Entity-Based Document Classification on the CORD - 19 Corpus

Gollam Rabby and Tomas Kliegr

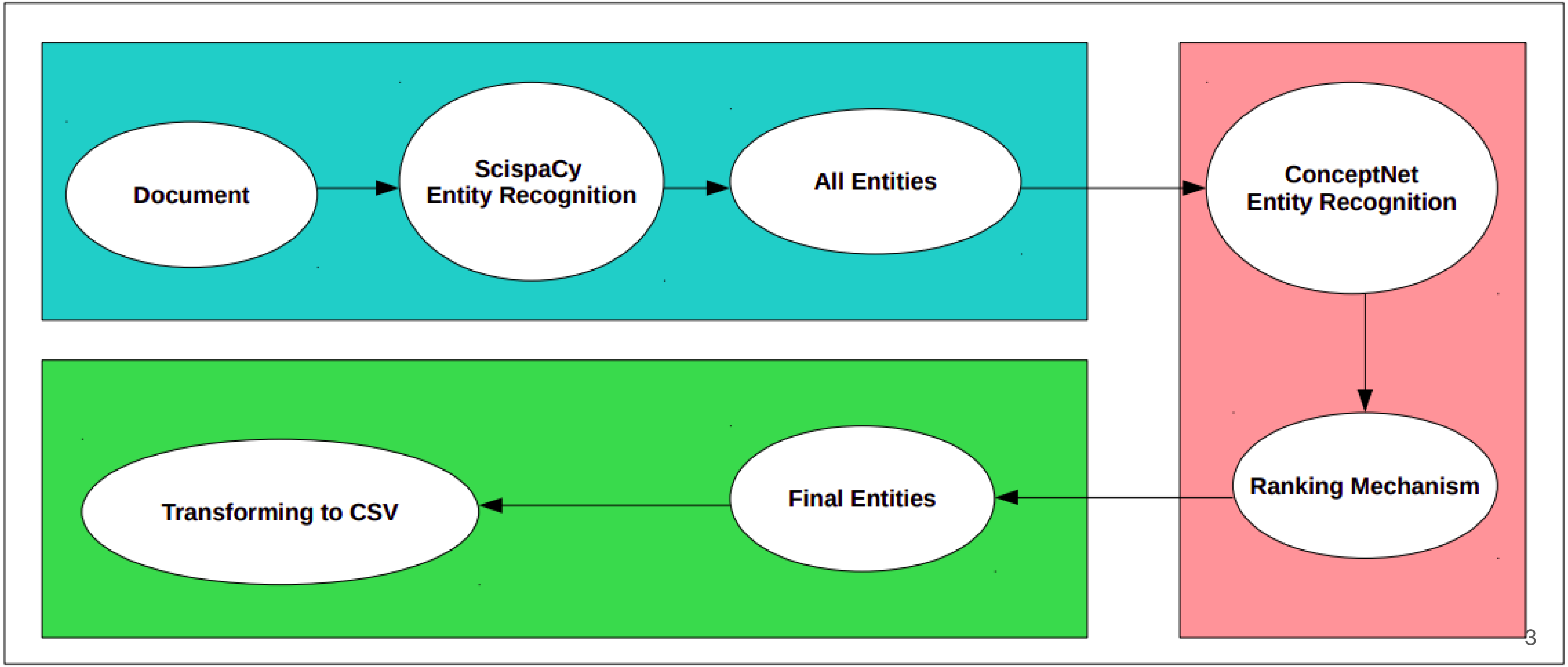
Department of Information and Knowledge Engineering, VSE University, Prague, Czech Republic

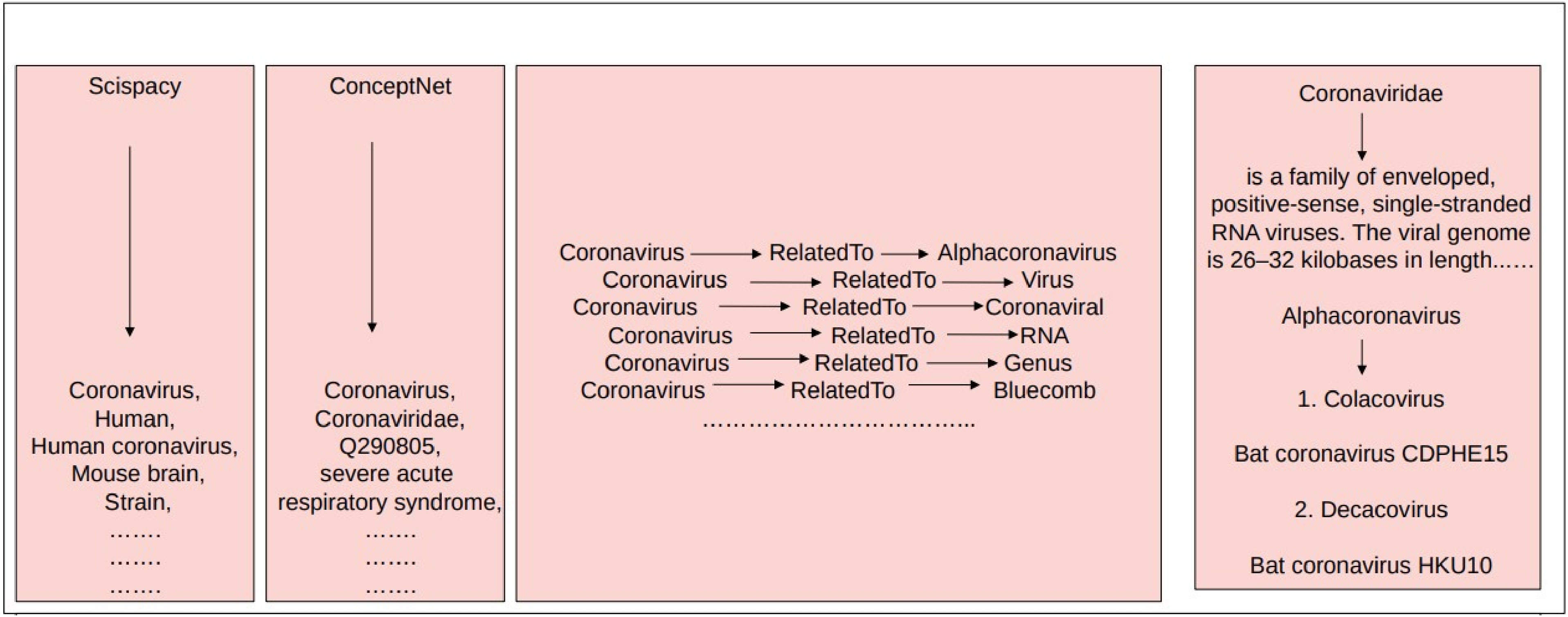
Project

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

Sample discovered rules

Project - Preprocessing 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

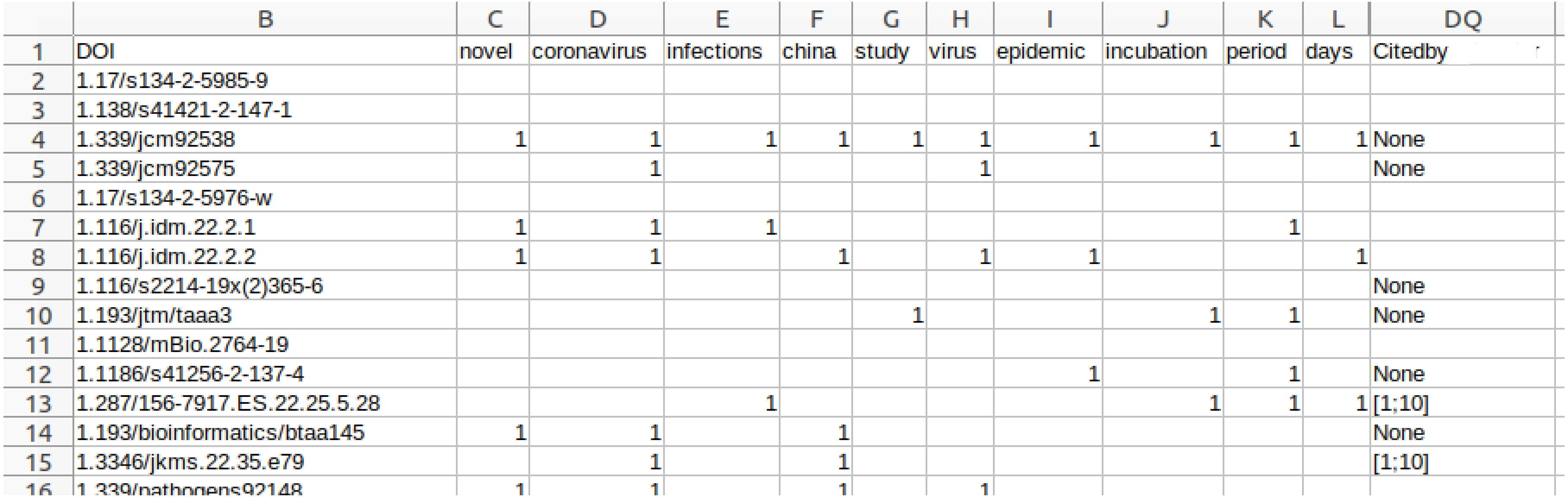
1. Colacovirus

Bat coronavirus CDPHE15

2. Decacovirus

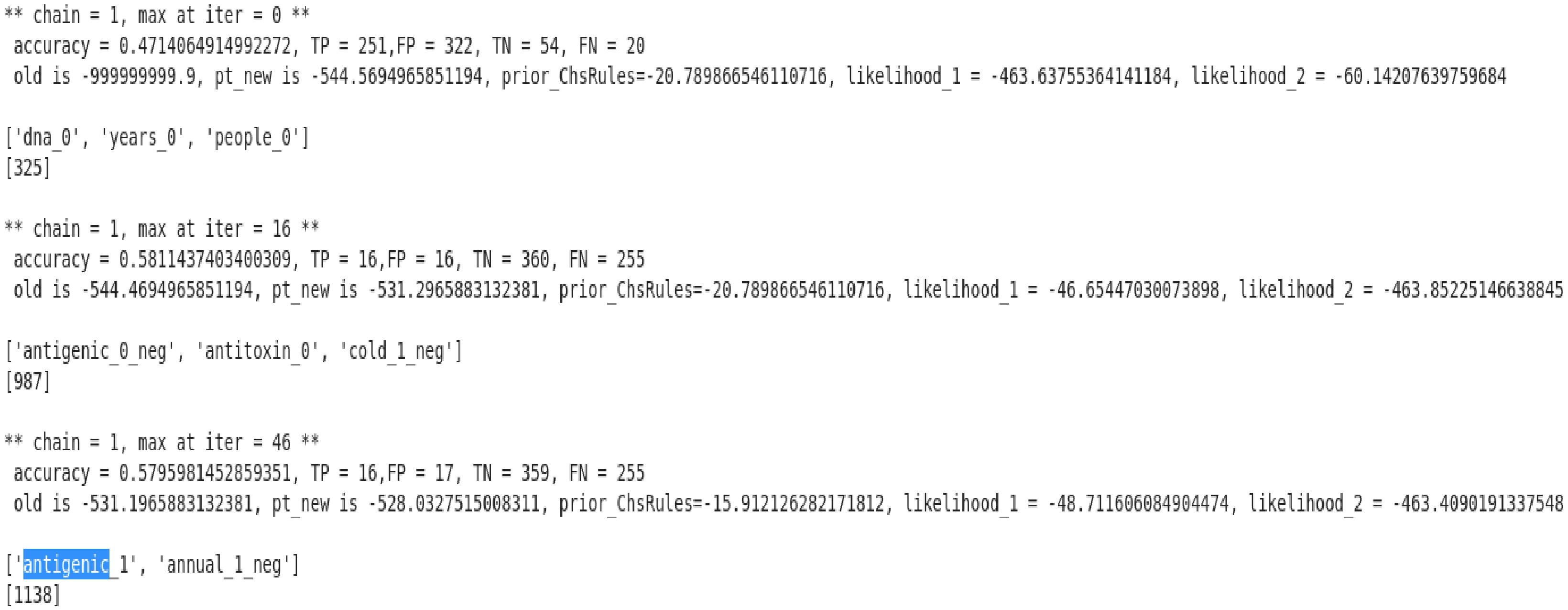
Bat coronavirus HKU10

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: Association rule mining with Bayesian Rule Set mining



```
** chain = 1, max at iter = 0 **
 accuracy = 0.4714064914992272, TP = 251,FP = 322, TN = 54, FN = 20
 old is -999999999.9, pt_new is -544.5694965851194, prior_ChsRules=-20.789866546110716, likelihood_1 = -463.63755364141184, likelihood_2 = -60.14207639759684

['dna_0', 'years_0', 'people_0']
[325]

** chain = 1, max at iter = 16 **
 accuracy = 0.5811437403400309, TP = 16,FP = 16, TN = 360, FN = 255
 old is -544.4694965851194, pt_new is -531.2965883132381, prior_ChsRules=-20.789866546110716, likelihood_1 = -46.65447030073898, likelihood_2 = -463.85225146638845

['antigenic_0_neg', 'antitoxin_0', 'cold_1_neg']
[987]

** chain = 1, max at iter = 46 **
 accuracy = 0.5795981452859351, TP = 16,FP = 17, TN = 359, FN = 255
 old is -531.1965883132381, pt_new is -528.0327515008311, prior_ChsRules=-15.912126282171812, likelihood_1 = -48.711606084904474, likelihood_2 = -463.4090191337548

['antigenic_1', 'annual_1_neg']
[1138]
```

Summary

Feedback appreciated

We would appreciate any pointers to code in the Jupyter notebook, particularly:

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?

Future Work

1.Building a knowledge graph (KG)
2. Prediction of missing triples in KG
3. Classification in KGs
4. Clustering of similar rules