「Kaggle における新型コロナウイルスのデータ分析のメタ分析」

学生 ID：1116181041
氏名：南端　尚樹

概要

実験の背景

Kaggle では、新型コロナウイルスのデータ分析のコンペティションが開催されており、多くの人がコンペティションに参加している。Kaggle ではコンペティションの参加者が独自の手法を用いてデータ分析を行っているため、同一のタスクにおいて様々なデータ分析結果が示されている。示されているデータ分析結果は様々であるが与えられたタスクは同じであるため、これらの結果をメタ分析することで与えられたタスクに対して有用な結果を得ることができると考え、この実験を行うことにした。

実験の目的

この実験では、同一タスクにおける結果の中なら二つの結果に対してメタ分析を行うことで様々な結果に対してメタ分析を行う方法を確立するとともに、メタ分析により新たな知見を獲得することを目的としている。この実験ではメタ分析として、それぞれの結果の中の論文を類似度で比較している。

理論

この実験では、「COVID-19 Open Research Dataset Challenge (CORD-19)」（https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge）のタスク「What do we know about COVID-19 risk factors?」の実行結果についてメタ分析を行う。今回は、「Risk Factor Excerpt Extraction」（https://www.kaggle.com/niksbuds7/risk-factor-excerpt-extraction）と「CoronaWhy.org -Task: Risk Factors」（https://www.kaggle.com/arturkiulian/coronawhy-org-task-risk-factors）の二つの結果に対してメタ分析を行い、論文の類似度の比較には Doc2Vec を用いる。BeautifulSoup を使用して上の二つのサイトからそれぞれの実行結果をスクレイピングしてくる。それぞれの結果は pandas の DataFrame 型で取得し、得られた DataFrame を整形して結合する。こうして図 1 のような DataFrame を得る。

	title	factor	URL
0	Bulk and single-cell transcriptomics identify ...	Smokimg	https://www.medrxiv.org/content/10.1101/2020.0...
1	Prevalence, Severity and Mortality associated ...	Smokimg	https://www.medrxiv.org/content/10.1101/2020.0...
2	Articles Clinical and epidemiological features...	Smokimg	https://www.thelancet.com/journals/laninf/arti...
3	Comorbid Diabetes Mellitus was Associated with...	Diabetes	https://www.medrxiv.org/content/10.1101/2020.0...
4	Articles Clinical and epidemiological features...	Diabetes	https://www.thelancet.com/journals/laninf/arti...

図 1：スクレイピングした実行結果をつなげたデータフレームの抜粋

その後、各行に対して URL のサイトからスクレイピングにより論文の本文を抽出して DataFrame に追加し、一行に一つの論文が入ったテキストファイルを出力する。出力したテキストファイルを Doc2Vec で扱えるオブジェクトに変換したのち、Doc2Vec を実行して各論文に対してほかの論文との類似度を得る。

結果

論文の類似度を求め、各ファクターの論文に対してほかの論文の中から類似度が上位三件のファクターを調べ、ファクター間の関係を調べた。以下の表はファクターごとの関連ファクターの上位三件を表示したものである。

表 1：Smoking と関連の高いファクター

Smoking	
Hypertension	3
Pollution	2
humidity	2

表 2：Diabetes と関連の高いファクター

Diabetes	
Hypertension	9
Diabetes	4
Tuberculosis	4

表 3：Pregnancy と関連の高いファクター

Pregnancy	
Hypertension	12
Diabetes	9
Tuberculosis	7

表 4：Hypertension と関連の高いファクター

Hypertension	
Hypertension	15
Diabetes	14
Tuberculosis	11

表 5：Tuberculosis と関連の高いファクター

Tuberculosis	
Hypertension	19
Diabetes	17
Tuberculosis	12

表 6：Heart Disease と関連の高いファクター

Heart Disease	
Hypertension	20
Diabetes	17
Tuberculosis	13

表 7：Old Age と関連の高いファクター

Old Age	
Hypertension	21
Diabetes	18
Pollution	16

表 8：humidity と関連の高いファクター

humidity	
Hypertension	22
humidity	20
pollution	19

表 9：pollution と関連の高いファクター

pollution	
pollution	25
humidity	25
Hypertension	22

表 10：population density と関連の高いファクター

population density	
pollution	29
humidity	29
Hypertension	22

この結果で気になったのは、humidity と Hypertension、humidity と pollution の関係である。ほかのファクター同士の関連は予想のつくものだったが、上の二つの関係は予想していないものだった。調べてみると、humidity と Hypertension は湿度が高いと発汗がしづらくなり、その結果血圧が上がるという関連があった。humidity と pollution は湿度が高いと PM2.5 の濃度が高くなりやすいという関連があった。

考察

結果からリスクファクターとして考えられるものの中にも、様々なファクターと関連しているものもあればほかのファクターとはあまり関連していないものもあるということがわかった。関連しあっているファクターの共通点を見つけ出し、それを新たなリスクファクターとすればより重要なリスクファクターが得られると考える。このようなリスクファクターやほかのファクターとあまり関連がないリスクファクターは、キーファクターとなる可能性が高いため、注目すべきであると考える。

参考文献

文章をベクトル化して類似文章の検索 – Qiita
(https://qiita.com/akira_/items/f9bb46cad6834da32367)
BeautifulSoup webscraping find_all（）：完全一致を見つける
(https://www.366service.com/jp/qa/27822e06f2361243e87a7f83ce9f34e4)
[Python]BeautifulSoup で属性を前方一致検索で指定して要素を取得する – Qiita
(https://qiita.com/d_m/items/f477c6665ec69dfaf594)