一、本周研究内容

1. 按照上周会议老师们的指导，对美国各州的 twitter 数据集进行了扩充，将 BERT 预测的从 756 扩充到了 4318，手工标注的从 29 扩充到 208。

2. 从约翰霍普金斯的疫情 github 上整理了美国 CDC 各州确诊人数数据。

二、项目实施当前状态

目前已经完成对后台数据的整合，下周需要最终建立一个疫情预警 score 的 model，整合进各类已有数据，并在 web app 上可视化。

三、本周成果

1. 扩充数据

- 按照 **50_us_states_all_data.csv** 文件里的美国各州的缩写，在上周数据的基础上进行了扩充。

```
[71]: len(bert_location_final)

[71]: 4318
```

```
[126]: len(twitter_location_final)

[126]: 208
```

```
twitter_with_geo_location.head(5)
```

	created_at	full_text	full_name	label	state
6	Mon Mar 02 11:39:44 +0000 2020	Coronavirus-19 update: we have 95 cases in Spa...	Barcelona, Spain	1.0	
43	Tue Mar 03 03:05:18 +0000 2020	2 confirmed cases of Coronavirus in GA. Discov...	Newnan, GA	1.0	Georgia
52	Thu Mar 05 15:03:13 +0000 2020	So my friend just told me one of his tenants (...	Queens, NY	1.0	New York
54	Wed Mar 04 07:46:44 +0000 2020	Anxious days ahead for patients, their familie...	Birmingham, England	1.0	
77	Sun Mar 01 13:05:51 +0000 2020	20 people have now died in South Korea. Approx...	National Institutes of Health (NIH)	1.0	New Hampshire

图一： 扩充后数据截图

- 也尝试了把城市对应到各州，但是发现效果不是特别好。通过查看结果 ，**发现原因是美国相同的城市名出现在不同州，而且和国外的很多城市重名。**因此这样处理会导致很多错误的归类，所以没有通过城市归类，仅仅是通过州的缩写进行了归类。

2. 整理美国各州确诊人数

从约翰霍普金斯大学的 Github repoistory 上 https://github.com/CSSEGISandData/COVID-19 获取美国各州的确诊人数。

四、本周问题

1. 通过机器学习/深度学习 分类到美国各州的 twitter 个数曲线图不是很有规律。

分析：由于分类到的与疫情爆发有关的 twitter 数量比较少，分散到每天，以及各州后，数量会更少，没有规律。与 raw twitter 个数不同，尽管随着时间推移，人们在 twitter 上讨论 covid19 的频率更高，但是不一定代表在 twitter 上报道自己等咳嗽症状并且认为自己是疑似的相关 twitter 会更多。所以没有规律影响不是很大，也在预计之中。重要的是下一步对现有数据进行有效利用，生成自定义的疫情预警 score regression model。

五、下周计划

最终建立一个疫情预警 score 的 model，整合进各类已有数据，并在 web app 上可视化。