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

Topic 3 companion sheet

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

This file contains the following supplementary information for Topic 3 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
statistical analysis (studies)	analysis->used (10.6%)	data->analysis->used (5.1%)	data->analysis->used->analyzed (1.2%)

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.
value	0.751242	0.008042	0.000188	0.319357 117	0.622856
rank	95	52	27		108

Based on the aforementioned measures, Topic 3 has been classified as a CROSS-CUTTING topic.

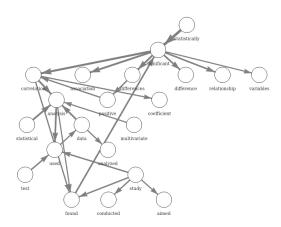


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

determine determine analyses

significance or used factors as ignificant or used factors or us

relationship statistical of analyzed study of difference statistically of the statistical of the stat

statistically statistically statistically statistical by statistic

strong strong strong separation with the strong separation strong separation strong separation strong separation studytest strong stron

coefficients difference difference difference difference sassociation analysis pearson total test used expensive correlationship determine

difference
assess relationship
statistically variables
differences correlation
statistically variables
differences correlation
statistically variables
differences correlation
statistically variables
difference
statistically variables
difference
assessingly
statistically variables
difference
assessingly
statistically variables
difference
assess relationship
populous
statistically variables
difference
assess relationship
populous
statistically variables
difference
statistically variables
difference
statistically variables
difference
statistically variables
differences correlation
differences correlati

Out-degree Betweenness PageRank

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

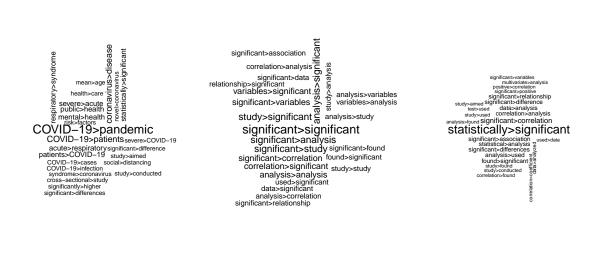


Figure 3: Top 25 bigrams (i.e., edges) by measure.

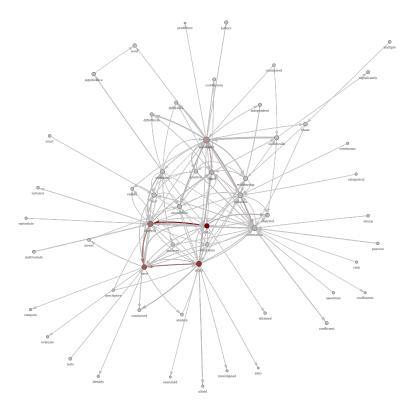


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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