## 一、本周研究内容

**1.** 项目 *"COVID-19 Twitter Mining and Text Classification"* 的数据全部标注完成，7064/7064，存储在文件 "**COVID19_twitter_dataset.csv**" 里。

**2.** 对整体数据集的 class 分布情况等进行了统计：总 7064 条 twitter, 其中 434 条有关 (label 1)，6630 条无关(label 0)。

## 二、项目实施当前状态

完成全部数据集标注，接下来是调研一下可否在数据集层面解决数据 class 分布不均匀的问题，以及能否使用文本增强技术对数据集扩充。

## 三、本周成果

**1. 数据集打标：**

完成了对全部数据集的人工打标。打标后的数据集存储在文件 '**COVID19_twitter_dataset.csv**' 里。

```
df_labeled_dataset
```

| | index | created_at | id_str | place | full_text | label |
|---|---|---|---|---|---|---|
| 0 | 474 | Tue Mar 03 19:58:00 +0000 2020 | 1234931050360406018 | {'id': '7929cea6bd5b32bd', 'url': 'https://api... | "Tokyo 2020 could be postponed amid coronaviru... | 0.0 |
| 1 | 663 | Wed Mar 04 13:32:10 +0000 2020 | 1235196337383247873 | {'id': '5ac77d85af64387f', 'url': 'https://api... | I don't know which i'd want less, Coronavirus ... | 0.0 |
| 2 | 972 | Sun Mar 01 05:58:23 +0000 2020 | 1233994975853060097 | {'id': 'e84c921cdb974f84', 'url': 'https://api... | @realDonaldTrump #BlaBlaDonaldTrump again. Wha... | 0.0 |
| 3 | 984 | Mon Mar 02 00:05:20 +0000 2020 | 1234268517865680896 | {'id': '9dafd05b1158873b', 'url': 'https://api... | Chucky Toad from Meet the Depressed is one of ... | 0.0 |
| 4 | 1114 | Thu Mar 05 06:00:33 +0000 2020 | 1235445072746795014 | {'id': '53b67b1d1cc81a51', 'url': 'https://api... | Nicola Sturgeon warns Scots of 'rapid rise' in... | 0.0 |
| 5 | 2094 | Tue Mar 03 18:10:25 +0000 2020 | 1234903975125635072 | {'id': '300bcc6e23a88361', 'url': 'https://api... | As a Washington based outlet, the question bea... | 0.0 |
| 6 | 2174 | Mon Mar 02 11:39:44 +0000 2020 | 1234443269397979136 | {'id': '1a27537478dd8e38', 'url': 'https://api... | Coronavirus-19 update: we have 95 cases in Spa... | 1.0 |
| 7 | 2195 | Wed Mar 04 03:44:26 +0000 2020 | 1235048429631369216 | {'id': '99e789320196ef6a', 'url': 'https://api... | If you've ever drank water from the garden hos... | 0.0 |
| 8 | 4188 | Thu Mar 05 14:47:40 +0000 2020 | 1235577725622120449 | {'id': 'dd9c0d7d7e07eb49', 'url': 'https://api... | @KingKong_SYMZ This is the only cure for #Coro... | 0.0 |
| 9 | 4443 | Wed Mar 04 09:03:33 +0000 2020 | 1235128737500475393 | {'id': '1e5cb4d0509db554', 'url': 'https://api... | UTC -7 \| KR — If coronavirus spreads to Nort... | 0.0 |
| 10 | 4455 | Thu Mar 05 03:29:06 +0000 2020 | 1235406960717189120 | {'id': '3df4f427b5a60fea', 'url': 'https://api... | I didn't think the coronavirus would actually ... | 0.0 |
| 11 | 5041 | Tue Mar 03 07:51:13 +0000 2020 | 1234748147739455488 | {'id': 'a3b39b40a6f077f5', 'url': 'https://api... | @ClayTravis Clay,\nThe #StockMarket is coming ... | 0.0 |
| 12 | 5042 | Tue Mar 03 10:06:22 +0000 2020 | 1234782158583144450 | {'id': '7ce215910e5ebe5c', 'url': 'https://api... | BBC News - Coronavirus: Greggs 'would pay staf... | 0.0 |
| 13 | 5887 | Fri Mar 06 06:54:05 +0000 2020 | 1235820932913958912 | {'id': '72600cca9a98ed11', 'url': 'https://api... | Coronavirus: how to protect yourself from the ... | 0.0 |
| 14 | 6796 | Mon Mar 02 04:24:59 +0000 2020 | 1234333858646102016 | {'id': '00cc0d5640394308', 'url': 'https://api... | You know an interesting fact about #CoronaViru... | 0.0 |
| 15 | 6874 | Fri Mar 06 18:31:07 +0000 2020 | 1235996348735873025 | {'id': '5c43cbdfce4d3247', 'url': 'https://api... | 🔴\nPeople are Dying because of animals eating ... | 0.0 |
| 16 | 6888 | Sun Mar 01 09:33:56 +0000 2020 | 1234049220845408256 | {'id': '72606cdcf2847dd4', 'url': 'https://api... | Fear not the Government has got #coronavirus u... | 0.0 |
| 17 | 7110 | Wed Mar 04 21:09:57 +0000 2020 | 1235311541496541184 | {'id': '1c69a67ad480e1b1', 'url': 'https://api... | #Coronavirus won't make Houstonians sick as mu... | 0.0 |
| 18 | 7572 | Thu Mar 05 09:16:30 +0000 2020 | 1235494386097668096 | {'id': '5c9d123437711a9d', 'url': 'https://api... | IMF provides $50bn to fight coronavirus outbre... | 0.0 |
| 19 | 7611 | Sun Mar 01 22:55:12 +0000 2020 | 1234250866011492352 | {'id': '5faafada28b440c3', 'url': 'https://api... | LookVice President Mike Pence defends Donald T... | 0.0 |
| 20 | 7749 | Thu Mar 05 10:57:03 +0000 2020 | 1235519688421056514 | {'id': '514d0719e0a80a43', 'url': 'https://api... | Abu Dhabi Crown Prince Mohammed bin Zayed has ... | 0.0 |

图一： 数据集部分真实数据截图

**2. 数据集分布情况:**

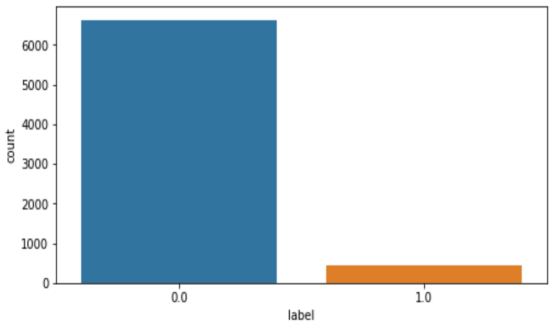
数据集描述：原始数据集是基于 twitter API 关键词 'coronavirus' 搜集到的 twitter 数据，搜集时间是在 2020 年 1 月份到 2020 年 3 月份。然后本项目去除掉转推和没有地理位置信息的 twitter 后，剩下了 7068 个 twitter，又去除了 4 个一样的 twitter，剩下 7064 个 twitter 进行人工标注。人工标注是将 twitter 分为 COVID-19 在本地区潜在爆发相关的和无关系的这两类。标注时候所采取的规则 (guidelines)为：twitter 中有

描述相关**症状**如咳嗽；twitter 中有提到**确诊** covid19 的相关内容；twitter 中提到有 covid19 在本地**疑似**传播的。如符合以上规则，则 label 为 ‘1’ 否则为 ‘0’ 。

数据集分布为：6630 个为 0，434 个为 1。

| | label | full_text |
|---|---|---|
| 0 | 0 | 6630 |
| 1 | 1 | 434 |

图二： 数据集的统计分布数据




图三： 数据集分布柱状图

## 四、下周计划

调研一下可否在数据集层面解决数据 class 分布不均匀的问题，以及能否使用文本增强技术对数据集扩充。