Authors
Tykhonov, V., Polishko, A, Kiulian A.

Who we are
CoronaWhy.org is a global AI/ML community created as a response to the fight against coronavirus spread in the world. It's building an Infrastructure for Open Science that can be distributed, scaled up in the future and reused for other important tasks like cancer research.

Our mission
The vision of the community is to build this Artificial Intelligence infrastructure completely from Open Source components and with publicly available ML models like scispacy developed by Allen AI and other organizations. All data should be published in a FAIR way in the database where all provenance information is also available and reliable, and supported by two different annotations services, Hypothesis for the verification of COVID-19 related papers and Doccano for Natural Language Processing annotations.

Challenges
The main challenge of this work is to get credibility and trust from all involved communities, especially from the medical experts as usually they don't have confidence in the results produced by people from other communities like Computer Science or Scientometrics. This research infrastructure effort should increase the involvement of the medical community in the analysis of COVID-19 research papers and datasets, the transparency of data and services can guarantee the reproducibility of all experiments. CoronaWhy community is using Harvard Data Commons as a foundation for all members to work together on the same problem and organizes efficient communication and collaboration through data exchange and reuse. This experience can be considered as a lesson for other research infrastructures dealing with coronavirus research both in Europe and worldwide

Results - Operating System for Open Science

CoronaWhy Common Research and Data Infrastructure is a distributed and scalable system that can be used for tasks completely different from COVID-19 like cancer or diabetes research. **Anyone can use it, anyone can join and contribute.**

Data preprocessing pipeline implemented on Jupyter notebook Docker with extra modules.
Dataverse as data repository to store data from automatic and curated workflows.
Elasticsearch has CORD-19 indexes on sections and sentences level with spacy enrichments. Other indexes: MeSH, Geonames, GRID.
Hypothesis annotation service is running allows to annotate CORD-19 papers.
Virtuoso triplestore with public SPARQL Endpoint to query COVID-19 Knowledge Graph
Other services: Colab, MongoDB, Kibana, BEL Commons 3.0, INDRA, Geoparser, Tabula

More at: https://github.com/CoronaWhy/covid-19-infrastructure