How to select data

Data:

\* What kind of data do you need to answer these questions? For what time period? From how many sources? How will you get it?

\* Are sources representative? Are they disproportionately of one form? Are all relevant time windows covered? Does the data represent all relevant groups, including those often marginalized?

\* If you are filtering, subsampling, or selecting from the original data, is the remaining subset representative? Can you describe how selective removal alters the data and the interpretation of the data? Are you losing anything that might be valuable at a later stage?

Yi Ting:

errors

df2['year']=df2['date'][0:4]

Haodong:

How to make it work with Chinese corpus

https://github.com/jsrpy/Chinese-NLP-Jieba/blob/master/jieba\_intro.ipynb

Target-level sentiment analysis

time-range of filtering

challenge: the number of sentimentw ords is fews

try out: dependency parses

last stuff in the week 4 02 notebok

use something more than sentiment to identify the AFFECT

bigger dictionaries ANEW

https://dr.ntu.edu.sg/bitstream/10356/145624/2/The%20annotated%20Lexicon%20of%20Chinese%20emotion%20words.pdf

action item for kokil: find more target-side sentiment analysis papers

Yifei

how to read more usernames

how to get more information

**Data concerns**

1. Why do you select these sources and not any other sources?

Cleaning data:

* good idea to first filter by keyword,
* then do topic modeling,
* and then look at sentiment

Suppose your question is about representative opinion: nationally representative:

* Think about creating a nationally or internationally representative sample.
  + Look at the location of social media users. And randomly select one person from each country, one person from each state in your country, one person from each district in your country……
    - Missing data
      * Still good for inter-group comparison
      * **Chapter 1-2 of Bit by Bit.**
    - Only 25% of people put geotag on their tweets
    - Using this information we can randomly sample, so that I have at least 30 people form each US county: <https://www.sciencedirect.com/science/article/abs/pii/S0747563220302958>
    - I did a similar thing with China data: <https://arxiv.org/pdf/2110.15726.pdf>
  + Probability-based sample. Run a survey on a nationally representative panel. Get their social media handles

Suppose your question is about popular opinion (a lot of people agree with it or talk about it)

* No sampling required. Just collect all the data that is, and say what is emerging as most popular. Of course, there might be biases, like some people may not talk, but to identify popular opinion approach is fine.

Suppose your question is about alternative opinions or minority opinions. Good coverage and diversity, not representative

* Also look at the edge cases. Suppose there are 50 edge cases, of which 10 are saying somewhat the same thing.
  + By grouping them, it starts to be less of an edge case and more of a minority opinion. Maybe you can group them to make them more meaningful.
  + Eg a lot of people are showing distrust in different ways, you can group them to a meaningful category maybe they become 5% of data.
* Undersample the popular opinion, oversample the alternative opinion