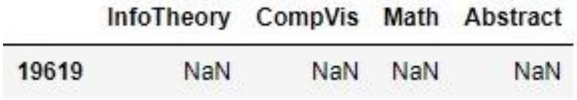
PART 1: Text Classification (Machine Learning and Neural Networks)

Pre - Processing

Both the ML Methods and Neural Network methods use the same pre-processing.

The steps are detailed below –

1. The dataset was checked for NULL values in any of the rows
   The Training Data set was clean but the Testing Data had a NULL row at index 19619. All NULLS and duplicates were removed from both the sets



	InfoTheory	CompVis	Math	Abstract
19619	NaN	NaN	NaN	NaN

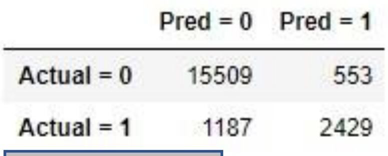
2. Punctuations were removed using regular expressions. Any character not either an alphabet or space were removed

3. Stop Words were removed from the text using regular expressions
   The stop words were obtained from the NLTK package

4. This cleaned data was then stored in a new CSV file for the NN method

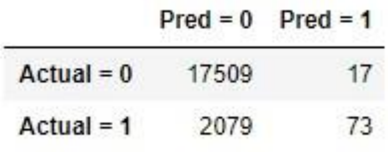
Neural Network Method

We are using **Recurrent Neural Networks** for this task. A bunch of different parameter were tried and tested. The confusion matrix from the NN method is shown below.



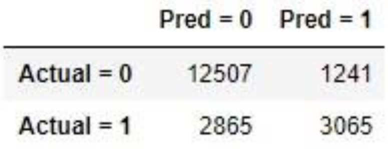
	Pred = 0	Pred = 1
Actual = 0	15509	553
Actual = 1	1187	2429

Info Theory



	Pred = 0	Pred = 1
Actual = 0	17509	17
Actual = 1	2079	73

Comp Vis



	Pred = 0	Pred = 1
Actual = 0	12507	1241
Actual = 1	2865	3065

Math

The accuracy of prediction is high for all 3 categories. For **Infor Theory, Comp Vis and Math** the accuracies are **91.1%, 89.3% and 79.1% respectively.**

But accuracy presents a very limited picture of the model. There are 3 other Metrics that are often used for evaluating the model. They are **Precision, Recall and Mathew's Correlation Coefficient**

Recall measures the fraction of actual positives that are returned or correctly classified. It tells you the ratio of articles correctly classified as positive to the total number of positive entries

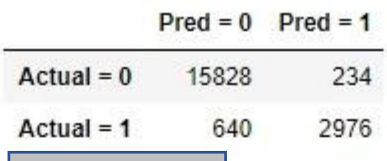
Precision measures the fraction of the articles marked positive that are actually positive. This metric tells us how many of the articles marked positive are actually positive.

For the NN method the Recall and Precision are 87% and 81% for Info Theory. Which is a good score. Math also has decent scores. But for Comp Vis the Precision is 85% but the Recall only 51%. This means those the algorithm marked as positive are actually positive 85% of the time but it is missing 49% of the positives which it is wrongly classifying as Negative.

MCC is the another that is used to evaluate a model. This is a more suitable metric to use because of the 4 metrics mentioned so far MCC is the only metric to consider the True Negative score. Hence giving a more suitable score. The MCC for NN methods is not very good. The scores are **68.8%, 15.2%, 47.3%** for **Info Theory, Comp Vis and Math** respectively.

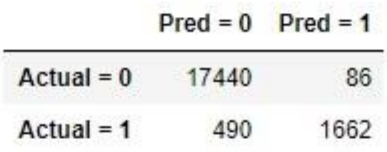
ML methods

For ML methods we used Logistic Regression (LR) and Linear SVC with TF IDF word vectors. The confusion matrix for the SVC method is shown below as it performed slightly better than Logistic Regression.



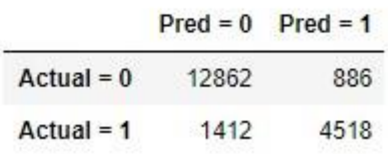
	Pred = 0	Pred = 1
Actual = 0	15828	234
Actual = 1	640	2976

Info Theory



	Pred = 0	Pred = 1
Actual = 0	17440	86
Actual = 1	490	1662

Comp Vis



	Pred = 0	Pred = 1
Actual = 0	12862	886
Actual = 1	1412	4518

Math

Both the methods performed much better than the NN method. But SVC was marginally better than LR.

For Info Theory and Comp Vis the accuracies are more than 95% and MCC more than 84%. These are very good scores. For Math the accuracy is 88% and MCC only 71%

SVM is good for high number of dimensions and semi structured data and it is also know for fewer overfitting issues.

Both the methods used the same pre-processing methods. The difference comes in the TF IDF vectors. A lot of parameters were tried, and the best was selected. For both the models the minimum document frequency was 2. That is words appearing only in 1 document are not considered.

Logistic Regression works well with max features. Setting max features at 10,000 helped LR while the scores for SVC dropped. For SVC the maximum document frequency was set at 0.5. That is word appearing in more than 50% of the corpus are removed.

NN Vs ML

We can see that both the ML methods performed better than the NN method. This is because the neural network was a simple implementation. There are many more parameter to experiment with that will give us better results.

Part 2: Topic Modelling using LDA

Two runs of LDA were tried with different pre-processing and model parameters

First Run

Pre-Processing

- Using a **regex tokenizer** we first convert the sentences into **tokens**. A **lower()** operation is then performed to convert all upper case characters **into lower case** characters.
- **Numbers** that are **by themselves** with **no alphabets** suffixed or prefixed are also **removed.**
- Words that are only a **single character** are **removed.** This gets rid of any stray characters that don't provide any information.
- Finally we are left with only the meaningful words in the document. These words are **lemmatized** using the **Word Net lemmatizer**.
- Using the **phrases method** in python we generate **bi grams** occurring in the document and add them to the token list if they occur **more than 10 times**.
- We then create a **dictionary representation** of the document corpus and **remove** all **words** occurring in **less than 20 documents** or in **more than 60%** of the documents.

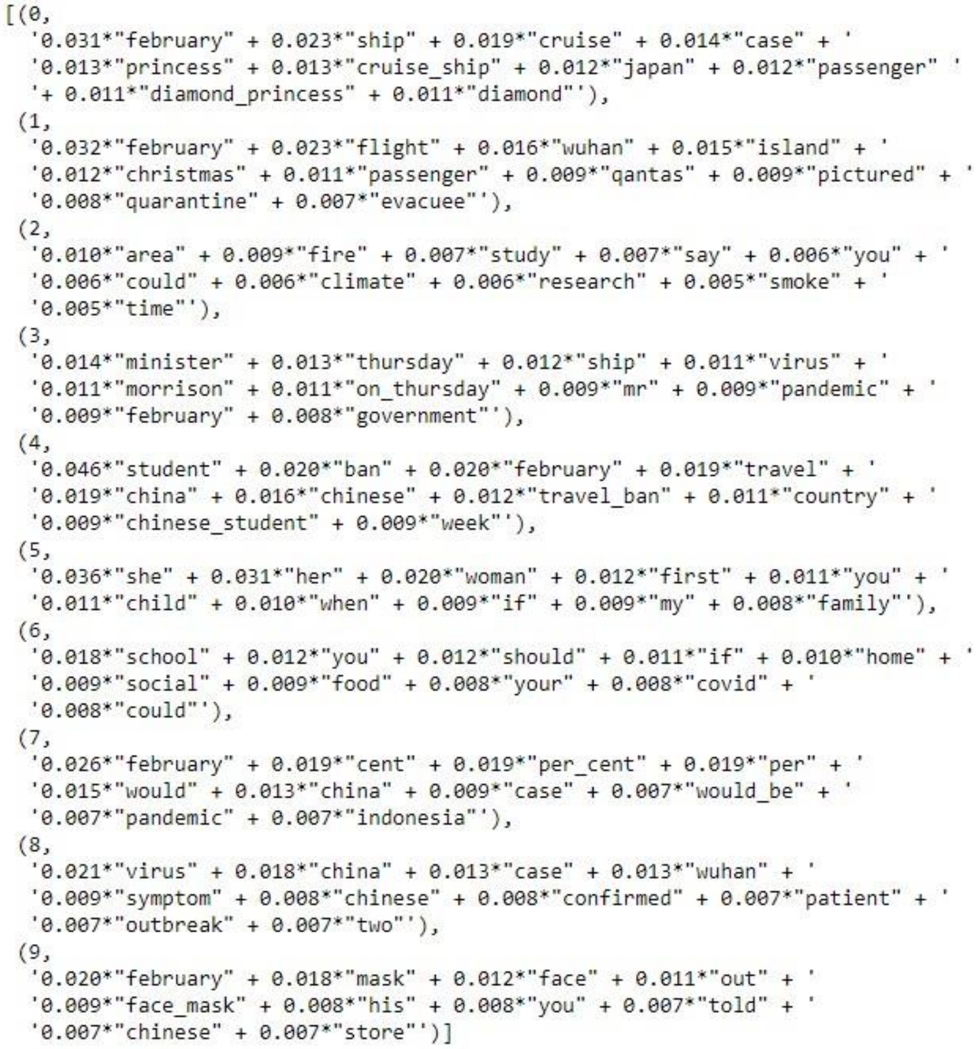
Model Parameters

Model parameters were adjusted based on intuition and the best results were frozen. For the **first model** we want **10 topics** recognized in the corpus. The Chunk size, passes and iterations are set to 2000, 30 and 500 respectively.

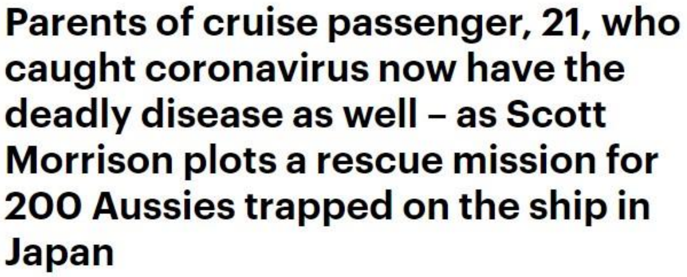
Observations

The figure in the next page shows the 10 topics generated. Each topic will be represented as a vector of words with weights attached to each word showing how much it is associated with the topic.

We can tell from a quick look that **most** of the topics are about **Corona Virus**. In fact, except for the topic at index 2, everything is about the on-going corona virus pandemic. A closer look will make it apparent that each of the Corona Virus topics have subtle differences.



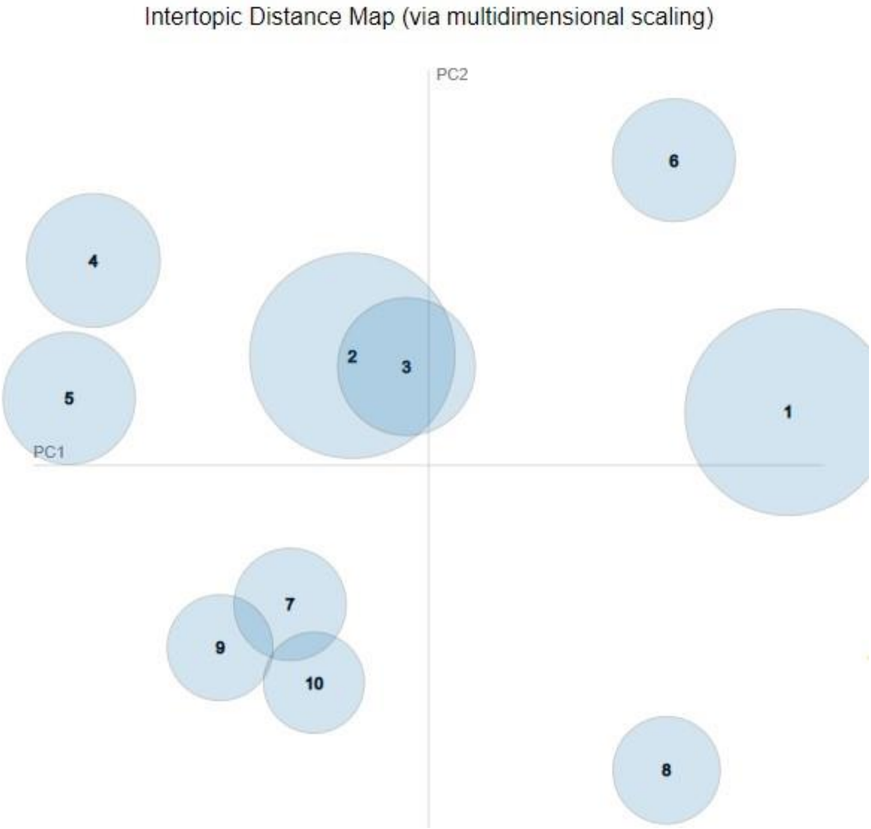
```
[(0,
  '0.031*"february" + 0.023*"ship" + 0.019*"cruise" + 0.014*"case" + '
  '0.013*"princess" + 0.013*"cruise_ship" + 0.012*"japan" + 0.012*"passenger" '
  '+ 0.011*"diamond_princess" + 0.011*"diamond"'),
 (1,
  '0.032*"february" + 0.023*"flight" + 0.016*"wuhan" + 0.015*"island" + '
  '0.012*"christmas" + 0.011*"passenger" + 0.009*"qantas" + 0.009*"pictured" + '
  '0.008*"quarantine" + 0.007*"evacuee"'),
 (2,
  '0.010*"area" + 0.009*"fire" + 0.007*"study" + 0.007*"say" + 0.006*"you" + '
  '0.006*"could" + 0.006*"climate" + 0.006*"research" + 0.005*"smoke" + '
  '0.005*"time"'),
 (3,
  '0.014*"minister" + 0.013*"thursday" + 0.012*"ship" + 0.011*"virus" + '
  '0.011*"morrison" + 0.011*"on_thursday" + 0.009*"mr" + 0.009*"pandemic" + '
  '0.009*"february" + 0.008*"government"'),
 (4,
  '0.046*"student" + 0.020*"ban" + 0.020*"february" + 0.019*"travel" + '
  '0.019*"china" + 0.016*"chinese" + 0.012*"travel_ban" + 0.011*"country" + '
  '0.009*"chinese_student" + 0.009*"week"'),
 (5,
  '0.036*"she" + 0.031*"her" + 0.020*"woman" + 0.012*"first" + 0.011*"you" + '
  '0.011*"child" + 0.010*"when" + 0.009*"if" + 0.009*"my" + 0.008*"family"'),
 (6,
  '0.018*"school" + 0.012*"you" + 0.012*"should" + 0.011*"if" + 0.010*"home" + '
  '0.009*"social" + 0.009*"food" + 0.008*"your" + 0.008*"covid" + '
  '0.008*"could"'),
 (7,
  '0.026*"february" + 0.019*"cent" + 0.019*"per_cent" + 0.019*"per" + '
  '0.015*"would" + 0.013*"china" + 0.009*"case" + 0.007*"would_be" + '
  '0.007*"pandemic" + 0.007*"indonesia"'),
 (8,
  '0.021*"virus" + 0.018*"china" + 0.013*"case" + 0.013*"wuhan" + '
  '0.009*"symptom" + 0.008*"chinese" + 0.008*"confirmed" + 0.007*"patient" + '
  '0.007*"outbreak" + 0.007*"two"'),
 (9,
  '0.020*"february" + 0.018*"mask" + 0.012*"face" + 0.011*"out" + '
  '0.009*"face_mask" + 0.008*"his" + 0.008*"you" + 0.007*"told" + '
  '0.007*"chinese" + 0.007*"store"')]
```



Parents of cruise passenger, 21, who caught coronavirus now have the deadly disease as well – as Scott Morrison plots a rescue mission for 200 Aussies trapped on the ship in Japan

For example, the first topic is focusing on the **Diamond Princess Cruise Ship**. The figure shows an article picked from the corpus (**ID 1448809605**) that talks about the cruise ship being stranded near Japan. **The term Japan also shows up in the topic.**

The figure below shows the inter-topic distances shown on a graph with two principle axes. These axes represent a trend that can be observed when moving along either of the axes from one topic to another.



We generated **10 topics** but a few of the topics are **bunched close together** and they seem to have formed their own groups. **Maybe there aren't as many topics as we are trying to create.** From the above graph we can deduce that **there are only 6 groups.** Topics **7, 9 and 10** for example are in the bottom left segment of the graph. This indicates a similarity in the topics. A quick look at the topics in the LDA Visualization in the jupyter file tells us that these topics are related to the Corona Virus pandemic.

4,5 and **2,3** are the two topic pairs that are on one side of the PC2 axis. 4 and 5 deal with **Corona Virus and international travel**. They mention the Princess ship and flights. 2 and 3 are focused on **local news**. They talk about **face masks, customers business and stores etc.**

Topics on the other side of the PC2 axis are completely different. 1 deals with the **Bush Fires** in January while 6 is about **general topics like family etc.**

There are a few issues with this model. For example we seem to have chosen a topic count that's too high. **There are words like she, her that offer no meaningful value.** We can try another run of LDA with different parameters and a lower topic count.

The Second Run

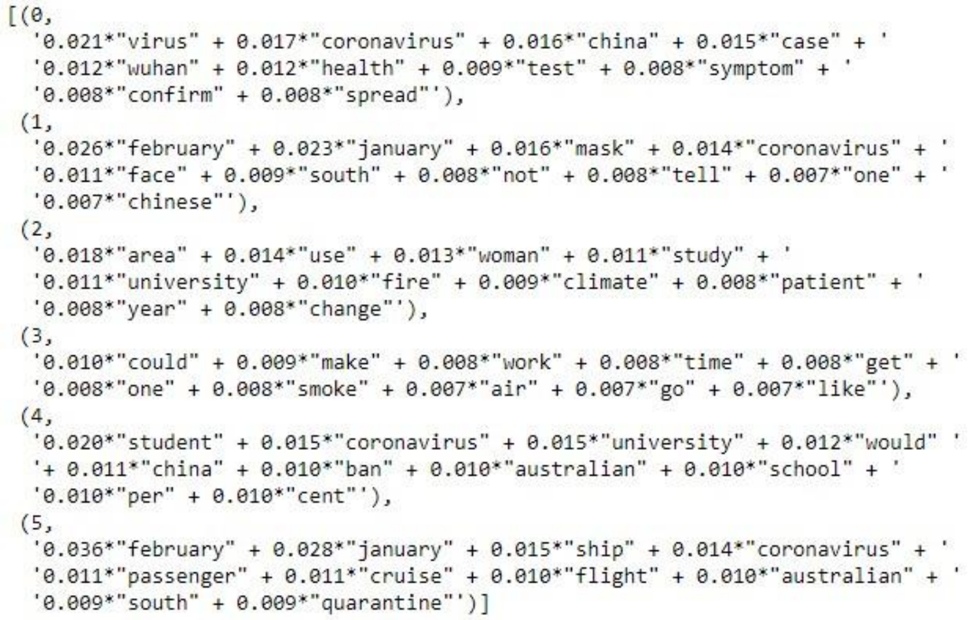
Pre-Processing

There were a few changes made this time. Like using the Spacy Lemmatizer instead of Word Net. The max document frequency was also changed to 0.7.

But the most important change made was the topic size. This time we went for 6 because those are the number of clusters we saw.

Other Parameters include chunk size at 2000, passes 30 and iterations at 1000

Observations

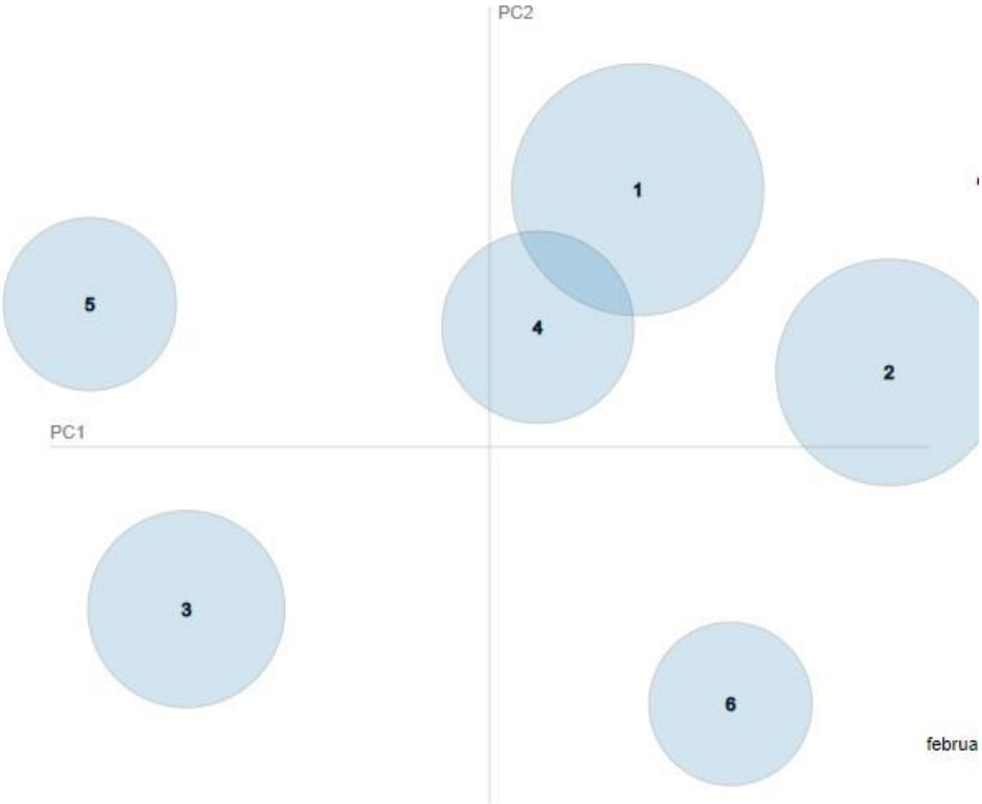


```
[(0,
  '0.021*"virus" + 0.017*"coronavirus" + 0.016*"china" + 0.015*"case" + '
  '0.012*"wuhan" + 0.012*"health" + 0.009*"test" + 0.008*"symptom" + '
  '0.008*"confirm" + 0.008*"spread"'),
 (1,
  '0.026*"february" + 0.023*"january" + 0.016*"mask" + 0.014*"coronavirus" + '
  '0.011*"face" + 0.009*"south" + 0.008*"not" + 0.008*"tell" + 0.007*"one" + '
  '0.007*"chinese"'),
 (2,
  '0.018*"area" + 0.014*"use" + 0.013*"woman" + 0.011*"study" + '
  '0.011*"university" + 0.010*"fire" + 0.009*"climate" + 0.008*"patient" + '
  '0.008*"year" + 0.008*"change"'),
 (3,
  '0.010*"could" + 0.009*"make" + 0.008*"work" + 0.008*"time" + 0.008*"get" + '
  '0.008*"one" + 0.008*"smoke" + 0.007*"air" + 0.007*"go" + 0.007*"like"'),
 (4,
  '0.020*"student" + 0.015*"coronavirus" + 0.015*"university" + 0.012*"would" '
  '+ 0.011*"china" + 0.010*"ban" + 0.010*"australian" + 0.010*"school" + '
  '0.010*"per" + 0.010*"cent"'),
 (5,
  '0.036*"february" + 0.028*"january" + 0.015*"ship" + 0.014*"coronavirus" + '
  '0.011*"passenger" + 0.011*"cruise" + 0.010*"flight" + 0.010*"australian" + '
  '0.009*"south" + 0.009*"quarantine"')]
```

The picture above and the one in the next page are from the second run. They show a much more appropriate spread of topics. They are again grouped into clusters.

Cluster 1 – Corona Virus

The first cluster is seen in the top right segment of the inter distance graph. These three topics (1,4,2 in the graph below) are shown in the index positions 0, 4 and 5. This cluster is related to Corona Virus. Topic 1 (index 0) talks about Corona virus and the origin in China. Topic 2 (index 5) is about the cruise ship related content. Topic 4 (index 4) is about student and the travel ban.



Cluster 2 – Corona Virus, local effects

The second cluster is seen in the right of the PC2 axis but below the PC1 axis. That is separate from the first cluster. This cluster (Topic 6, index 1) also deals with Corona Virus but the news here is more local. Words like face mask, Victoria, Queensland, business

Cluster 3 – Bush Fires

The 3rd cluster takes us across the PC2 axis. There are 2 topics (topics 5 and 3 at indexes 2 and 3 respectively) here both related to the bush fires. They talk about research university and climate change etc.

Are all top topic words comprehensible sets of words?

The first run had a lot of words like she, her etc that don't offer any value. The second run was much better due to the more suitable number of topics.

What sorts of news mentioning Monash University is there, and why is it mentioning the university

Overall, **2 major topics** stand out. **Coronavirus and Bush Fires**. But not all of them mentioned Monash in the same context. We've **identified 3 usages for Monash**. **Monash Hospital** is one that is mentioned a lot in the **corona virus related articles**. Monash University is also mentioned in some articles talking about **medical research** or when someone working at the university gives in interview. Monash university is also mentioned in the context of **travel bans and students getting stranded in china**.

Finally, there is one other mention of Monash. **Area of Monash** is used to refer to the suburbs around the university campuses. Some even use the phrase city of Monash.

How does the topic modelling present the topics and any advantages or shortcomings of topic modelling for the role.

The PCA analysis breaks the documents down. As we move from one end of a dimension to the other we see a shift in the topic context. Like the Corona Virus topics. Topic Modeling does a very good clustering job.

The visualization is also interactive. When a word is selected the graph adjusts the sizes of the circles to reflect the distribution of the selected word in each of the topics. For example when Fire is highlighted most of the circles disappear and only the topics covering bush fires are shown.