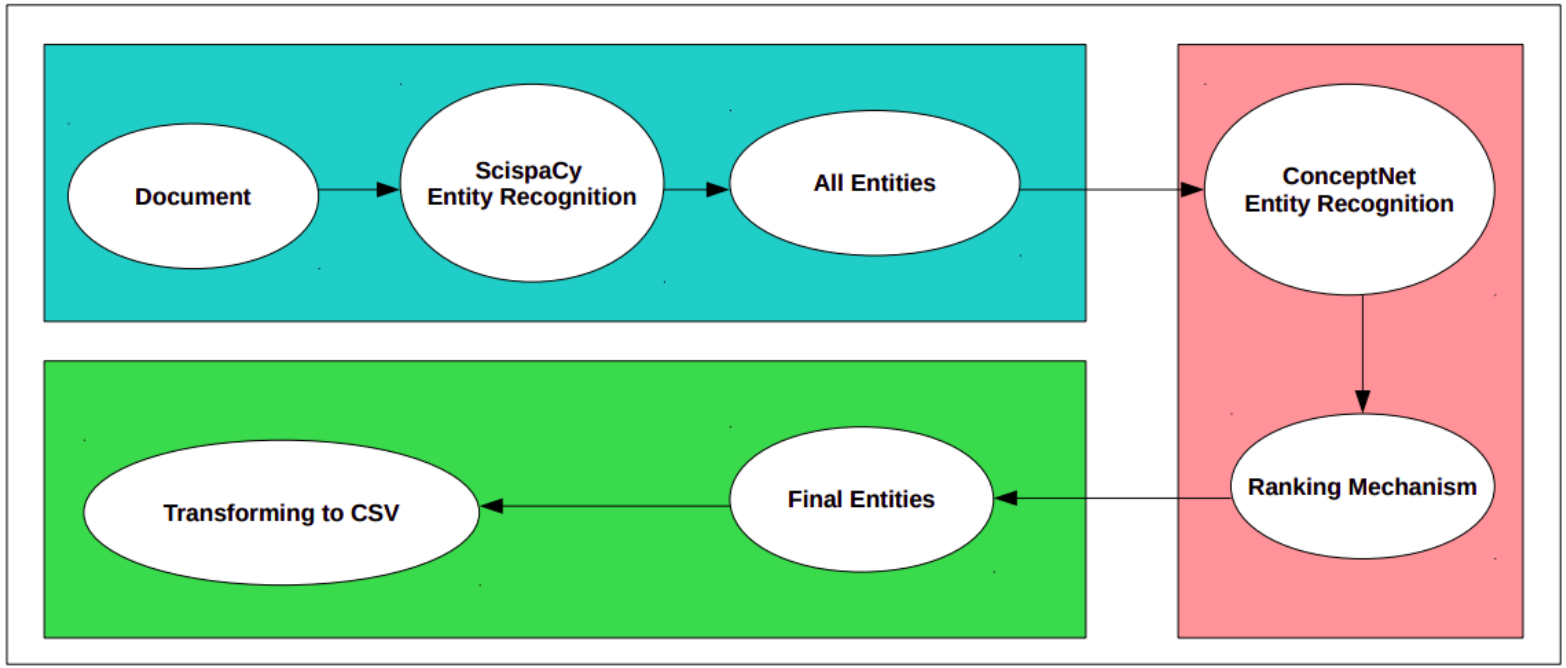
Entity-Based Document Classification on the CORD - 19 Corpus

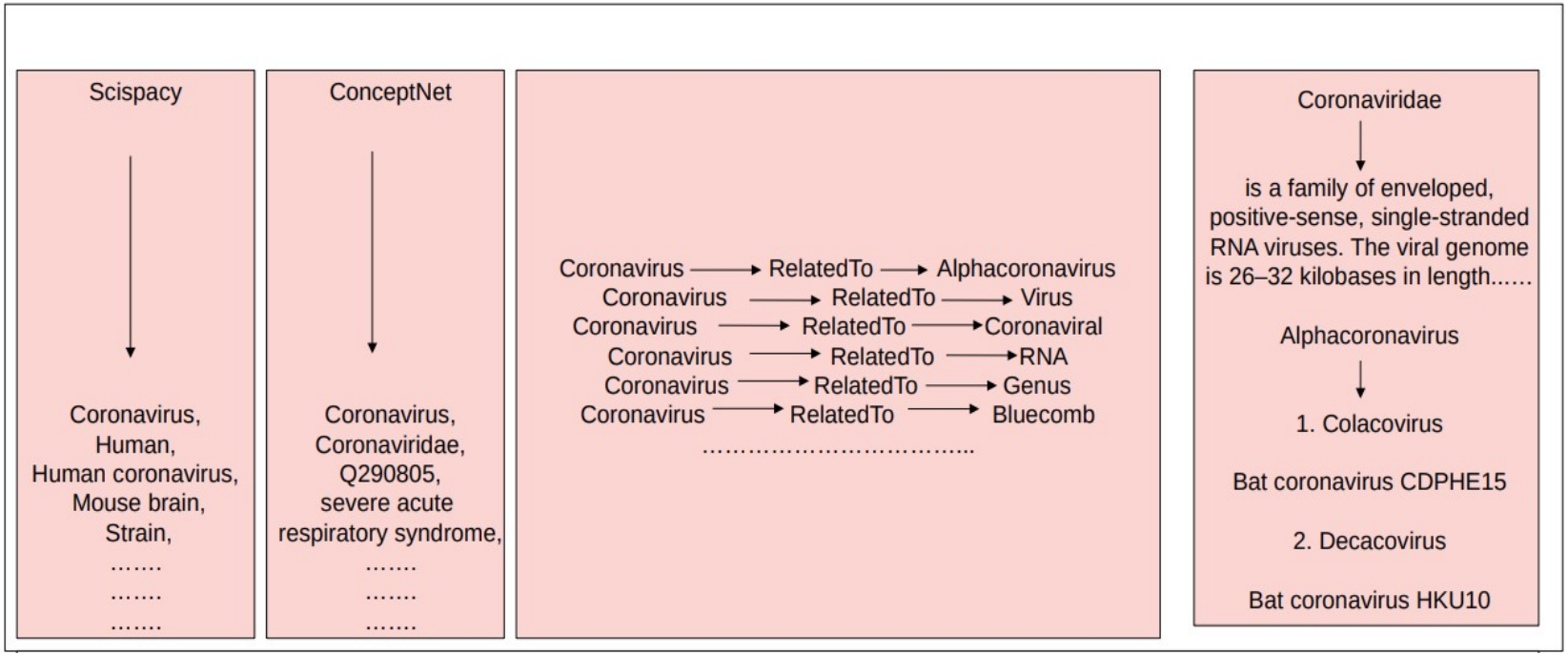
Gollam Rabby and Tomas Kliegr
Department of Information and Knowledge Engineering, VSE University, Prague, Czech Republic

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 explainable machine learning tools/algorithms, we could find which combination of concepts (e.g. chemical substances) is predictive of paper success.

Pre-processing data

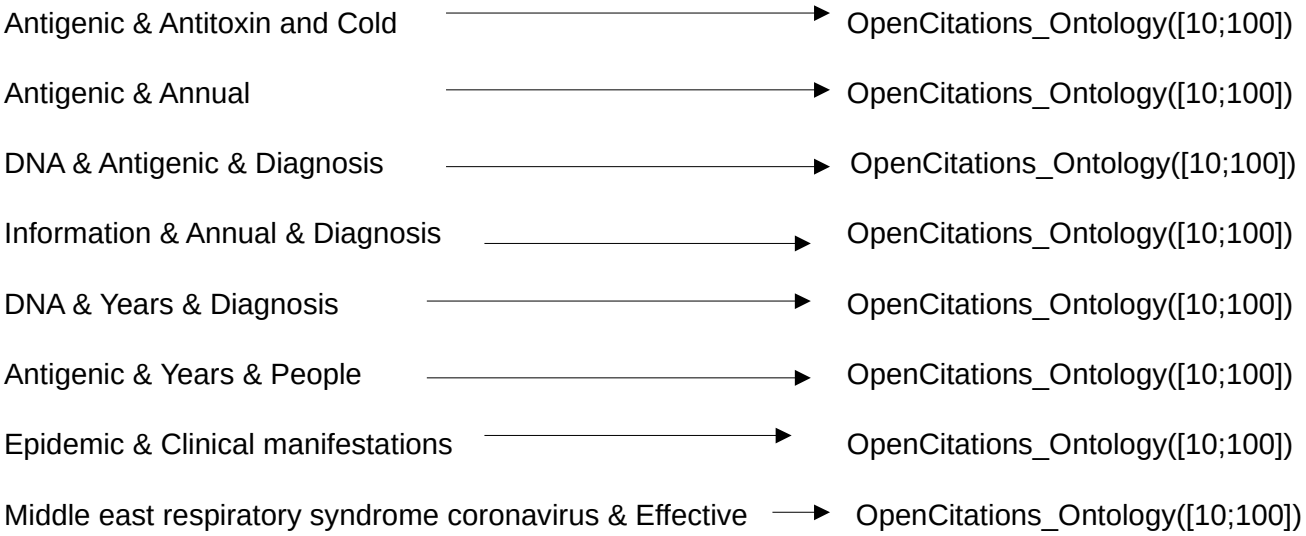



Scispacy
Coronavirus,
Human,
Human coronavirus,
Mouse brain,
Strain,
.....
.....
.....
ConceptNet
Coronavirus,
Coronaviridae,
Q290805,
severe acute
respiratory syndrome,
.....
.....
.....
Coronavirus → RelatedTo → Alphacoronavirus
Coronavirus → RelatedTo → Virus
Coronavirus → RelatedTo → Coronaviral
Coronavirus → RelatedTo → RNA
Coronavirus → RelatedTo → Genus
Coronavirus → RelatedTo → Bluecomb
.....
Coronaviridae
is a family of enveloped, positive-sense, single-stranded RNA viruses. The viral genome is 26–32 kilobases in length.....
Alphacoronavirus
1. Colacovirus
Bat coronavirus CDPHE15
2. Decacovirus
Bat coronavirus HKU10

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/pathogens92148	1	1		1		1					

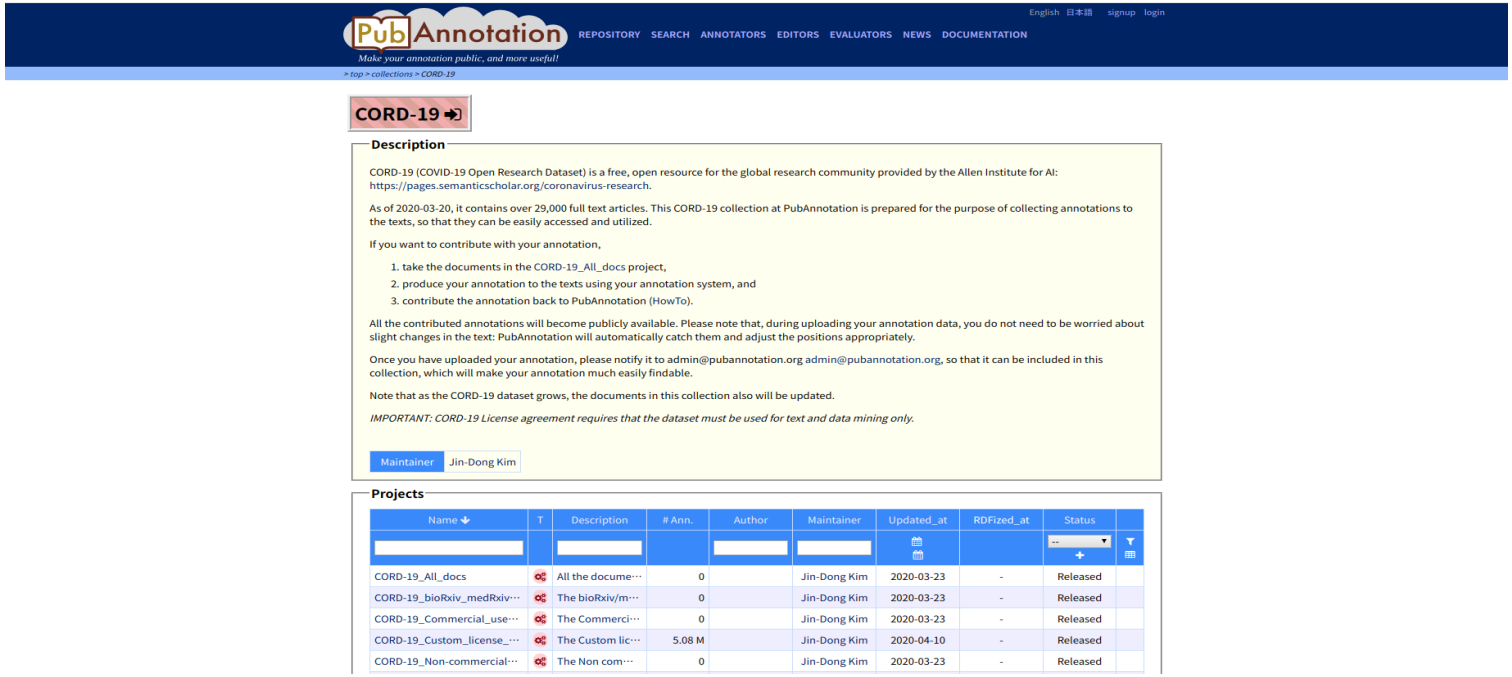
Mining Tabular: Example results of rule mining (Bayesian Rule Set mining)



Summary

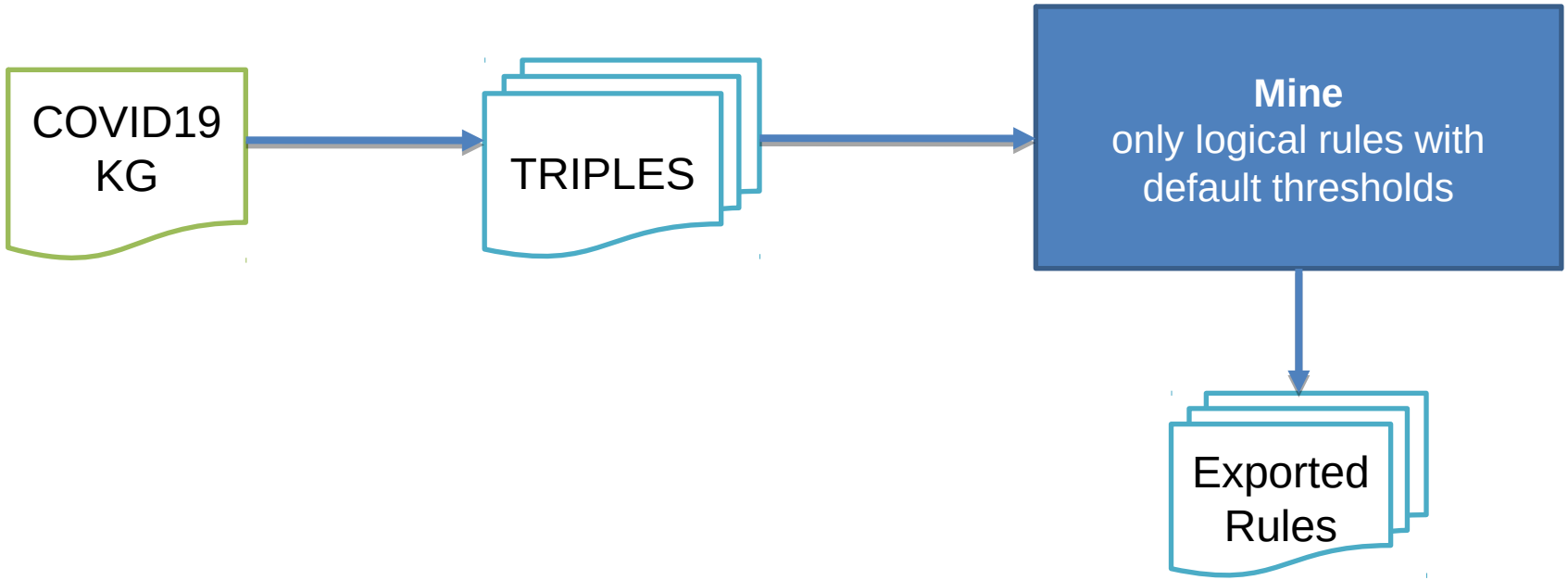
1. Ways to generate higher quality entities, assign weights to entities, remove uninteresting entities. Currently, we have experimented with Scispacy, ConceptNet, Scispacy with ConceptNet and TF-IDF model.
2. We use a number of citations (OpenCitations Ontology) as a proxy of the significance of results 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



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

Future Work - Mining data with RDFRules



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: 0.048374021155779 61, 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


THANK YOU,
FRONTLINERS!