LING 575 : Value Sensitive Data Processing

Exercise 2

Manisha Singh

1. Select an NLP dataset
   I am selecting the "CORD-19 (COVID-19 Open Research Dataset)" dataset.

2. Select a documentation practice
   I have selected data statement for documentation of CORD-19.

3. Define a set of factors that you think are most relevant to include in a 2-page documentation for the dataset, and motivate why you chose them.
   Below is the table that shows the important factors in case of CORD-19.
   I have prioritized the factors. The logic behind prioritization is every factor is important, but it will be helpful to know which factor is more important over the other.

Factors	Priority P1 – Highest P4 – lowest	Reason of Importance
Header:	P1	Dataset Title: For CORD-19 : Knowing the correct name of the dataset is important to search the dataset in any repository. Dataset Curator(s) : For CORD- 19 : Knowing the author let me to discover the correlation : S2ORC dataset was also created by the same person and to great extent it helps me to understand the usage of S2ORC in the creation CORD-19 dataset. It also helps to reach out the person who created in case of a questions. Dataset Version: For CORD-19: The schema of the dataset for the version released on 2020-03-13, is different than the version released on 2020-05-26. Hence knowing the right version and corresponding schema is important , else it can give rise to compatibility issues. Dataset Citation: FOR CORD-19 : This is important for release of the datastatement.
	P4	Data Statement Author/Version/Citation : FOR CORD-19: It is important but in current state when I am writing the first datastatement it is not much of importance. However should be important if one creates the next datastatement after few years and need to refer this datastatement. Hence, I have given a lower priority.
Executive Summary: Curation Rationale, Language and	P1	For CORD-19: This answered my question, whether I want to pursue this dataset for writing my datastatement. In quick glance I came to know the quantitative , contents and space requirement for the dataset. Since I am interested in information extraction and this dataset in executive summary states this, helped me to decide if I want to write the datastatement for this dataset.

Quantitative information		
Curation Rationale: The reason, logic, and constituents of the dataset	P2	For CORD-19: This section states about the field and content of the dataset accompanied with the reason behind to select the field. For instance in CORD-19 dataset the metadata.csv fields has 'Abstract' and 'Title' fields. By doing the pattern search gives the first list of the papers in a particular domain. So, apart from stating the usability of dataset it also states the schema of the dataset.
Documentation for source dataset : Link to the data source on which the current dataset is based	P2	For CORD-19: The dataset , is collection of the pdf files in json format and the corresponding metadata. The language of the pdf files becomes the language of the metadata which is directly extracted from the source. Also, it is important to know the source for further quality check, dependency, regeneration of the CORD-19 dataset.
Language Varieties: Language used in the dataset.	P3	FOR CORD-19:Although it is important part of the datastatement, but CORD-19 case the metadata file had en-US and all the papers mostly are in en-US. However it was not clear to me how to find other language present in around 1015768 files, I have lowered the priority level.
Speaker Demographic: Detail about the speaker like (age, gender, and other specific information)	P4	For CORD-19: The reason for lower priority is, it was not feasible as of now to find the speaker information due to huge number of the academic research papers in the dataset. The dataset is based on scientific experiments and hence there is less scope of deviation of the text based on person writing the text, although the way of writing would definitely vary.
Speech situation and text characteristics	P3	FOR CORD-19: Text characteristics is an academic research writing. Academician /Scientist with various medical background have written the scripted form of the text.
Preprocessing and data formatting	P1	For COVID-19: This is core of the dataset. The data is collected based on specific filters. Then all the pdf files are parsed with a schema as in S2ORC.json and then formatted. At the same time a metadata is created which consolidates the metadata for all the papers. Cord_19_embediings : This is collection of the precomputed SPECTRE document embeddings for each CORD-19 paper. So, the CORD-19 data is the result of all the preprocessing and formatting and hence it is very important aspect of the data.
Capture quality: Description of quality issues in data capture.	P2	For CORD-19: The Other section – highlights the data quality issues in data capture. One major issue is not capturing all the research papers regarding COVID -19. The data was extracted based on certain filter criteria("*COVID" OR "COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome")* , what if these filter criteria is not exhaustive list to get all the research papers.

Limitation: Challenges that could not be fully addressed.	P2	For CORD-19: the limitation during data extractions and preprocessing did affect the resulting dataset. However the tradeoff was to keep the data than to lose the information by deleting the duplicate rows. Certain design choices were also made to have a sustainable reproducible system in order to avoid overhead in refreshing the dataset with recent published academic papers.
Metadata	P1	For CORD-19: This is a high priority section because papers in CORD-19 and academic papers more broadly are made available under a variety of copyright licenses. So, the curator of a dataset like CORD-19 must pass on best-to- knowledge licensing information to the end user. Hence the priority level is high for this section.
Disclosure and ethical review	P1	For CORD-19: CORD-19 is the collaborative effort between organizations . This work was supported in part by NSF Convergence Accelerator award 1936940, ONR grant N00014-18-1-2193, and the University of Washington WRF/Cable Professorship. This section is of high importance because without proper organizational funding it will be difficult to get the resources from the organizations.
Other	P2	For CORD-19, the author added the FAQ's which to my understanding is very important to understand the logic behind the data quality issues, the structure of the metadata and some duplicates value. Hence those FAQ's I have added in this section. I have added a P2 level priority because it is really important for the user to go through the FAQ if one wants to work with the dataset.

4. Fill out a data statement in light of the aspects you've picked.
   Below is the data statement for CORD-19 dataset.

Contents

Data Statements for CORD-19

1. HEADER

Dataset Title

- The title of the datasets is CORD- 19 and stands for : COVID-19 Open Research Dataset

Dataset Curator(s) [name, affiliation]

- Name: Wang, Lucy Lu and Lo, Kyle and Chandrasekhar, Yoganand and Reas, Russell and Yang, Jiangjiang and Burdick, Doug and Eide, Darrin and Funk, Kathryn and Katsis, Yannis and Kinney, Rodney Michael and Li, Yunyao and Liu, Ziyang and Merrill, William and Mooney, Paul and Murdick, Dewey A. and Rishi, Devvret and Sheehan, Jerry and Shen, Zhihong and Stilson, Brandon and Wade, Alex D. and Wang, Kuansan and Wang, Nancy Xin Ru and Wilhelm, Christopher and Xie, Boya and Raymond, Douglas M. and Weld, Daniel S. and Etzioni, Oren and Kohlmeier, Sebastian
- Affiliation : Allen Institute for AI, IBM Research, Microsoft Research, National Library of Medicine, Kaggle, Chan Zuckerberg Initiative, Georgetown University and University of Washington

Dataset Version [version, date]

Latest release : [cord-19_2022-04-28.tar.gz, 2022-04-28]
Further note on Version and dates:

- Previous version and releases
- At the time of writing this datastatement, CORD-19 is released weekly.
- Planned final release : 2022-06-02
- While CORD-19 was initially released on 2020-03-13, the current schema is defined based on an update on 2020-05-26. Older versions of CORD-19 will not necessarily adhere to exactly the schema defined in this README.
  Please reach out for help *(email)*,on this if working with old CORD-19 versions.

Dataset Citation and DOI:

- Dataset Citation:

```
@inproceedings{wang-etal-2020-cord,
    title = "{CORD-19}: The {COVID-19} Open Research Dataset",
    author = "Wang, Lucy Lu  and Lo, Kyle  and Chandrasekhar, Yoganand
and Reas, Russell  and Yang, Jiangjiang  and Burdick, Doug  and Eide,
Darrin  and Funk, Kathryn  and Katsis, Yannis  and Kinney, Rodney
Michael  and Li, Yunyao  and Liu, Ziyang  and Merrill, William  and Mooney,
Paul  and Murdick, Dewey A.  and Rishi, Devvret  and Sheehan, Jerry  and
Shen, Zhihong  and Stilson, Brandon  and Wade, Alex D.  and Wang,
Kuansan  and Wang, Nancy Xin Ru  and Wilhelm, Christopher  and Xie,
Boya  and Raymond, Douglas M.  and Weld, Daniel S.  and Etzioni, Oren
and Kohlmeier, Sebastian",
    booktitle = "Proceedings of the 1st Workshop on {NLP} for {COVID-19} at
{ACL} 2020",
    month = jul,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1"
}
```

- DOI : https://doi.org/10.48550/arXiv.2004.10706

Data Statement Author(s) [name, affiliation]

[Manisha Singh, University of Washington]

Data Statement Version [version, date]

[v1, 2022-05-03]

Data Statement Citation

Manisha Singh. (2022). Data Statement for the CORD-19(COVID-19 Open Research Dataset). Version 1. University of Washington (UW).

2. EXECUTIVE SUMMARY

CORD-19 is a corpus of academic papers about COVID-19 and related coronavirus research to support text mining and NLP research. This resource is a large and growing collection of publications and preprints on Covid-19 and related historical coronaviruses such as SARS and MERS. The language is en-US. For more refer to section Language Variety.

The dataset comprises of below files.

- changelog: A text file summarizing changes between this and the previous version.
- cord_19_embeddings.tar.gz : 768-dimensional document embedding.
- Metadata.csv: Total number of rows (CORD UID): 1015768
- Document_parses: Consist of 377111 pd_json and 295705 pmc_json files

3. CURATION RATIONALE

This dataset was created to make the repository of all the machine-readable papers related to COVID-19 and its related diseases. The dataset was created so that AI based techniques in information retrieval and NLP can be leveraged to extract useful information from existing and ongoing covid literature and hence make way to produce effective treatment and management policy for COVID-19. Metadata are harmonized and deduplicated, and document files are processed to extract full text.

3.1 Data collection logic

All the papers from (PubMed Central (PMC), World Health Organization (WHO) Covid-19 Database and bioRxiv and medRxiv preprint servers) which matches the particular search pattern in below query was included in the datasets.

"COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"

3.2 Internal organization of the dataset and data constituents

- The internal organization of the dataset consist of below files tagged with a datestamp:

```
|-- datestamp/
    |-- changelog
    |-- cord_19_embeddings.tar.gz
    |-- document_parses.tar.gz
    |-- metadata.csv
```

- The files in each version are:
  - changelog: A text file summarizing changes between this and the previous version.
  - cord_19_embeddings.tar.gz: A collection of precomputed SPECTER document embeddings for each CORD-19 paper
  - document_parses.tar.gz: A collection of JSON files that contain full text parses of a subset of CORD-19 papers
  - metadata.csv: Metadata for all CORD-19 papers.

3.2.1 Internal organization each file and data constituents.

- changelog: A text file summarizing changes between this and the previous version. Below is small snippet , how data looks in the change log.

  2022-04-28

  ---CHANGES---

  No major changes.

  ---SUMMARY---
  total metadata rows: 1022888
  CORD UIDs (new: 7096, removed: 29)

  Full text:
  PDF - 381435 json (new: 4659, removed: 336)
  PMC - 299482 json (new: 3777)

- Cord_19_embeddings.tar.gz:

  When cord_19_embeddings.tar.gz is uncompressed, it is a 769-column CSV file, where the first column is the cord_uid and the remaining columns correspond to a 768-dimensional document embedding.

  For example: *ug7v899j,-2.939983606338501,-6.312200546264648,-1.0459030866622925,5.164162635803223,-0.32564637064933777,-2.507413387298584,1.735608696937561,1.9363566637039185,0.622501015663147,1.5613162517547607,...*

- Document_parses.tar.gz:

  When document_parses.tar.gz is uncompressed, it is a directory: This is the json file for the pdf documents.

  ```
  |-- document_parses/
      |-- pdf_json/
          |-- 80013c44d7d2d3949096511ad6fa424a2c740813.json
          |-- bfe20b3580e7c539c16ce4b1e424caf917d3be39.json
          |-- ...
      |-- pmc_json/
          |-- PMC7096781.xml.json
          |-- PMC7118448.xml.json
          |-- ...
  ```

- Metada.csv : consist of the fields as shown below

Column Id	Column description	data type	Example values
cord_uid	A str-valued field that assigns a unique identifier to each CORD-19 paper. This is not necessarily unique per row, which is explained in the FAQs	string	*02tnwd4m*

sha	A List[str]-valued field that is the SHA1 of all PDFs associated with the CORD-19 paper. Most papers will have either zero or one value here (since either have a PDF or we don't), but some papers will have multiple. For example, the main paper might have supplemental information saved in a separate PDF. Or might have two separate PDF copies of the same paper. If multiple PDFs exist, their SHA1 will be semicolon-separated (e.g. '4eb6e165ee705e2ae2a24ed2d4e67da42831ff4a; d4f0247db5e916c20eae3f6d772e8572eb828236')	string	*6b0567729c2143a66d737eb0a2f63f2dce2e5a7d*
source_x	A List[str]-valued field that is the names of sources from which we received this paper. Also semicolon-separated. For example, 'ArXiv; Elsevier; PMC; WHO'. There should always be at least one source listed	string	*PMC*
title	A str-valued field for the paper title	string	*Nitric oxide: a pro-inflammatory mediator in lung disease?*
doi	A str-valued field for the paper DOI	string	*10.1186/rr14*
pmcid	A str-valued field for the paper's ID on PubMed Central. Should begin with PMC followed by an integer.	string	*PMC59543*
pubmed_id	An int-valued field for the paper's ID on PubMed.	integer	*11667967*
license	A str-valued field with the most permissive license we've found associated with this paper. Possible values include: 'cc0', 'hybrid-oa', 'els-covid', 'no-cc', 'cc-by-nc-sa', 'cc-by', 'gold-oa', 'biorxiv', 'green-oa', 'bronze-oa', 'cc-by-nc', 'medrxiv', 'cc-by-nd', 'arxiv', 'unk', 'cc-by-sa', 'cc-by-nc-nd'	string	*no-cc*
abstract	A str-valued field for the paper's abstract	string	*Inflammatory diseases of the respiratory tract are commonly associated with elevated production of nitric oxide (NOâ€¢) and increased indices of NOâ€¢ -dependent oxidative stress. Although NOâ€¢ is known to have anti-microbial, anti-inflammatory and anti-oxidant properties, various lines of evidence support the contribution of NOâ€¢ to lung injury in several disease models. On the basis of biochemical evidence, it is often presumed that such NOâ€¢ -dependent oxidations are due to the formation of the oxidant peroxynitrite, although alternative mechanisms involving the phagocyte-*

			derived heme proteins myeloperoxidase and eosinophil peroxidase might be operative during conditions of inflammation. Because of the overwhelming literature on NOâ€¢ generation and activities in the respiratory tract, it would be beyond the scope of this commentary to review this area comprehensively. Instead, it focuses on recent evidence and concepts of the presumed contribution of NOâ€¢ to inflammatory diseases of the lung.
publish_time	A str-valued field for the published date of the paper. This is in yyyy-mm-dd format. Not always accurate as some publishers will denote unknown dates with future dates like yyyy-12-31	string	*8/15/2000*
authors	A List[str]-valued field for the authors of the paper. Each author name is in Last, First Middle format and semicolon-separated.	string	*Vliet, Albert van der; Eiserich, Jason P; Cross, Carroll E*
journal	A str-valued field for the paper journal. Strings are not normalized (e.g. BMJ and British Medical Journal can both exist). Empty string if unknown.	string	*Respir Res*
mag_id	Deprecated, but originally an int-valued field for the paper as represented in the Microsoft Academic Graph.	integer	
who_covidence_id	A str-valued field for the ID assigned by the WHO for this paper. Format looks like #72306.	string	
arxiv_id	A str-valued field for the arXiv ID of this paper.	string	
pdf_json_files	A List[str]-valued field containing paths from the root of the current data dump version to the parses of the paper PDFs into JSON format. Multiple paths are semicolon-separated. Example: document_parses/pdf_json/4eb6e165ee705e2ae2a24ed2d4e67da42831ff4a.json; document_parses/pdf_json/d4f0247db5e916c20eae3f6d772e8572eb828236.json	string	*document_parses/pdf_json/6b0567729c2143a66d737eb0a2f63f2dce2e5a7d.json*
pmc_json_files	A List[str]-valued field. Same as above, but corresponding to the full text XML files downloaded from PMC, parsed into the same JSON format as above	string	*document_parses/pmc_json/PMC59543.xml.json*

url	A List[str]-valued field containing all URLs associated with this paper. Semicolon-separated.	string	*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC59543/*
s2_id	A str-valued field containing the Semantic Scholar ID for this paper. Can be used with the Semantic Scholar API (e.g. s2_id=9445722 corresponds to http://api.semanticscholar.org/corpusid:9445722	string	

4. DOCUMENTATION FOR SOURCE DATASETS

CORD-19 integrates papers from several sources as shown in the figure below. The source of the figure below is in the introduction section of the paper here

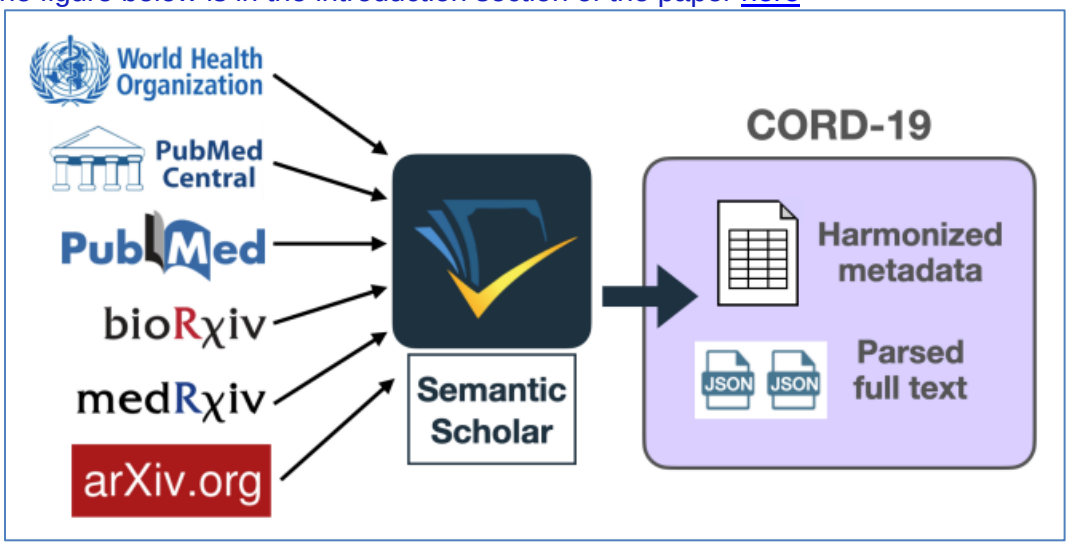

Figure 1: Data collection follow

CORD-19 integrates papers from several sources

- PubMed Central (PMC)
- World Health Organization (WHO) Covid-19 Database : a collection of hand-curated papers about Covid-19. Available here
- bioRxiv and medRxiv preprint servers
- *Semantic Scholar API*: Metadata, paper abstracts, and citation information for papers we index are available through our API. Documentation here.
- *PubMed Central*: The National Library of Medicine (NLM) continues to collaborate with publishers to make COVID-19 and coronavirus-related publications and associated data immediately accessible in PubMed Central (PMC) in human- and machine-readable forms. Available here.

Papers in CORD-19 are sourced from PubMed Central (PMC), PubMed, the World Health Organization's Covid-19 Database, and preprint servers bioRxiv, medRxiv, and arXiv. The PMC Public Health Emergency Covid-19 Initiative (here) expanded access to COVID-19 literature by working with publishers to make coronavirus-related papers

discoverable and accessible through PMC under open access license terms that allow for reuse and secondary analysis. BioRxiv and medRxiv preprints were initially provided by CZI, and are now ingested through Semantic Scholar along with all other included sources. Publishers such as Elsevier(here)and Springer Nature (here) to provide full text coverage of relevant papers available in their back catalog.

5. LANGUAGE VARIETIES

Although there is not a directly stated about the language variety , but below is my conclusion about the language variety in each file. This is just my logical analysis. Since size of the dataset is very big, it was not feasible for me to do through analysis.
Below files are computer generated at the institute situated in US, so it should be *en-US.*
Exception can be for the columns where the value is directly from the text(pdf) file.

- changelog: A text file summarizing changes between this and the previous version. This should have language as *en-US*
- cord_19_embeddings.tar.gz: A collection of precomputed embeddings for each CORD-19 paper. This should have language as *en-US.*
- metadata.csv: Metadata for all CORD-19 papers. This file can be mix of the *en-US* and other language varieties , depending on the research paper.
- document_parses.tar.gz: Language variety for the text in the file will depends on the language of the paper. The json tags can be en-US.

6. SPEAKER DEMOGRAPHIC

N/A

7. ANNOTATOR DEMOGRAPHIC

N/A

8. SPEECH SITUATION AND TEXT CHARACTERISTICS

Research paper collect from all over the world till date 2022-04-28 while metadata generated at Allen institute(US). The modality of the dataset is “written.” Since it is a research paper and computer-generated value I will categorize as scripted/edited. Academic research related to COVID-19 and related historical coronaviruses such as SARS and MERS. Collection of metadata and structured full text papers.

9. PREPROCESSING AND DATA FORMATTING

The preprocessing pipeline is done after collecting all the pdfs as show in data collection from all the sources. Below are the processing section for different forms of data object. Source document here

9.1 Processing metadata:

The initial collection of sourced papers suffers from duplication and incomplete or conflicting metadata. Following operations are performed to harmonize and deduplicate all metadata:

1. Cluster papers using paper identifiers:
   Cluster the papers if they overlap on any of the following identifiers:{doi, pmcid, pubmed_id, arxiv_id, who_covidence_id, mag_id}. If two papers from different sources have an identifier in common and no other identifier conflicts between them, they are assigned to the same cluster. Each cluster is assigned a unique identifier CORD_UID, which persists between dataset releases. No existing identifier, such as DOI pmc ID, is sufficient as the primary CORD-19 identifier.
2. Select canonical metadata for each cluster.
   For each cluster the canonical entry is selected to prioritize the availability of document files and the most permissive license.
   For example, between two papers with PDFs, one available under a CC license and one under a more restrictive COVID-19 specific copyright license, the CC-licensed paper entry as canonical. If any metadata in the canonical entry are missing, values from other members of the cluster are promoted to fill in the blanks.
3. Cluster filtering
   The entries which are not papers are removed from the dataset on ad hoc manner.

9.2 Processing full text:

Below are steps carried to extract the data pdfs.

1. First step is to parse all PDFs to TEI XML using GROBID. Documentation is here.
2. Second step is to parse all TEI XML files to S2ORC JSON using files Documentation is here.
3. Third step is as part of the post processing, clean up the links between inline citations and bibliography entries. For some file from PMC source JATS XML a custom parser was used to generate target like S2ORC JSON. Source documentation is here.
4. S2ORC-doc2json: This library is used to process PDFs and PubMed JATS XML into the format released in CORD-19. This library can be adapted to produce your own versions of the dataset. Source code and instructions for using the library is here.

9.3 Table parsing:

Table parsing: Below tools were used for table extraction from pdf.

Smart Document Understanding: is part of IBM Watson Discovery is used for parsing table from pdf. Documentation here.

Table understanding is part of IBM Watson Discovery is used to annotate the parsed table with additional semantic information such as row and column header and table caption. Documentation here.

Global Table Extractor (GTE) which uses a specialized object detection and clustering technique to extract table bounding boxes and structures. Documentation here.

For further reading on the table parsing go to section (Appendix A : Table parsing result) and section (2.4 Table parsing) of the paper here

9.4 Processing Cord-19-embeddings:

CORD-19-embeddings file was created using SPECTER. This was created using paper titles and abstract of each paper.

10. CAPTURE QUALITY

Below are the sections stating the data quality issue and the reason behind the issues. Some were due to design selection, while some due to missing data in the source dataset.

10.1 Same cord_uid appear in multiple rows:

This is a very tricky issue, and we have not decided on the best way forward. To explain, let's take example cord_uid=hox2xwjg. Examining their respective rows in the metadata file, we see that they are the same paper, but sent from different sources (Elsevier, PMC). The Elsevier row has DOI and PDF, but the PMC row doesn't. Furthermore, the PMC ID, publication date, and URL for each of these rows is different.
Technically all of this data is representative of paper hox2xwjg, so we don't want to remove any of it. But combining them into one cluster would require a schema change to the data, which would break a lot of people's code. Hopefully this is not too big an issue because there are only a small percentage of papers affected, but know that this issue exists and we're debating what's the best way forward.
Source documentation is here

10.2 No abstract in PMC JSONs:

Abstracts in the metadata.csv file are "gold" provided directly from publishers or digital archives. Because PMC is very consistent at providing us "gold" abstracts, we do not bother with parsing the PMC XMLs for abstract text (it's already in the metadata.csv). As such, the PMC JSONs do not contain abstracts. This is not the case for PDF JSONs. We often obtain PDFs through crawling, and in this manner, we would not have "gold" abstracts provided to us. As such, we still opt to parse the PDF for abstract text, which is why that field exists.
Source documentation is here

10.3 Title/authors in the JSON look different than metadata file:

The most likely reason is PDF parsing errors. Occasionally, publishers will have different metadata from what is actually displayed on the PDF itself (e.g. slight differences in

author names). We encourage users to use fields in the metadata file by default and only fall back on the JSON when it is missing.
Source documentation is here

10.4 JSON missing certain metadata:

The JSONs are only meant for representing the full text of the PDF in a structured, machine-readable format. Many metadata fields like dates and venues don't commonly appear on the PDF. Please defer to the metadata file for all such fields since these come from the publishers directly.
Source documentation is here

10.5 Multiple PDF JSONs:

We view these as different attempts/views to represent the same paper/document. Some are going to be higher quality than others. Treat these are separate representations of the same document – you can choose to use one, both, neither (i.e. just use the metadata fields). On average, we believe the PMC JSONs are cleaner than the PDF JSONs but that's not necessarily true.
Source documentation is here

10.6 Same 'sha' for different cord_uid:

Let's take a look at examples cord_uid=d9v5xtx7 and cord_uid=8avkjc84. They both share PDF sha=5d0d0bd116976e1412c10a84902894999df4a342. These are two papers we sourced from Elsevier. If you follow the URLs, you'll notice that they actually retrieve the same PDF despite different having different DOIs. This is an upstream error from the publisher, which we can't necessarily do anything about. Hopefully the number of these cases is small.
Source documentation is here

11. LIMITATIONS

Below are the few limitations that came across as challenge during CORD-19 dataset creation. These challenges were specially due to the fact that the curator wanted to make the processing pipeline reusable and loose less information during formatting. Few of the challenges are discussed below. For detail documentation refer the 'Design decision and challenges ' section of the paper here.

11.1 Challenge in keeping the data up to date:

In order to keep pace with growing literature and keep CORD-19 updated, it was very important that processing pipeline consistent and reproducible. That is, the metadata and full text parsing results must be reproducible, identifiers must be persistent between releases, and changes or new features should ideally be compatible with previous versions of the dataset. Hence the design decision for processing pipeline was made

accordingly. So, may not be the best preprocessing pipeline but the goal should be reproducible.

11.2 Various data formats from multiple sources:

Since the source for the CORD-19 data was diverse the metadata format was also diverse. Papers from different sources must be integrated and harmonized. Each source has its own metadata format, which must be converted to theCORD-19format,while addressing any missing or extraneous fields. So, the design of the processing pipeline must be flexible to update the new source with different metadata structure.

11.3 Clean canonical metadata:

The metadata file may contain duplicate ids. Since the papers source was diverse in nature , it was possible that a same paper is stored at multiple location. Now, effort was made to remove the duplicate papers by using conservative clustering algorithm . However in case of conflict the paper was not removed as it better to have two similar papers than losing an important literature source.

11.4 Machine readable full text:

The machine-readable text is represented S2ORC JSON format. Although conversion between pdf or xml to json is not perfect, but due to the reusable standard structure requirement of the CORD-19 the curators considered S2ORC JSON format.

11.5 Observe copyright restrictions:

Although most papers in cord-19 has open accesses license, however the provision on these open licenses differ across papers. For instance some papers may grant read/consume but restrict the redistribution for commercial purpose. So, it is the curator of the dataset to pass the best knowledge on licensing information to end user. The license is updated under in the Metadata schema.

11.6 Issue in handling tables, figures, and equations:

Many papers in CORD-19 include HTML table parses. These table parses are available in the document parse files under ref_entries of type table. Note: not all tables will have HTML parses. These parses leverage IBM Watson Discovery capabilities (more details can be found in our paper).

Figure images are currently not available. We're currently looking into how to best support these. As for equations, we do not do anything special here – the symbols are treated as text and should be included in the text blobs.
Source documentation is here

12. METADATA

12.1 License:

Below are the license details for the CORD-19 dataset.

Dataset license link : https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/2020-03-13/COVID.DATA.LIC.AGMT.pdf

Open access license: The PMC Public Health Emergency Covid-19 Initiative expanded access to Covid-19 literature by publishers to make coronavirus-related papers discoverable and accessible through PMC under open access license terms that allow for reuse and secondary analysis.

Covid-19 open access licenses: Publishers, such as Elsevier and Springer Nature to provide full text coverage of relevant papers available in their back catalog; these papers are made available under special Covid-19 open access licenses.

Open licenses include :

- Creative Commons (CC),
- publisher-specific COVID-19 licenses
- identified as open access through DOI lookup in the Unpaywall database

12.2 How to Cite:

When referring to the dataset in general, cite the paper associated mentioned in the section here
The paper was accepted to the NLP-COVID workshop at ACL 2020. See the reviews on OpenReview: here

12.3 Subscribe to notifications:

Subscribe to notifications about CORD-19 here

12.4 Errata:

For any question or concerns please email lucyw@allenai.org and kylel@allenai.org

13. DISCLOSURES AND ETHICAL REVIEW

Funding : This work was supported in part by NSF Convergence Accelerator award 1936940, ONR grant N00014-18-1-2193, and the University of Washington WRF/Cable Professorship.

14. OTHER

Below are some sections that in my view are useful in reference to CORD-19 datasets.

14.1 Project using CORD-19:

Google tracking sheet systems and demos consist of project that use CORD-19. For incomplete data or project not updated use Google Form or email.

14.2 Kaggle challenge

The Kaggle site, has a challenge based on the CORD-19. The reason I am mentioning because the task attached in target table is a great way to look at the dataset and tasks that can be done using CORD-19 dataset.

14.3 Reference

For any further reference , please visit the paper here.

15. GLOSSARY

- COVID-19 : Coronavirus disease 2019 (COVID-19) is the official name given by the World Health Organization (WHO) to the disease caused by SARS-CoV-2, the new coronavirus that surfaced in Wuhan, China in 2019 and spread around the globe.
- S2ORC: A dataset of millions of full text papers processed in the same way as CORD-19, but covering many different fields of science. Not regularly updated; intended for offline research, like model development. Available here.
- LitCovid: NLM continues to update its LitCovid dataset of COVID-19 related publications to facilitate text mining. Available here.

About this document

> Include this information about the document verbatim at the end of your data statement. If you adapt the data statement template, include a note about your changes here.

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 2 Schema. The template was prepared by Angelina McMillan-Major, Emily M. Bender, and Batya Friedman and can be found at http://techpolicylab.uw.edu/data-statements.

References

1. Data Statement guide : https://techpolicylab.uw.edu/data-statements/
2. The COVID-19 Open Research Dataset paper : CORD-19: The COVID-19 Open Research Dataset - ACL Anthology
3. SPECTER : https://arxiv.org/abs/2004.07180
4. WHO database : https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov
5. Semantic Scholar API : https://www.semanticscholar.org/product/api
6. PubMed Central : https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/
7. PMC Public Health Emergency Covid-19 Initiative: https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/
8. Publishers such as Elsevier: https://www.elsevier.com/connect/coronavirus-information-center
9. Springer Nature: https://www.springernature.com/gp/researchers/campaigns/coronavirus
10. GROBID: https://github.com/kermitt2/grobid
11. JATS XML : https://jats.nlm.nih.gov/
12. S2ORC-doc2json : https://github.com/allenai/s2orc-doc2json
13. IBM Watson Discovery : https://www.ibm.com/cloud/watson-discovery
14. Dataset license link: Microsoft Word - COVID.DATA.LIC.AGMT (2).docx (ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com)
15. Creative Commons: https://creativecommons.org/
16. Publisher-specific COVID-19 licenses: https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/
17. Unpaywall: https://unpaywall.org/
18. OpenReview : https://openreview.net/forum?id=0gLzHrE_t3z
19. Kaggle site : https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
20. Project Google Form : https://forms.gle/s48a9RFoyBxxV9J7A