Risk Factors:

COVID-19

Devon-Eric Lee

University of Hawaii Maui College

**CoronaWhy.org-Task: Risk Factors: Arthur Kiulian**

In this submission this notebook extracts factors were identified as being the most important by the medical community linked to COVID-19. Under demographic factors list there is a 10% of research papers on individuals 60 plus, 8% of research papers on Men and Women at risk. 3% of papers on neonates or newborns, children or babies and mice or rats. 11% of research papers on Infants, Race/Ethnicity at 21%. Sexual orientation at 1%, pregnant women at 8% and family composition at 7%. Finally, is Day Care attendance at 2% and crowding and siblings at 4%. The submission also includes the percentage of papers published regarding prior diseases that aid in the risk factor, environmental factors, genetic factors, lifestyle factors, and socioeconomic factors. This submission also includes the stages in detail beginning with produce N-gram synonyms. This stage utilizes an assortment of linguistic/medical packages to extend the initial input and search for synonymous/similar N-grams (unigrams, bigrams, and trigrams) in the CORD-19 database. Next is searching for N-grams in papers which this stage involves searching for the papers containing these N-grams inside the Abstract, Result, and Method sections of each paper. Following by, quality assessment - medical annotators where in this stage, the output papers from the previous stage undergo a manual examination to secure their place as relevant articles in the search. Continually this submission utilized a degree of automation, and probability. This submission established initial task exploration, it established knowledge base and ultimate list of all risk factors. The realization that they needed to prioritize and filter out factors that are underrepresented in the CORD-19, prioritization based on MD input, semantic exploration of n-gram approach, filtering out most relevant papers, establishing steps for annotators to determine paper relevancy, and producing final list of the best papers based on crowdsourced human/medical input.

**[Assystem] COVID MUST DIE Risk Factors Analysis: Aleksei Iancheruk**

In this submission, it aids scientist get valuable information from research papers within a short period of time. This submission displays how to properly centralize loading and preprocessing of large volumes of data by making use of developed methods and multiprocessing compute. How to explore large quantities of data using topic modelling for extracting relevant topics out of raw text, automate the search of these topics, their visualization and interpretation, how to make use of word embeddings as a robust NLP techniques that will allow us to get similar words based on their distributions but also complex semantic relations between them. How to use word embeddings for Query expansion to power search engines. How to build simple yet powerful search engines to get a list of relevant articles according to user's query, and how to conceive complex graph representations of data to search for strong and weak connections between them but also cluster them based on their importance. A V4 dataset is being used to provide the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. The approach used in this submission for discovering risk factors of COVID-19 is data loading using DataHandler class that will read metadata and raw JSON files. Preprocessing that is used to clean text and prepare it for future use. This step includes language detection, tokenization, lemmatization, stopwards deletion, deaccent and other types of normalization. Then topic modeling which is data exploration techniques used to extract topic from documents if they are correspondent relevant words using classic BOW and TFIDF. Search engine by training of Glove embedding on a subset of document, query expension, and basic search engine using TFIDF. In addition to graph generation and advanced visualizations, most least connected articles and articles clustering.

**CORD-19 Analysis with Sentence Embeddings: David Mezetti**

In this submission this project builds an index over the CORD-19 dataset to assist with analysis and data discovery. A series of tasks were explored to identify relevant articles and help find answers to key scientific questions on several COVID-19 research topics. It begins by installing the full source code from GitHub, then building SQLite articles database. The raw CORD-19 data is stored across a metadata.csv file and json files with the full text. This project uses SQLite to aggregate and store the merged content. The ETL process transforms the csv/json files into a SQLite database. The process iterates over each row in metadata.csv, extracts the column data and ensures it is not a pure duplicate (using the sha hash). This process will also load the full text if available. The next process is tagging where articles are tagged based on keyword matches. The only tag at this time is COVID-19 and articles are tagged with this if the article text contains any of the following regular expressions. Then additional metadata is parsed out of the article to derive information on the study design. Next is design type, which the text is analyzed to determine a design type for the backing study in the article using a machine learning model. The model has a pre-defined vocabulary and features are a count of each of these defined keywords. A Random Forest Classifier is then trained using the feature set and is used to predict study design labels. Moving to attribute type and extraction, using machine learning model which has a combination of features including a TF-IDF vector of the text elements and Natural Language Processing (NLP) elements. The NLP features are built from entity, part of speech and dependency labels extracted with scispacy. Scispacy has been pretrained on medical articles and has good detection on articles in this dataset. A Logistic Regression Classifier is then trained using the feature set and is used to predict attribute labels. Next is adding grammar labels, building the index which takes each COVID-19 tagged, then exploring the data.

References

Alekseiiancheruk. (2020, April 16). [Assystem] COVID MUST DIE Risk Factors Analysis. Retrieved from <https://www.kaggle.com/alekseiiancheruk/assystem-covid-must-die-risk-factors-analysis>

Arturkiulian. (2020, April 16). CoronaWhy.org - Task: Risk Factors. Retrieved from <https://www.kaggle.com/arturkiulian/coronawhy-org-task-risk-factors>

Davidmezzetti. (2020, May 2). CORD-19 Analysis with Sentence Embeddings. Retrieved from <https://www.kaggle.com/davidmezzetti/cord-19-analysis-with-sentence-embeddings>