CS410: Text Information Systems
Final Project Proposal

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

Introduction

What is your free topic?

Our free topic is "Mining and analyzing heat waves-related topics using news or social media."

Please give a detailed description. What is the task? Why is it important or interesting?

As suggested by the name, the tasks consist of (1) locating topics related to "heat waves" via news or social media (depending on their accessibility), and (2) performing sentiment analysis.

This study is important and interesting because heat waves (extremely high-temperature events) are among the most damaging climate extremes to human and natural systems on a global scale. It is likely that heat wave events will become more intense, more frequent, and last longer, particularly in urban areas and in the context of climate change. Understanding the relevant topics and public opinion can help us better prepare for heat waves. Our research will benefit from text data.

What is your planned approach? What tools, systems or datasets are involved?

Approach: Our approach includes text retrieval, data cleaning, and text mining/analysis. Text retrieval is concerned with retrieving text data such as news or social media (e.g., tweets), whereas text mining and analysis focus on identifying relevant topics and performing sentiment analysis.

Tools and datasets: We will utilize the Twitter API (subject to approval) to access text data and other Python packages (such as TextBlob) or other tools for text processing and analysis.

What is the expected outcome? How are you going to evaluate your work?

The expected outcome will be topics (relevant to heat waves) and opinions from the news or social media (e.g., tweets). Given the nature that Natural Language Processing is challenging to evaluate, here we will evaluate our work based on common sense (e.g., is it reasonable?) and some domain expertise.

Programming Language

Which programming language do you plan to use?

The majority of this project will be written in Python. It is possible, however, that we will use additional programming languages to complement Python for various tasks.

Workload Justification

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

We have N=4, corresponding to at least 80 hours (at least 20 + 20 + 20 + 20 + 10 = 90 hours). The appendix provides additional information regarding the justification.

Appendix: DETAILS OF THE JUSTIFICATION

Data sources searching and evaluation [at least 20 hours].

As this will be our initial step in tackling the problem, we intend to retrieve relevant text data using Twitter API (or other information sources and methods). There will be some waiting time because some of them require permission/API keys. After the application(s) is/are approved, we will need to learn and test the query engine and our own queries, as well as finalize the queries we will use to obtain all the required data for our next steps.

[Alternative/Optional] Web crawler development, testing, and evaluation [at least 20 hours]

In case anything related to the data source doesn't work the way we expected, we will have to develop our own web crawler to retrieve the necessary information. The benefit of having our own crawler is that we can specify the information and keywords to be crawled, but the crawling process could be challenging, especially for websites with anti-crawler functions. During this phase, additional adjustments and tests will be conducted to ensure that we have the necessary data and are not banned from the source.

Data cleaning, exploration, and evaluation; Data curation [at least 20 hours]

This would be the most challenging step. In this step, we will look deeper into our data set to determine what we should keep and what should be removed. We also need to make all the data into a unified, expected format. Then, we will evaluate our data to determine the next step. If there are insufficient data after the cleaning step, we must return to the first two steps.

If the data size is acceptable, we will begin general exploration to familiarize ourselves with the data before feeding it into models.

Eventually, we will be required to prepare our data for our models and subsequent analyses. However, preparation could occur concurrently with model development and data analysis; as a result, the data will better meet our requirements.

Topic model (clustering) for heat waves-related text [at least 20 hours]

In accordance with the initial steps presented, we will use text clustering methods to perform Topic modeling in order to identify topics related to the main topic "heat waves." In this step, we may employ Hierarchical Agglomerative Clustering (HAC) or K-means, among other Similarity-based Clustering techniques (we have not yet decided). Then, we will conduct direct and indirect evaluations of clustering to compare the performance of various clustering methods.

Eventually, with the optimal clustering strategy, we can achieve our objective and identify the topics that are most pertinent to the main topic of "heat waves."

Sentiment analysis [at least 20 hours]

The purpose of this step is to conduct a sentiment analysis in order to gain insight into people's opinions, reactions, and feelings in social media or news about the heat waves (time periods TBD, depending on data availability). Due to the complexity and difficulty of implementing sentiment analysis, we will likely devote a considerable amount of time to testing and assessing the available algorithms and/or tools. Throughout the time period, we will begin with *graded sentiment analysis* and, if possible (depending on our progress), then dive deeper into *emotional sentiment analysis*. If the first two analyses go smoothly, we may attempt more complex analyses, such as *aspect-based* and *multilingual sentiment analysis*. If necessary, we will need to spend time optimizing the algorithms in order to speed up the execution process.

Organization and integration [at least 10 hours]

Before submitting our work, this is our final step to clean everything up and prepare a clean package. This includes but is not limited to documenting/organizing source codes.