Method for Producing Curated and Tagged COVID-19 Reference Sets

1. We downloaded April 17th edition of CORD-19 and downloaded a fresh export from ClinicalTrials.gov (searching for COVID-19 trials).
2. We searched for and removed duplicates using DistillerSR's Duplicate Detection Engine
    - We searched for duplicates against the Cord_uid field to remove already collected CORD-19 references. This removed 49,925.
    - We searched for duplicates against Pubmed_id, and removed 975 more.
    - We ran a standard duplicate detection search using the Extreme Precision setting to remove any additional duplicates.
3. We then used a DistillerSR AI classifiers to identify and Tag references relating to COVID-19, as well as to Tag references that are related to systematic reviews.
4. We then had a human review the references Tagged by the COVID-19 classifier as a quality check. Humans then screened approximately 200 additional references using Continuous AI Prioritization to look for anything that might have been missed.
5. Finally, we use the AI Audit tool in DistillerSR to flag anything that may have been misclassified.
6. The references were then exported out of DistillerSR for use by the research community.

Reference Sets

- **Full Reference Set**: deduplicated with the above-mentioned Tags added.
- **COVID-19 Tagged Reference Set:** contains all references that were screened both by the AI and human screeners. This set includes references tagged as not specific to COVID-19.
- **COVID-19 Only Reference Set**: contains only references that refer to COVID-19. This set has been human verified.