一、本周研究内容

研究内容：

1. 解决了上周发现数据乱码的问题。

2. 根据打标的 guidelines，对数据集进行了打标，进度 1000/7068。

3. 扩展打标的 guidelines 到了 3 条，简单概括为：**有相关症状的**，**有确诊的或因病死亡的**，**有疑似可疑或社区传播的** tweet 都分类为 class “1”。

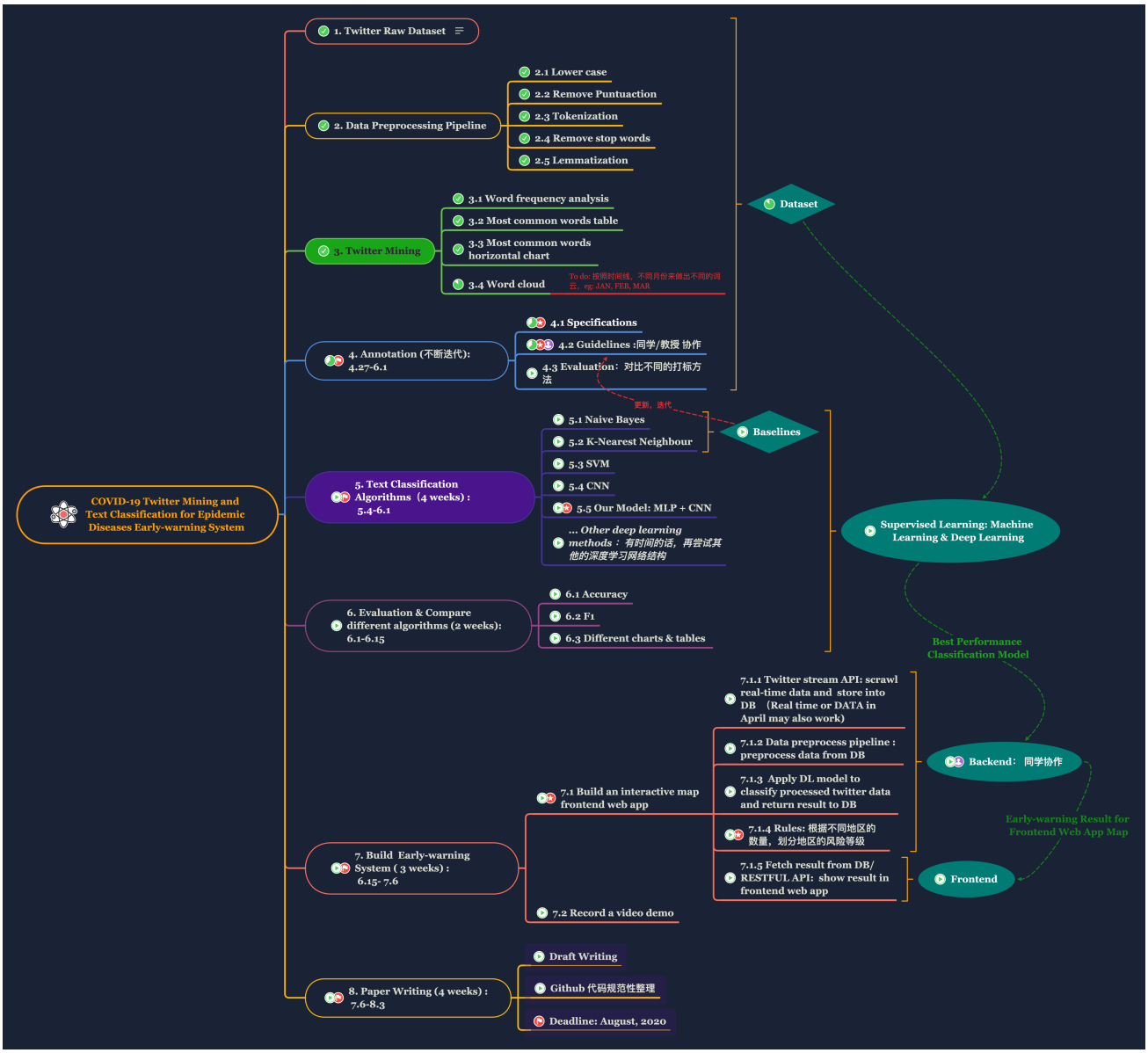
二、项目实施当前状态

项目进度实施情况：

目前在不断迭代和调整打标的 guidelines。

项目整体进度完成情况：

项目整体情况如下图所示，目前完成了 twitter raw dataset, data preprocessing 和 twitter mining，正在进行 annotation 和 text classification algorithms。



图一： 项目整体规划思维导图

三、本周成果

1. 针对上周提的建议以及本周的实验，对打标方案的 guidelines 进行了更新：
Guidelines:

- According to WHO official website [1] : COVID-19 Symptoms including *fever, dry cough, tiredness, aches, pains, nasal congestion, sore throat, diarrhea.* Hence the first guideline is to label tweet data ‘1’ if it describes such **symptoms** in the tweet. The describtion can be **formal** (include keywords) or very **colloquial**. For example:

    - ***“She looked up Due To Me Coughing ..My Girlfriend Didn't came colser to me...#coronavirus #COVID19india https://t.co/wO78BqaOY5”***
    - ***“I've been fighting a cold. Maybe I have the coronavirus. Lol”***
    - ***” with each passing day as i wake up more and more ill and my body a little more sore then it was the previous night, im beginning to accept that coronavirus may have just taken over me. “***
    - ***” Someone told me u normally dont run a high fever with pneumonia plus I've had a headache the whole time. She said it's possible i have it just tested too soon. My worry, if you had coronavirus how would u know since they obviously have no tests? They asked her if she cond/ “***
    - ***“Absolutely dying of a cold. I guess that’s what I get for cracking jokes about getting coronavirus”***
    - ***” @Boreeeeennnnnn I'm being cheeky and speculating. Not confirmed. But ironic how he has a "cold" after helping people with #coronavirus “***
    - ***“I think I’ve caught a cold, unless it’s Coronavirus... 🦠”***
    - ***“Okay, when I'm sick with the cold or #manflu my butt is in bed whilst I annoy my wife! Who's going to the mall, restaurants and other public spaces???? #CoronaOutbreak #coronavirus #Covid_19 #SanAntonioTX https://t.co/A5hGQAFS2P”***
    - ***“Me: *clears throat* Hot Cheeto Girl: You got dat coronavirus Me dying of a sore throat and allergies: no I just have a sore throat Hot Cheeto girl: sure you do”***
    - ***“So I’m just gonna be awake tho, cool. Btw, I don’t believe I have coronavirus, a mf sick but not dying and hella medicated.”***
    - ***“traveling in this day and age is tough. Got back from Vegas with a jetlag cold and was convinced I had coronavirus for 24 hours”***

- According to CDC official website [2], two kinds of tests are available for COVID-19 which are viral test and antibody test. The second guideline is to label tweet data '1' if the tweet indicates the user or someone in his community/city/state has been **tested positive** for either viral test or antibody test. Also, it should include someone who is **dead** from COVID-19. For example:

    - *"2 confirmed cases of Coronavirus in GA. Discovered just hours ago. With the new test kits now being used more widely case numbers are on the rise very quickly."*
    - *"#Telangana reports their first confirmed case of #coronavirus. https://t.co/vxEcwy2e7w #Sakal #SakalNews #viral #ViralNews #SakalMedia #news #coronavirusindia #CoronaVirusUpdate #CoronaVirusUpdates https://t.co/dIawdCqFSw"*
    - *"#NYC #CORONAVIRUSNYC 🧯 The wife, son, daughter and a neighbor of a #WestchesterCounty man who tested positive for the novel #coronavirus have also contracted the virus, according to New York Gov. Andrew #Cuomo. https://t.co/0Ae59rb8bT"*
    - *"Brazil🇧🇷 #CoronaVirusUpdates #Coronavirus positive test ⏩ 2 Suspected cases ⏩ 252 #coronavirusbrasil #healthcare #Coronavid19 #Pandemia #CDC"*
    - *"An additional seven cases of coronavirus have been diagnosed in the Republic o f Ireland, bringing the total number of cases to 13 here. There are now 16 cases in total on the island of Ireland #coronavirus "*
    - *"Music teacher tests positive for coronavirus https://t.co/kCIHl7q3xt via @MailO nline"*
    - *Covid-19: Dublin secondary school to close for two weeks after pupil confirmed as first case in Rep (via @thejournal_ie) https://t.co/uLYraQcQ0y*
    - *" Thoughts? San Francisco:: San Francisco Health Officials Confirm First Two Cases Of Coronavirus https://t.co/oXOVqm6kT7 #Coronavirus #Local #News @HainesF orSF 2020 "*
    - *"@MortalcoyleMi @HopefulGardenr Psst. We had our first death from coronavirus today in Washington state. Walls don't work against viruses, especially when the test kits don't work and there aren't enough masks and respirators."*
    - *"Not a hoax as #Trump said in SC rally. Illinois officials say patient has tested positive for #coronavirus https://t.co/lbgsYxqRo8"*
    - *"Confirmed cases of #Coronavirus in Africa. Please note data has been compiled by Africa CDC(Centre for Disease Control & Prevention). "https://t.co/cbsaSBxVGy"*

- The third guideline: Someone **may have COVID-19(suspect) or someone was infected or there is community spread**, but not corfirmed. For example:

    - *"So my friend just told me one of his tenants (not at his site) May have the coronavirus but hasn't been confirmed yet. "*
    - *"#BREAKING: @CDCgov is working to confirm another 'presumptive positive' #coronavirus case in #Florida. This would be the 3rd #COVID19 case in the state. 247 people are being monitored in FL right now #CoronaOutbreak #Coronavirusflorida @10NewsWTSP"*
    - *"The Coronavirus is in Lexington Kentucky... Someone pls help..."*
    - *" Coronavirus cases are spreading in Switzerland https://t.co/65NbugnyO3 "*
    - *"@mizzylizzy No...there is now community spread in Vancouver. https://t.co/7wz0kniB2n"*
    - *" @markp8888 @GBooth74 Sadly Coronavirus has even spread up here. Back to the Big Smoke on Sunday sadly😕 "*
    - *" guys whys toni girls family really have the coronavirus "*
    - *"Four coronavirus have been suspect in Odisha,Do not panic this extra forward message in social medias🙏Today when I woke up in the morning, I saw that I also had such forward messages, don't be afraid of them all🙏@Sasta24 @VertigoWarrior @Lochness4000 #coronavirusinindia"*
    - *" Continuing to monitor the situation at Life Care Center in Kirkland. I saw 2 people transported by ambulance around 4am. The center says if someone shows symptoms for #coronavirus they need to be transported to a hospital for testing. They can't test at the nursing home. #KING5 "*
    - *"Coronavirus: Italian-Nigerian footballer infected https://t.co/7rexGWibKN #Nigeriatunes"*
    - *" A letter sent to the @PlymouthSch community warns that a student who just got back from Italy last month was hospitalized with flu-like symptoms. We're tracking this potential case of #coronavirus on @boston25 at 5 and 6:30 https://t.co/yZAswzS2dd "*

2. 代码部分： 上周发现了一个问题，在数据集上有一些数据显示乱码，如下图所示：

RT @nypost: First case of coronavirus confirmed in Manhattan https://t.co/DHZ5uROtAa https://t.co/6NfATbDFT6
RT @thehowie: Thread: #COVID19 #Coronavirus updates & data.
RT @tonypierson2: How worried are you by the coronavirus??
RT @IbrahimLudwick: Somebody who got the coronavirus in China got a lung transplant to save their miserable life.
RT @AKasingye: When friends bump into each. Even #coronaVirus can,Äôt come into your way. @DoreenNasaasira. Photo credit: David Lubz @933kfm,Ä¶
#Covid19usa is highlighting problems in #health systems. | Carl Gibson https://t.co/zlWCiTL2tc
How can philanthropy plan for, respond to, and support communities affected by the #COVID-19 #Coronavirus? Join @funds4disaster this Thursday to learn ab
RT @Anna_Rothschild: To my friends who are reporting on coronavirus: are there advisories yet for people with severe asthma? I've checked a,Ä¶

图二： 数据集中有乱码

本周通过研究发现，这是因为 Twitter 公司对转推的 tweet 进行了 characters 的个数限制，转推的 tweet 最多有 140 个 characters，因此这些 tweets 是在抓取的时候就被截断了，不是乱码问题。结合本项目是为了研究用户本人是否有症状以及是否有附近的人为确诊和疑似，因此转推的信息用处不大，可以将转推的过滤掉以提升数据质量。这样这个问题并不会影响项目的数据质量。

按照这个思路，我们需要的数据是与 coronavirus 有关的，英文的，不是转推的，而且有地理位置的 tweet 数据。本项目原始数据集为 ***2576357 条 tweets***, 只选取英文且不是转推的 tweets 后为 ***295399 条 tweets***，再从其中选取有地理位置的，之后剩下的有效数据为 ***7068 条 tweets***。

目前已经使用如上 guidelines 完成了 **1000 多条 tweets** 的手动打标，目标是完成对 7068 条全部的数据打标，然后用 baseline 方法对其进行评估。

1.6 Drop retweets data

```
[94]: df = df[~df.full_text.str.startswith('RT')]

[60]: df.shape

[60]: (295399, 5)
```

图三： 过滤掉转推的 tweets

1.7 Drop data without location info

```
[95]: df = df[df['place'].notnull()]

[96]: df.shape

[96]: (7068, 5)
```

图四： 过滤掉没有地理位置的 tweets 只剩 7068 条有效 tweets

[159]:

	index	created_at	id_str	place	full_text	label
300	104119	Mon Mar 02 20:59:36 +0000 2020	1234584162411175938	{'id': 'e4c447e00985824a', 'url': 'https://api.twitter.com/1.1/geo/id/e4c447e00985824a.json', 'place_type': 'city', 'name': "O'Fallon", 'full_name': "O'Fallon, MO", 'country_code': 'US', 'country': 'United States', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[-90.772734, 38.71256], [-90.632554, 38.71256], [-90.632554, 38.846753], [-90.772734, 38.846753]]]}, 'attributes': {}}	Considering the grocery store chaos every time there is a forecast for 2" of snow in St. Louis, decided to do my #coronavirus panic shopping early 😇 #BePrepared #STL https://t.co/fdFPzwzHXz	0.0
301	104282	Sun Mar 01 20:36:36 +0000 2020	1234215986250674176	{'id': '0fa2e45c48f0ae2a', 'url': 'https://api.twitter.com/1.1/geo/id/0fa2e45c48f0ae2a.json', 'place_type': 'city', 'name': 'Loveland', 'full_name': 'Loveland, CO', 'country_code': 'US', 'country': 'United States', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[-105.176024, 40.3529092], [-104.973792, 40.3529092], [-104.973792, 40.4658383], [-105.176024, 40.4658383]]]}, 'attributes': {}}	Great 2 minute video to understand the coronavirus. https://t.co/1YUvXudr2S	0.0
302	104737	Mon Mar 02 22:03:05 +0000 2020	1234600138410164225	{'id': '3078869807f9dd36', 'url': 'https://api.twitter.com/1.1/geo/id/3078869807f9dd36.json', 'place_type': 'city', 'name': 'Berlin', 'full_name': 'Berlin, Germany', 'country_code': 'DE', 'country': 'Germany', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[13.088304, 52.338079], [13.760909, 52.338079], [13.760909, 52.675323], [13.088304, 52.675323]]]}, 'attributes': {}}	@ClaudiaZettel und beide haben Drachen 🚜 \n\nprediction: Coronavirus is blamed for disrupting negotiations, no-deal 2.0 arrives as 2021 does	0.0
303	104822	Tue Mar 03 01:02:40 +0000 2020	1234645334258835456	{'id': '6ffcf3b0b904bbcb', 'url': 'https://api.twitter.com/1.1/geo/id/6ffcf3b0b904bbcb.json', 'place_type': 'admin', 'name': 'Kentucky', 'full_name': 'Kentucky, USA', 'country_code': 'US', 'country': 'United States', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[-89.57151, 36.497129], [-81.964971, 36.497129], [-81.964971, 39.147359], [-89.57151, 39.147359]]]}, 'attributes': {}}	Fkn coronavirus 😡😡 messing up all that travel plans . I'm upset	0.0
304	105467	Wed Mar 04 10:59:18 +0000 2020	1235157867470045185	{'id': '624467bb324bd06d', 'url': 'https://api.twitter.com/1.1/geo/id/624467bb324bd06d.json', 'place_type': 'city', 'name': 'Firozabad', 'full_name': 'Firozabad, India', 'country_code': 'IN', 'country': 'India', 'contained_within': [], 'bounding_box': {'type': 'Polygon', 'coordinates': [[[78.299168, 26.983558], [78.720936, 26.983558], [78.720936, 27.509683], [78.299168, 27.509683]]]}, 'attributes': {}}	Precautions do not afraid of #coronavirus https://t.co/Nd8OZs8nvP	0.0

图五： 打标后的数据集格式，包含 tweet 的创建时间，tweet id 号，tweet 的地理位置信息，tweet 的全文以及我们人工 Label 的结果

四、下周计划

1. 继续对数据集进行标注以及迭代，然后不断更新打标的 guidelines。

2. 代码上实现机器学习 baseline 算法 NB, KNN, SVM。

Reference

[1] Q&A on coronaviruses (COVID-19). (2020). Retrieved 26 April 2020, from https://www.who.int/news-room/q-a-detail/q-a-coronaviruses

[2] Testing for COVID-19. (2020). Retrieved 29 April 2020, from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html