Automatic Extraction of Risk Factors from COVID-19 Literature

Francis Wolinski

Yotta Conseil, France

francis.wolinski@yotta-conseil.fr

The novel coronavirus, named *SARS-CoV-2*, and the corresponding disease *COVID-19* appear to have disconcerting and devastating effects on patients. Reports on risk factors and co-morbidities are spread in many publications, and do not give an analytic view with actionable information. Being able to collect automatically such risk factors from literature, to visualize their importance, and to provide relevant insights is at stake.

We developed a software project to cope with this problem. It is able to extract automatically from the CORD-19 corpus [1] all diseases referenced in the International Classification of Diseases (ICD-11) maintained by the WHO [2], as well as the diseases which might be considered as risk factors. The extractions rely on a keyword processing third-party library [3], based on Trie dictionary data structure, and suitable to extract complex disease names and their synonyms, as well as risk factors expressions.

Within this project information extraction consists in producing items with a document reference from CORD-19 and one-to-many disease codes from ICD-11 considered possibly as risk factors. The software is currently extracting 4K diseases, including 2K risk factors, out of 60K scientific papers. In addition, it is able to make incremental updates of the extractions by processing upcoming documents, since the CORD-19 dataset is updated daily.

The analytic exploitation of the extractions leads to compute several indicators: shares, occurrences, and document frequency of encountered diseases. These indicators can be computed at any level of the taxonomy provided by ICD-11 from single leaves to top-level branches. Providing insights along with the structure of the diseases' repository is a key point in this project, since it meets the domain's nomenclature.

A graphical exploration dedicated to each computed indicator is proposed to users. The shares of diseases are represented with stacked bar charts in the flat space of diseases. This synthetic view provides a mean to check globally which branches are more related to risk factors. The occurrences of diseases are displayed in tree map graphics along with the ICD-11 taxonomy. This representation provides a kind of fisheye view: it enables to highlight risk factors through the graphical deformation of the taxonomy. Beyond these meaningful graphics, it is also possible to access to the document frequency of diseases independently of their lexical forms.

The final outcome of the project is a dashboard available on the web, see VIDAR-19 [4]. It includes also a dedicated search engine which enables the user to navigate from risk factors to the original documents. The extracted risk factors are compatible with those referenced in general information [5] and literature review [6] which validate our approach. Designed initially for the *COVID-19*, the software could be directly used to process any corpus of scholarly articles dealing with other pathologies to extract their specific risk factors.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni and S. Kohlmeier, "CORD-19: The COVID-19 Open Research Dataset," *ArXiv e-prints*, 2020, arXiv:2004.10706.

[2] World Health Organization, "International classification of diseases for mortality and morbidity statistics (11th revision)." `https://icd.who.int/browse11/l-m/en`, 2018.

[3] V. Singh, "Replace or retrieve keywords in documents at scale," *ArXiv e-prints*, 2017, arXiv:1711.00046.

[4] F. Wolinski, "Visualization of diseases at risk in the covid-19 literature," 2020, arXiv:2005.00848.

[5] Centers for Disease Control and Prevention, "Groups at higher risk for severe illness." `https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/groups-at-higher-risk.html`, 2020.

[6] M. Wadman, J. Couzin-Frankel, J. Kaiser, and C. Matacic, "How does coronavirus kill? clinicians trace a ferocious rampage through the body, from brain to toes," *Science*, doi:10.1126/science.abc3208.