

iLCM – A Virtual Research Infrastructure for Large-Scale Qualitative Data

Andreas Niekler¹, Arnim Bleier², Christian Kahmann¹, Lisa Posch^{2,3}, Gregor Wiedemann¹, Kenan Erdogan², Gerhard Heyer¹, Markus Strohmaier^{4,2}

¹Faculty of Mathematics and Computer Science, University Leipzig, Augustusplatz 10, 04109 Leipzig, Germany {aniekler, kahmann, wiedemann, heyer}@informatik.uni-leipzig.de ²Department of Computational Social Science, GESIS – Leibniz Institute for the Social Sciences Unter Sachsenhausen 6-8 50667 Cologne, Germany {firstname.lastname}@gesis.org ³Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria ⁴HumTec Institute, RWTH Aachen University Theaterplatz 14, 52062 Aachen, Germany markus.strohmaier@humtec.rwth-aachen.de

Computational social science (CSS) is the interdisciplinary study of socio-cultural phenomena through new kinds of data and technologies. One of its central objectives is the **extraction of useful and interpretable knowledge from potentially large behavioral digital datasets**.

In CSS, unstructured data (usually text) and structured data (e.g. metadata, survey data or log/sensor data) are both important sources of information. Through the digitization of social processes, they are available in large quantities. With the iLCM infrastructure, we want to offer analysis capabilities for reusable and reproducible research with both data types.



Users of computer-assisted text analysis from these disciplines face non-trivial technical and methodological challenges, especially when it comes to the requirement to analyze large amounts of text and reproducibility of research. The iLCM infrastructure provides solutions to these challenges.

The iLCM architecture consists of two major components, the Leipzig Corpus Miner (LCM) to preprocess and analyze large amounts of unstructured text data, and the ORC component to further process structured data extracted from text or imported from external sources. Analysis scripts on structured data, so-called notebooks, will be published in a central repository. The LCM provides a browser-based Graphical User Interface (GUI) targeting end-users from applied disciplines who are primarily interested in the application of generic text mining methods on their data. The Open Research Computing (ORC) environment runs scripts in active executable documents, so-called "notebooks". By this, the ORC component allows analyzing structured data either extracted from texts by the LCM or imported from external database.

Leipzig Corpus Miner (LCM)

The LCM is not a stand-alone program, but rather a **server infrastructure** comprising a number of components including a **document database** (MariaDB), an **NLP pipeline** for preprocessing text data (spaCy), **a full-text index** (Solr), a **collection of text mining processes** and finally a **web application GUI** (R Shiny).



Eas	urah Doculto	Time Serier	Document View Facets							
Sea		Time Series								Simple Detaile
1288 Do id ∳	cuments were	e found!	title	author	date ≑ tol	ken ≑	keyword.and.context	ວ k	eep	Simple search. Terms may be c + = AND # = OR
32330	13.736834	News	No More Pain for UK Banks in 2017 BoE Tests, but Brexit Risks Ahead	By REUTERS	2017-11- 28	680	time since 2014, but the Bank of England warned of pain ahead if there is no Brexit deal and said the	•		- = NOT Keyword brexit
13525	13.744796	News	UK Opposition Look to Force Government Into Publishing Brexit Dossiers	By REUTERS	2017-10- 31	264	release its assessment of the impact Brexit will have on 58 different sectors of the economy.The	•		Use raw text
54235	13.766976	News	Farage Meets Barnier, Warns of 'No Deal' on Brexit	By REUTERS	2018-01- 08	251	opt for a hard Brexit if the European Union does not offer an agreement on financial services.Farage	•		Search
42774	13.781782	News	Bank of England's Brexit Views in Focus as Rates Set to Stay on Hold	By REUTERS	2017-12- 14	517	LONDON — The Bank of England's views on Brexit will be scrutinised by investors on Thursday when	¥		Save Collection
39189	13.781782	News	Brexit Deal a Much-Needed Win for UK's Embattled Theresa May	By THE ASSOCIATED PRESS	2017-12- 08	530	LONDON — Few people will be more relieved at Friday's long-sought U.KEU deal on Brexit than	¥		Collection Name:
17345	13.796283	News	Bank of England's Carney Sees Slower Growth Without Brexit Deal	By REUTERS	2017-11- 05	311	a future trading deal with the European Union after Brexit, Bank of England Governor Mark Carney	¥		Save Collection
52038	13.838283	News	Tony Blair Tells UK Voters: Time Is Running Out to Stop Brexit Folly	By REUTERS	2018-01- 04	637	out to reverse Brexit, a folly that he said would torpedo Britain's remaining clout and be regretted	•		
49922	13.845696	News	Departing UK Infrastructure Czar Vows to Fight Brexit	By THE ASSOCIATED PRESS	2017-12- 30	126	Brexit policy.Andrew Adonis said Saturday he will campaign next year to keep Britain inside the Europ	ean 💌		
38237	13.884807	News	UK Government Still Undecided on Final Brexit Aims: Hammond	By REUTERS	2017-12- 06	508	LONDON — Britain's government has not yet decided what it wants from a final Brexit deal because	۲		
52283	13.960379	News	Tony Blair Warns British Voters: Time Is Running Out to Stop Brexit Folly	By REUTERS	2018-01- 04	548	out to reverse Brexit, a folly he said would torpedo Britain's remaining clout and be regretted for	•		
Page:									129	
		14	27 40	53	66				129	

LCM Document Search. Screenshot of the document search interface.

Open Research Computing (ORC)

The ORC component extends the iLCM with an **editor environment for program scripts** which enables the **creation of scripts along with their documentation**. GESIS will host a public instance of the ORC environment. This ORC instance is then **extended with a repository on which notebooks can be published** by users. The repository not only **secures the longterm availability** of published notebooks, it also serves as a notebook gallery that **users can browse and search via keywords and metadata**.

	m		

Probability: 0.1501	Name for Subcollection
better oter on the must free	brexit_police_attacks
change politics let hard	Save Subcollection
be vet cial remain	

LCM Topic Model Results. Screenshot of a topic model visualization.



LCM Word Context / Co-occurrence Results. Screenshot of word context and co-occurrence graph visualizations.

Gesis Leibniz Institute for the Social Sciences	Contact	<u>About</u>	
GESIS Notebooks	(beta)		



Project information

The project iLCM is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in the funding line for virtual research environments (FKZ/project number: 324867496). Funding period: 02/2017–01/2020

Research institutions:

for the Social Science

GESIS – Leibniz Institute for the Social Sciences Department of Computational Social Science, Dr. Arnim Bleier

University Leipzig Faculty of Mathematics and Computer Science Department of Natural Language Processing Prof. Dr. Gerhard Heyer



Home Your Server Binder			
	8 binder		
executable environment, ma	Jupyter notebooks? With Binder, op aking your code immediately reprodu		
Build and launch a repository			
GitHub repo or URL			
GitHub repository name or link			
Git branch, tag, or commit	Path to a notebook file (optional)		
master	Path to a notebook file (optional)	File 🗸	launch
Copy the URL below and share your B	linder with others:		
Fill in the fields to see a URL fo	r sharing your Binder.		Ê

Instance of an ORC-environment hosted by GESIS. It is based on Jupyter Notebooks and binder (https://binderhub.readthedocs.io/en/latest/).

© 2018 – iLCM Projekt, Uni Leipzig / GESIS Köln ilcm.informatik.uni-leipzig.de