheels . Max Berns , the movie 's producer and a longtime friend of Hooper 's , warns Roger that the film is already over budget and ca n't afford the $ 100,000 Hooper wants to perform the rocket car jump . Roger tells Max he wants his ending , including the rocket car jump ; he does n't care what else has to be cut as long as he gets that gag . Tony , sent to negotiate a lower price for the stunt , is taken on a wild ride in a tuned - up stunt car by Hooper , who refuses to lower the price for the gag . Tony agrees to pay him and Ski the $ 100,000 due to the danger involved in jumping a car 335 feet and an uncertain landing . Jocko suffers a stroke , but he denies the gravity of his condition . Seeing the elder stuntman in the hospital motivates Sonny to promise Gwen that he will quit the business after the film wraps . Sonny learns from his friend and assistant , Cully (James Best) , that Roger is downsizing the crew to pay for the final gag . Cully , unaware that Hooper has kept the rocket car stunt and his health status secret from Gwen , speaks openly about both in front of her . She is horrified . Hooper later tells Roger that he is backing out of the gag , but Max convinces him to reconsider — no qualified stuntman is available , or willing , to replace him . Gwen threatens to leave Sonny if he agrees to do the gag . Sonny reluctantly decides that he has cost too many people their jobs , and he must go through with it . Sonny and Ski perform the first part of the gag perfectly . Arriving at the bridge , they find that they have lost fuel pressure in the rocket engine ; it 's below the minimum needed to make the jump . The two stuntmen argue as Roger rails at them over the radio , and decide to attempt the jump anyway . The rocket car clears the gorge but overshoots the prepared landing area and lands hard on the far side . Ski emerges from the car on his own , but the impact was more of a shock to Sonny 's system . Gwen tearfully pushes her way through the gathering crowd as the chief engineer extracts Sonny from the rocket car . Sonny slowly comes out of his daze and takes Gwen in his arms . As Sonny , Ski , Gwen , Cully and Jocko view the bridge lying in the river and the gorge the rocket car had jumped , Roger comes up to them and asks to speak to Hooper . He wants to apologize to the stuntman for the grief he gave him during filming , but comes across as an egomaniac justifying himiself . Sonny 's response is to knock Roger out with a single punch . He , Gwen , Ski , Cully and Jocko then triumphantly walk off the set .

Figure 66: Overview of the topic distribution for a chosen document in a movie dataset

15.3 Details Term Frequency extraction

- **Frequency plot**: In the *Frequency Plot* the time series counts for the chosen words are plotted. The time-series data can be downloaded by selecting the relevant words and then clicking on *Download Time Series*. The user can change the following settings: relative or absolute counts, time interval, counts on the document or word level and the words themselves. It is possible to plot and compare the time series data for multiple words



Parameters

relative or absolute: relative

Timeintervall: Day

Word- or Documentlevel: Word

□ use regexp?

Words: contest (154) fan (149) coaches (143) winners (137)

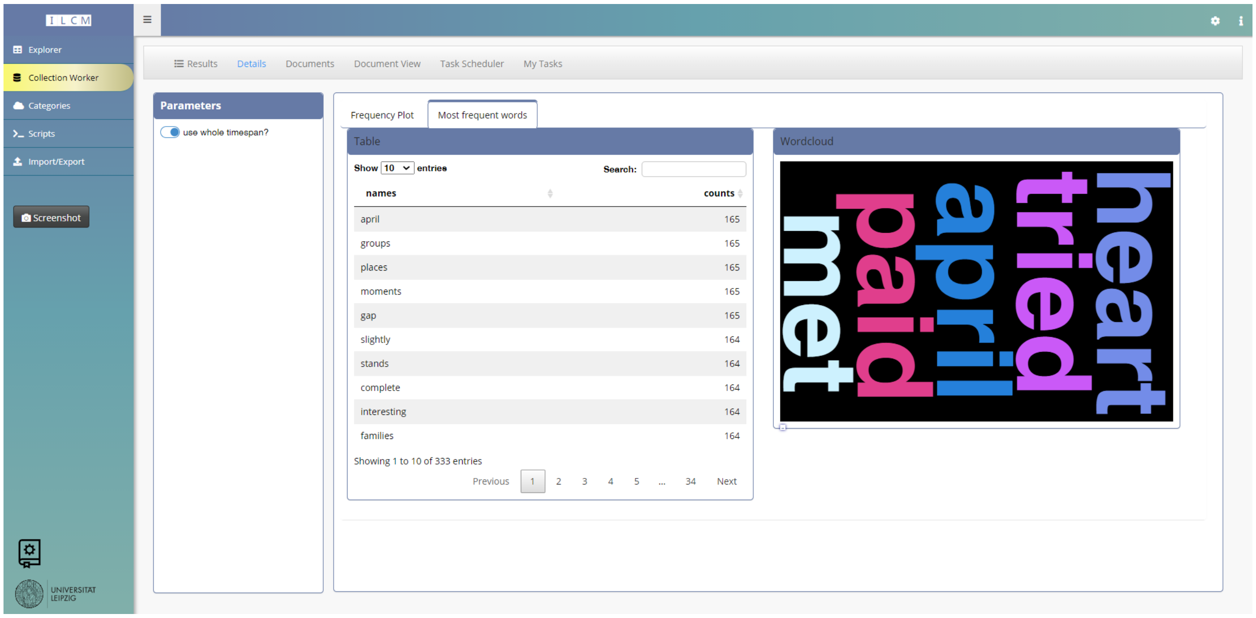
Select terms in the list below
The Time Series Plot will be exported according to the graph settings

TERMS

□ contest □ fan □ coaches □ winners

Figure 67: Frequency extraction results for 'sports competition' context

- **Most frequent words**: Here, the user will see the most frequent words of the collection. They have the option to enable the usage of the whole timespan. So they can choose a specific time interval and a point in time for the visualization. The system will then plot a data table with all the relevant words and their frequency counts and a word cloud with the most frequent words.



Parameters

● use whole timespan?

Table

Show 10 entries Search:

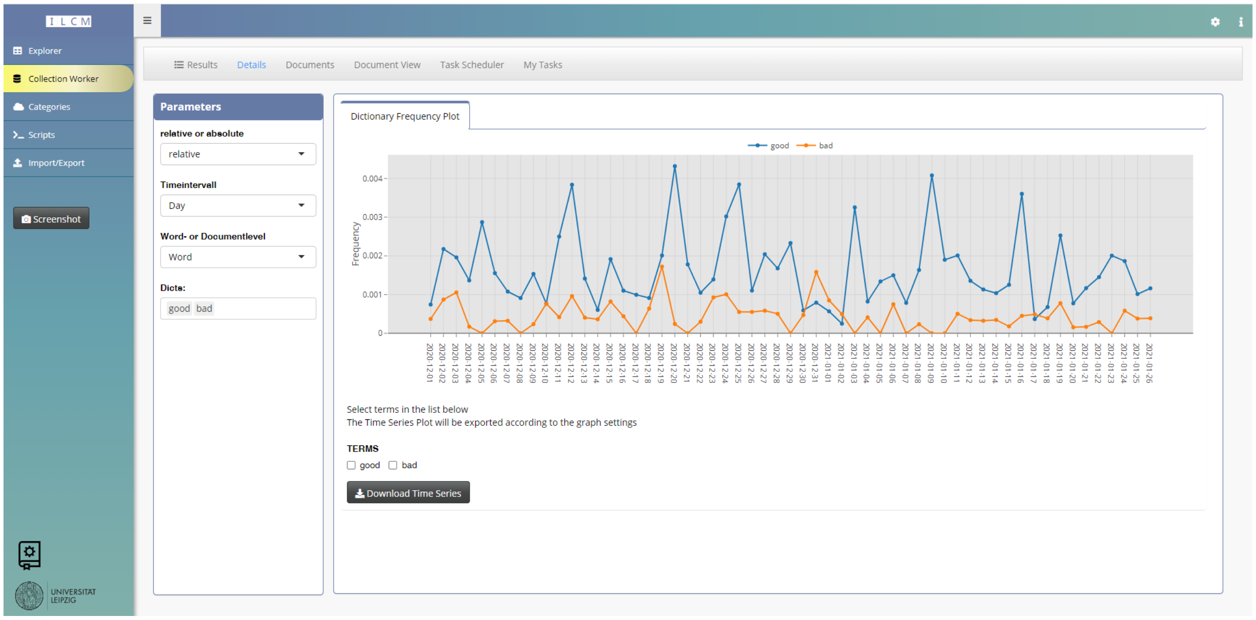
names	counts
april	165
groups	165
places	165
moments	165
gap	165
slightly	164
stands	164
complete	164
interesting	164
families	164

Showing 1 to 10 of 333 entries

Figure 68: Most frequent words in a news related collection

15.4 Details on Dictionary Frequency Extraction

- **Frequency Plot**: In the *Frequency Plot*, the chosen dictionaries' time-series counts are plotted. The time-series data can be downloaded by selecting the dictionaries of interest and then clicking on *Download Time Series*. The user can change the following settings: relative or absolute counts, time interval, counts on document or word level. Its possible to plot and compare the time series data for multiple dictionaries



relative or absolute: relative

Timeintervall: Day

Word- or Documentlevel: Word

Dicts: good bad

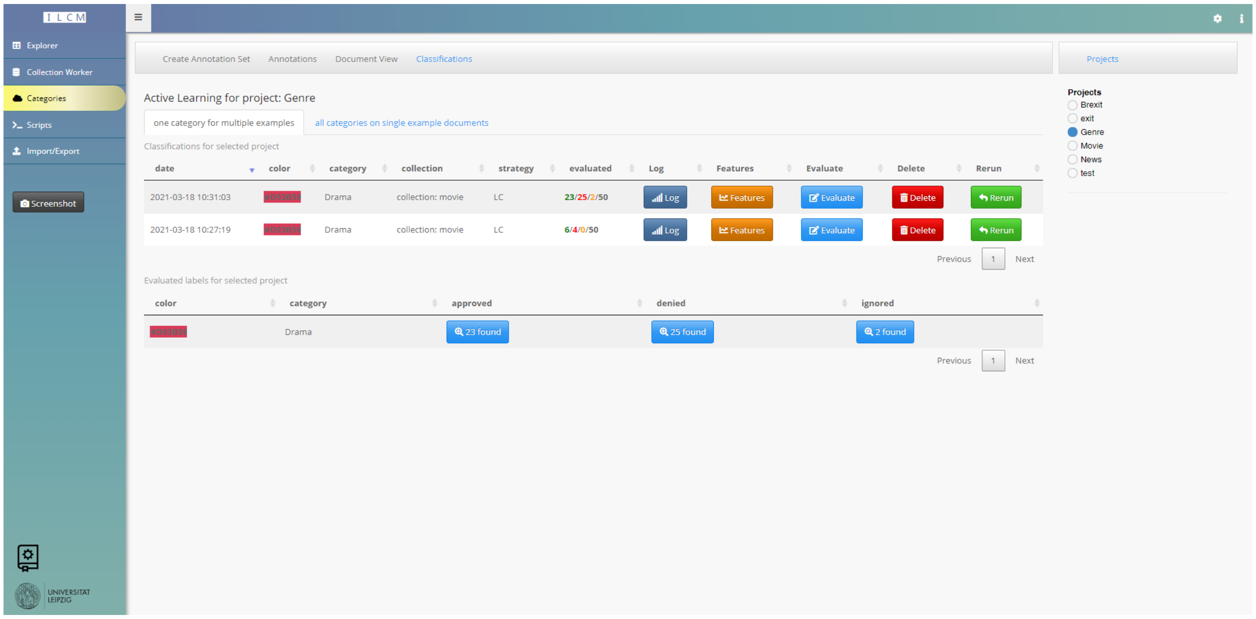
Select terms in the list below
The Time Series Plot will be exported according to the graph settings

TERMS
□ good □ bad

Figure 69: Dictionary frequency extraction example

15.5 Classification

When selecting a result from the *Classification* task, the user will be led to the *Categories* tab and the *Classifications* subtab. There they will find a list of all classifications to a selected project. They have the chance to evaluate the active learning example so and start the evaluation new in order to see a potential improvement. Additionally, they can check log and feature information from the classifications and delete a particular classification. On the other hand, the user can also see all categories on a single document example. Therefore the user needs to select training sets for the projects.



Active Learning for project: Genre

one category for multiple examples | all categories on single example documents

Classifications for selected project

date	color	category	collection	strategy	evaluated	Log	Features	Evaluate	Delete	Rerun
2021-03-18 10:31:03	#D93B59	Drama	collection: movie	LC	23/25/2/50	Log	Features	Evaluate	Delete	Rerun
2021-03-18 10:27:19	#D93B59	Drama	collection: movie	LC	6/4/0/50	Log	Features	Evaluate	Delete	Rerun

Previous 1 Next

Evaluated labels for selected project

color	category	approved	denied	ignored
#D93B59	Drama	23 found	25 found	2 found

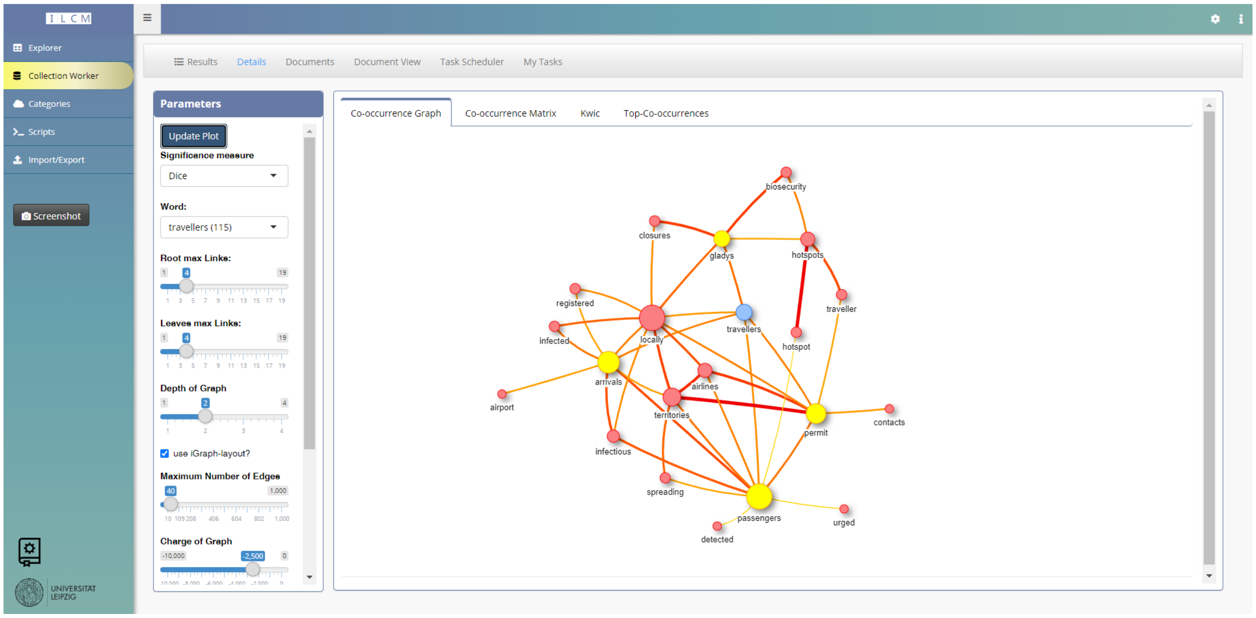
Previous 1 Next

Projects
○ Brexit
○ exit
● Genre
○ Movie
○ News
○ test

Figure 70: Classification tasks to find movie data corresponding to the genre 'Drama'

15.6 Details Co-occurrence

- **Co-occurrence graph**: The co-occurrence graph aims to visualize the chosen words' calculated co-occurrence behavior. For this purpose, the user has to choose a word (network root, blue-coloured dot) and a significance measurement for calculating the network (e.g. dice, mutual information, ...). Then they can set the maximum number of links departing from the root. Next, they can set the maximum number of links for all other nods of the graph. With the parameter *Depth of Graph*, the number of iterations whereby new nodes will be added to the current leaves can be set. The higher this number, the more nodes the graph will have. For stability and clearness purposes, the user can also limit the number of links plotted to a certain number. With the *Charge of Graph* parameter, the user can manipulate the gravitation of the nodes (more details: visNetwork). It is possible to switch to a hierarchical layout by ticking the corresponding checkbox. Also, the user can either use straight edges or curved ones (smooth edges parameter). The node size can be determined as well by 3 measures: degree. Betweenness and centrality. The last parameter for the co-occurrence plot is the significance threshold. The values of the links reflect the calculated significance values. Only the links will be displayed with a higher value than the threshold with setting a certain threshold. When hovering over a node, the corresponding value to its size will be shown. After changing the plot parameters, the user has to hit the *Update Plot* button to see the graph's changes. When working with a depth of 4, recalculating the plot might take a few minutes. The user can also move the nodes around the canvas by clicking and moving them to the new position.



Significance measure: Dice

Word: travellers (115)

Root max Links: 4 (1–19)

Leaves max Links: 4 (1–19)

Depth of Graph: 2 (1–4)

☑ use iGraph-layout?

Maximum Number of Edges: 40 (10–1,000)

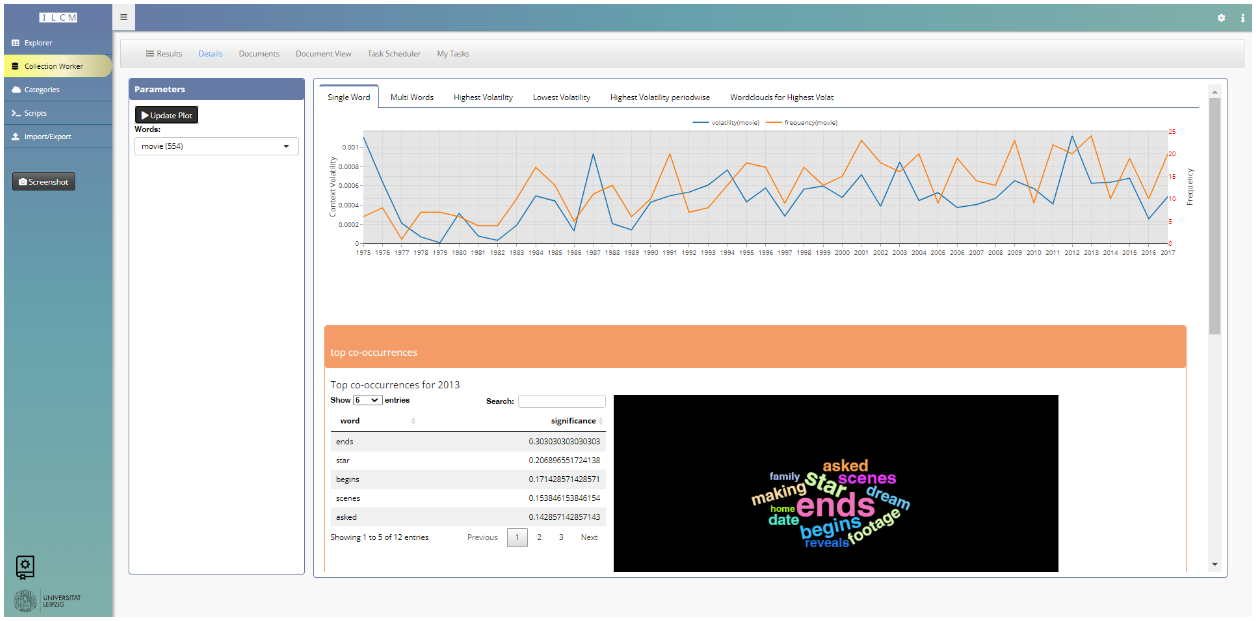
Charge of Graph: -2,500 (-10,000–0)

Figure 71: Co-occurrecne graph for the term 'travellers' whithin news reports from 2020 to 2021

- **Co-occurrence matrix**: In the *Co-occurrence Matrix*, the calculated significance (of a selected measurement) is displayed using a heatmap. The included words depend on the settings in the parameters section. The user can select an area in the plot to focus on a specific area.
- **Kwic**: In this subtab, the keyword and its context are displayed. The word can be chosen in the parameter setting. It is also possible to select more than one word. The number of shown examples and the size of the context needs to be defined.
- **Top-Co-Occurrences**: Here, the user will find an overview of the co-occurrence between at least two chosen words with different significance measurements. Therefore the user needs to select the first word for the calculation and an optional second one. If the second word textbox stays empty, the co-occurrence between all other words and the first selected one will be shown. In the parameter box, the user can set a maximum number of the shown occurrences, and they have the opportunity to download the results calculated with dice, mutual information or log-likelihood significance measurement.

15.7 Details Context Volatility

- **Single word**: In this subtab, the context volatility and the frequency for a single word are plotted. The user has to select a word in the parameters section and click *update plot*. In the time series plot, the orange line indicates the frequency of the selected word. The blue line shows the calculated context volatility values. Below the times series plot, there are 2 boxes. In the upper one, the most significant co-occurrences for the selected word and a specific point in time are displays in a table and a corresponding wordcloud. The user can specify the point in time by clicking on the time series plot to correspond to the times they are interested in. The two points in time are the last points the user has clicked on in the time series plot. Most divergent means those words, who have changed their significance value the most. Green words in the word cloud indicate words with a higher significance value on the last clicked point in time than in the one before. Pink words show those words that reveal more minor significances.



top co-occurrences

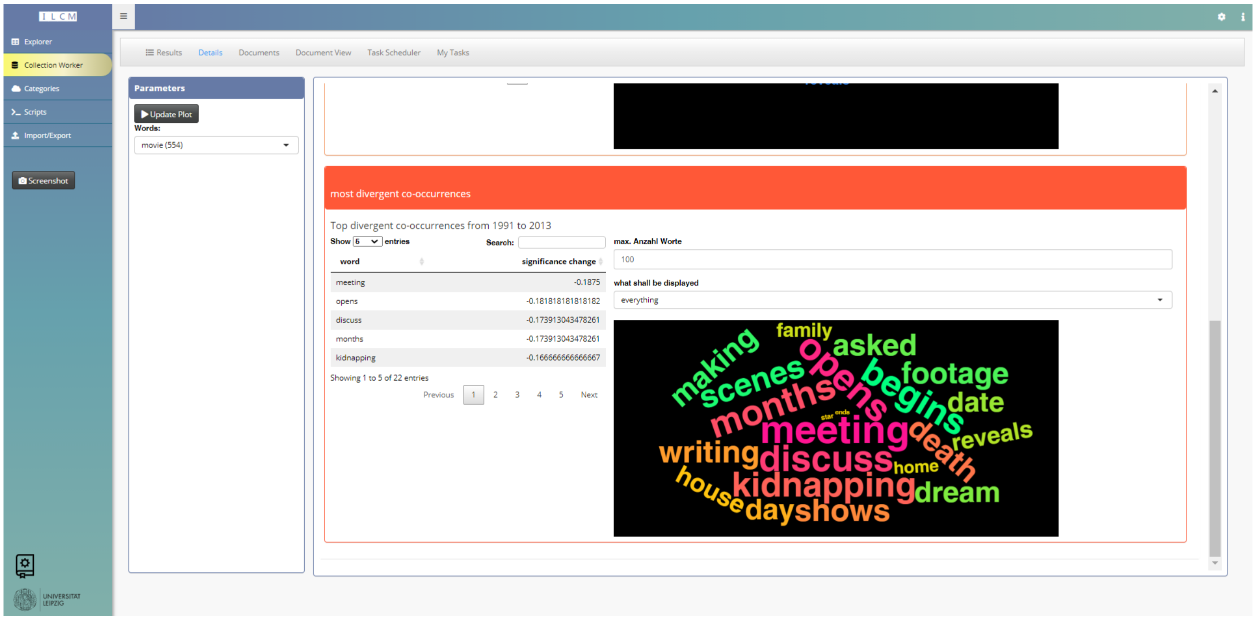
Top co-occurrences for 2013

Show 5 entries   Search:

word	significance
ends	0.303030303030303
star	0.206896551724138
begins	0.171428571428571
scenes	0.153846153846154
asked	0.142857142857143

Showing 1 to 5 of 12 entries   Previous 1 2 3 Next

Figure 72: Context volatility for the word 'movie'



most divergent co-occurrences

Top divergent co-occurrences from 1991 to 2013

Show 5 entries   Search:

word	significance change
meeting	-0.1875
opens	-0.181818181818182
discuss	-0.173913043478261
months	-0.173913043478261
kidnapping	-0.166666666666667

Showing 1 to 5 of 22 entries

Previous 1 2 3 4 5 Next

max. Anzahl Worte: 100

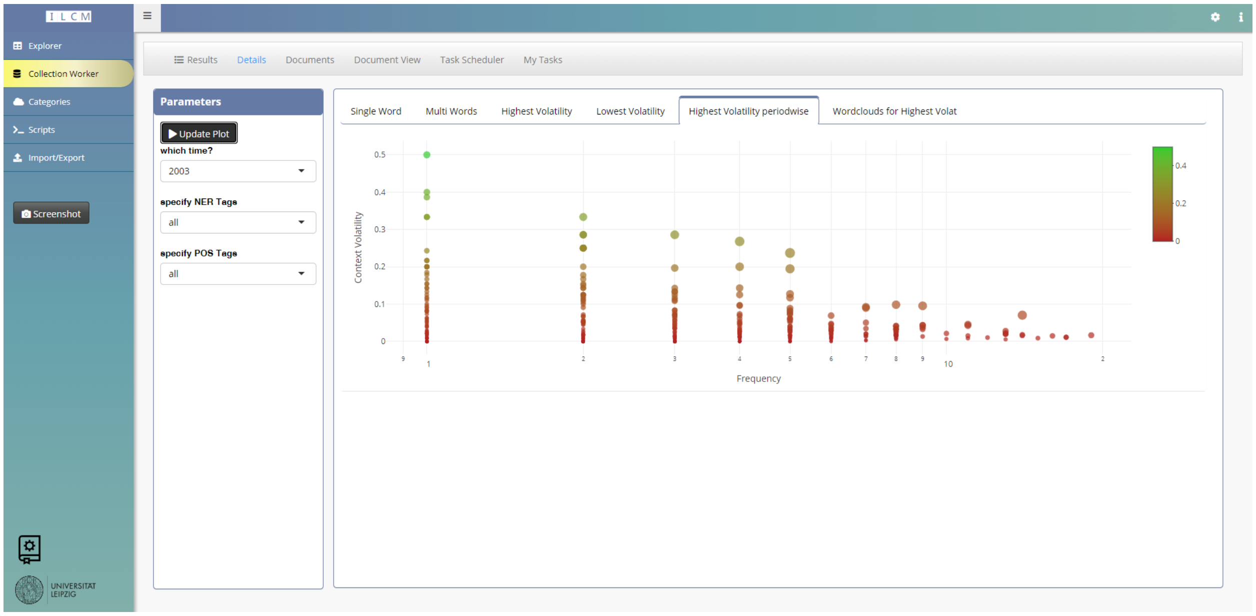
what shall be displayed: everything

Figure 73: most divergent co-occurrences for the word 'movie'

- **Multi words**: In this subtab, multiple word's context volatility can be prepared. Therefore the user can select multiple words from the select input on the parameters settings.
- **Highest volatility**: In the *Highest Volatility* subtab, a table with the average word frequencies over time and the overall average context volatility is shown. In the parameters section, a maximum and a minimum frequency can be specified. The results are ordered to decrease

by their context volatility.

- **Lowest volatility**: In the *Lowest Volatility* subtab, a table with the average word frequencies over time and the overall average context volatility is shown. In the parameters section, a maximum and a minimum frequency can be specified. The results are ordered increasing by their context volatility.
- **Highest volatility period wise**: The subtab *Highest Volatility Periodwise* enables the user to identify the words with the highest context volatility values for a specific point in time. This point in time can be specified in the parameters section. Also, in the parameters section, the user can limit the shown words to a specific NER- or POS-tag. In the graph, the words are plotted concerning their frequency (x-axis) and their context volatility (y-axis). When hovering a point in the plot, the corresponding word is shown with the calculated frequency and context volatility.



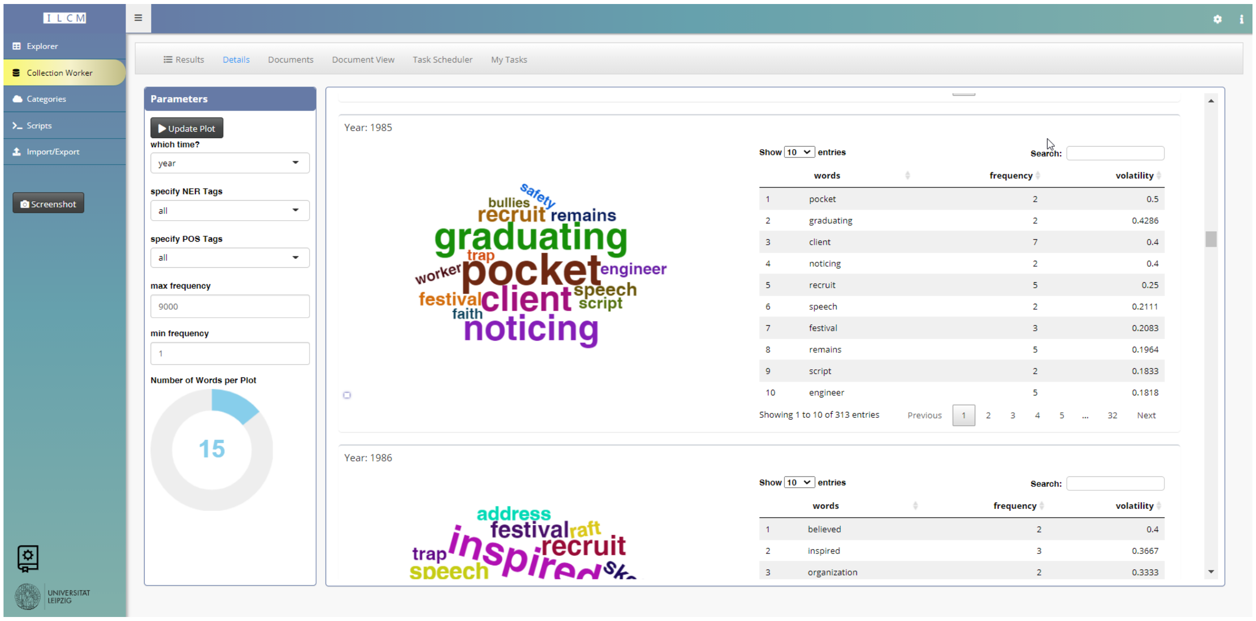
which time? 2003

specify NER Tags all

specify POS Tags all

Figure 74: Most volatile words periodwise

- **Wordclouds for highest volatility**: In the subtab *Wordclouds* for Highest Volatility, the most volatile words are displayed at a specific point in time. Nevertheless, this time, the results are shown in the wordcloud for each point in time available. On the right to each wordcloud, there is a table with the exact values calculated. In the parameters section, the user can change the time intervals depending on whether the chosen time interval in the parameters of the *Task Scheduler* was set today, month or year. If "day" was the chosen time interval, the periods could be aggregated to month and years in the visualization. If the choice was "month", The periods can be month or years. When "year" was the chosen option, it was impossible to split the data back into data for months or days—furthermore, it is again possible to restrict the words to a specific NER- or POS-tag. In addition to that also a minimum and a maximum frequency can be defined. The last options declare the maximum number of words included in each wordcloud.



which time? year

specify NER Tags: all

specify POS Tags: all

max frequency: 9000

min frequency: 1

Number of Words per Plot: 15

Year: 1985

Show 10 entries Search:

	words	frequency	volatility
1	pocket	2	0.5
2	graduating	2	0.4286
3	client	7	0.4
4	noticing	2	0.4
5	recruit	5	0.25
6	speech	2	0.2111
7	festival	3	0.2083
8	remains	5	0.1964
9	script	2	0.1833
10	engineer	5	0.1818

Showing 1 to 10 of 313 entries Previous 1 2 3 4 5 ... 32 Next

Year: 1986

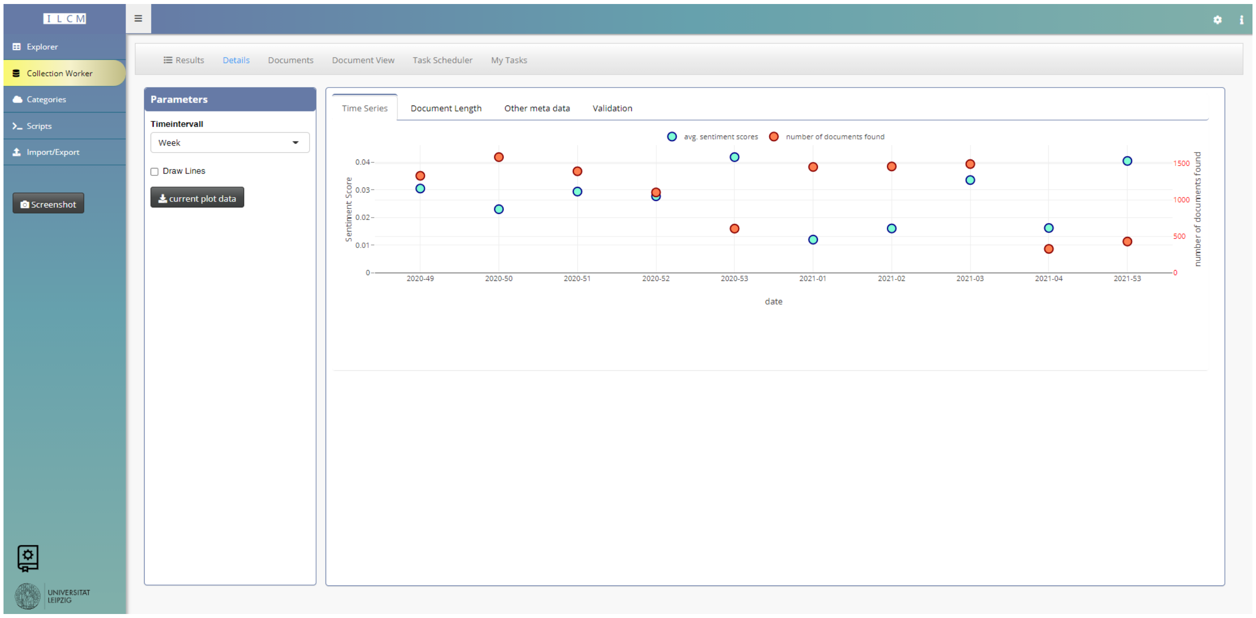
Show 10 entries Search:

	words	frequency	volatility
1	believed	2	0.4
2	inspired	3	0.3667
3	organization	2	0.3333

Figure 75: Most volatile words periodwise - wordclouds

15.8 Details Sentiment Analysis

- **Time Series**: In this part, the user will see the average sentiment scares along with the number of found documents over time. Therefore the user can select a specific time interval. The user can also decide if the plot should only show points or points connected with a line. On top of that, the user has the chance to download the current plotted data.



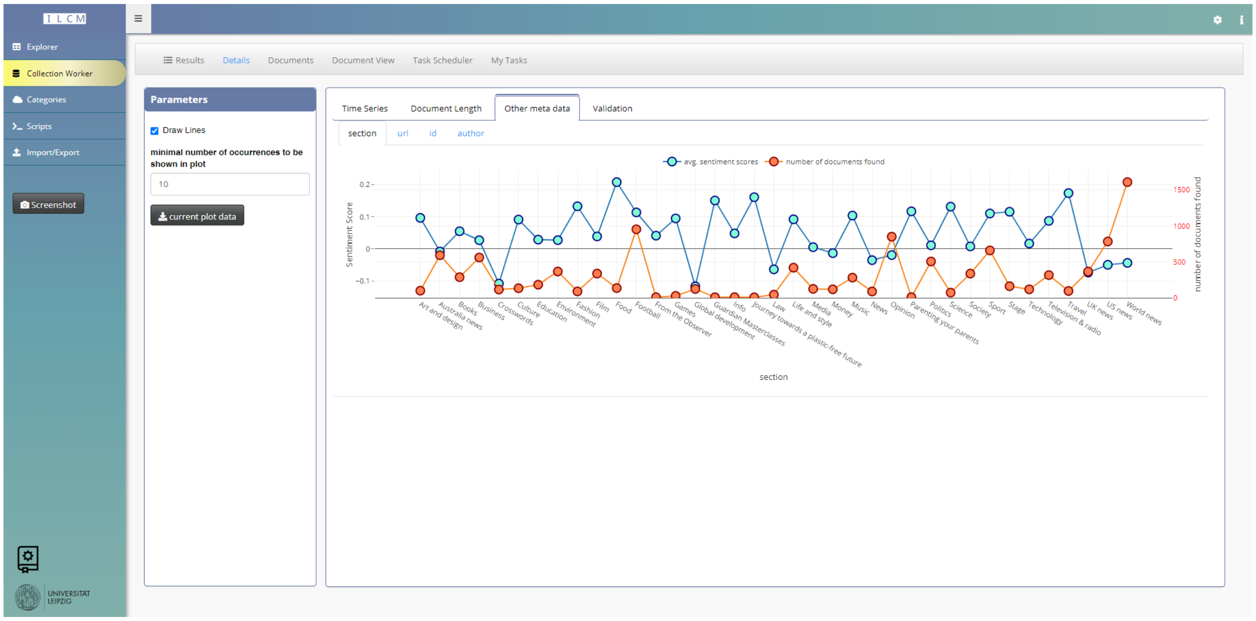
Timeintervall: Week

□ Draw Lines

Figure 76: sentiment scores and number of found documents whithin a weekly time intervall

- **Document length**: In this subsection, the user will see a sentiment score in correlation with the number of found documents and the number of words per document. The user can change the maximal number of breaks which will make the plot more accurate. They can, again, select to draw lines between the shown dots or download the current plot.

- **Other metadata**: Here, the user gets the chance to see the different sentiment score corresponding to the used collection's metadata. Corresponding to the metadata, several subtabs will be shown each for one specific piece of metadata. For each part, the user can download the plot and specify the minimal number of occurrence to be shown in the plot.



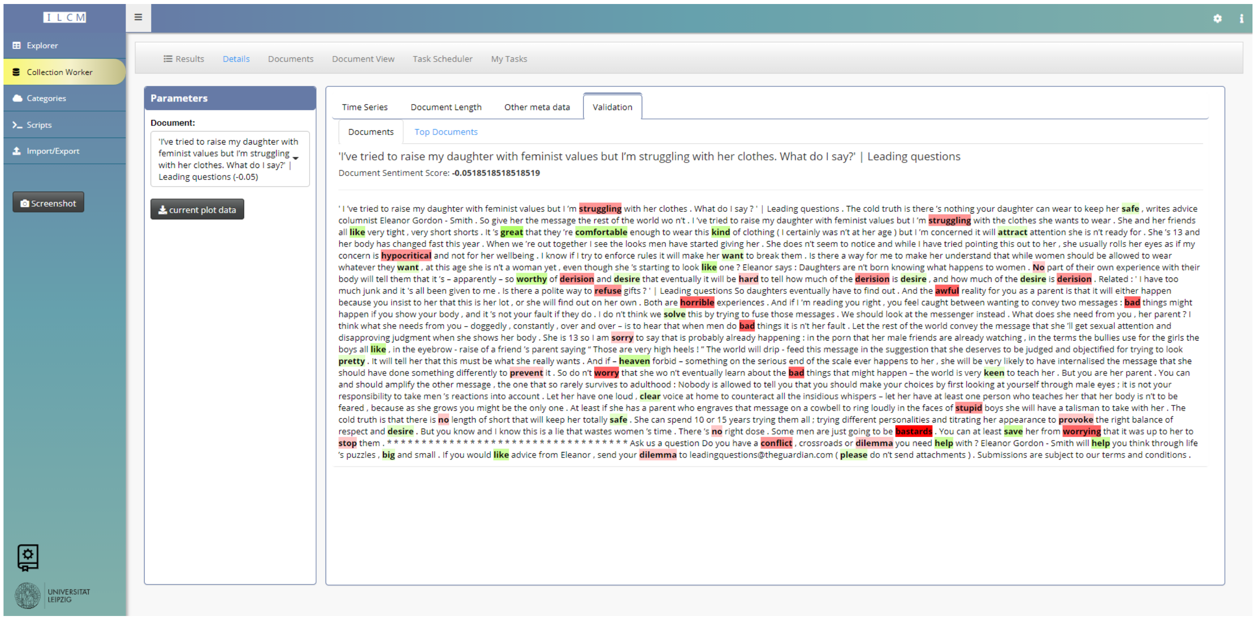
☑ Draw Lines

minimal number of occurrences to be shown in plot

10

Figure 77: Sentiment scores corresponding to the categories of news articles

- **Validation**: In this subsection the user has the chance to validate the results of sentiment analysis. Therefore they can either show single documents with the found sentiments highlighted (red = negative, green = positive) or see a list of the top documents with the most positive and negative documents.



Document: 'I've tried to raise my daughter with feminist values but I'm struggling with her clothes. What do I say?' | Leading questions (-0.05)

'I've tried to raise my daughter with feminist values but I'm struggling with her clothes. What do I say?' | Leading questions

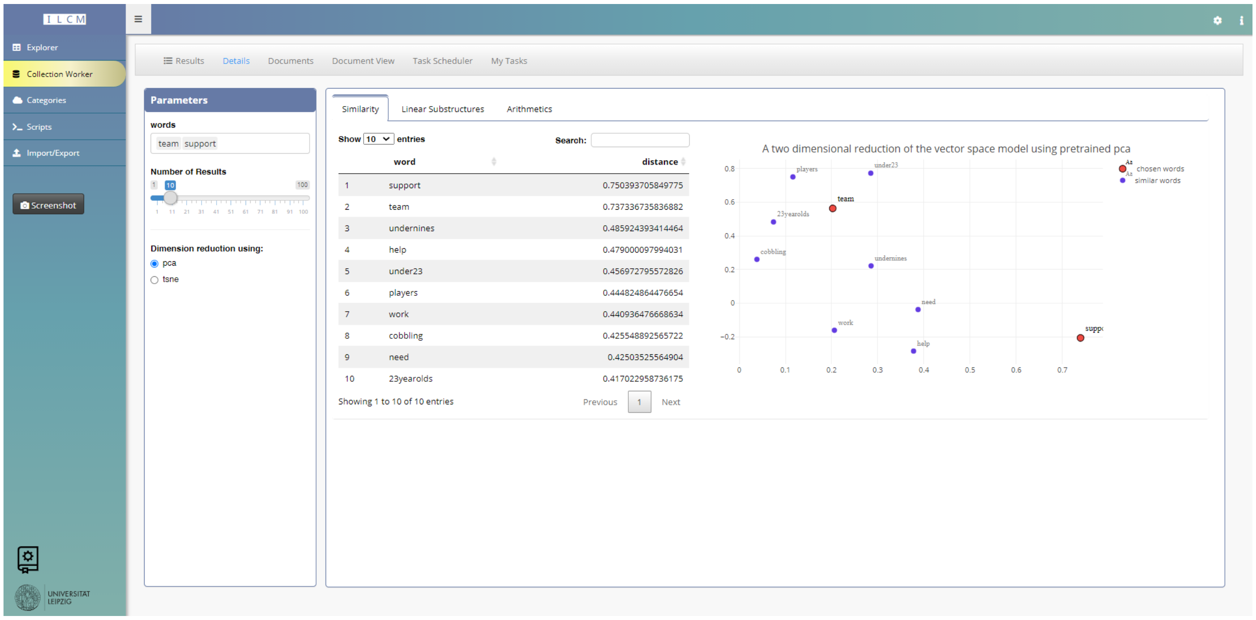
Document Sentiment Score: **-0.0518518518518519**

' I 've tried to raise my daughter with feminist values but I 'm **struggling** with her clothes . What do I say ? ' | Leading questions . The cold truth is there 's nothing your daughter can wear to keep her **safe** , writes advice columnist Eleanor Gordon - Smith . So give her the message the rest of the world wo n't . I 've tried to raise my daughter with feminist values but I 'm **struggling** with the clothes she wants to wear . She and her friends all **like** very tight , very short shorts . It 's **great** that they 're **comfortable** enough to wear this **kind** of clothing (I certainly was n't at her age) but I 'm concerned it will **attract** attention she is n't ready for . She 's 13 and her body has changed fast this year . When we 're out together I see the looks men have started giving her . She does n't seem to notice and while I have tried pointing this out to her , she usually rolls her eyes as if my concern is **hypocritical** and not for her wellbeing . I know if I try to enforce rules it will make her **want** to break them . Is there a way for me to make her understand that while women should be allowed to wear whatever they **want** , at this age she is n't a woman yet , even though she 's starting to look **like** one ? Eleanor says : Daughters are n't born knowing what happens to women . **No** part of their own experience with their body will tell them that it 's – apparently – so **worthy** of **derision** and **desire** that eventually it will be **hard** to tell how much of the **derision** is **desire** , and how much of the **desire** is **derision** . Related : ' I have too much junk and it 's all been given to me . Is there a polite way to **refuse** gifts ? ' | Leading questions So daughters eventually have to find out . And the **awful** reality for you as a parent is that it will either happen because you insist to her that this is her lot , or she will find out on her own . Both are **horrible** experiences . And if I 'm reading you right , you feel caught between wanting to convey two messages : **bad** things might happen if you show your body , and it 's not your fault if they do . I do n't think we **solve** this by trying to fuse those messages . We should look at the messenger instead . What does she need from you , her parent ? I think what she needs from you – doggedly , constantly , over and over – is to hear that when men do **bad** things it is n't her fault . Let the rest of the world convey the message that she 'll get sexual attention and disapproving judgment when she shows her body . She is 13 so I am **sorry** to say that is probably already happening : in the porn that her male friends are already watching , in the terms the bullies use for the girls the boys all **like** , in the eyebrow - raise of a friend 's parent saying " Those are very high heels ! " The world will drip - feed this message in the suggestion that she deserves to be judged and objectified for trying to look **pretty** . It will tell her that this must be what she really wants . And if – **heaven** forbid – something on the serious end of the scale ever happens to her , she will be very likely to have internalised the message that she should have done something differently to **prevent** it . So do n't **worry** that she wo n't eventually learn about the **bad** things that might happen – the world is very **keen** to teach her . But you are her parent . You can and should amplify the other message , the one that so rarely survives to adulthood : Nobody is allowed to tell you that you should make your choices by first looking at yourself through male eyes ; it is not your responsibility to take men 's reactions into account . Let her have one loud , **clear** voice at home to counteract all the insidious whispers – let her have at least one person who teaches her that her body is n't to be feared , because as she grows you might be the only one . At least if she has a parent who engraves that message on a cowbell to ring loudly in the faces of **stupid** boys she will have a talisman to take with her . The cold truth is that there is **no** length of short that will keep her totally **safe** . She can spend 10 or 15 years trying them all ; trying different personalities and titrating her appearance to **provoke** the right balance of respect and **desire** . But you know and I know this is a lie that wastes women 's time . There 's **no** right dose . Some men are just going to be **bastards** . You can at least **save** her from **worrying** that it was up to her to **stop** them . * Ask us a question Do you have a **conflict** , crossroads or **dilemma** you need **help** with ? Eleanor Gordon - Smith will **help** you think through life 's puzzles , **big** and small . If you would **like** advice from Eleanor , send your **dilemma** to leadingquestions@theguardian.com (**please** do n't send attachments) . Submissions are subject to our terms and conditions .

Figure 78: Sentiment analysis of a news aerticle

15.9 Details on Vector Space Representation

- **Similarity**: In this section, the user can compare the selected words' vector representation. Therefore the plot will show a coordinate system and presents the selected words as red dots and other not selected words within this space as blue dots. In the parameter section, the user can select the number of shown points within the coordinate system and a reduction method. On the left, the user will also see a list of the distance between the shown points.



words: team support

Number of Results: 10

Dimension reduction using:
● pca
○ tsne

Show 10 entries Search:

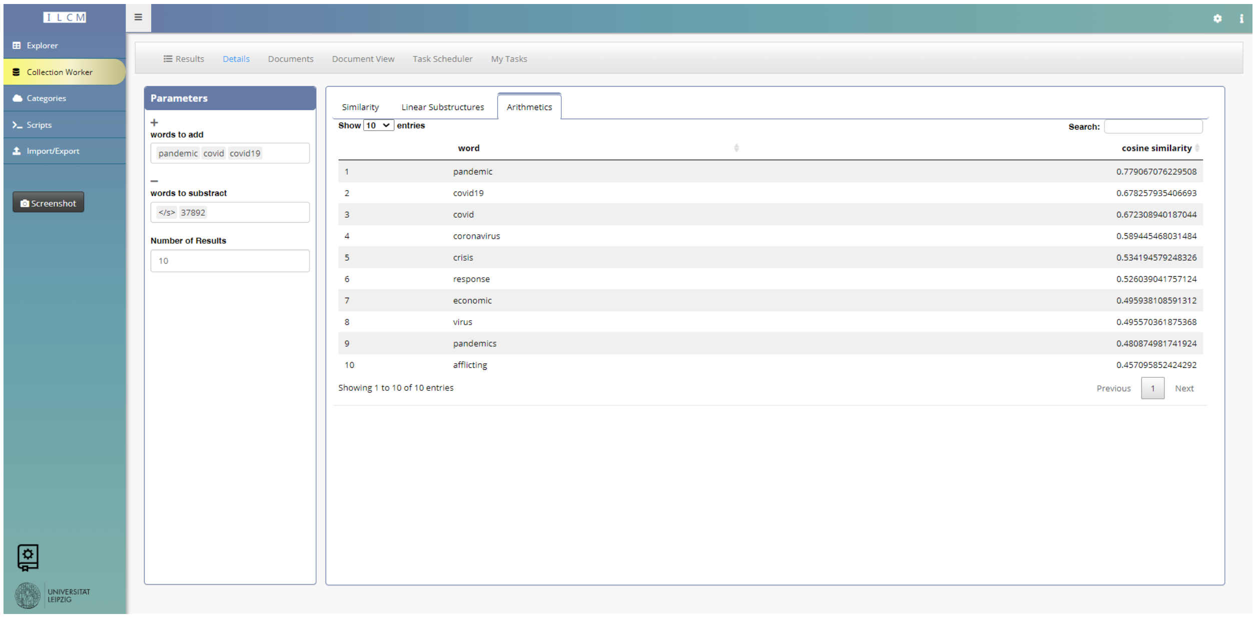
	word	distance
1	support	0.750393705849775
2	team	0.737336735836882
3	undernines	0.485924393414464
4	help	0.479000097994031
5	under23	0.456972795572826
6	players	0.444824864476654
7	work	0.440936476668634
8	cobbling	0.425548892565722
9	need	0.42503525564904
10	23yearolds	0.417022958736175

Showing 1 to 10 of 10 entries Previous 1 Next

Figure 79: Vizulisation of the vetor space representation

- **Linear Substructures**: Here, the user can select a certain number of word pairs or triples plotted in the coordinate system. This will give the user a vague idea of how close the selected words are related when calculating the vector space representation with PCA or tsne.

- **Arithmetics**: In this section, the user can select specific words (at least one) for which the cosine similarity to other words will be calculated. This calculation is shown in a data table. The user can also set the number of relevant words (the ones with the nearest similarity to the selected words) they wish to see along with a list of words for which the cosine similarity should not be calculated (e.g. for stopwords).



Parameters

words to add: pandemic covid covid19

words to substract: </s> 37892

Number of Results: 10

Show 10 entries

Search:

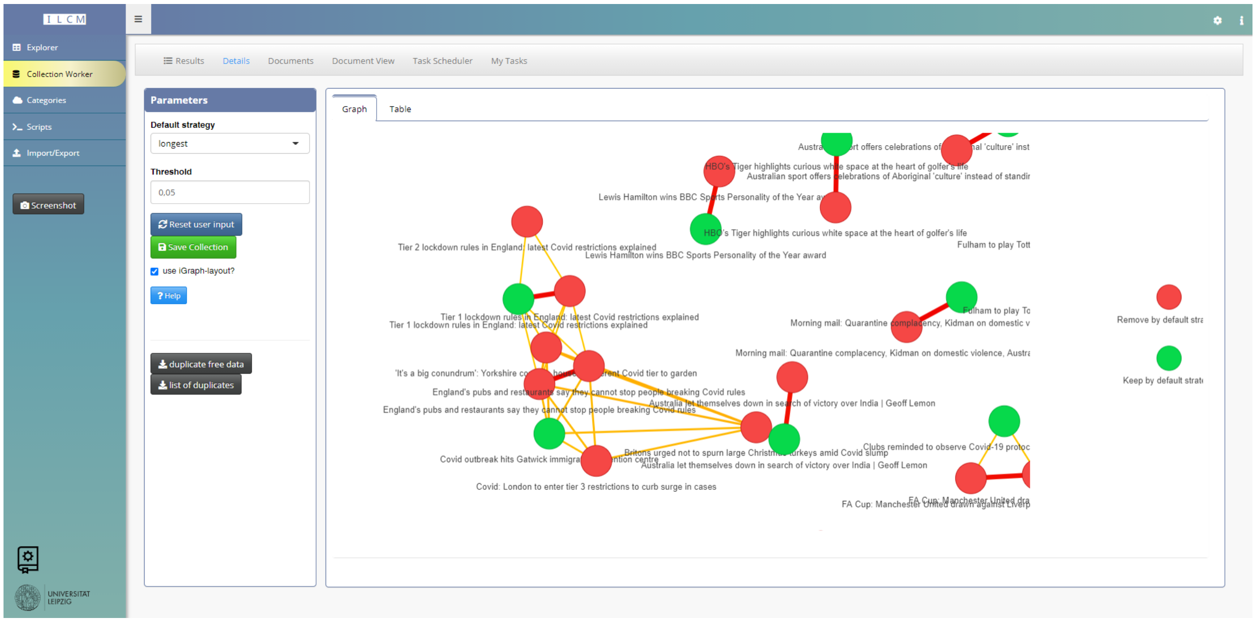
	word	cosine similarity
1	pandemic	0.779067076229508
2	covid19	0.678257935406693
3	covid	0.672308940187044
4	coronavirus	0.589445468031484
5	crisis	0.534194579248326
6	response	0.526039041757124
7	economic	0.495938108591312
8	virus	0.495570361875368
9	pandemics	0.480874981741924
10	afflicting	0.457095852424292

Showing 1 to 10 of 10 entries

Figure 80: Cosine simularity of chosen words in a news copus

15.10 Details Document Deduplication

- **Graph**: This section will show the user a graph consisting of all documents as nodes. There will be an edge between two nodes if they are in some way identical. In the parameter box, the user can set the default strategy of placing the nodes in the graph and a threshold defining when two documents are similar. The user has the chance to download either a list of duplicates or duplicate free data. They can also reset their input or save the current collection of words. Within the graph, the user can zoom in in order to see the names of the nodes.


I L C M
Explorer
Collection Worker
Categories
Scripts
Import/Export
Screenshot
UNIVERSITÄT LEIPZIG
Results
Details
Documents
Document View
Task Scheduler
My Tasks
Parameters
Reset user input
Save Collection
? Help
duplicate free data
list of duplicates
Graph
Table
Austra ort offers celebrations of al 'culture' inst
HBO's Tiger highlights curious wh e space at the heart of golfer's life
Australian sport offers elebrations of Aboriginal 'culture' instead of standir
Lewis Hamilton wins BBC Sp rts Personality of the Year a
HBO's Tiger highlights curious white space at the heart of golfer's life
Fulham to play Tott
Tier 2 lockdown rules in England: latest Covid restrictions explained
Lewis Hamilton wins BBC Sports Personality of the Year award
Tier 1 lockdown rules in England: latest Covid restrictions explained
Tier 1 lockdown rules in England: latest Covid restrictions explained
Fulham to play To
Morning mail: Quarantine complacency, Kidman on domestic v
Remove by default stra
Morning mail: Quarantine complacency, Kidman on domestic violence, Austra
Keep by default strat
'It's a big conundrum': Yorkshire co house erent Covid tier to garden
England's pubs and restaurants say they cannot stop people breaking Covid rules
Australia let themselves down in search of victory over India | Geoff Lemon
England's pubs and restaurants say they cannot stop people breaking Covid rules
Clubs reminded to observe Covid-19 protoc
Britons urged not to spurn large Christmas turkeys amid Covid slump
Covid outbreak hits Gatwick immigra ntion centre
Australia let themselves down in search of victory over India | Geoff Lemon
Covid: London to enter tier 3 restrictions to curb surge in cases
FA Cup: Manchester United drawn against Liverp
FA Cup: Manchester United dra

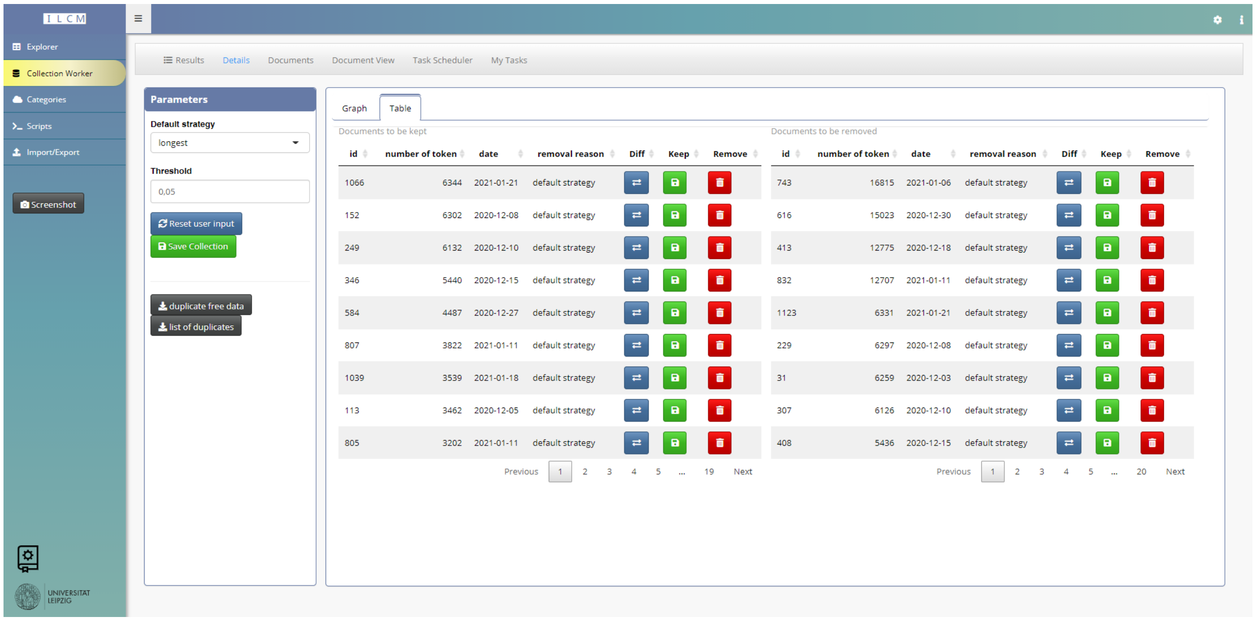
Default strategy
longest

Threshold
0,05

☑ use iGraph-layout?

Figure 81: Part of a news document graph, zoomed in to an example of simular documents

- **Table**: Here, the user will find two data tables. The first one will show a list of documents the system would keep (to prevent duplicates), and in the second one, the user will find documents the system would delete from the collection. The user can now scroll through the lists and either remove or keep a document in his collection by clicking on the corresponding button in the document row. They can also check manually if the document is similar to another one. Therefore they need to click on the button in the *Diff* column of a document row and select via the given document-id-list a document they want to compare the selected one with.



Default strategy: longest

Threshold: 0,05

Documents to be kept

id	number of token	date	removal reason	Diff	Keep	Remove
1066	6344	2021-01-21	default strategy	⇄	💾	🗑
152	6302	2020-12-08	default strategy	⇄	💾	🗑
249	6132	2020-12-10	default strategy	⇄	💾	🗑
346	5440	2020-12-15	default strategy	⇄	💾	🗑
584	4487	2020-12-27	default strategy	⇄	💾	🗑
807	3822	2021-01-11	default strategy	⇄	💾	🗑
1039	3539	2021-01-18	default strategy	⇄	💾	🗑
113	3462	2020-12-05	default strategy	⇄	💾	🗑
805	3202	2021-01-11	default strategy	⇄	💾	🗑

Previous 1 2 3 4 5 ... 19 Next

Documents to be removed

id	number of token	date	removal reason	Diff	Keep	Remove
743	16815	2021-01-06	default strategy	⇄	💾	🗑
616	15023	2020-12-30	default strategy	⇄	💾	🗑
413	12775	2020-12-18	default strategy	⇄	💾	🗑
832	12707	2021-01-11	default strategy	⇄	💾	🗑
1123	6331	2021-01-21	default strategy	⇄	💾	🗑
229	6297	2020-12-08	default strategy	⇄	💾	🗑
31	6259	2020-12-03	default strategy	⇄	💾	🗑
307	6126	2020-12-10	default strategy	⇄	💾	🗑
408	5436	2020-12-15	default strategy	⇄	💾	🗑

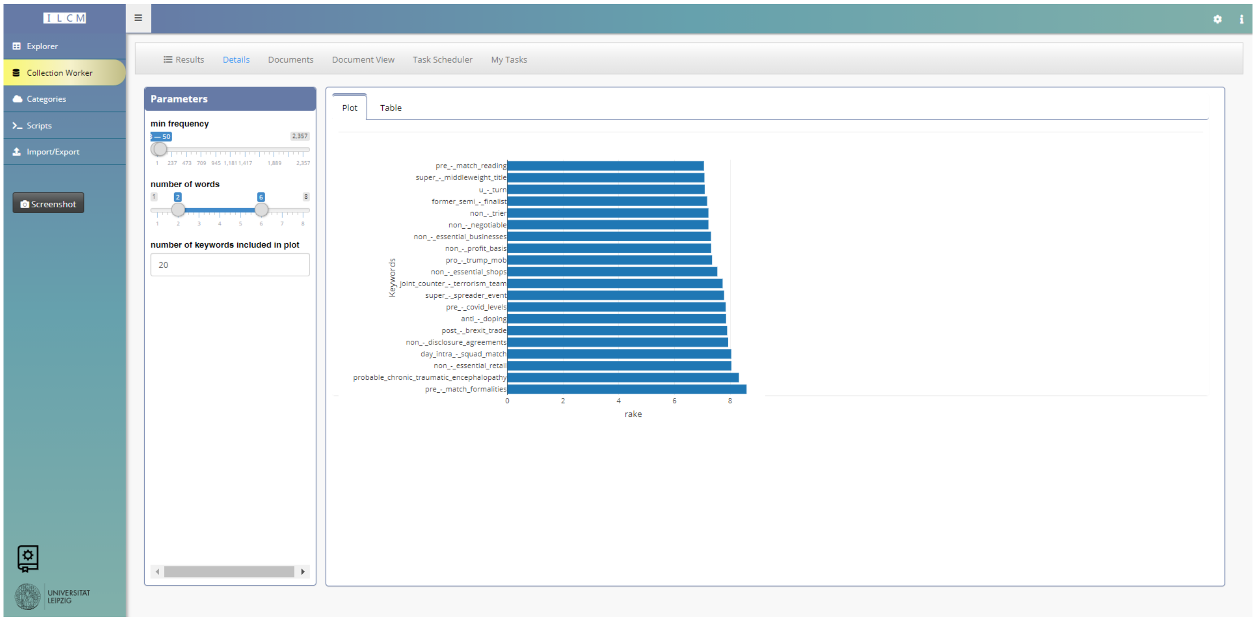
Previous 1 2 3 4 5 ... 20 Next

Figure 82: Data tables of documents to keep and delete from a collection according to theyre similarity to each other

15.11 Details Keyword Extraction

The parameter in this section will depend on the corresponding method but in some sort they will always allow the user to specify the frequency each word needs to fit in, in order to appear in the plot along with a parameter to specify the number of words/phrases the user wants to see.

- **Plot**: Here the user will see a plot of selected keywords or key-phrases and their frequency. The plot can be changed by setting the parameters for a frequency interval or a limit for the number of worda within key-phrases, etc.



min frequency: 3 — 50 (1 – 2,357)

number of words: 2 – 6 (1 – 8)

number of keywords included in plot: 20

Figure 83: Plot for the keyword extraction with the rake method

- **Table**: In this subsection, the user will see the extracted keywords in a data table. With the parameter, they can again specify which specific kind of keywords they wish to see. They also can copy this list of keywords or download it as a csv-file or in excel-format.

15.12 Details Syntactic Parsing

- **Search**: Here, the user can get an overview from syntactic parsing by selecting a specific subject, predicates and objects he can specify the sentence results they wish to see. The changing of these parameters will result in a change of the list of documents. The user can select a specific sentence in the data table and see the calculated syntax tree. This sentence will be extracted from the documents in the collection.

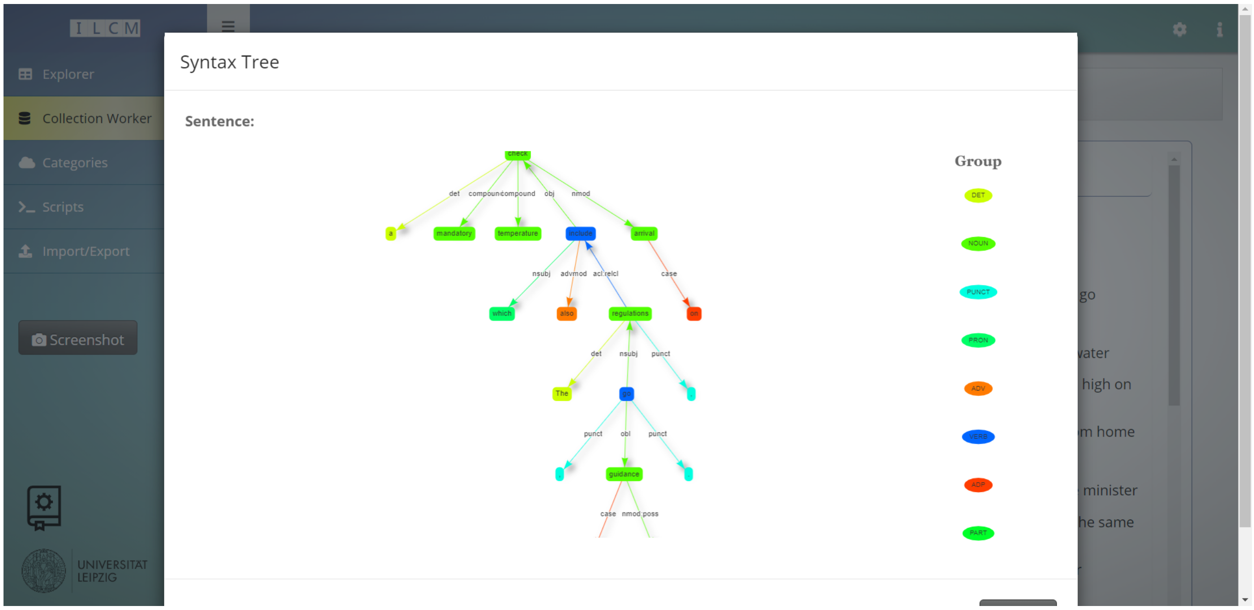


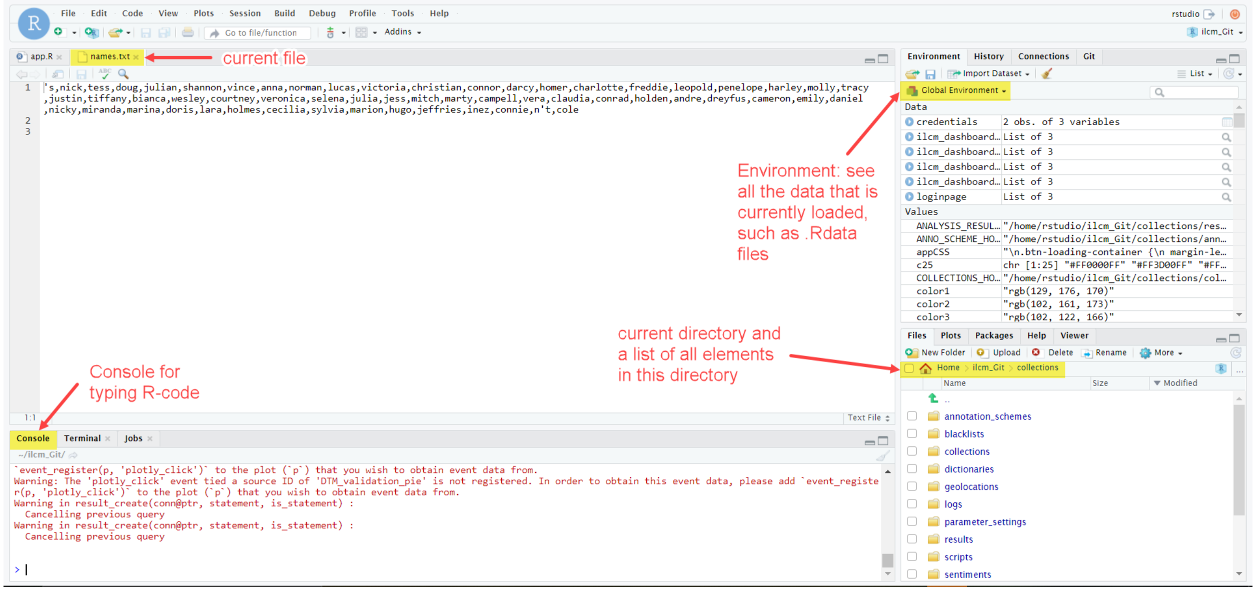
Figure 84: Syntax tree for a sentence containing the selected words 'regulation' and 'include'

16 Closing iLCM and restarting

The iLCM can be stopped using the starndard docker command for managing running containers (docker stop). All results, collections, annotations, blacklists, and essential data to the database and Solr will also be available when restarting the images. This is secured with the use of docker volumes. The stop and the start commands are using the name of the running container. The name can be seen using 'docker ps'.

17 Working with RStudio

While using the local version of iLCM the user can open the whole iLCM-tool in RStudio. This allows the experienced user to get a quick overlook of all the data used for performing the NLP-Tasks. After importing the iLCM project the user can see a couple of folders. In the folder *collections* the user will find all personalized data from the iLCM, such as *annotation-schemes*, *blacklists*, *whitlists*, and so on. Clicking through these folders the user will find *Rdata* or *CSV* files labeled like in the iLCM program. In the folder *logs* the user can additionally find the the log-entries of all started task separated into *failed*, *finished*, and *running*. This is similar to the lists the user will see under the *Collection Worker* tab in the *My Task* subtab. To change existing files in RStudio just open them by clicking on the name in the directory list then the selected file will open and the user can change the content of the file respecting the format of the file, don't forget to save the changes after editing the files. Keep in mind that *.Rdata* files will not open as a file but be loaded into the global environment of the RStudio distribution. To change these kind of files use the console and write the corresponding R-commands. After the changes are made the user can search for the *app.R* file in the iLCM directory and open the file and click on the play button that will appear on the top right in order to start the iLCM tool.



current file

Environment: see all the data that is currently loaded, such as .Rdata files

current directory and a list of all elements in this directory

Console for typing R-code

```
's,nick,tess,doug,julian,shannon,vince,anna,norman,lucas,victoria,christian,connor,darcy,homer,charlotte,freddie,leopold,penelope,harley,molly,tracy
,justin,tiffany,bianca,wesley,courtney,veronica,selena,julia,jess,mitch,marty,campell,vera,claudia,conrad,holden,andre,dreyfus,cameron,emily,daniel
,nicky,miranda,marina,doris,lara,holmes,cecilia,sylvia,marion,hugo,jeffries,inez,connie,n't,cole
```

```
`event_register(p, 'plotly_click')` to the plot (`p`) that you wish to obtain event data from.
Warning: The 'plotly_click' event tied a source ID of 'DTM_validation_pie' is not registered. In order to obtain this event data, please add `event_registe
r(p, 'plotly_click')` to the plot (`p`) that you wish to obtain event data from.
Warning in result_create(conn@ptr, statement, is_statement) :
  Cancelling previous query
Warning in result_create(conn@ptr, statement, is_statement) :
  Cancelling previous query

> 
```

Data	
credentials	2 obs. of 3 variables
ilcm_dashboard…	List of 3
ilcm_dashboard…	List of 3
ilcm_dashboard…	List of 3
ilcm_dashboard…	List of 3
loginpage	List of 3
Values	
ANALYSIS_RESUL…	"/home/rstudio/ilcm_Git/collections/res…
ANNO_SCHEME_HO…	"/home/rstudio/ilcm_Git/collections/ann…
appCSS	"\n.btn-loading-container {\n margin-le…
c25	chr [1:25] "#FF0000FF" "#FF3D00FF" "#FF…
COLLECTIONS_HO…	"/home/rstudio/ilcm_Git/collections/col…
color1	"rgb(129, 176, 170)"
color2	"rgb(102, 161, 173)"
color3	"rgb(102, 122, 166)"

..
□ annotation_schemes
□ blacklists
□ collections
□ dictionaries
□ geolocations
□ logs
□ parameter_settings
□ results
□ scripts
□ sentiments

Figure 85: Overview of Rstudio

18 For Questions

If you are facing any questions, feel free to contact kahmann@informatik.uni-leipzig.de. For more information please visit the iLCM website.

19 Problens and Feature Requests

Any problems and feature requests can be reported via the Github-Page.