

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

Topic 20 companion sheet

G. Minello

C.R.M.A. Santagiustina

M. Warglien

This file contains the following supplementary information for Topic 20 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 |
|------------------------|-------------------|-------------------------|--------------------------------|
| instrumental diagnosis | chest->CT (42.1%) | chest->CT->scan (10.6%) | chest->CT->scan->images (8.3%) |

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 | 0.611599 | 0.008853 | 0.000712 | 0.081059 | 0.585982 |
| rank | 42 | 88 | 106 | 80 | 76 |

Based on the aforementioned measures, Topic 20 has been classified as a SPECIALIZED topic.

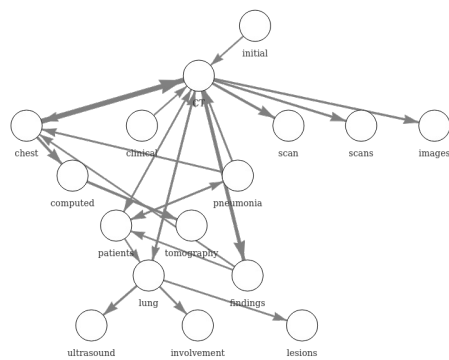


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

A word cloud of terms related to chest CT scan findings. The words are arranged in a circular pattern around the central text 'CT scan'. The words include: abnormalities, involvement, images, signs, computed, findings, suspected, scans, cases, confirmed, lesions, disease, consolidation, diagnosis, ultrasound, patients, clinical, scan, lung, chest, tomography, pneumonia, opacities, and radiological. The words are in various sizes and orientations, with 'CT scan' being the largest and most central.

A word cloud visualization showing various terms associated with COVID-19 research. The most prominent words are "patients" and "pneumonia". Other significant words include "scans", "images scan", "lung", "findings", "chest", "tomography", "computed", "peripheral", "diagnosed", "clinical", "disease", "lesions", "ultrasound", "type", "mixed", "score", "cases area", "coronavirus", and "involvement". The words are arranged in a circular pattern, with their size indicating their frequency or importance in the dataset.

A word cloud of medical terms related to pneumonia. The most prominent words are 'pneumonia', 'findings', 'chest', 'CT', 'patients', 'diagnosis', 'tomography', 'scans', 'ultrasound', 'disease', 'images', 'normal', 'underwent', 'cases', 'lesions', 'involvement', 'radiological', 'lung', 'clinical', 'scan', 'computed', 'confirmed', and 'consolidation'.

3

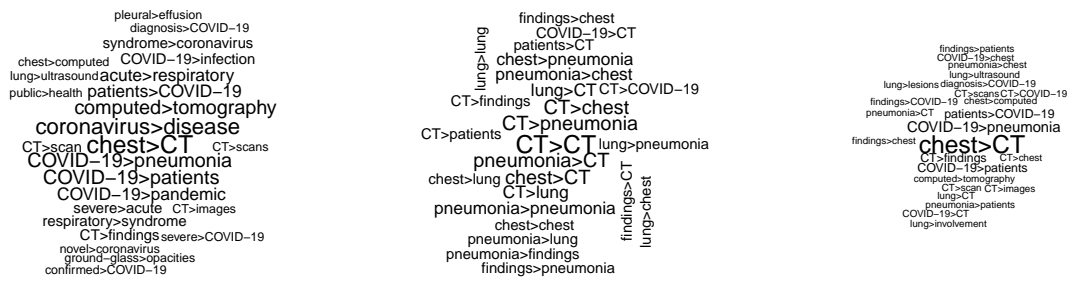


Figure 3: Top 25 bigrams (i.e., edges) by measure.

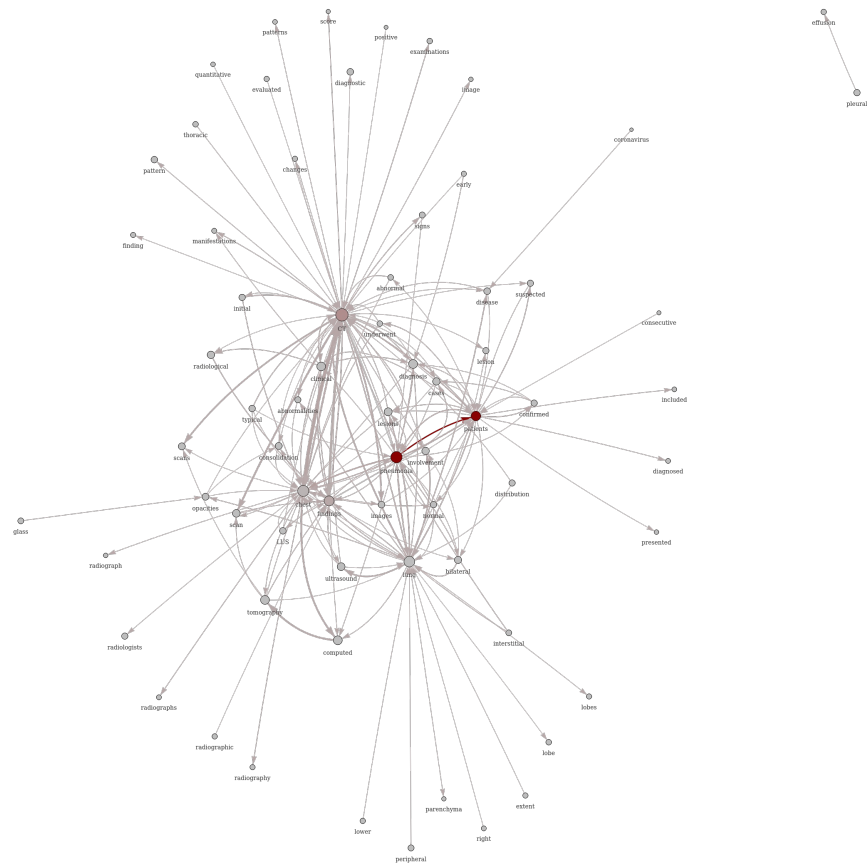


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness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