LDA2Net: Digging under the surface of COVID-19 topics in literature

Topic 39 companion sheet

G. Minello C.R.M.A. Santagiustina M. Warglien

This file contains the following supplementary information for Topic 39 of the manuscript "*LDA2Net*: Digging under the surface of COVID-19 topics in scientific literature":

- Human label and automatic n-gram label proposals (Table 1)
- Summary measures (Table 2)
- Network of top 25 bigrams (Figure 1)
- Wordclouds of top 25 words by node relevance measure (Figure 2)
- Wordclouds of top 25 bigrams by edge relevance measure (Figure 3)
- Filtered (0.99 percentile) topic network (Figure 4)

Table 1: Human and automatic label proposals. Automatic label candidate for largest word community of the topic. In parenthesis: absolute frequency of the walk out of a sample of size 1000.

Human label	2-gram label	3-gram label	4-gram label
data	data->collection (12.7%)	data->collection->analysis (6%)	data->collection->sources->used (1.9%)

Here follows the set of topic-specific measures that have been used to classify the topic and to analyse its structural properties (see manuscript for details):

Table 2: Summary measures

JSD	Mean propensity	Variance propensity	Modularity	Barrat Clustering Coeff.
 0.750685	0.008148	0.000216	0.000246	0.619882
94	56	36	47	105

Based on the aforementioned measures, Topic 39 has been classified as a CROSS-CUTTING topic.

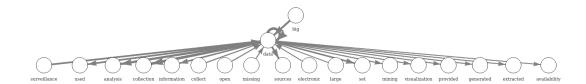


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

developed
collect electronic
potools database
electronic
p

obtained
electronic
provided
p

gathered surveillance
generated obtained
obtained
generated generated
provides
Collection multiple
sources extracted
information inform
visualization
availability

provided
extracted to record
information baccollecting
or collect to be sources to be sources open to be sources open to be source surveillance to missing set to the source surveillance to the source surveillan

availability surveillance per surveillan

extracted developed visualization toolssurveillance use de sources identify system source of collection believe de source information Collection believe multiple provides

Out-degree Betweenness PageRank

Figure 2: Top 25 unigrams (i.e., nodes) by measure.



Figure 3: Top 25 bigrams (i.e., edges) by measure.

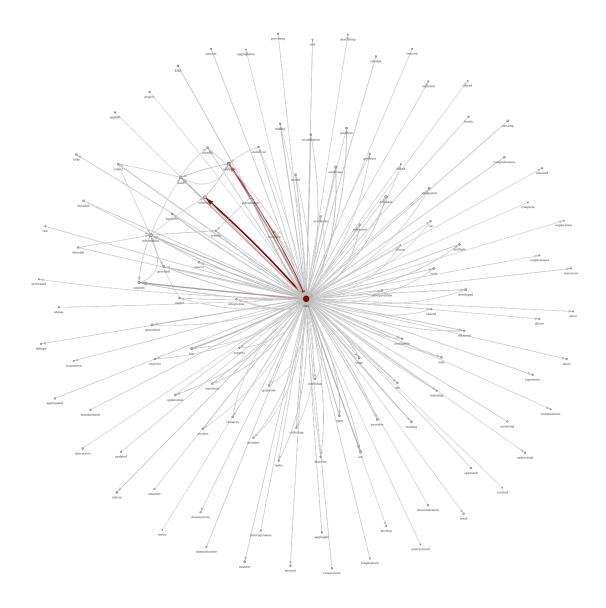


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness centrality. Isolated nodes have been removed after filtration.