We build a score-based search criteria to generate this dataset of online tweets around sensitive covid-19-related content against the Asian population.

The search function looks for special keywords (listed below) in the tweets and assigns each tweet a score based on the number of keywords in it. For example, if there is a tweet that has 5 keywords in it, it will get a score of 5. and if there is only 1 of the keyword in a given tweet, it will get a score of 1. It is important to mention that duplicate keywords are counted only once; i.e., the score is based on unique distinct keywords. Also, only the tweets which contained one of the words {china|chinese|Asian|cpp} have been included in the query.

It's good to mention that before calculating the score for each tweet, both the tweet body and keywords have undergone preprocessing steps:

1. The keywords were lemmatized

2. Each tweet body was preprocessed for noise removal (see the code associated with this file for more details)

3. Each tweet body was lemmatized

Note:
1. Some keywords were treated as pairs of words such as "wet-market", "eat-animal", and "bio-weapon".
2. The code for extracting dataset version 2.0 is listed in the same directory with the name "extract_tweets.py"

Keywords

```
Cultural Stigmatization
==================
bat, animal, eating, eat, snake, dog, rat, soup, cat, wild, critter, swine,
pig, meat
market, wet
kill, killer, killed, killing
madeinchina
plague, fatality, threat
guilty
pay
boycott
```

```
Political Stigmatization
==================
madeinchina
lie, lied, liar, cover, covering, silenced, trust, trusted, cover-up,
coverup, cover, conceal, concealed, deception, transparency, hoax
biological, weapon, bio
war, warfare, biowarfare
lab, leaked, deliberately
stealing
dictator, tyranny
communist, posed, threat, fear
guilty
guard, army, military
murderer
boycott
imprisoned
sanction
genocide

Uncertain list:
crippling, died, censorship, freedom, fight, fighting, fuck, fucking, shit,
bastard, evil, sewage, racism, racist, suffering, blame, blamed, threaten
```

The final list of keywords used in code

```
bat
animal
eating
eat
snake
dog
rat
soup
cat
wild
critter
swine
pig
meat
wet
market
kill
killer
killed
killing
madeinchina
```

plague
fatality
threat
guilty
pay
boycott
lie
lied
liar
cover
covering
silenced
trust
trusted
coverup
conceal
concealed
deception
transparency
hoax
biological
weapon
bio
war
warfare
biowarfare
lab
leaked
deliberately
stealing
dictator
tyranny
communist
posed
threat
fear
guilty
guard
army
military
murderer
boycott
imprisoned
sanction
genocide
crippling
died

censorship
freedom
fight
fighting
fuck
fucking
shit
bastard
evil
sewage
racism
racist
suffering
blame
blamed
threaten