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

Topic 105 companion sheet

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

This file contains the following supplementary information for Topic 105 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
pandemic in selected countires	italy->spain (19.7%)	italy->spain->one (16.1%)	italy->spain->one->france (7.8%)

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 (0.650692	0.007870	0.000190	0.145092	0.579160
	57	40	28	92	64

Based on the aforementioned measures, Topic 105 has been classified as a CROSS-CUTTING topic.

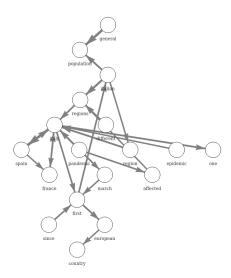


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.



lompardy general since brazil since brazil since brazil regions or italy march pandemic different affected region european beginning

beginning
casecountry
brazil european
sinceitalian
eginstitaly tally tally tally tally
francespain
state march
region
different
epidemic



measures be region italian puip spain country france one since one first number pandemic pandemic pandemic pandemic peuropean adopted peuropean adopted peuropean adopted peginning

pandemic
state different
sincepopulation
region spain france
italian italy country
april first cone level
data regions beriod
number epidemic

Out-degree Betweenness PageRank

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

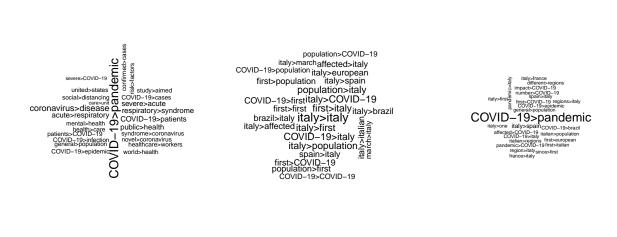


Figure 3: Top 25 bigrams (i.e., edges) by measure.

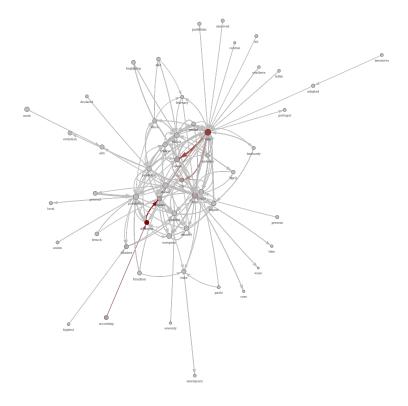


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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