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

Topic 4 companion sheet

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

This file contains the following supplementary information for Topic 4 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
mortality rate related to infection	number->cases (17.9%)	number->cases->deaths (3.3%)	number->cases->deaths->per (1.4%)

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	0.785059	0.009139	0.000319	0.000000	0.584476
rank	100	100	64	22	72

Based on the aforementioned measures, Topic 4 has been classified as a CROSS-CUTTING topic.
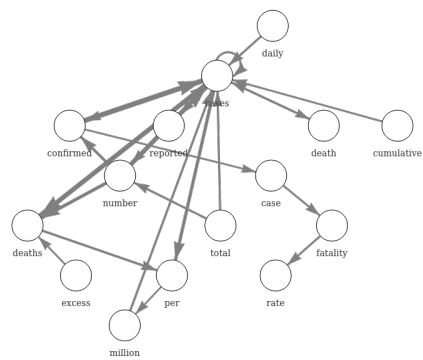
Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

LDA probability

Degree

In–degree

Out–degree

Betweenness

PageRank

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

cases>COVID–19
risk>factors   case>fatality
number>deaths   number>cases
health>care social>distancing
cases>deaths   severe>acute
mortality>rate   acute>respiratory
novel>coronavirus public>health
coronavirus>disease
COVID–19>pandemic
COVID–19>cases
confirmed>cases
COVID–19>patients
respiratory>syndrome
syndrome>coronavirus
united>states confirmed>COVID–19
COVID–19>infection
patients>COVID–19
fatality>rate

number>number
deaths>deaths
cases>per case>cases
rate>cases
cases>total   cases>confirmed
fatality>cases cases>number
cases>daily deaths>cases
cases>cases
cases>deaths total>cases
number>cases
confirmed>cases cases>fatality
cases>rate cases>death
cases>case   death>cases
number>deaths million>cases
deaths>number
daily>cases

excess>deaths
cumulative>cases
cases>cases
daily>cases
reported>cases
cases>number
cases>COVID–19
number>confirmed
deaths>per number>deaths
cases>deaths
confirmed>cases
number>cases cases>death
COVID–19>cases total>cases
cases>per million>cases
total>number
cases>confirmed
death>cases cases>reported
deaths>cases fatality>rate
case>fatality

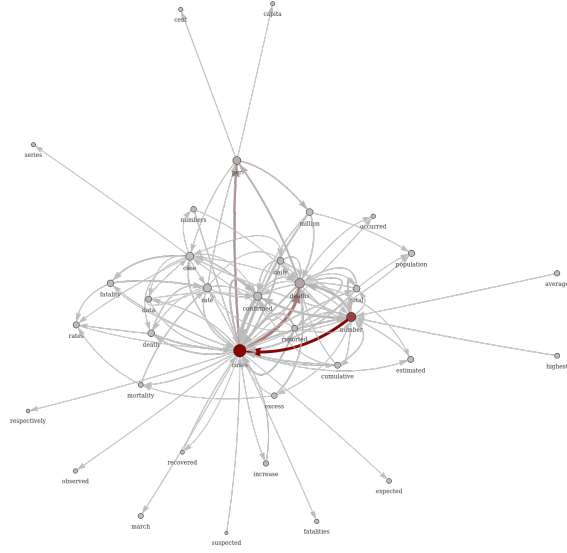Figure 3: Top 25 bigrams (i.e., edges) by measure.

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.