

# Entity-Based Document Classification on the CORD - 19 Corpus

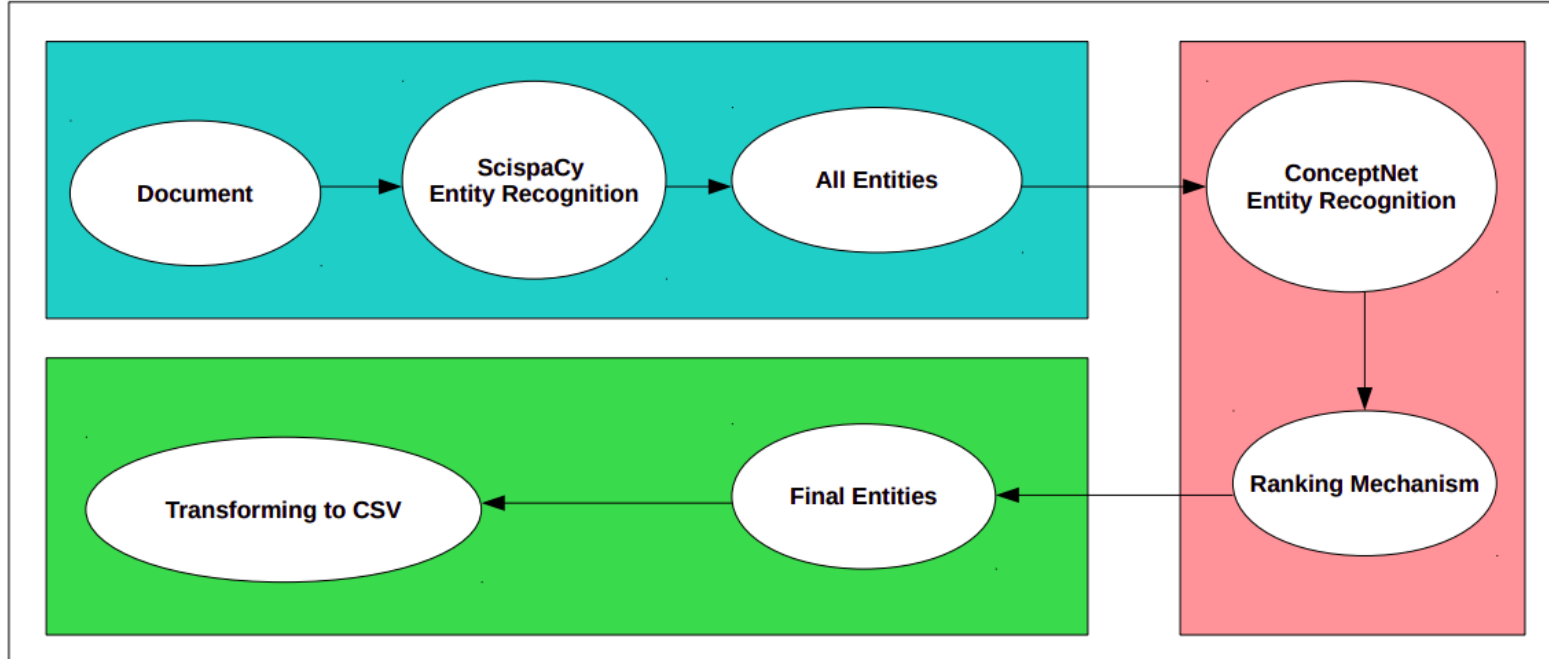
Gollam Rabby and Tomas Kliegr

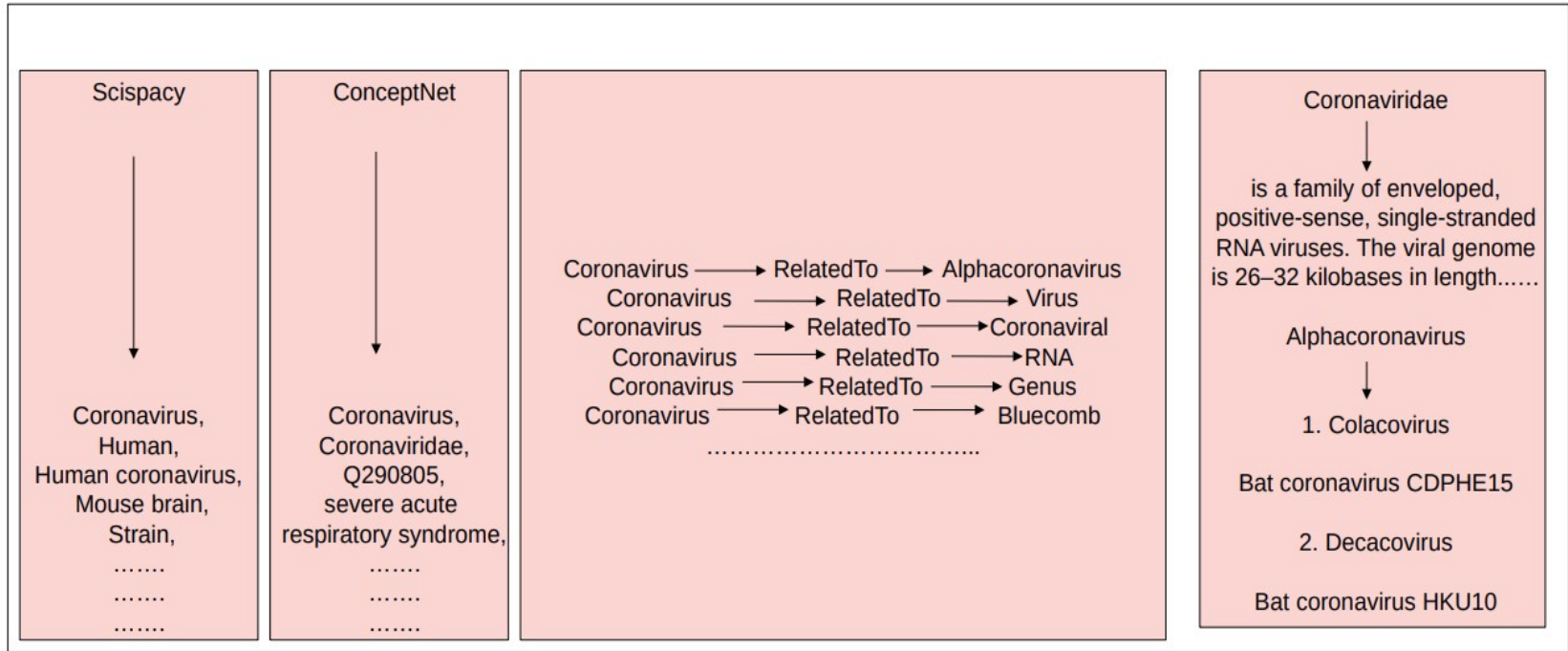
Department of Information and Knowledge Engineering, VSE University, Prague,  
Czech Republic

# Project

1. Transforming CORD-19 to a flat csv file (e.g. with resources corresponding to columns) to which standard rule learning tools/algorithm can be applied.
2. The task is to predict the (academic) success of a paper (as measured by citations).
3. Since we use an "explainable" machine learning tool/algorithm, we could find which combination of concepts (e.g. chemical substances) is predictive of paper success.

# Project - Pre-processing data





# Mining from Tabular: Result of preprocessing

	B	C	D	E	F	G	H	I	J	K	L	DQ
1	DOI	novel	coronavirus	infections	china	study	virus	epidemic	incubation	period	days	Citedby
2	1.17/s134-2-5985-9											
3	1.138/s41421-2-147-1											
4	1.339/jcm92538	1	1	1	1	1	1	1	1	1	1	None
5	1.339/jcm92575		1				1					None
6	1.17/s134-2-5976-w											
7	1.116/j.idm.22.2.1	1	1	1							1	
8	1.116/j.idm.22.2.2	1	1		1		1	1				1
9	1.116/s2214-19x(2)365-6											None
10	1.193/jtm/taaa3					1			1	1		None
11	1.1128/mBio.2764-19											
12	1.1186/s41256-2-137-4							1			1	None
13	1.287/156-7917.ES.22.25.5.28			1					1	1	1	[1;10]
14	1.193/bioinformatics/btaa145	1	1		1							None
15	1.3346/jkms.22.35.e79		1		1							[1;10]
16	1.339/nathons92148	1	1		1		1					

# Mining Tabular: Example results of rule mining (Bayesian Rule Set mining)

Antigenic & Antitoxin and Cold	→	OpenCitations_Ontology([10;100])
Antigenic & Annual	→	OpenCitations_Ontology([10;100])
DNA & Antigenic & Diagnosis	→	OpenCitations_Ontology([10;100])
Information & Annual & Diagnosis	→	OpenCitations_Ontology([10;100])
DNA & Years & Diagnosis	→	OpenCitations_Ontology([10;100])
Antigenic & Years & People	→	OpenCitations_Ontology([10;100])
Epidemic & Clinical manifestations	→	OpenCitations_Ontology([10;100])
Middle east respiratory syndrome coronavirus & Effective	→	OpenCitations_Ontology([10;100])

# Summary

---

1. Ways to generate higher quality entities, assign weights to entities, remove uninteresting entities. Currently, we have experimented with Scispacy, ConceptNet, Scispacy with ConceptNet and TF-IDF model.
2. We use a number of citations (OpenCitations Ontology) as a proxy of the significance of results reported in the paper. Do you have a better suggestion?
3. Use SBRL, CORELS and Random Forest for finding the combination of concepts from research papers.

# Future Work - PubAnnotation

PubAnnotation

Make your annotation public, and more useful!

[Home](#) [Collections](#) [CORD-19](#)

English 日本語 [signup](#) [login](#)

REPOSITORY SEARCH ANNOTATORS EDITORS EVALUATORS NEWS DOCUMENTATION

CORD-19

Description

CORD-19 (COVID-19 Open Research Dataset) is a free, open resource for the global research community provided by the Allen Institute for AI: <https://pages.semanticscholar.org/coronavirus-research>.

As of 2020-03-20, it contains over 29,000 full text articles. This CORD-19 collection at PubAnnotation is prepared for the purpose of collecting annotations to the texts, so that they can be easily accessed and utilized.

If you want to contribute with your annotation,

1. take the documents in the CORD-19\_All\_docs project,
2. produce your annotation to the texts using your annotation system, and
3. contribute the annotation back to PubAnnotation (HowTo).

All the contributed annotations will become publicly available. Please note that, during uploading your annotation data, you do not need to be worried about slight changes in the text: PubAnnotation will automatically catch them and adjust the positions appropriately.

Once you have uploaded your annotation, please notify it to [admin@pubannotation.org](mailto:admin@pubannotation.org) [admin@pubannotation.org](mailto:admin@pubannotation.org), so that it can be included in this collection, which will make your annotation much easily findable.

Note that as the CORD-19 dataset grows, the documents in this collection also will be updated.

*IMPORTANT: CORD-19 License agreement requires that the dataset must be used for text and data mining only.*

Maintainer Jin-Dong Kim

Projects

Name	T	Description	# Ann.	Author	Maintainer	Updated_at	RDFized_at	Status	
CORD-19_All_docs	CC	All the docume...	0		Jin-Dong Kim	2020-03-23	-	Released	
CORD-19_bioRxiv_medRxiv...	CC	The bioRxiv/m...	0		Jin-Dong Kim	2020-03-23	-	Released	
CORD-19_Commercial_use...	CC	The Commerci...	0		Jin-Dong Kim	2020-03-23	-	Released	
CORD-19_Custom_license...	CC	The Custom lic...	5.08 M		Jin-Dong Kim	2020-04-10	-	Released	
CORD-19_Non-commercial...	CC	The Non com...	0		Jin-Dong Kim	2020-03-23	-	Released	

[1] <http://pubannotation.org/collections/CORD-19>



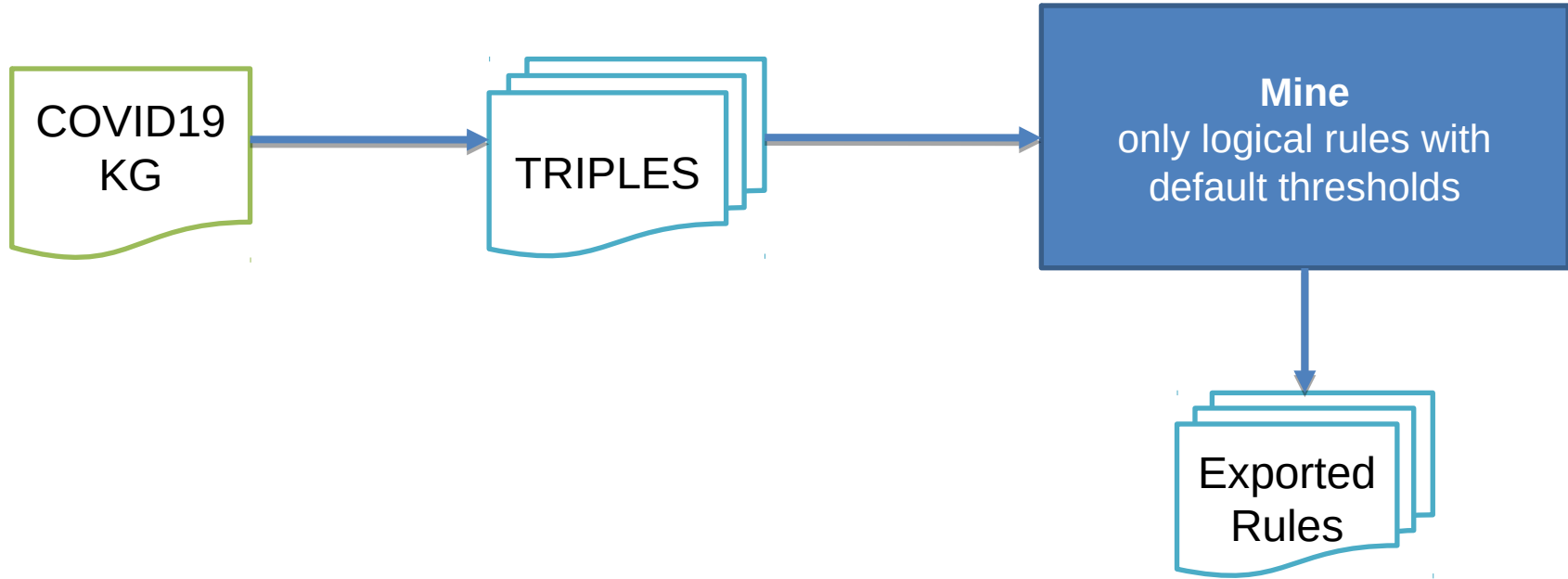
# Future Work – Demo Results

(?a <<https://www.ica.org/standards/RiC/ontology#publishedBy>> <<http://dbpedia.org/resource/Elsevier>>) ^ (?b <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> <<http://dbpedia.org/ontology/ChemicalCompound>>) -> (?a <<http://idlab.github.io/covid19#hasConcept>> ?b) | support: 81987, headCoverage: 0.04837402115577961, headSize: 1694856

(?a <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> <<http://purl.org/spar/fabio/JournalArticle>>) ^ (?b <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> <<http://dbpedia.org/ontology/ChemicalCompound>>) -> (?a <<http://idlab.github.io/covid19#hasConcept>> ?b) | support: 161969, headCoverage: 0.09556505095418136, headSize: 1694856

(?a <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> <<http://purl.org/spar/fabio/Work>>) ^ (?b <<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>> <<http://dbpedia.org/ontology/ChemicalCompound>>) -> (?a <<http://idlab.github.io/covid19#hasConcept>> ?b) | support: 178542, headCoverage: 0.10534346280745975, headSize: 1694856

# Future Work - Mining data with RDFRules



**THANK YOU,  
FRONTLINERS!**

