"Ningbo 3315 Project"

一、本周研究内容

研究内容：

1. 继续优化前端可视化的部分。

项目 GitHub 地址：https://github.com/yiming95/3315project

网页链接：https://yiming95.github.io/3315project/

2. 大数据和自然语言处理部分: 建立 Twitter 数据集，对数据集进行预处理。

二、项目实施当前状态

项目进度实施情况：

1. 3315 DEMO 前端可视化的部分基本完成。

2. 大数据和自然语言处理部分，建立了 Twitter 数据集，对 Twitter 数据集进行了预处理。

项目整体进度完成情况

1. 3315 DEMO: 可视化 DEMO 采取前后端分离的开发方式，目前的开发都基于前端的可视化部分。仍需要建立数据库，进行后端部分的开发，并针对前端部分 UI/UX 进行修改与补充。

2. 大数据和自然语言处理部分: Twitter 数据没有 label，下周内容是手动给 twitter 数据打标签，然后接下来进行模型训练。

三、本周成果

1. 在 3315 DEMO 的预警部分，增加了 1 个 环形图 (doughnut chart) 和 1 个横向条形图表 (horizontal bar chart)。如图一所示，当用户点击"数据详情"时，在 table 下面会显示出两个新加的 charts。

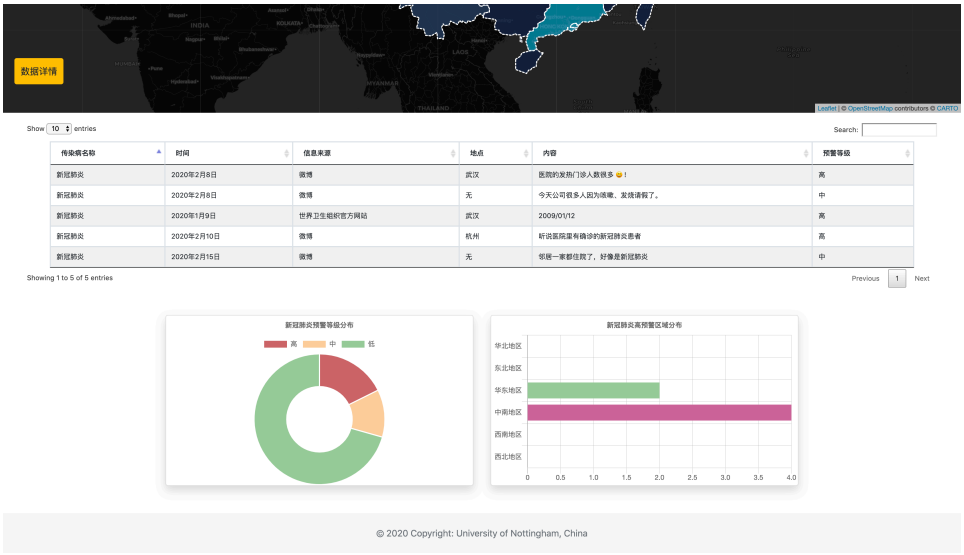

图 1： 新增环形图和横向条形图表

环形图是对上面的预警地图进行归纳。用来表征全国 34 个省、自治区和直辖市中，预警等级的分布。红色(6 个) 代表预警等级为高，黄色（4 个）为预警等级为中，绿色（24 个）表示预警等级为低。在环形图右边的 horizontal bar 是对预警等级为高的地区进行分类，由图可知，预警等级为高的 6 个地区中有 2 个位于华东地区，4 个位于中南地区。

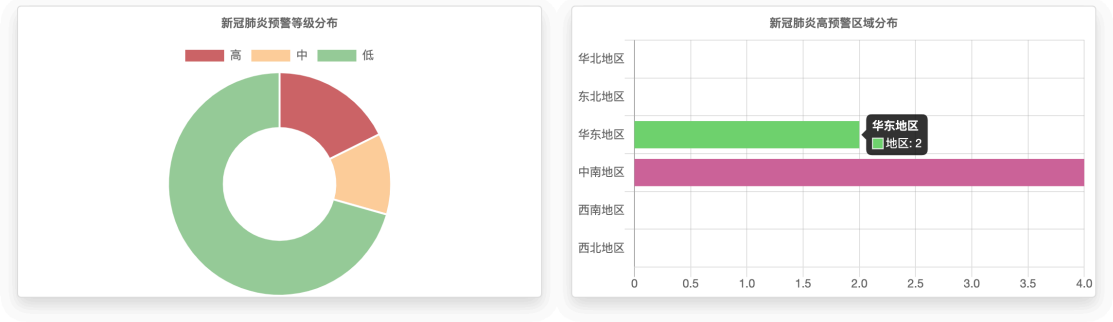

图 2：环形图和横向条形图表

2. 在大数据和 NLP 方向，使用 Christian et al. 整理的 twitter 数据集 [1]。该数据集使用 Twitter API 持续收集 tweets，搜集时间从 1 月 22 日开始，GitHub 仓库地址为 https://github.com/lopezbec/COVID19_Tweets_Dataset。

然而由于 Twitter 公司的 Term & Service，该数据集只能存储 tweet_id 并且有 twitter developer account 才能获得 tweets 数据。因此需要在本地通过代码进一步将 tweet_id 转为 tweet。相关代码运行如下图：

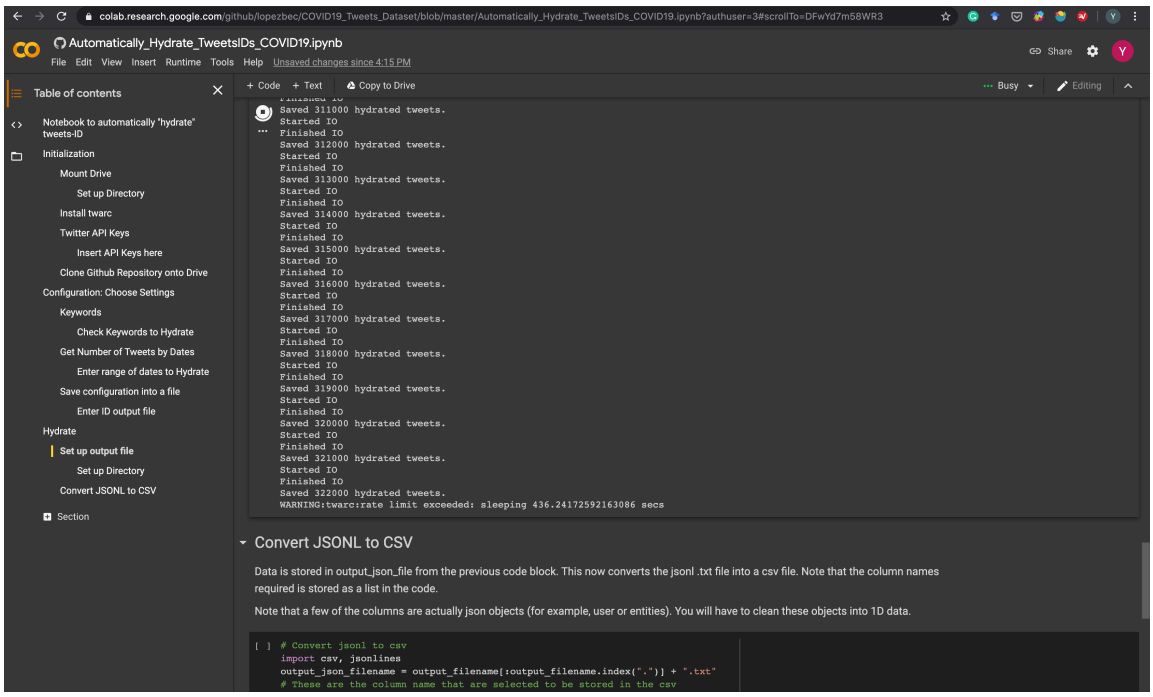

图 3：处理 Twitter 数据集（从 tweet_id 到 tweet）

处理后的 Twitter 数据集存储为 CSV 格式，如下图所示。

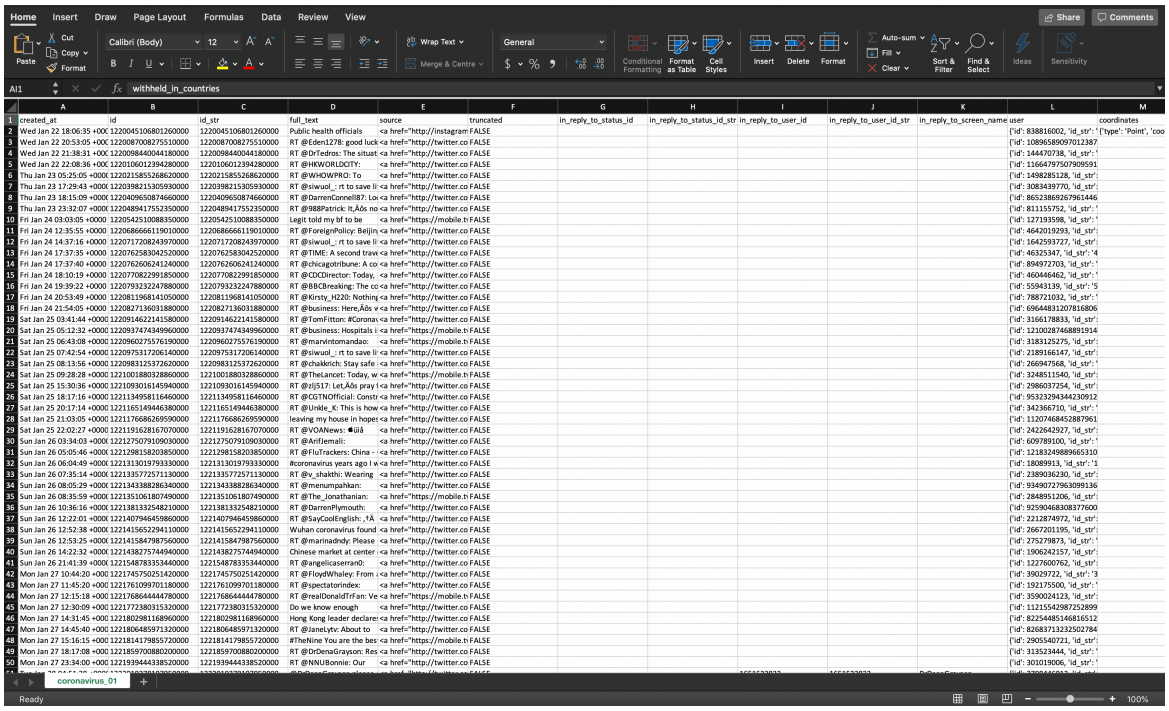

图 4：处理后的 Twitter 数据集 CSV 格式

3. 对 Twitter 数据（full_text 列）进行预处理，包括转化小写(lower case)，删除标点符号(remove puntuaction)等。具体代码如下所示。

Load dataset

Dataset is cited from the paper "UNDERSTANDING THE PERCEPTION OF COVID-19 POLICIES BY MINING A MULTILANGUAGE TWITTER DATASET", GitHub repository is https://github.com/lopezbec/COVID19_Tweets_Dataset.

[4]:
```
# load csv data
data = pd.read_csv('coronavirus_01.csv',encoding='utf-8')
df = pd.DataFrame(data)

# only select some important fileds
df_new = df[['created_at','id_str','full_text','coordinates','place','retweet_count','entities','retweeted','lang']]
df_new.sample(10)
```

[4]:

	created_at	id_str	full_text	coordinates	place	retweet_count	entities	retweeted	lang
293102	Sun Jan 26 10:06:56 +0000 2020	1221373951336538114	RT @new_prykm: This video shows how South Kore...	NaN	NaN	44187	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en
162451	Fri Jan 31 13:18:22 +0000 2020	1223234063856627713	RT @PhilippineStar: President Duterte has agre...	NaN	NaN	483	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en
142173	Fri Jan 31 18:19:16 +0000 2020	1223309787833798656	Virologist reveals the science behind fight to...	NaN	NaN	18	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en
300805	Tue Jan 28 04:21:07 +0000 2020	1222011697419837441	RT @Grandpa: coronavirus, Kobe, now this. bad ...	NaN	NaN	34728	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en
91403	Wed Jan 29 02:42:48 +0000 2020	1222349343241445376	RT @SJPFISH: 🔥 China Temporary bans Wildlife tr...	NaN	NaN	456	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en
68631	Wed Jan 29 17:33:36 +0000 2020	1222573522867802112	RT @DrTedros: I was struck by the determinatio...	NaN	NaN	307	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en
100297	Sun Jan 26 15:08:59 +0000 2020	1221449963428663296	RT @hanipersian: Military medical teams arrive...	NaN	NaN	8794	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en
77580	Thu Jan 30 20:57:19 +0000 2020	1222987177899855872	RT @Robertfrank615: I showed this video to a t...	NaN	NaN	757	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en
133380	Tue Jan 28 12:08:03 +0000 2020	1222129204726321152	RT @Grandpa: coronavirus, Kobe, now this. bad ...	NaN	NaN	34728	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en
88474	Thu Jan 23 02:24:43 +0000 2020	1220170466259652610	RT @Jordan_Sather_: I saw a headline that said...	NaN	NaN	929	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en

图 5：使用 Pandas 加载数据集

Preprocessing

Lower case, Remove Puntuaction, Remove StopWords

[13]:
```
## Lower case
def remove_uppercase(text):
    text_lowercase = ' '.join(x.lower() for x in text.split())
    return text_lowercase
```

[16]:
```
df_new['full_text'].apply(lambda x: remove_uppercase(x))
```

[16]:
```
0         public health officials have confirmed the fir...
1         rt @eden1278: good luck everyone #coronavirus ...
2         rt @drtedros: the situation with new #coronavi...
3         rt @hkworldcity: 🦠🇭🇰 *cough* *cough* *cough* *...
4         rt @whowpro: to reduce risk of #coronavirus, p...
5         rt @siwuol_: rt to save life #coronavirus http...
6         rt @darrenconnell87: looking forward to scotti...
7         rt @988patrick: it's not a joke . share guys a...
8         legit told my bf to be careful and wear a mask...
9         rt @foreignpolicy: beijing's successful attemp...
10        rt @siwuol_: rt to save life #coronavirus http...
11        rt @time: a second travel-related case of coro...
12        rt @chicagotribune: a coronavirus case has bee...
13        rt @cdcdirector: today, @cdcgov confirmed a 2n...
14        rt @bbcbreaking: the coronavirus which first e...
15        rt @kirsty_h220: nothing to worry about, but w...
16        rt @business: here's where cases of coronaviru...
17        rt @tomfitton: #coronavirus is yet another rea...
18        rt @business: hospitals in wuhan are turning a...
19        rt @marvintomandao: hey, we have a better idea...
20        rt @siwuol_: rt to save life #coronavirus http...
```

图 6：对文本全部小写

[20]:
```
## Remove Puntuaction
def remove_punctuaction(text):
    text_nopunct = "".join([char for char in text if char not in string.punctuation])
    return text_nopunct
```

[21]:
```
df_new['full_text'].apply(lambda x: remove_punctuaction(x))
```

...

[22]:
```
df_new.head(5)
```

[22]:

	created_at	id_str	full_text	coordinates	place	retweet_count	entities	retweeted	lang	full_text_clean
0	Wed Jan 22 18:06:35 +0000 2020	1220045106801266688	public health officials have confirmed the fir...	{'type': 'Point', 'coordinates': [-84.33137373...	{'id': '00c26afc77ee7aaa', 'url': 'https://api...	0	{'hashtags': [], 'symbols': [], 'user_mentions...	False	en	public health officials have confirmed the fir...
1	Wed Jan 22 20:53:05 +0000 2020	1220087008275513344	rt @eden1278: good luck everyone #coronavirus ...	NaN	NaN	9823	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en	rt @eden1278: good luck everyone #coronavirus ...
2	Wed Jan 22 21:38:31 +0000 2020	1220098440044187648	rt @drtedros: the situation with new #coronavi...	NaN	NaN	423	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en	rt @drtedros: the situation with new #coronavi...
3	Wed Jan 22 22:08:36 +0000 2020	1220106012394283008	rt @hkworldcity: 🦠🇭🇰 *cough* *cough* *cough* *...	NaN	NaN	1857	{'hashtags': [{'text': 'FreeHongKong', 'indice...	False	en	rt @hkworldcity: 🦠🇭🇰 *cough* *cough* *cough* *...
4	Thu Jan 23 05:25:05 +0000 2020	1220215855268626433	rt @whowpro: to reduce risk of #coronavirus, p...	NaN	NaN	820	{'hashtags': [{'text': 'coronavirus', 'indices...	False	en	rt @whowpro: to reduce risk of #coronavirus, p...

图 7：对文本去除标点

四、下周计划

1. 针对产品 UI/UX 的反馈和建议对前端部分进行修改与补充。

2. 对 Twitter 数据集进行进一步处理，进行人工打标 (label) 以为后续的 classification model 做准备。

Reference

[1] Lopez, Christian E., Malolan Vasu, and Caleb Gallemore. "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset." *arXiv preprint arXiv:2003.10359* (2020).