一、本周研究内容

1. 对 Twitter 预警系统以及 3315 计划的 Web app，还有 CXR 数据集建立项目写了需求。

2. 针对上周五开会中，陈茁教授所提出的 twitter 数据集中的地理位置信息问题，找到了解决方案。通过使用 twitter 用户 profile 中设定的地理位置信息，数据集新增了 214081 条 tweets。

3. 研究了 semi-supervised learning 的方法，预计下周通过 semi-supervised learning 的方法对 214081 条 unlabeled 数据进行标注。

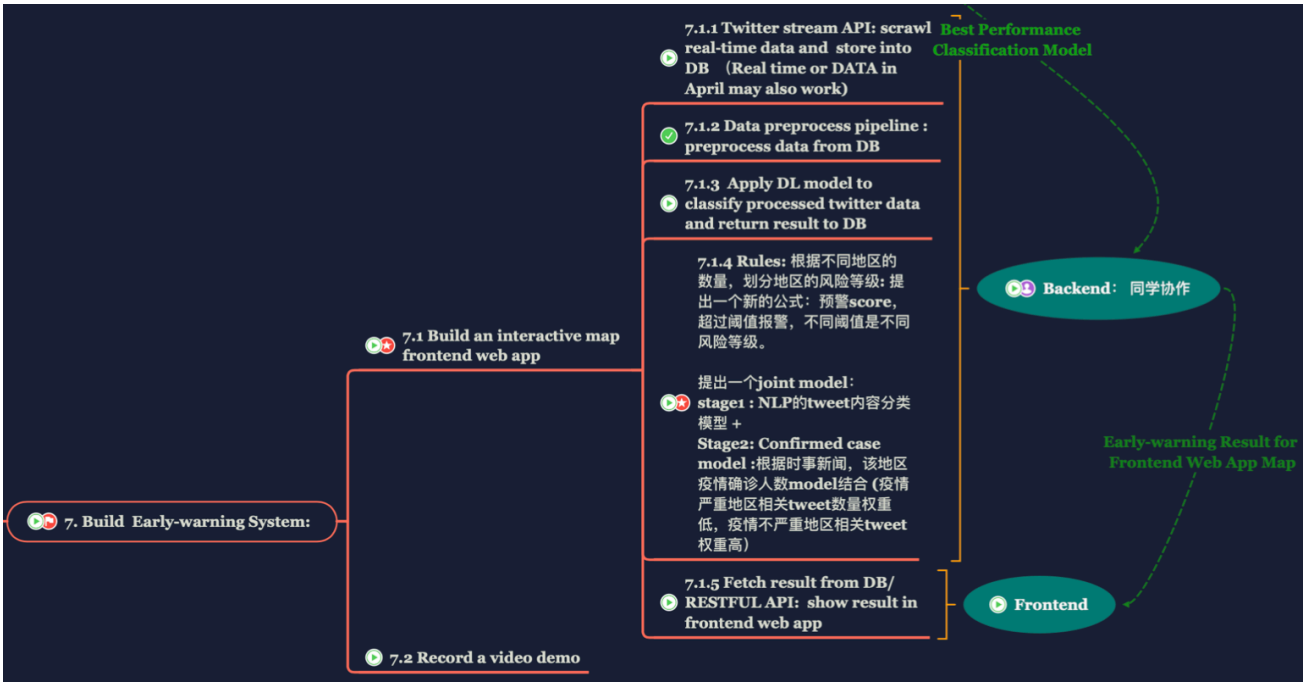
二、项目实施当前状态

数据集方面本周进行了调整，通过使用 twitter profile geo location 对数据集进行了扩充。下一步是用 semi-supervised learning 将扩充后的数据集进行 label，然后将处理过后的数据集，重新在之前实现好的 KNN, SVM，DPCNN 以及 Fine tune BERT 上进行实验，获取最终的实验结果。

三、本周成果

1. 预警 Web App 需求

预警 Web App 需求中主要将 twitter 预警系统的具体需求列了出来，包括：Twitter API 部分，数据处理部分，分类器，预警规则制定以及前端可视化地图展示。目前 BERT 在分类器上已经取得了特别好的效果，下一步重点是完成对预警规则的制定。



图一： Twitter 预警系统思维导图

2. Twitter 数据集扩充

针对陈茁教授在会上提出的建议，经过研究，发现 twitter 数据的地理位置信息包括两种：第一种为 tweet 本身的地理位置信息 (tweet location)，第二种为用户在账号中设置的地理位置信息 (profile location)。本项目

之前手动打标的 7000 多条数据是基于第一种 tweet 本身的地理位置信息的，这次扩充的是第二种用户的账号地理位置信息。

扩充后的数据集如下图所示，用户 profile 中的地理位置现在位于 location column 里。通过 profile 扩充了 214081 条 tweets。

```
df_location.head(10)
```

	created_at	created_at	id_str	id_str	full_text	place	lang	lang	user	location
29	Mon Mar 02 20:57:31 +0000 2020	Sat Nov 04 04:47:20 +0000 2017	1234583636890046465	926672186512891904	#Covid19usa is highlighting problems in #healt...	NaN	en	None	{'id': 926672186512891904, 'id_str': '92667218...	Ottawa, Ontario
31	Mon Mar 02 21:57:41 +0000 2020	Thu Jul 30 17:58:53 +0000 2009	1234598781372502027	61552834	How can philanthropy plan for, respond to, and...	NaN	en	None	{'id': 61552834, 'id_str': '61552834', 'name':...	Atlanta, GA
38	Tue Mar 03 10:01:46 +0000 2020	Wed May 06 08:47:09 +0000 2009	1234781001706065920	38142380	Coronavirus deaths in US rise to six as Trump ...	NaN	en	None	{'id': 38142380, 'id_str': '38142380', 'name':...	London
41	Tue Mar 03 12:52:23 +0000 2020	Tue Nov 29 04:50:13 +0000 2016	1234823936719847424	803461037517213696	@VinnieTortorich @drdrew Have that person run ...	NaN	en	None	{'id': 803461037517213696, 'id_str': '80346103...	San Diego, CA
77	Thu Mar 05 07:46:02 +0000 2020	Tue Apr 13 14:18:12 +0000 2010	1235471619239350272	132535895	#Trump is in over his head...and he knows it\n...	NaN	en	None	{'id': 132535895, 'id_str': '132535895', 'name...	Philadelphia, PA.
111	Mon Mar 02 01:52:06 +0000 2020	Wed Oct 10 17:18:24 +0000 2018	1234295384677273600	1050073072336887815	NBC's Chuck Todd pressed Mike Pence to 'name s...	NaN	en	None	{'id': 1050073072336887815, 'id_str': '1050073...	Texas, USA
119	Mon Mar 02 21:32:19 +0000 2020	Tue May 02 16:53:25 +0000 2017	1234592395091570688	859450768805347329	@nitashatiku Oh nice. We doing Tide Challenge ...	NaN	en	None	{'id': 859450768805347329, 'id_str': '85945076...	Oakland, CA
127	Tue Mar 03 12:51:42 +0000 2020	Tue Jan 21 09:57:20 +0000 2014	1234823768549228544	2302837495	We think that someone wears a mask is to stop ...	NaN	en	None	{'id': 2302837495, 'id_str': '2302837495', 'na...	Lille, France
159	Thu Mar 05 17:45:08 +0000 2020	Thu Aug 30 18:15:14 +0000 2012	1235622388051648512	792119786	British airline Flybe collapses as coronavirus...	NaN	en	None	{'id': 792119786, 'id_str': '792119786', 'name...	Georgia
170	Fri Mar 06 07:56:10 +0000 2020	Wed Aug 19 10:31:13 +0000 2009	1235836557312921600	66966263	The superhero iin need now. \nMeet Stalin Raj,...	NaN	en	None	{'id': 66966263, 'id_str': '66966263', 'name':...	Bengaluru, Karnataka

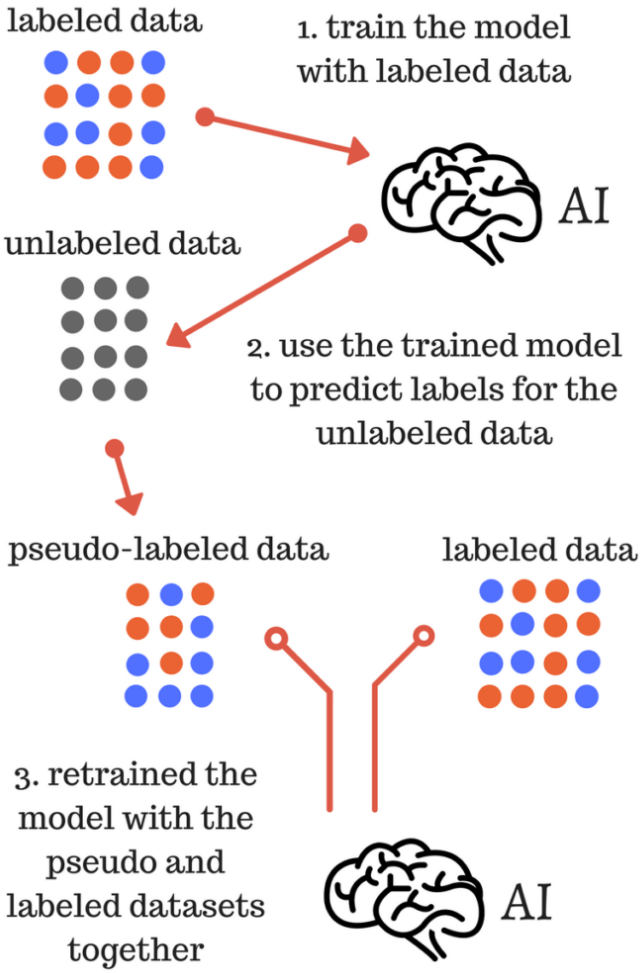
```
df_location.shape
```

```
(214081, 10)
```

图二： 扩充后的 Twitter 数据集

3. Semi-supervised learning（Pseudo labeling）

现在的数据集组成是 7064 条 labeled tweets 和 214081 unlabeled tweets。我们已经通过 labeled tweets 训练了 classifier，所以接下来就是通过训练好的 classifier 去 predict 214081 unlabeled tweets。然后 predict 后的结果 pseudo-labeled data 和 labelled data 组合成新的数据集，再重新训练 classifier，获取最终的模型。



图三： Semi-supervised learning（图片来源：https://datawhatnow.com/pseudo-labeling-semi-supervised-learning/ ）

四、下周计划

下周通过 semi-supervised learning 完成对全部数据集的标注。