一、本周研究内容

研究内容：

1. 研究打标方案：调研了相关文献，针对本项目的 COVID19 的 Twitter 数据集提出自己的方案。

2. 实验代码进展：预处理部分更新，text-mining 部分更新。

二、项目实施当前状态

项目进度实施情况：

在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，提出了打标方案，以及对部分数据进行了打标。

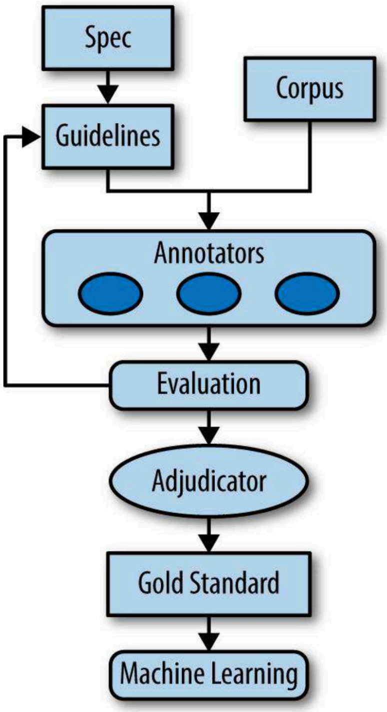
项目整体进度完成情况

大数据和自然语言处理部分：下周内容是写一个 baseline 方法，然后评估打标的质量。

三、本周成果

1. 打标的相关文献：

通过阅读教材“Natural Language Annotation for Machine Learning”[1] 学习了 NLP 数据的打标流程，高质量的打标是要通过不断的迭代以及调整 guidelines。



图一： NLP 数据打标流程

通过阅读文献 [2]，作者在文中建立了一个与本项目类似的系统，但使用 twitter 是来判别是否 health-related，作者手动打标是根据他们设定的 1026 个与健康有关的关键词，然后针对每个关键词手动标注 10 个有关的 tweets。

2. 针对本项目提出的打标方案：

Specifications: The goal of the project is to build a Epidemic Diseases Early-warning System through COVID-19 Twitter Mining and Text Classification. As many users use Twitter to talk about public heatlh topics and sometimes they also share information about the self-reporting of an illness (such as COVID19). We label the tweet data ‘1’ if the tweet can be treated as a COVID-19 self-reporting, otherwise we label ‘0’.

Guidelines:

- According to WHO official website [3] : 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 contains such keywords in health related context, but not news report.
- Other similar expressions about these symptoms.
- Twitter that indicated the user or her/his family member or friend has been tested positive.
- ...

3. 从打标数据中找到的一些例子：

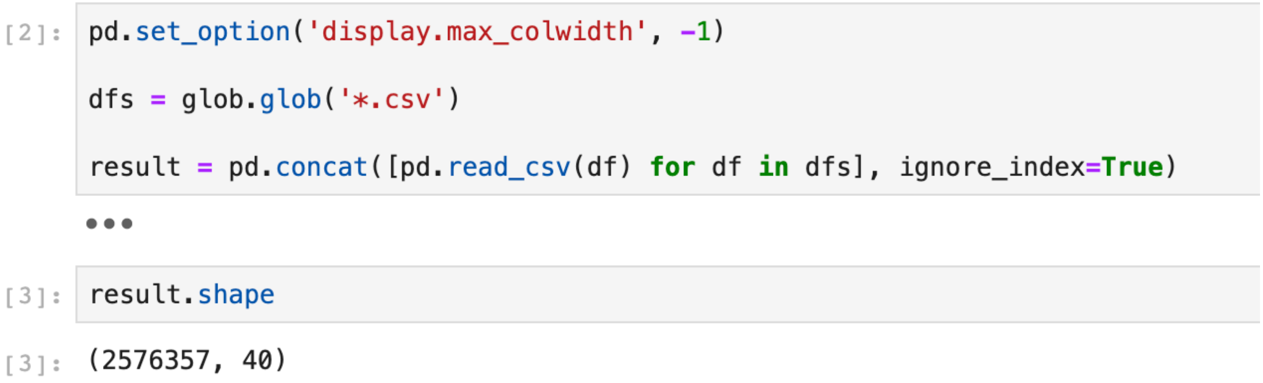
- RT @CriticalCezanne: He had fever of 39 degrees for 4 consecutive days so went to hospital to get medicine. However, hospital had no medici…
- Coronavirus outbreak: 3 still quarantined at hosp

4. 打标中要注意的问题以及局限和难点：

- 如果标注的噪声太大或者标签边界太过模糊（**大量标注错误，或标注规则写的太松、太模糊，**导致人都分不清某几个类别之间的区别），很可能再复杂的模型都在这份数据集上无法收敛。
- 如果标签与内容有非常直接的映射关系（**类别太过具体或标注规则写的太死**），例如只有出现某些关键词才可以贴为某个标签。则会导致一个非常简单的模型都会表现非常好，那这个模型学到的知识基本是没有什么实际意义的，没有了使用数据驱动的机器学习和深度学习的需要。

5. 代码进展：

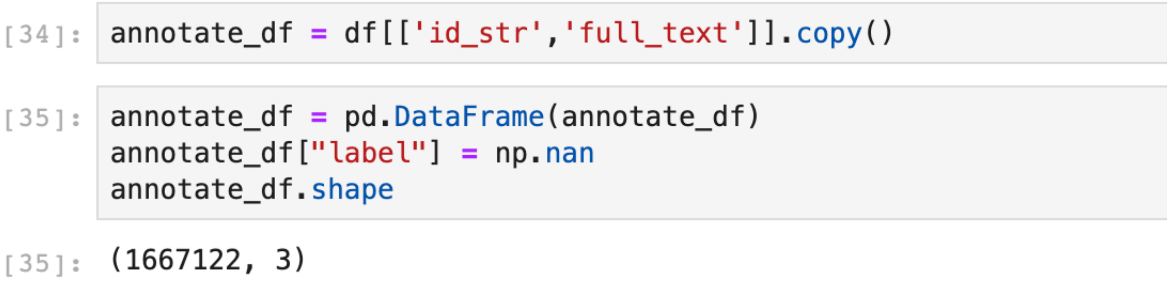
5.1 原始数据集一共有 **2576357 条** tweets。



```
[2]: pd.set_option('display.max_colwidth', -1)

     dfs = glob.glob('*.csv')

     result = pd.concat([pd.read_csv(df) for df in dfs], ignore_index=True)
     •••

[3]: result.shape

[3]: (2576357, 40)
```

图二： 原始数据集

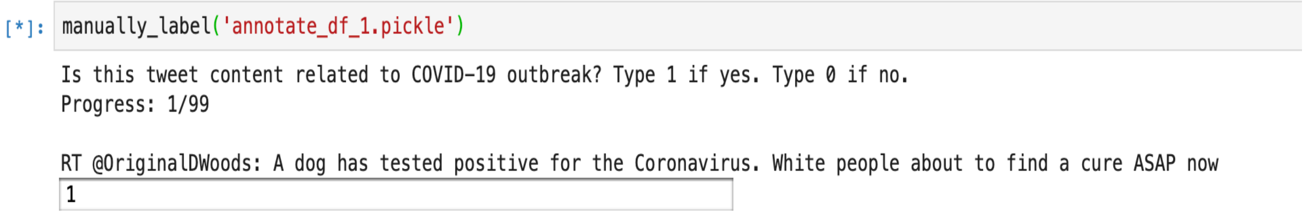
5.2 原始数据集中 twitter 为英文的一共有 **1667122 条** tweets。



```
[34]: annotate_df = df[['id_str','full_text']].copy()

[35]: annotate_df = pd.DataFrame(annotate_df)
      annotate_df["label"] = np.nan
      annotate_df.shape

[35]: (1667122, 3)
```

图三： 英文数据集

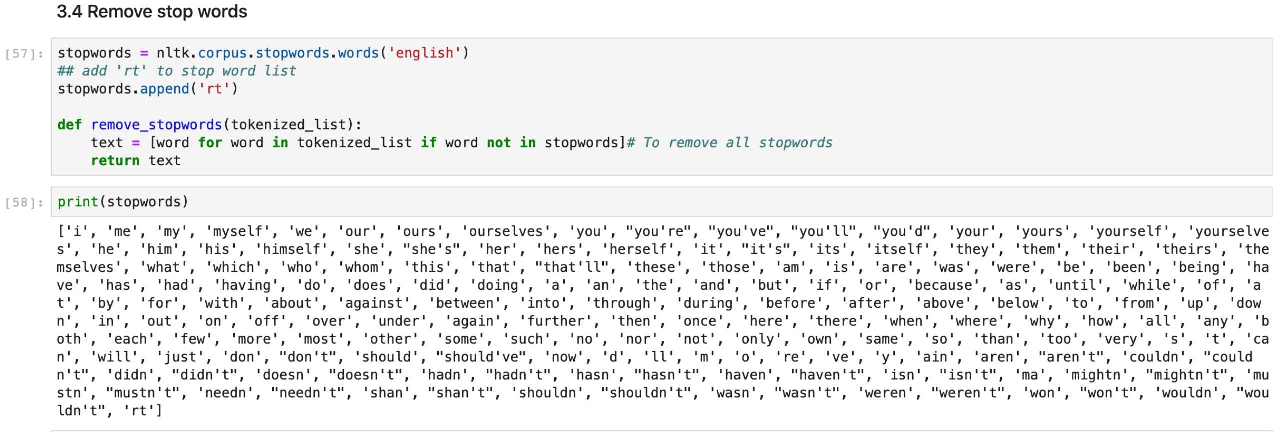
5.3 手动打标可交互代码实现。



```
[*]: manually_label('annotate_df_1.pickle')

     Is this tweet content related to COVID-19 outbreak? Type 1 if yes. Type 0 if no.
     Progress: 1/99

     RT @OriginalDWoods: A dog has tested positive for the Coronavirus. White people about to find a cure ASAP now
     1
```

图四： 手动打标

5.4 文本预处理中，通过分析文本，发现包含 ‘rt’ 比较多，在去停词 list 增加了 ‘rt’ 。

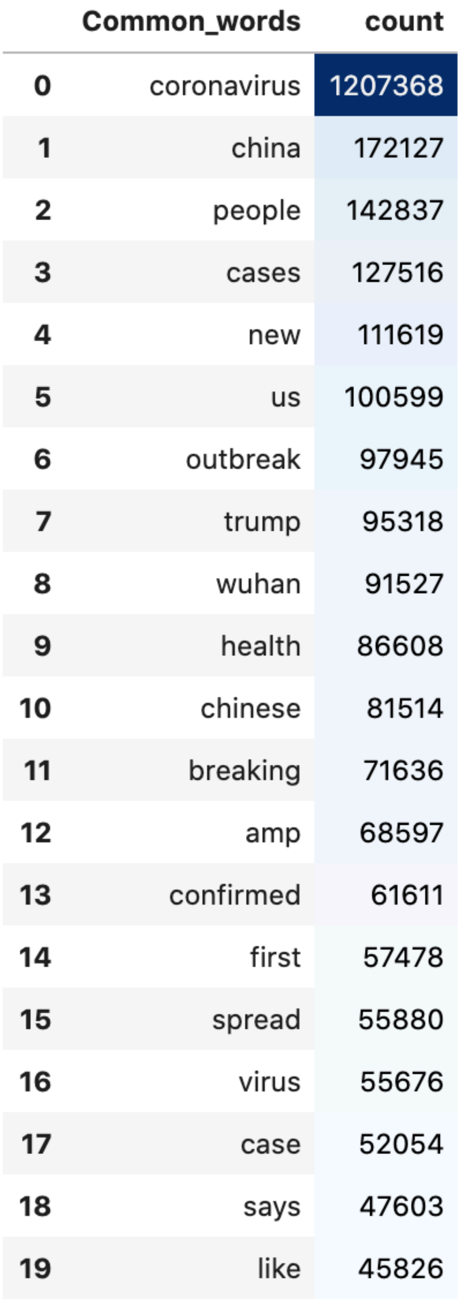


3.4 Remove stop words

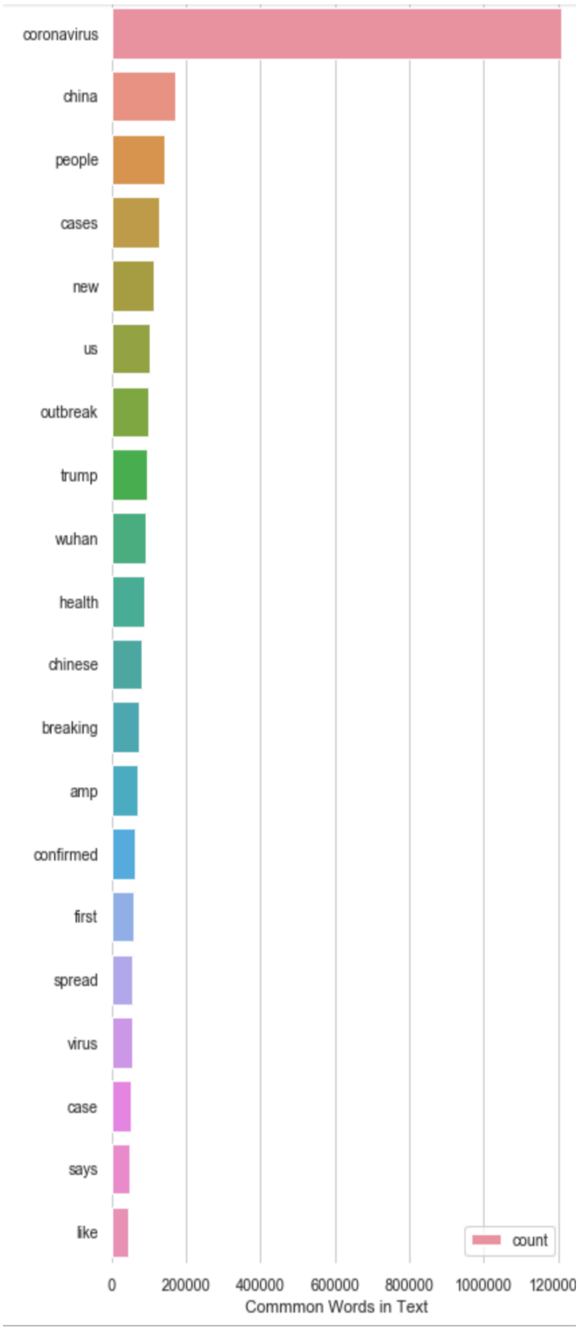
```
[57]: stopwords = nltk.corpus.stopwords.words('english')
      ## add 'rt' to stop word list
      stopwords.append('rt')

      def remove_stopwords(tokenized_list):
          text = [word for word in tokenized_list if word not in stopwords]# To remove all stopwords
          return text

[58]: print(stopwords)

      ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselve
      s', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'the
      mselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'ha
      ve', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'a
      t', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'dow
      n', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'b
      oth', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'ca
      n', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "could
      n't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mu
      stn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wou
      ldn't", 'rt']
```

图五： 去停词

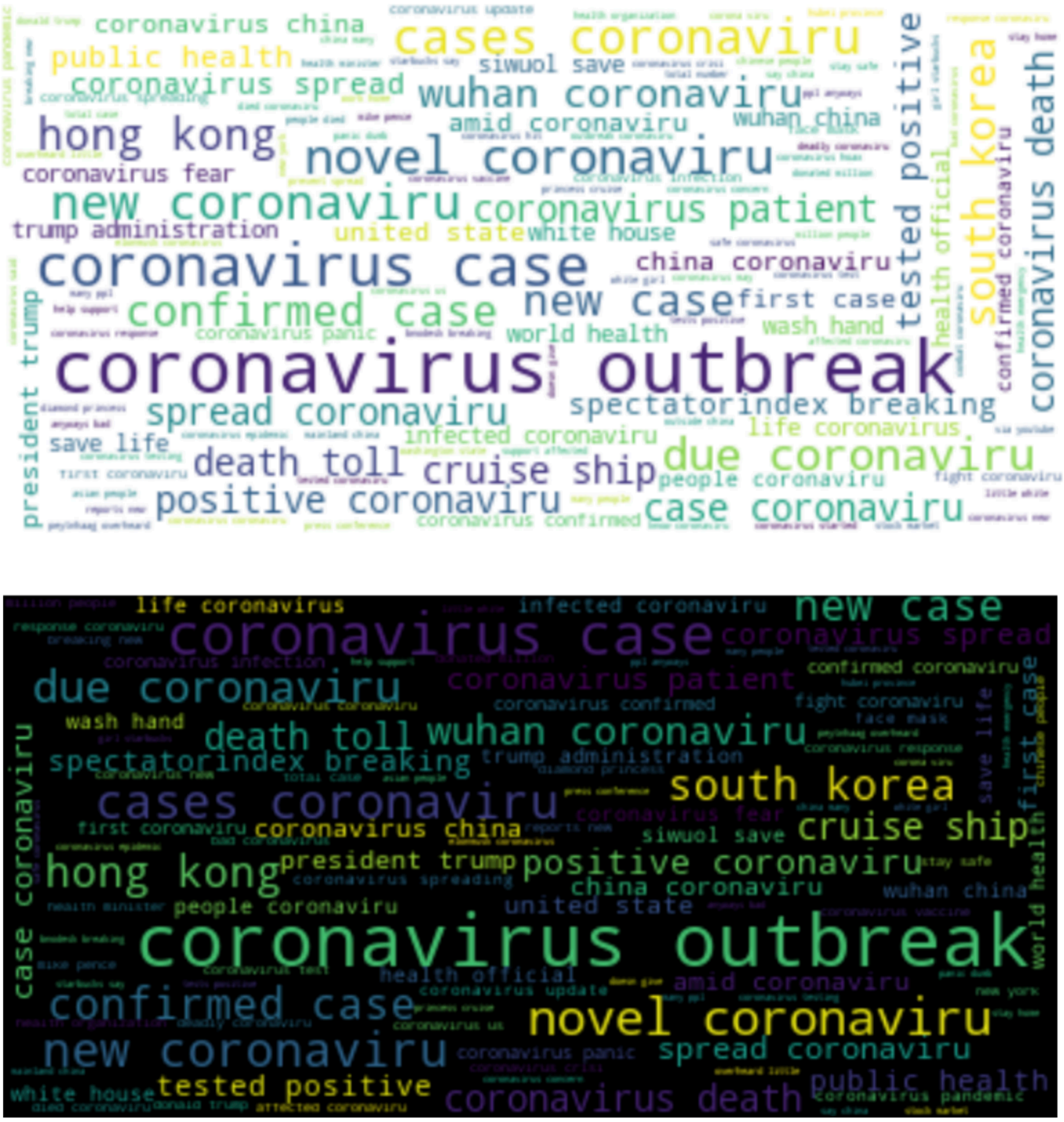
5.5 Tweet Text mining 部分，找出了在文本中最常见的 top20 词，绘制了 table, horizontal chart 和 word cloud。



	Common_words	count
0	coronavirus	1207368
1	china	172127
2	people	142837
3	cases	127516
4	new	111619
5	us	100599
6	outbreak	97945
7	trump	95318
8	wuhan	91527
9	health	86608
10	chinese	81514
11	breaking	71636
12	amp	68597
13	confirmed	61611
14	first	57478
15	spread	55880
16	virus	55676
17	case	52054
18	says	47603
19	like	45826

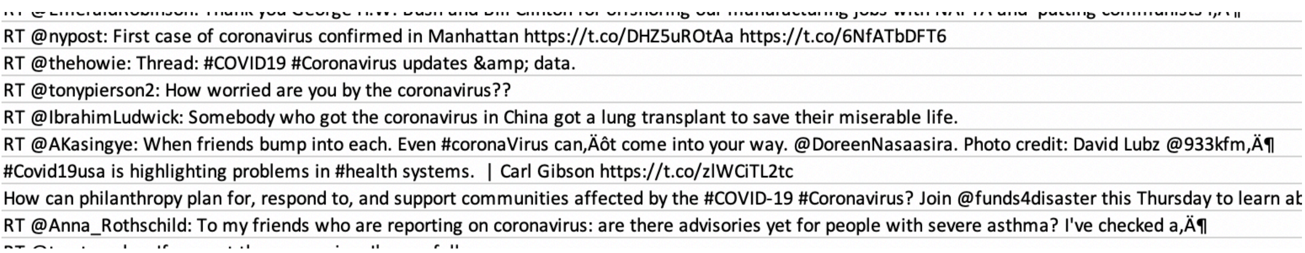


图六： COVID19 数据集中 Top 20 常见词



图七： 用 COVID19 数据集生成词云

6. 发现的问题：原始数据集中无法正确解码 emoji 的表情，造成乱码以及文本缺失。



RT @EmeraldRobinson: Thank you George H.W. Bush and Bill Clinton for offshoring our manufacturing jobs with NAFTA and 'putting communists i,Ä¶
RT @nypost: First case of coronavirus confirmed in Manhattan https://t.co/DHZ5uROtAa https://t.co/6NfATbDFT6
RT @thehowie: Thread: #COVID19 #Coronavirus updates &amp; 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¶

图八： Raw data 乱码问题

四、下周计划

1. 尝试解决发现的数据集无法正确解码的问题。

2. 尽快标注好数据集，使用小规模数据集或使用 semi-supervised learning 的方法。

3. 写一个 baseline 方法，然后对标注好的数据集进行迭代，测试在数据集上的表现，来评估打标的质量。

Reference

[1] Pustejovsky, J. (2013). *Natural Language Annotation for Machine Learning*. OREILLY.

[2] Șerban, O., Thapen, N., Maginnis, B., Hankin, C. and Foot, V., 2019. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. Information Processing & Management, 56(3), pp.1166-1184.

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