CS410: Text Information Systems
Project Progress Report

Name: Chengyuan Ai (*Captain*), Linhan Yang, Yixuan Zhang, Zhonghua Zheng
NetID: ai13**,** linhany2, yixuan21, zzheng25
GitHub Username: acy925, Jupiter1995, yixuan21, zzuiuc
GitHub Link: https://github.com/acy925/CS410CourseProject
Guideline: Theme 5: Free Topics

As we described in our project proposal, we have five tasks for this project, including "1 - data sources searching and evaluation," "2 - data cleaning, exploration, and evaluation; data curation," "3 - topic model (clustering) for heat waves-related text," "4 - sentiment analysis," and "5 - organization and integration."

1) Which tasks have been completed?

We have completed "1 - data sources searching and evaluation" and part of "2 - data cleaning, exploration, and evaluation." First, we applied for and were granted Academic Research access to the Twitter API, which provides us with a means of accessing tweets.

Next, the data collection script was done. Using this script, we can quickly modify the query and retrieve any Twitter data we need. We have already retrieved the first batch set of data using the hashtag "#heatwaves" (about 15,000 tweets) and cleaned them.

For the data cleaning and data curation, all the normal punctuations, web page addresses, newline characters, and extra spaces are removed. However, these are not the final results/formats. If any requirements are uncovered during the analysis step, we will likely return to the cleaning/curation stage.

2) Which tasks are pending?

Tasks 3 - 5 are pending. However, we have begun our literature reviews of the various algorithms and Python packages that will be utilized.

3) Are you facing any challenges?

Thus far, we have faced a number of challenges, and likely there are more challenges for the next steps.

First, it is not easy to decide what should be included in the query when retrieving tweets. There are numerous options available in the Twitter official API, but we must experiment with each of them to determine the type of data we require. Finding relevant keywords is one of the most painful steps, because using simple keywords increases the likelihood that the final result will contain numerous false positives.

After many attempts, we decided to design a query based mostly on Twitter hashtag(s) and time period, as hashtags may practically ensure results that are relevant to our interests if the correct hashtag is used. However, the use of hashtags may cause us to miss a large number of tweets.

It is difficult to decide what actions to take during the data cleaning/curation process. Obviously, web page addresses should be eliminated from the text, as they contribute nothing to our research; nevertheless, we must also check additional aspects such as emojis and usernames. We've decided to maintain the majority of things as-is for the time being, as we can always remove them quickly, and we're trying to find something interesting to do with them.

Now we are facing challenges with the selection of algorithms and Python packages for the remaining tasks.