COVID-19 / SARS-CoV-2 関連のプレプリントに関する分散表現データセット

— 2020 年 05 月 17 日版 —

小柴 等 *, 林 和弘 *

Distributed representation dataset for COVID-19 / SARS-CoV-2 related preprints

— 17 May 2020 Dataset —

KOSHIBA Hitoshi, HAYASHI Kazuhiro

投稿区分: Dataset: csv format

初版: 2020.05.17

* 文部科学省 科学技術・学術政策研究所
100-0013 東京都千代田区霞ヶ関 3 丁目 2-2

1　説明

プレプリントにおける COVID-19 / SARS-CoV-2 に関連する研究動向を把握するため，収集・整備したデータセットを公開するものです.

1.1　形式

形式は UTF8 でエンコードされたシンプルな CSV です.
カンマ (,) は各フィールド内には出現せず，フィールドの区切りのみに使われます.
1 行目はヘッダ行です.
フィールドの並び，内容は表 1 の通りです.

項番	フィールド名	型	DESC
1	id	int	ID
2	pps_kind	str	PPS の種別.　arXiv, bioRxiv, chemRxiv, medRxiv, SSRN, SSRN Lancet の 6 種
3	url	str	DOI など，アイテムの公開場所
4	pub_date	str	記事の投稿日（DOI に記載のものとは異なる場合がある）
5	topic_cls	int	分散表現を K-means++ で 16 分割した場合のクラスタ ID
6	dim_001	float	分散表現値（1 次元目）
⋮	⋮	⋮	⋮
106	dim_100	float	分散表現値（100 次元目）

表 1　各フィールドの詳細

1.2　収録内容

arXiv, bioRxiv, chemRxiv, medRxiv, SSRN という 5 種類のプレプリントサーバにおける，2020 年 05 月 17 日までの COVID-19 関連記事について,

- プレプリントサーバ種別
- URL もしくは DOI
- 投稿日
- 分散表現における座標値 (100 次元)
- 分散表現を 16 分類した際のクラスタ番号

を整理し，CSV 形式で搭載しています.
プレプリントサーバ種別に関連し，SSRN のうち，後述する The Lancet 関連のプレプリントは SSRL Lancet として別に整理しています.
投稿日については，各プレプリントサーバで確認できる投稿日と DOI から取得できる投稿日が異なることも多いため，基本的にはプレプリントサーバの内容を採用しています.

1.3　収集対象

arXiv, bioRxiv, chemRxiv, medRxiv, SSRN のうち，chemRxiv 以外は COVID-19 / SARS-CoV-2 に関連する記事について，独自にとりまとめたリンク集を生成しています（図 1 参照）.

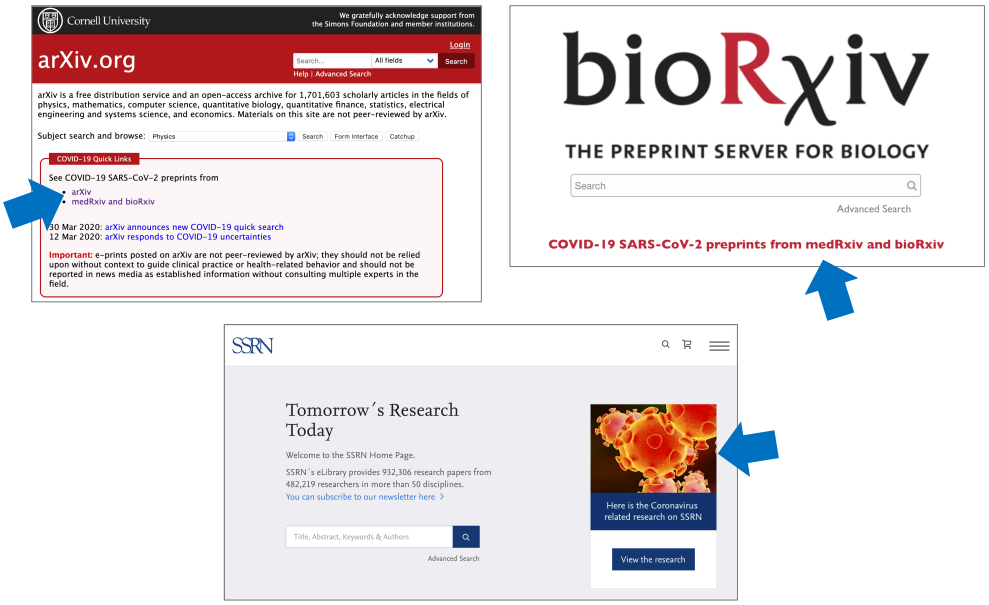

図 1　収集対象

chemRxiv については，データ収集を行った 2020 年 5 月 25 日時点で，chemRxiv 独自のリストは見当たりませんが，論文等文献検索サービスである Dimensions [1)]がとりまとめて公開している COVID-19 関連のデータセット[2)]に chemRxiv 上の記事が出てくるため，これを利用しています[3)].

ここで，SSRN については図 2 に示すとおり，“Preprints with THE LANCET” との表示がついた記事も散見されます.

Lancet は医学系の著名雑誌の一つであり，相対的に人文社会系のプレプリントが多いと考えられる SSRN の中では異質と言えます．そこで本報告においては，“Preprints with THE LANCET” との表示がついた記事について，これを “SSRN Lancet” と切り分けて扱います.

収集した記事の中には 2018 年など古い情報も見られたため，記事を 2020 年 1 月以降のものに限定し，結果として，それぞれ表 2 に示した記事数を得ました.

1.4　分散表現

収集したプレプリントのタイトル，概要をもとに，FastText[Bojanowski17, Joulin16] を用いて 100 次元の単語分散表現を作成しました.

[1)] https://app.dimensions.ai/

[2)] https://dimensions.figshare.com/articles/Dimensions_COVID-19_publications_datasets_and_clinical_trials/11961063

[3)] chemRxiv は figshare というシステムを採用しているが，Dimensions，figshare の運用母体は両方とも “Digital Science” 社.

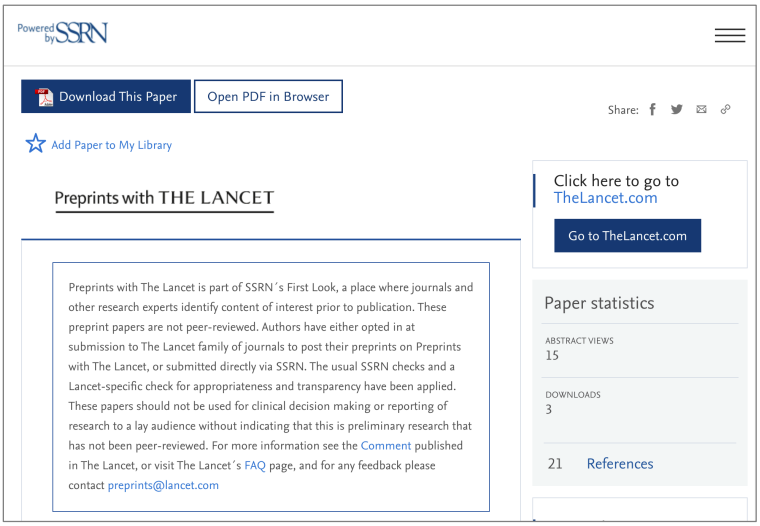

図 2　SSRN における Lancet のプレプリント

PPS	Num	PPS	Num
arXiv	936	medRxiv	2837
bioRxiv	716	SSRN	612
chemRxiv	175	SSRN Lancet	496

表 2　PSS ごとの記事数 (2020 年 5 月 25 日時点)

その上で，各プレプリントのタイトル，概要に出現する単語の分散表現を線形加算して，正規化したものをプレプリントの分散表現として採用しました．

この 100 次元のプレプリント分散表現も，本データセットに記載しています．

1.5　クラスタリング

プレプリントの分散表現（100 次元）に対し，k-means++ を用いて 16 分割を行いました．

1.6　利用上の注意点

収集時時点のデータをできるだけ正確に反映するよう心がけていますが，データの抜け漏れが発生していたり，その後取り下げなどで状況が変化した可能性もあります．

これらのデータセットを利用した上での不利益等に際して，作成者らは一切の責任を負いません．

また，URL/DOI に紐付いたプレプリントを利用する際には，各プレプリントやプレプリントサーバの利用規約に従ってください．

2　ライセンス等

URL/DOI に紐付いた各プレプリントの著作権等は，各プレプリントやプレプリントサーバに帰属します．

本データセットに関しては，別途データベースとしての著作権が付属し，Creative Commons の “CC BY 3.0” の条件でご利用いただけます．

参考文献

[Arthur07] Arthur, David and Vassilvitskii, Sergei：K-means++: The Advantages of Careful Seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*，pp,1027–1035, 2007. http://dl.acm.org/citation.cfm?id=1283383.1283494

[Bojanowski17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.: Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017. arXiv:1607.04606

[Joulin16] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T.: FastText.zip: Compressing text classification models, *arXiv preprint*, 2016. arXiv:1612.03651