COVID-19 in Academia

Mingming Cui

University of Chicago

Abstract

By utilizing computational techniques, two corpora of scholarly papers and abstracts were analyzed. Results showed that the content of each corpus was distinguishable both at the within-corpus and between-corpus level. Dynamic modeling revealed that the topics and semantic meanings of words changed over time.

Keywords: COVID-19, academic, scholar, science

COVID-19 in Academia

The coronavirus disease 2019 (COVID-19) pandemic has been raging for more than a year. What have the academics been researching? Is their connotation positive, negative, or neutral? Can we summarize some topics from their papers? Do the semantic meanings of their words change over time? Do they choose words differently? To address these questions, I use computational techniques to analyze two corpora: one for the social sciences, and the other for the natural sciences. Given that there are many more scholarly articles published and available in the natural sciences than in the social sciences, I manually downloaded 55 economic papers from National Bureau of Economic Research (NBER) and 17 psychological papers from the Journal of Applied Psychology (JAP). All of these papers study COVID-19 and its impact on economy/psychology. I extracted the titles, abstracts, and body text as the corpus for the social sciences. By contrast, I scraped and extracted the titles and short abstracts from 1959 Google Scholar entries (483 in 2019, 836 in 2020, and 640 in 2021) as the corpus for the natural sciences.

1. Word Clouds

To learn about the content of the papers, drawing some word clouds based on titles is the fastest way. As the Figure 1 shows, the psychological papers mainly concern with COVID-19 and career (e.g., job, work, and occupation) and psychological states (e.g., anxiety and stress). Figure 2 shows that economists care about the relationships between COVID-19 and labor, market, and economics. We can also see that scholars in the natural sciences study the nature of COVID-19 and its treatment from Figure 3.

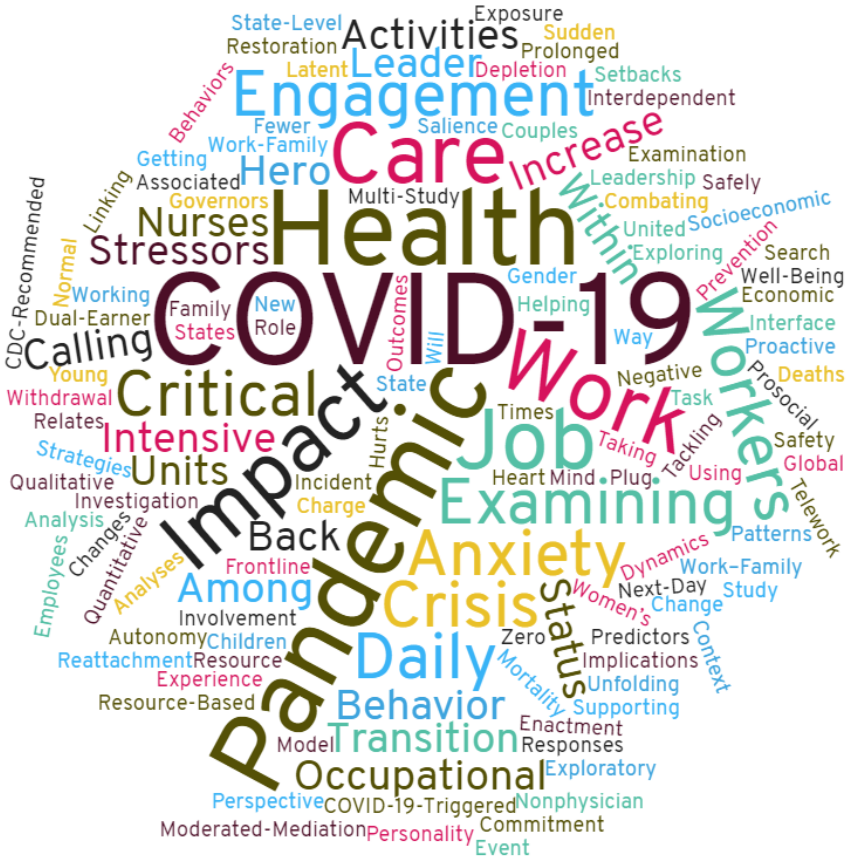


Figure1. Word Could of Psychological Papers

Figure 2. Word Could of Economical Papers

Figure 3. Word Could of Papers in Natural Sciences

2. Sentiment Analysis

Sentiment analysis is informative for evaluating the connotation of papers. I used the **textblob** module to analyze the sentiment of my corpora. Specifically, the polarity score is 0.07493156434782125 for the psychological papers, and 0.0728681077814532 for the economical papers, and 0.10677599837734063 for the titles and abstracts in the natural sciences. The subjectivity score is 0.3916037609531709 for the psychological papers, and 0.3848585802602516 for the economical papers, and 0.4166789151770683 for the titles and abstracts in the natural sciences. Overall, researchers use neutral and moderately subjective words.

3. Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence computes the relative entropy between two distributions--how they differ in bits. As such, I computed the KL divergence between the social

and natural sciences corpora to learn about their difference. The social-natural divergence is 1.0003624288133506 and the natural-social divergence is 1.0638245866284137, meaning that the two corpora are distinguishable.

4. Word Vectorization Analysis

In this section, I tried to analyze how the same words differentiate from each other in the two corpora. By vectorizing words, I found that “job” is similar to “performance”, “task”, “insecurity”, and “workplace” in the social sciences corpus, and similar to “disorder”, and “investigate” in the natural sciences corpus. “leader” is similar to “commitment”, “goal”, “regulation”, and “potential” in the social sciences corpus, and similar to “government”, “community”, “policy” and “lockdown” in the natural sciences corpus. Besides, I also visualized the most popular 50 words in each corpus and their distances. Overall, natural scientists study COVID-19 and its syndrome and treatment, while social scientists focus on COVID-19 and aspects of our life. See the details below.

- Most similar words to “**vaccine**” in the social (left) and natural (right) sciences.

```
[('infectiousness', 0.9439640045166016), [('model', 0.9998939037322998),
 ('effective', 0.9404677152633667),       ('effective', 0.9998881816864014),
 ('cure', 0.9244608283042908),            ('potential', 0.9998874664306641),
 ('know', 0.9193854331970215),            ('strategy', 0.999883234500885),
 ('extreme', 0.9165095686912537),         ('include', 0.9998825788497925),
 ('quarantine', 0.9129801392555237),      ('drug', 0.9998806715011597),
 ('isolation', 0.9125193357467651),       ('understand', 0.9998775720596313),
 ('trace', 0.9110430479049683),           ('early', 0.9998756647109985),
 ('lockdowns', 0.9108986258506775),       ('reduce', 0.9998752474784851),
 ('contract', 0.9098805785179138)]        ('control', 0.9998699426651001)]
```

- Most similar words to “**treatment**” in the social (left) and natural (right) sciences.

```
[('belief', 0.9609736800193787),
 ('vary', 0.9371767640113831),
 ('differential', 0.9222626686096191),
 ('interpret', 0.9134619235992432),
 ('additional', 0.9041493535041809),
 ('differ', 0.9014716148376465),
 ('overall', 0.8998055458068848),
 ('scenario', 0.8960479497909546),
 ('hold', 0.8951650857925415),
 ('trajectory', 0.8904632329940796)]
```

```
[('low', 0.9998729228973389),
 ('relate', 0.9998582601547241),
 ('understand', 0.9998553395271301),
 ('vaccine', 0.9998540878295898),
 ('patient', 0.9998528957366943),
 ('present', 0.999847412109375),
 ('severity', 0.9998459815979004),
 ('find', 0.9998447895050049),
 ('trial', 0.9998437166213989),
 ('march', 0.9998431205749512)]
```

- Most similar words to "**job**" in the social (left) and natural (right) sciences.

```
[('performance', 0.7767320871353149),
 ('perform', 0.7687323093414307),
 ('task', 0.7589665055274963),
 ('student', 0.7334562540054321),
 ('experience', 0.7280638217926025),
 ('insecurity', 0.7277714014053345),
 ('couple', 0.7221989631652832),
 ('woman', 0.7134095430374146),
 ('skill', 0.7128334045410156),
 ('workplace', 0.7101960778236389)]
```

```
[('disorder', 0.9986975193023682),
 ('et', 0.9986754655838013),
 ('implement', 0.9986667633056641),
 ('search', 0.9986576437950134),
 ('numb', 0.9986507892608643),
 ('adopt', 0.9986448287963867),
 ('use', 0.9986392259597778),
 ('ct', 0.998621940612793),
 ('role', 0.9986189603805542),
 ('investigate', 0.9986182451248169)]
```

- Most similar words to "**leader**" in the social (left) and natural (right) sciences.

```
[('commitment', 0.9728037714958191),
 ('indirectly', 0.9477722644805908),
 ('knowledge', 0.944723904132843),
 ('safety', 0.9422106742858887),
 ('contribute', 0.9403423070907593),
 ('exacerbate', 0.9397084712982178),
 ('progress', 0.9394020438194275),
 ('goal', 0.9372594356536865),
 ('regulation', 0.9357172846794128),
 ('potential', 0.9339116811752319)]
```

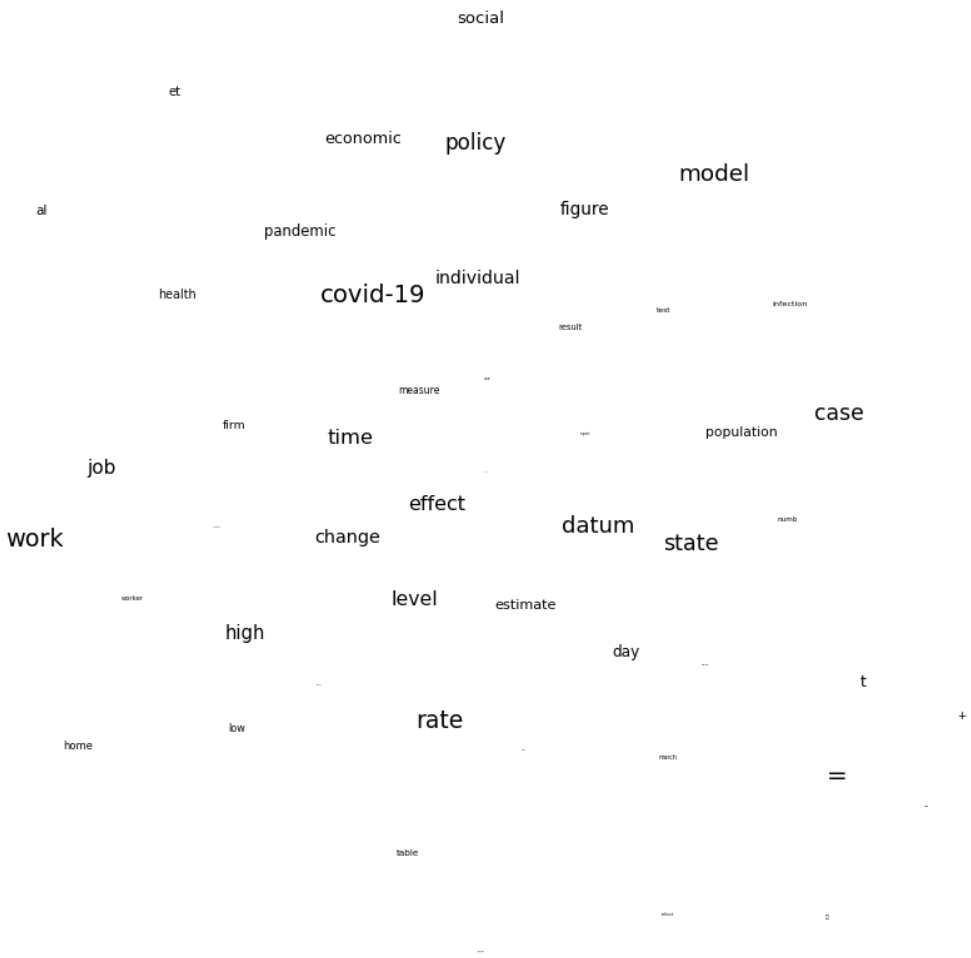
```
[('level', 0.9997876286506653),
 ('time', 0.9997591972351074),
 ('university', 0.9997567534446716),
 ('government', 0.9997553825378418),
 ('work', 0.9997516870498657),
 ('group', 0.9997513294219971),
 ('learn', 0.9997508525848389),
 ('community', 0.9997485876083374),
 ('lockdown', 0.9997460246086121),
 ('policy', 0.9997444152832031)]
```

- Most similar words to "**economy**" in the social (left) and natural (right) sciences.

```
[('shock', 0.9052436351776123),
 ('chain', 0.898902177810669),
 ('pharmaceutical', 0.8619531393051147),
 ('mitigate', 0.8553270101547241),
 ('externality', 0.8548004627227783),
 ('equilibrium', 0.8537750244140625),
 ('temporary', 0.8504856824874878),
 ('trade', 0.8500582575798035),
 ('consequence', 0.8491602540016174),
 ('affect', 0.8470218777656555)]
```

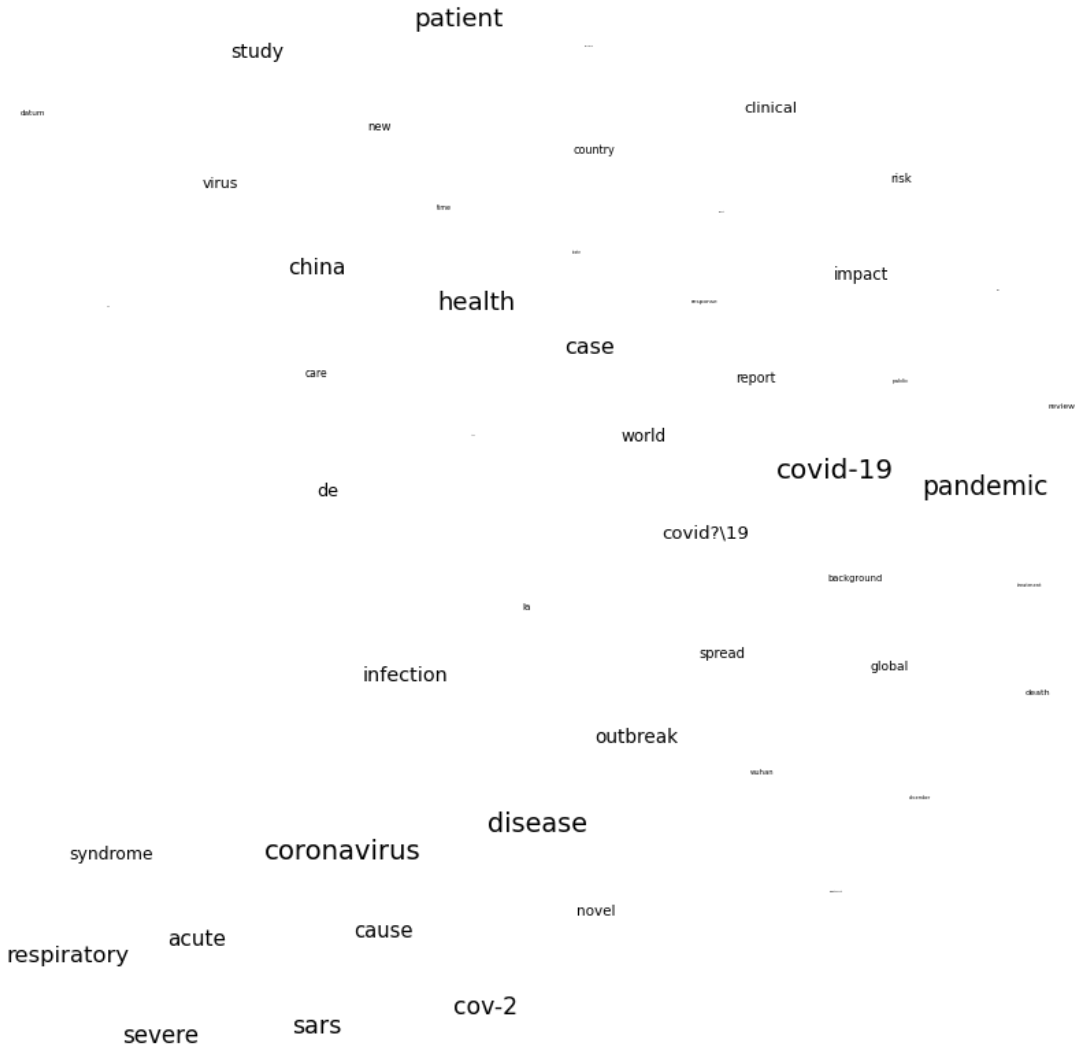
```
[('impact', 0.9998507499694824),
 ('crisis', 0.9998484253883362),
 ('society', 0.9998253583908081),
 ('work', 0.9998248815536499),
 ('hospital', 0.999822199344635),
 ('lockdown', 0.9998174905776978),
 ('practice', 0.9998120665550232),
 ('university', 0.9998114109039307),
 ('time', 0.9998106956481934),
 ('service', 0.9998049736022949)]
```

- Visualization of popular words in the social sciences corpus.



-

- Visualization of popular words in the natural sciences corpus.

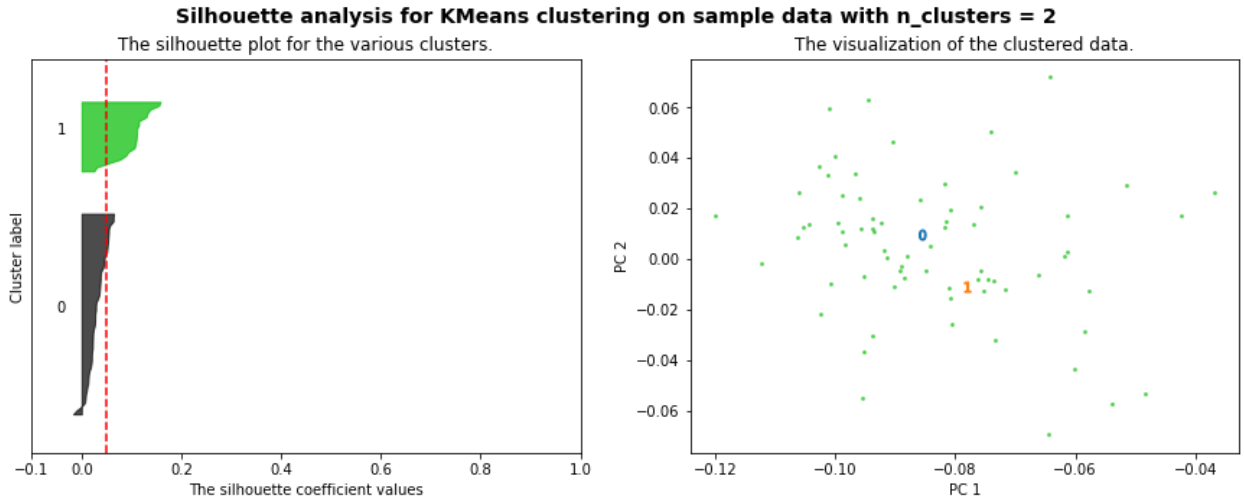


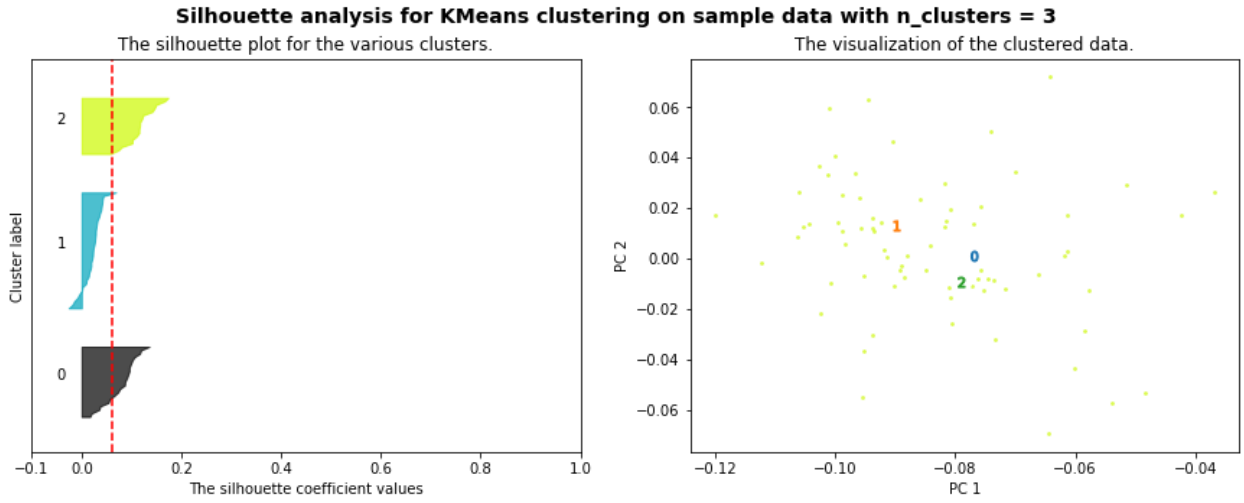
-

5. K-means Clustering

To learn more about the structures within the two corpora, I used k-means to cluster the words. Silhouette analysis revealed some information of the corpora.

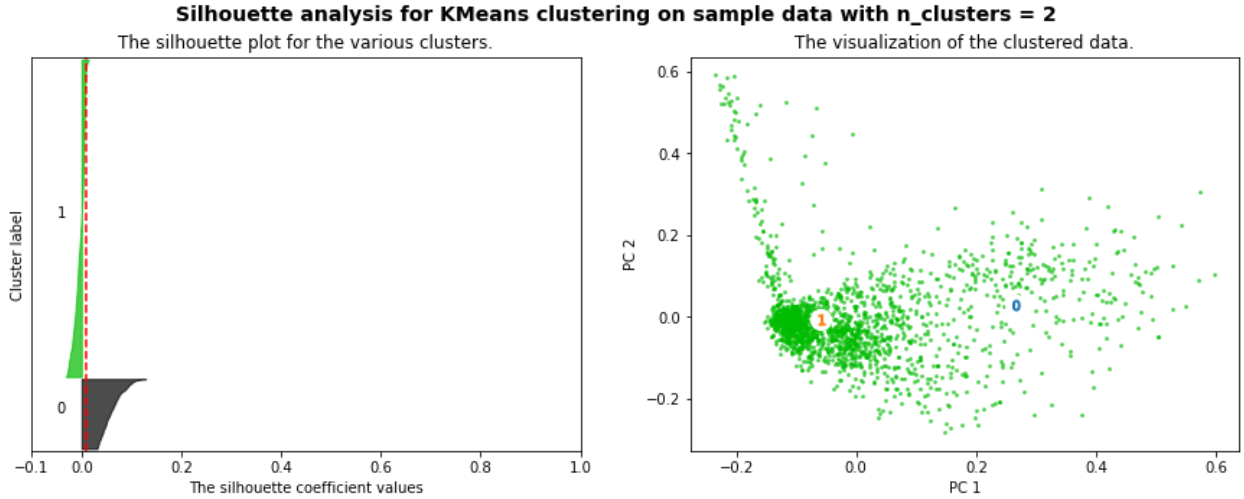
- Silhouette analysis for the social sciences corpora.

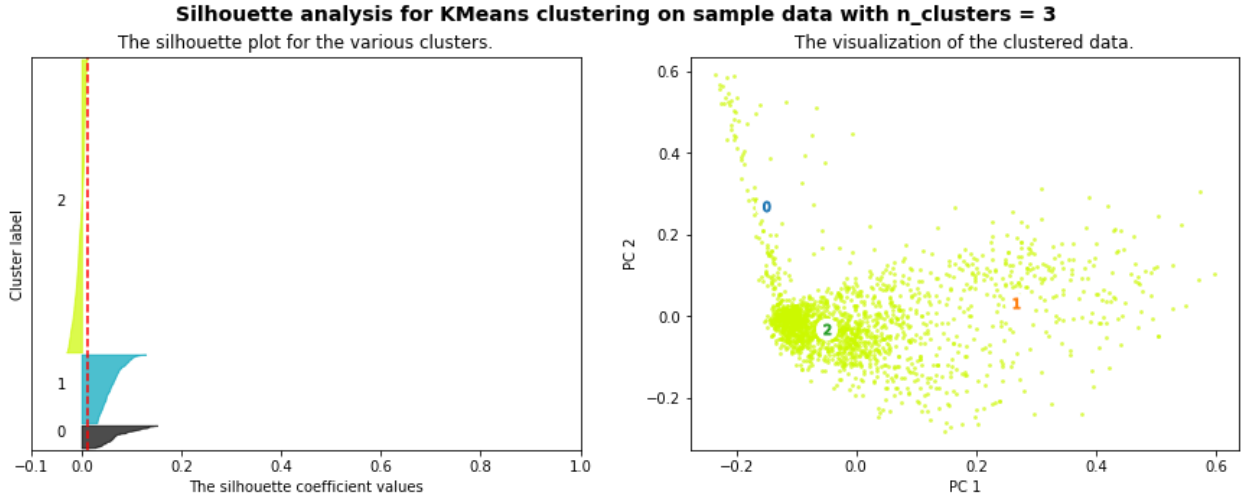
- 

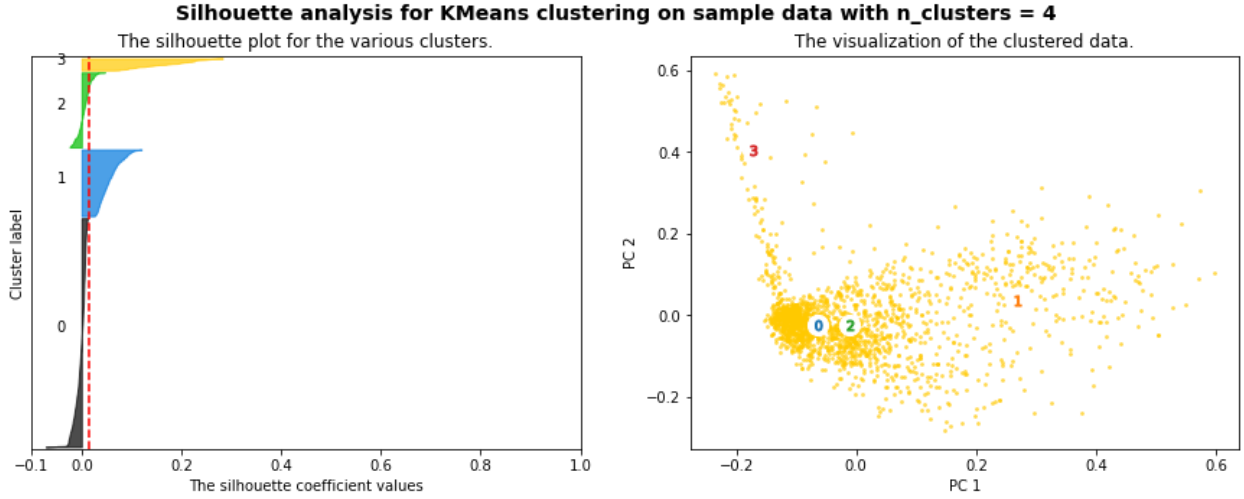
- 

Hence, I set two clusters for the social sciences corpus. The top terms for the first cluster are “job”, “anxiety”, “psychology”, “employee”, and “engagement.” Intuitively, the first cluster deals with psychology. The top terms for the second cluster are “policy”, “employment”, “panel”, and “labor,” which indicate that the second cluster is economical. Therefore, the k-means clustering divided the social sciences corpus into psychological and economical groups.

- Silhouette analysis for the natural sciences corpora.

- 
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
PC 2
PC 1

- 
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
PC 2
PC 1

- 
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
PC 2
PC 1

Hence, I also set two clusters for the natural sciences corpus. The top terms for the first cluster are “pandemic”, “disease”, “patients”, “coronavirus”, and “health.” Intuitively, the first cluster deals with COVID-19 itself. The top terms for the second cluster are “respiratory”, “acute”, “severe”, and “syndrome,” which indicate that the second cluster is the syndromes of COVID-19.

6. Topic Modeling

To delve deeper into the topics of the corpora, I employed **Gensim** to extract topics. Because the corpora are both of COVID-19, the extracted five topics are not distinguishable.

- Top words for the five topics of the social sciences corpus.

	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
0	rate	covid-19	=	=	covid-19
1	=	=	work	work	work
2	work	rate	covid-19	covid-19	model
3	–	policy	state	rate	=
4	t	work	model	datum	rate
5	model	datum	time	job	time
6	datum	model	datum	state	datum
7	covid-19	case	change	case	level
8	state	firm	case	time	effect
9	policy	effect	effect	figure	individual

- Top words for the five topics of the natural sciences corpus.

	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
0	covid-19	de	covid-19	covid-19	covid-19
1	pandemic	la	disease	coronavirus	coronavirus
2	coronavirus	covid-19	coronavirus	disease	disease
3	case	coronavirus	pandemic	pandemic	sars
4	disease	covid?\19	patient	health	patient
5	health	sars	severe	patient	cause
6	outbreak	outbreak	health	infection	respiratory
7	covid?\19	en	sars	china	cov-2
8	global	le	respiratory	outbreak	china
9	cov-2	cov-2	acute	severe	pandemic

-

7. Dynamic Topic Modeling

In this section, I chose to set up two topics for the natural sciences corpus and track the changes. For both topics, words like crisis (2019: 0.004726211224326175; 2020: 0.0048143407065999655; 2021: 0.004873021391956587), impact (2019: 0.006030995432705853; 2020: 0.006297414325254963; 2021: 0.0064823056876690205), and health (2019: 0.003995426408708622; 2020: 0.004120430568105463; 2021: 0.004170162839912912) are trending, meaning that the crisis and health problems caused by COVID-19 are escalating.

- The dynamics of topic 1: 2019 (upper), 2020 (middle), and 2021 (lower).

```
[('covid-19', 0.06428464962404917),
 ('coronavirus', 0.027254611441029808),
 ('disease', 0.02432902692950473),
 ('patient', 0.013227578385411444),
 ('pandemic', 0.012390558732265263),
 ('sars', 0.010489491673089726),
 ('health', 0.010097726048698887),
 ('respiratory', 0.009111447115459295),
 ('cov-2', 0.009055990726999847),
 ('case', 0.009021078516874852),
 ('severe', 0.008798192169837536),
 ('acute', 0.008205517396569361),
 ('china', 0.008172925714357054),
 ('cause', 0.00787177668801646),
 ('infection', 0.007664184090606179),
 ('outbreak', 0.007250528971536949),
 ('covid?\\19', 0.0067780641057692105),
 ('syndrome', 0.006722534570332845),
 ('study', 0.0062623970194557515),
 ('novel', 0.006223719335553649)]]
```

```
[('covid-19', 0.0638124287465592),
 ('coronavirus', 0.026969213165491154),
 ('disease', 0.024146529032677923),
 ('patient', 0.013237142599127186),
 ('pandemic', 0.01224390454015414),
 ('sars', 0.01039322933544439),
 ('health', 0.01033173903961803),
 ('severe', 0.009510829427422055),
 ('respiratory', 0.009037157011617634),
 ('cov-2', 0.008976071817818567),
 ('case', 0.008927837771974173),
 ('acute', 0.008156272769640045),
 ('china', 0.00807723726407815),
 ('cause', 0.007792473477521388),
 ('infection', 0.007611490603074289),
 ('outbreak', 0.0071779500172514344),
 ('covid?\\19', 0.006824205845398747),
 ('syndrome', 0.006669123360485533),
 ('study', 0.006282581951873635),
 ('novel', 0.006141570236040950 5)],
```

```
[[('covid-19', 0.06346346053327592),
  ('coronavirus', 0.02673126389515641),
  ('disease', 0.02401511538115626),
  ('patient', 0.01330413138984548),
  ('pandemic', 0.012128038912016657),
  ('sars', 0.010308268450830911),
  ('health', 0.00986025848576333),
  ('severe', 0.009223478843966168),
  ('respiratory', 0.008958171431812089),
  ('cov-2', 0.008903325532338982),
  ('case', 0.00882721770759777),
  ('acute', 0.008104757633943211),
  ('china', 0.007967753332214873),
  ('cause', 0.007711476302335946),
  ('infection', 0.007567542583296424),
  ('outbreak', 0.007113328763442282),
  ('covid?\\19', 0.006904749978605111),
  ('syndrome', 0.006592741782246652),
  ('study', 0.006369879843431602),
  ('clinical', 0.006069589891108472)],
```

- The dynamics of topic 2: 2019 (upper), 2020 (middle), and 2021 (lower).

```
[('covid-19', 0.046322261608395095),
 ('pandemic', 0.016420132572796182),
 ('de', 0.01636708598028615),
 ('la', 0.009145930256156301),
 ('impact', 0.006030995432705853),
 ('e', 0.004992934877059675),
 ('crisis', 0.004726211224326175),
 ('en', 0.004254720878231465),
 ('di', 0.004182274110766479),
 ('health', 0.003995426408708622),
 ('global', 0.003981191806455716),
 ('social', 0.003860794888898123),
 ('covid', 0.00349350440310812),
 ('article', 0.0034025119263356796),
 ('time', 0.003223772288278957),
 ('economic', 0.0032192136361130076),
 ('world', 0.00321826324729171),
 ('medical', 0.003173588334713717),
```

```
[('covid-19', 0.04660667569843402),
 ('pandemic', 0.01656140913626898),
 ('de', 0.014408436877613584),
 ('la', 0.008500917267636205),
 ('impact', 0.006297414325254963),
 ('crisis', 0.0048143407065999655),
 ('e', 0.004764436467061105),
 ('global', 0.004163796548585984),
 ('health', 0.004120430568105463),
 ('en', 0.004090482350693483),
 ('di', 0.003984883267667977),
 ('social', 0.003921407729428352),
 ('article', 0.0034478729289488847),
 ('covid', 0.0034288331732737556),
 ('economic', 0.0033368208656752627),
 ('time', 0.003329696401301124),
 ('medical', 0.003277169388507299),
 ('world', 0.003260283503466654),
 ('new', 0.003042821886329991),
 ('learn', 0.0030156952330389125)],
```

```
[[('covid-19', 0.046723443618186804),
  ('pandemic', 0.01666185694081015),
  ('de', 0.013147735282885771),
  ('la', 0.00805385891530807),
  ('impact', 0.006482305687669020 5),
  ('crisis', 0.004873021391956587),
  ('e', 0.004615595950721479),
  ('global', 0.004246331244818657),
  ('health', 0.004170162839912912),
  ('social', 0.003980834727989875),
  ('en', 0.003974607484036784),
  ('di', 0.0038445581715665614),
  ('time', 0.0034011534113324473),
  ('article', 0.0033993178606789775),
  ('economic', 0.0033862981987582675),
  ('covid', 0.0033821781033152687),
  ('world', 0.0032863856979313067),
  ('medical', 0.003206116218973925),
  ('new', 0.0030615651854937033),
  ('learn', 0.003056385709554595)],
```

8. Linguistic Change Analysis

The term “COVID-19” is impactful so that the words surrounding it may change semantically over time. In the natural sciences corpus, The word “america” changed a lot. A possible reason for that is the COVID situation in America influenced how the academics used “america.” The word “real” also changed. I would assume that there was quite amount of fake information and news so that “real” was influenced.

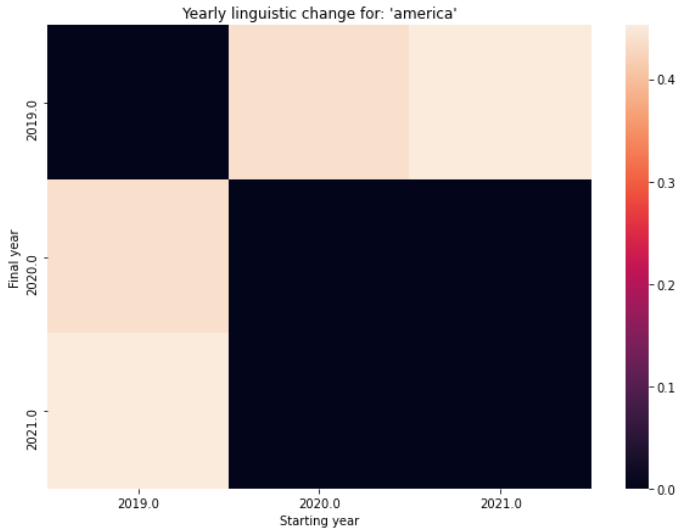
- The 20 most divergent words:

```
[('question', 0.9772264361381531),
 ('america', 0.8928688168525696),
 ('real', 0.8140999674797058),
 ('period', 0.8040522336959839),
 ('model', 0.7849266529083252),
 ('introduction', 0.7819625735282898),
 ('well', 0.7156000733375549),
 ('era', 0.6698452234268188),
 ('threat', 0.6512431502342224),
 ('link', 0.6267843842506409),
 ('non', 0.6138435006141663),
 ('search', 0.5444912910461426),
 ('document', 0.5067097544670105),
 ('help', 0.4897412657737732),
 ('assessment', 0.4890210032463074),
 ('intensive', 0.48097503185272217),
 ('aspect', 0.4807313084602356),
 ('hospitalize', 0.47300034761428833),
 ('lung', 0.4536757469177246),
 ('therapy', 0.44734495878219604)]
```

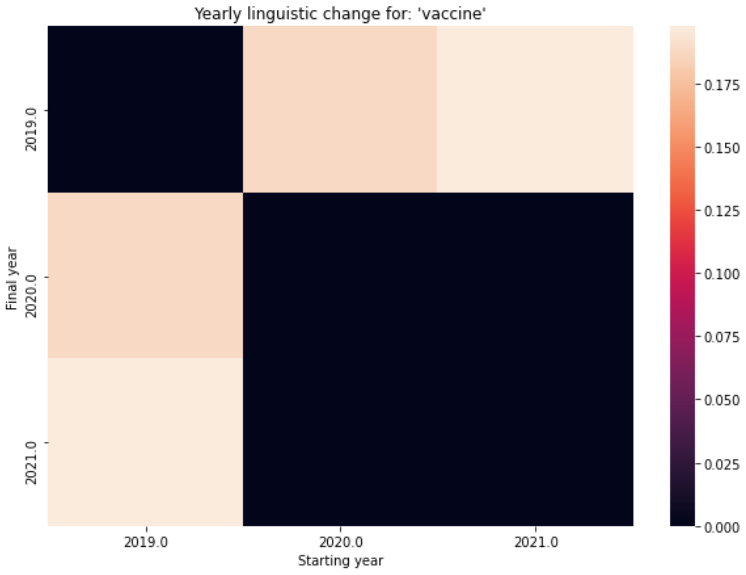
- The 20 least divergent words:

```
[('disease', 0.022487878799438477),
 ('syndrome', 0.02231299877166748),
 ('practice', 0.021638453006744385),
 ('medical', 0.021146655082702637),
 ('student', 0.02090078592300415),
 ('acute', 0.020557105541229248),
 ('care', 0.02038121223449707),
 ('state', 0.019770681858062744),
 ('sars', 0.019412517547607422),
 ('cause', 0.019277334213256836),
 ('country', 0.019088447093963623),
 ('system', 0.018129825592041016),
 ('new', 0.01715749502182007),
 ('world', 0.017048776149749756),
 ('case', 0.015927553176879883),
 ('health', 0.01433563232421875),
 ('severe', 0.01371544599533081),
 ('coronavirus', 0.011662721633911133),
 ('pandemic', 0.010997951030731201),
 ('covid-19', 0.003960847854614258)]
```

- Yearly linguistic change for "**america**"



- Yearly linguistic change for "**vaccine**"



9. Conclusion

Drawing on computational techniques, I found that psychologists study COVID-19 and psychological states, and economists study COVID-19 and labor, market, and economics. Scholars in the natural sciences study the nature of COVID-19 and its treatment. Overall, researchers use neutral and moderately subjective words. The social and natural corpora are distinguishable. Within each corpus, there are two clusters. Academics realized that the crisis and health problems caused by COVID-19 are escalating. Impacted by COVID-19, some words changed semantically.