Felipe Alamos
Bhargavi Ganesh
11/05/19

UML Project: Preliminary Results

Introduction

Our project attempts to answer the following question: Is Twitter data a reliable measure of public opinion? To that end, we have decided to compare the topics that emerge from Twitter data to responses to a global survey. Our project uses the issue of climate change as a case study for investigating this question.

The motivation for our project stems from the growing desire of social scientists to use Twitter data to understand public opinion, despite limited validation of the results. Twitter data is popular largely because it is a large data source, with a much larger sample size than any poll or survey. However, given that writing tweets is a choice, it is reasonable to expect that the population of Twitter users may be skewed on demographic characteristics. In fact, a survey on U.S. Twitter users conducted by the Pew Research Center suggested that U.S. Twitter users tended to be on average younger than the general public. They also note that approximately 10% of Twitter users in their sample were responsible for 80% of the tweets produced. Critics of using Twitter data to measure public opinion may note that twitter data mainly consists of the opinions of a smaller group of impassioned Twitter users.

In order to put these critiques to the test, we decided to use traditional surveys to validate the results of topic modelling on Twitter data. Traditional surveys typically have rigorous survey sampling techniques which ensure that the data is a reflection of the population at large. Therefore, we make the assumption that survey data is a proper source to use to validate our results.

Data Collection

Twitter Data

We have collected Twitter data using the python library GetOldTweets3. The Python library scrapes Twitter's advanced search mechanism and allows us to specify the countries that we would like to include, the keywords to search on, the timeframe we are interested in, and the radius we would like to search within[1]. We chose to use this Python library because the official

[1] Scraping Twitter data with this kind of queries is legal according to Twitter terms of use.

Twitter API only allows developers to download tweets that were posted 5-7 days prior to accessing the API. We decided that it would be more interesting to see the trends in topic modeling over time, and have the ability to match up the timeframe of the tweets and the date range in which the survey results were conducted.

We created a config file to store our inputs for the various different Twitter search fields mentioned above. An extract of our config file can be seen below:

```
countries: ["england"]
keywords: ["climate change"]
since_date: "2010-01-01"
until_date: "2011-12-31"
within_radius: "1000mi"
```

To start, we searched for the keyword “climate change” in England, within the timeframe of 2010 and 2011, and a radius of 1000 miles. We specified a date range between 2010 and 2011 because the survey questions were asked in that timeframe. We pulled a total of 100 tweets for this first iteration of results.

We have done basic pre-processing of the tweets, including removing stop words, extra characters, etc. At the moment, we have also removed the keyword from our results in order to focus on the subtopics and not the topic we searched for. We may have to do some more data cleaning in the future. For the survey data, we have not done any particular data cleaning yet but may do so in the future.

Survey Data

For the survey data, we have chosen to use the 2010 International Social Survey Programme’s survey module on the environment. The survey contains 50,437 observations. The countries included in the survey are Argentina, Australia, Austria, Belgium, Bulgaria, Canada, Chile, Taiwan, Croatia, Czech Republic, Denmark, Finland, France, Germany, Iceland, Israel, Japan, Korea, Latvia, Lithuania, Mexico, Netherlands, New Zealand, Norway, Philippines, Portugal, Russia, Slovak Republic, Slovenia, South Africa, Spain, Sweden, Switzerland, Turkey, Great Britain, and the United States.

The survey asks a series of detailed questions on attitudes towards environmental issues. Two examples of survey questions are below:

1) Here is a list of some different environmental problems. Which problem, if any, do you think is the most important for your country as a whole?
2) Which problem, if any, affects you and your family the most?

The potential survey responses to these questions are as follows: air pollution, chemicals and pesticides, water shortage, water pollution, nuclear waste, domestic waste disposal, climate change, and genetically modified foods, and other. The survey respondent is asked to rank each of these issues. Additionally, the surveys contain specific questions on climate change, asking if the rising temperatures due to climate change are extremely dangerous, very dangerous, a little dangerous, or not dangerous.

In our analysis, we hope to design a reasonable methodology of comparing the subtopics that show up in our twitter results with the topic rankings present in the survey. We are considering two broad approaches to this issue, which are discussed in the “Future Analysis Plan" section, under “Comparison of Data Sources”.

Topic Modelling

The modeling technique we chose to use for the topic modelling of tweets was Latent Dirichlet Allocation, also known as LDA.

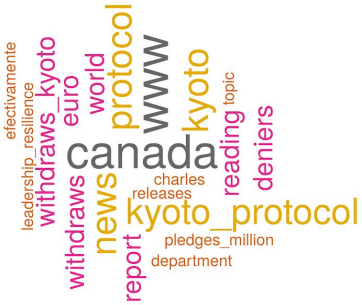
To generate the topics, we took the following steps:

- We created a document term matrix with our pre-processed words, which consists of tweets in rows and terms in columns, where each cell shows a term from each tweet.
- We explored term frequency and document frequency for each term.
- We fit an LDA model to our data. For this, we explicitly determined how many topics should be selected, so we ran it 20 times, with k= number of topics = 1...20.
- For each of the calculated models, we computed the coherence score, so as to choose the best number of topics from 1 to k. The coherence score is a numerical representation of whether words in a given topic make sense when they are put together. Hence, the coherence score gives us the quality of the topics being produced.
- Based on the previous step, we chose a number of topics k such that we maximize coherence. In our case, k=6.
- We finally created word cloud visualizations for each of the topics.

A visual representation of the subtopics that emerged can be seen below (and also at this link):

Topic 1:

Topic 2:



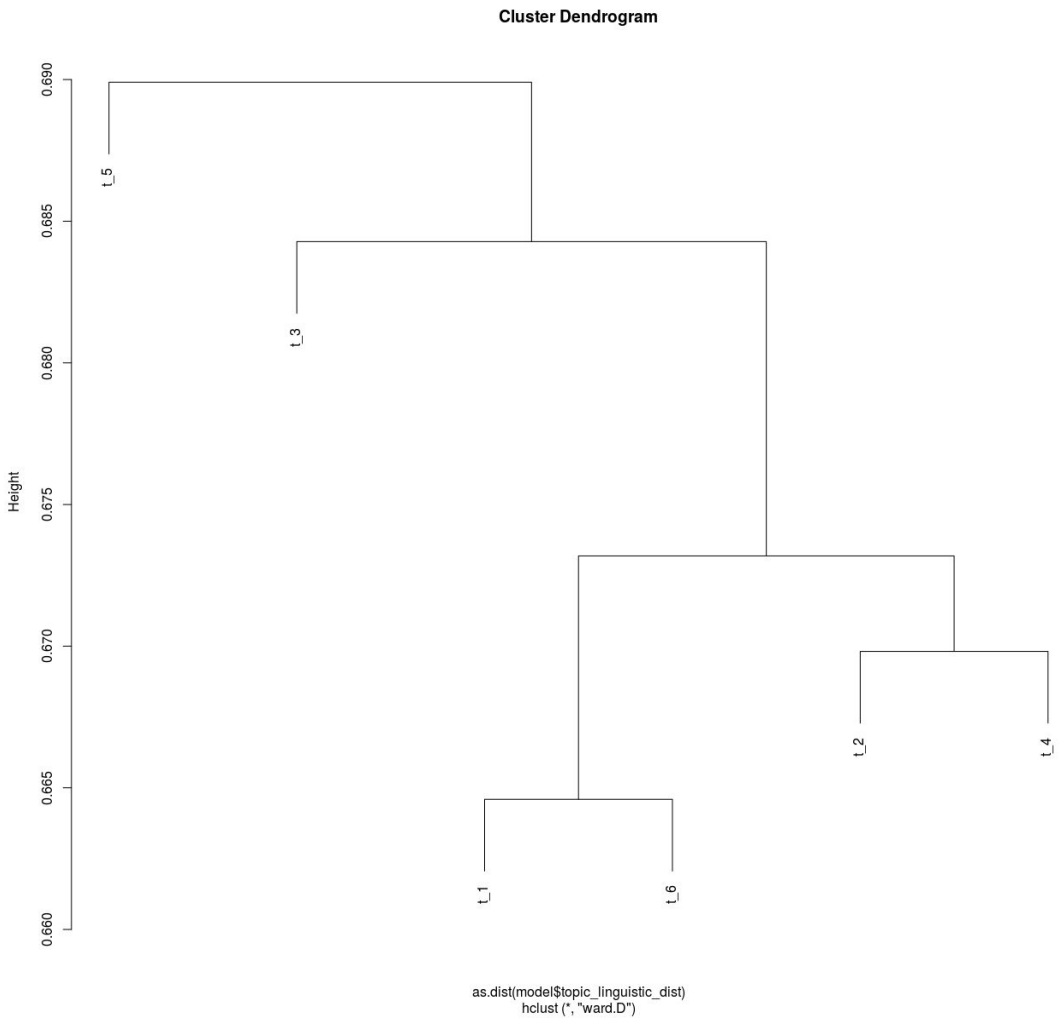
Topic 3:

Topic 4:

Topic 5:

Topic 6:

It is also interesting to study how similar or close to each other the topics are. For this purpose we can use the Hellinger distance, which computes the distance between 2 probability vectors.The dendrogram below shows the similarity or proximity between different topics, and we can observe, for example, that topic 5 is the most different one to the others, and that 1 and 6 are the closest/most similar to one another.



The subtopics that emerged from our first run of the model are somewhat interesting, particularly as we see that Canada withdrawing from the Kyoto protocol was relevant at the time. It is important to note, however, that these results do not provide a direct connection to the survey results. We discuss this in detail below.

Future Analysis Plan

Complete Generation of Data

We plan to complete our data generation process based on the code we have written to extract tweets. The following steps need to be taken in order to complete this process.

Select Countries

We are still considering which countries we would like to compare. Our survey data source has a wide range of countries included. We had considered starting with the U.S., France, England, and Germany, and then expanding from there. We selected these countries on the basis that they were the countries most pivotal in the process of the UN Paris Climate Accord talks, and therefore have an obvious vested interest in the topic.

Determine Sample Size of Tweets

We are still deciding how many tweets to pull for each of our subsamples. We would like to pull as many as we can computationally manage but we are not sure what that number is quite yet.

Complete Data Cleaning

We expect to continue the data cleaning process to understand if there are any words creating noise in our twitter data, and whether we need to remove those. We may also have to remove accounts that are bots, or do not contain meaningful text/dialogue.

Comparison of Data Sources

The main point we need to tackle in the rest of our analysis is determining a sound methodology for linking the survey results with the twitter topic models we have created. More broadly, we are considering the following two options:

Option 1

Using multidimensional scaling to build an underlying dimension of “climate preferences”

Under this option, we can make use of the specific questions within the survey which measure sentiments towards climate change. By scaling our twitter data, we would be able to compare extreme preference choice options in the survey to a scale of twitter topics that have emerged from the LDA topic modelling process.

Option 2

Using EDA and unsupervised techniques (like clustering and PCA) to compare general trends in frequencies of words used.

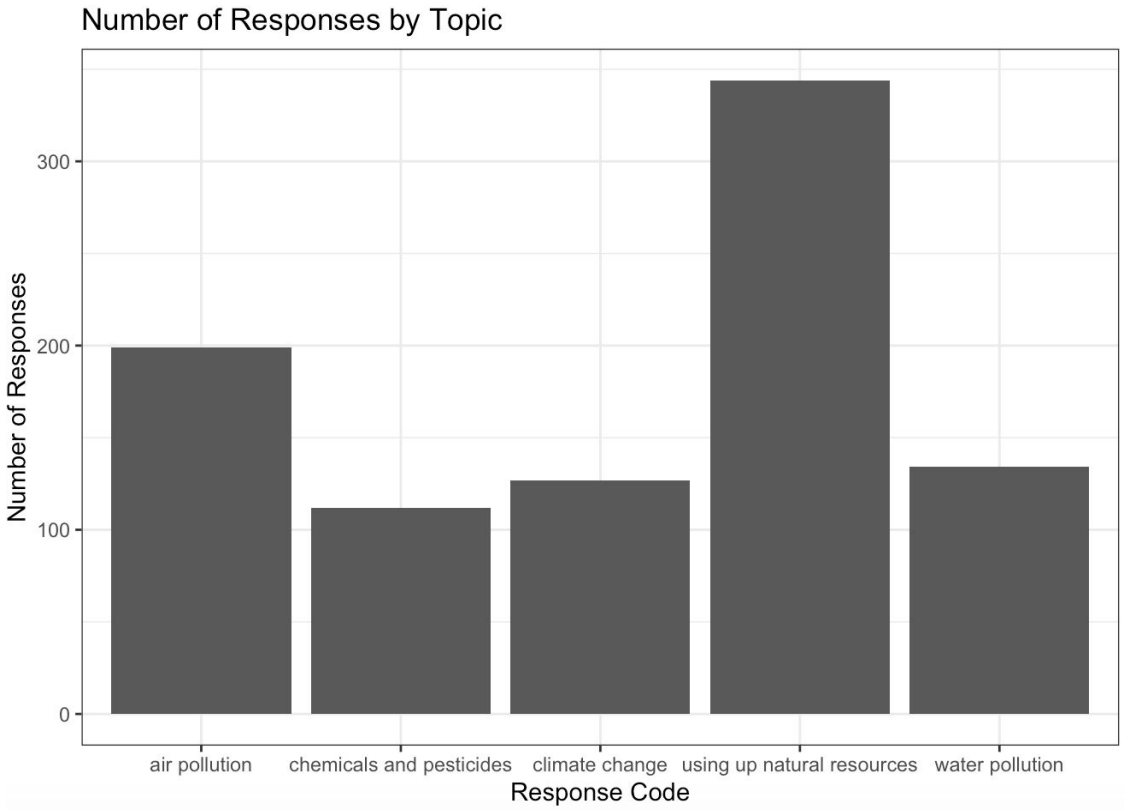
If we use this method, we may choose to use the survey questions asking respondents to rank the environmental issues that are the most important in their country, and compare these responses to our twitter topic models. This approach would require us to possibly widen our Twitter keyword search to environmental problems at large, to more directly compare ranking of topics in the surveys to the subtopics that show up most often in our topic models. One thing we

may have to consider is how to evaluate answers to the survey that did not fit into the specific topics asked.

Below, as a proof of concept, we conducted a preliminary EDA on the ranked survey questions in the United States.

Preliminary EDA on Survey

As a proof of concept, we did some simple exploratory data analysis on the survey to see which topic came up as the most important environmental problem. The results below represent the top topics selected by U.S. respondents.



As we see above, in the U.S., respondents were most concerned about using up natural resources, air pollution, and water pollution, and climate change was just the 4th most popular topic. Note: the above topics are just the top 5 topics (with most respondents).

References

We used this example as a general guideline to conduct topic modelling on twitter data.