Narrative

Wen Li Teng

12/4/2019

Introduction

The *Foreign Relations of the United States* (FRUS) is the official documentary historical record of major U.S. foreign policy decisions and significant diplomatic activity. The volumes of FRUS contain documents from Presidential libraries, Departments of State and Defense, National Security Council, Central Intelligence Agency, Agency for International Development, and other foreign affairs agencies as well as the private papers of individuals involved in formulating U.S. foreign policy. This project webscrapes all documents concerning Indonesia from FRUS.

Main Challenges

The main challenges of the project were to (1) webscrape the collections into a dataset; (2) clean and geocode the dataset; (3) visualize the data.

Solutions

1. Webscraping the Dataset

The webscraping process was broken into two parts: (1) collecting the links from the pages of search results and (2) collecting the metadata from each document associated with each link. Both parts of the webscraping relied on the packages tidyverse, stringr, rvest, lubridate, and purrr, as well as the browser plugin CSS Selector Gadget.

The first part of the webscraping process was straightforward. The Office of the Historian, fortunately, uses a consistent format for the pages of its search results for a query such as "Indonesia", as input into its search engine. A vector of these search pages was created using a for loop and a function to modify each URL as a string. Then, each page was scraped for the document links within it.

The second part of the webscraping process was time-consuming. Given that there were around 4,800 documents, the webscraping strained the capacities of the computer and internet connection on which it was run. An attempt to scrape the actual texts of the documents was abandoned after it caused repeated problems with the connection to Office of the Historian website. At a rate of just under 3 documents per second, the metadata (title, location, and date) were finally collected after half an hour.

2. Cleaning and Geocoding the Dataset

Given that there was substantial variation in how the metadata were formatted, the dataset had to be cleaned before it could be used for any plotting. A function was created to format the titles, most of which had random blank spaces or line breaks in them. This function was then mapped to the dataframe of links. A similar function was created to format the locations, which also had similar issues. The dates were the least easy to clean. In addition to random blank spaces or line breaks, some dates were missing months or days, or had more than one day within the same date. As it would have been inefficient to go through all 4,800 rows to identify the variations, an effort was made to work with the rare similarities shared by the

dates. In general, a month-day-year format was followed. Months were generally spelt in full. As such, days, months, and years were extracted and then recombined as a string, free of the clutter of the original data. The cleaning process resulted in a dataset of titles, locations, and dates.

This dataset was then fed into the geocoding process. The geocoding process required learning about a new package, tidygeocoder. This package contains a function (geo_cascade) that accepts a string of characters and returns the latitude, longitude, and "geo_method" of a given location. However, geo_cascade requires precise and modern-day spellings of locations. It would not accept misspellings such as "Toyko" for "Tokyo." These misspellings were rife throughout the data. As such, the unique() function was used to obtain a vector of every location, spelled correctly or otherwise, from the data. These strings had to be manually replaced with the updated spellings. A function was created to process the output of geo_cascade and prepare it for mapping. Tidygeocoder was nearly as slow as the webscraping process to return the coordinates of each location. As such, the vector of unique locations (which was much shorter than the full dataset) was used to obtain the coordinates. These were then joined onto the full dataset. In preparation for mapping the data, the locations were sorted roughly into five geographical regions: Africa, Asia and the Pacific, Europe, North America, and South America. The approximate coordinate boundaries of these regions, as they appear on a map, were used to sort the locations into the various regions. These regions were then appended to the dataframe, bringing the dataframe's columns to title, location, date, lat, long, and region.

A short script was written to add a column of decades (e.g. "1940" for the year "1945"). Attempts to achieve this using lubridate were unsuccessful. A simple function ("floor_decade") was written to mimic lubridate's floor_date function.

3. Visualizing the Data

This project aimed to visualize the data in three ways: (1) a line graph; (2) a series of bar graphs; (3) a series of maps.

The plotting of the line graph was the most easy of the three. This merely required the creation of a dataframe containing a count of the total number of articles as grouped by year. Accordingly, years were plotted on the x axis and the number of documents, on the y axis. Aesthetic modifications were minimal in this graph.

The plotting of the series of bar graphs proved more challenging. The creation of a series of graphs over decades necessitated the use of a for loop (for each decade). A choice was made to create separate data objects to avoid cluttering the loop. These data objects were namely a count of (1) the number of articles per decade per region and (2) the number of documents per decade per location. Further objects were created to find (3) the top 5 locations per decade and (4) top 10 locations overall. The objects were then called upon in the loop to produce the series of bar graphs. Aesthetic modifications included adding on labels showing the absolute number of articles, as well as rotating the x-axis labels to make them more readable.

The production of the maps was the most difficult task of this project. Much time had to be spent to learn about four new packages: (1) rnaturalearth, (2) rnaturalearthdata, (3) sf, and (4) viridis. The first three packages were chosen as an alternative to ggmap. Though ggmap would have been much easier to use with ggplot2, ggmap requires the creation of an API with Google and the provision of a debit or credit card account (with possible billing after the end of a trial period). The three packages above do not have such requirements. An online tutorial (see credits) was especially helpful for learning how to use the four packages in conjunction. (1) rnaturalearth and (2) rnaturalearthdata supplied the geometric information that outlined each country on the map. (3) sf was necessary to encode the spatial data. (4) viridis offered color palettes for the continuous variable - the percentages of documents from a region out of a given decade - to be used in the maps. Of these, (1) rnaturalearth and (2) rnaturalearthdata were the most hard to understand. The tutorial showed the creation of a large dataset ("world") from both packages. However, such a dataset was unwieldly for any binding, joining, mapping, or looping operations. It caused significant lags. A decision was made to create a subsetted version of the raw data ("mini_world"), which was cleaned to standardize its elements with the FRUS dataset. A three-part loop was created to plot the maps. First, the loop created a dataframe of four columns: region, decade, n (count), and percentage (of the total number of documents in

that decade). Second, the loop made a copy of "mini_world" and then joined the newly-created dataframe onto it. Third, the loop plotted a heat map of the percentages with viridis as its color palette. Overall, five heat maps were generated to show how the weight of diplomatic coverage shifted from region to region over time.

Additionally, a global scatter plot was created to show the overall distribution of locations in the dataset. The plot layered all the coordinates in the dataset over a map (created using "world"). This was colored by viridis.

Lastly, five region maps were plotted using a for loop. This also relied upon the mapping and spatial packages. The scatter plots show text labels for each location that appeared within a region, throughout the entire period captured in the data. Once again, cooordinate boundaries had to be set up to frame the x and y-axes appropriately for the region maps.

Results

Line Graphs

The line graph shows that the volume of documents decreases over time, with some occasional spikes in activity. This should not be taken directly as an indicator of the overall level of US interest in Indonesia waning over time. Rather, this is likely to be a product of the collection method of FRUS. FRUS contains only declassified documents. As such, more recent documents on Indonesia may not have been declassified yet. The spikes of activity might be attributable to changes in Indonesia's domestic situation or in the international context of the Cold War. The interest of the United States in Indonesia might have increased during the late 1940s and early 1950s because Indonesia was engaged with a military and diplomatic conflict with the Netherlands, its former colonizer. Coverage may have increased again in the 1950s because of constitutional changes and regional rebellions. It might have increased again in the 1960s when the Sukarno regime was replaced by Suharto's. These causal explanations would require a closer look at the relevant texts of each document.

Bar Graphs

Top 10 Graph

The top 10 graph shows a clear pattern in reporting. Unsurprisingly, Washington DC, where most of the foreign affairs agencies covered in FRUS are based, was the origin of most of the documents on Indonesia. Unsurprisingly as well, Jakarta, the capital of Indonesia, takes the second place. Major centers of international diplomatic activity follow: New York (where the United Nations is based), the Hague (home of the International Court of Justice), Paris, London, and Geneva take the next few positions. Rounding up the list are Asian cities that played a prominent role in Cold War diplomacy and politics - Bangkok, New Delhi, and Saigon.

Top 5 Graphs

The top 5 graphs of each decade shows interesting patterns as well. Washington DC remains as the top source location for the documents. However, Jakarta slips quickly from its second-place position. By the 1970s, it had been overtaken by New York and Paris. Not a single document from the 1980s came from Jakarta. Once again, this is likely to do with the declassification process for FRUS. Documents from Jakarta may be more sensitive in nature, given that they are likely to be written by US diplomats observing and commenting on Indonesian events around them. As such, these documents may await future inclusion in FRUS. Other points of note include the Hague's significant contribution in the 1940s and the Hague's subsequent disappearance from the top 5 lists. Indonesia's history explains this. Indonesia's negotiations for its independence from the

Netherlands took place at the Hague and were concluded by the by the late 1940s. Also, the emergence of Bangkok in the 1960s and 1970s might be explained by the creation of the Association of Southeast Asian Nations (ASEAN) in 1967. ASEAN is headquartered in the Thai capital. These preliminary guesses await confirmation through the texts of the documents.

Maps

Map of Overall Locations

The map of overall locations reveals a rather uneven geographic distribution of US diplomatic coverage on Indonesia. Documents on Indonesia are concentrated in Southeast Asia, East Asia, Europe, the Middle East, and the coasts of North America. Yet, documents on Indonesia also come from surprising places in Latin America and the middle of the Pacific (Wake Island and Pearl Harbor). A glaring omission is any documents from Africa. This might be a product of how US foreign affairs agencies are structured. Offices that focus on Africa might be less likely to work on issues relating to Indonesia. Otherwise, it might be because there were fewer large-scale international conferences or gatherings held in Africa, or because Indonesia did not participate in these events. This might be surprising given Indonesia's involvement in the Non-Aligned Movement, which brought together Asian and African states.

Regional Maps

The regional maps allow one to see the above trends in closer detail. The regional maps shows the major diplomatic centers in each region - Washington and New York in North America; Geneva, Paris, and the Hague in Europe; Tokyo, Beijing, and Jakarta in Asia and the Pacific. The emptiness of the Africa map is particularly stark - the entire continent is bare, save for the top, which is fringed by Middle Eastern cities. Unexpected reporting locations make their appearance. One wonders why there was at least one document of Indonesia from the middle of Utah.

Heat Maps

The heat maps show how the weight of reporting shifted over time. The 1940s map show a fairly even distribution of reporting between Asia (likely as a result of the sources from Jakarta) and North America (with the bulk coming from Washington DC). No sources came from South America during this time. The 1950s shows that all regions contributed at least one source. However, Asia was starting to contribute fewer sources than North America and Europe was starting to contribute even less. By the 1970s, most of the regions had receded as North America came to dominate reporting on Indonesia. By the 1980s, there were no more sources from Africa. These trends are likely to be most reflective of the collection policies of FRUS, as distinct from the actual amount of reporting from the US foreign affairs agencies.

Conclusion

This project shows how US diplomat coverage of Indonesia may have evolved over time. The methods of this project can apply to any other country or search query input into the FRUS. An interesting comparison would be to compare Indonesia with its regional neighbors - Vietnam, the Philippines, and Thailand, for example - to see if the US' interest in Southeast Asia shifted from Indonesia to elsewhere. Another avenue for further exploration is to webscrape the actual texts of the documents and conduct text analysis on them. The problem of slow webscraping might perhaps be overcome by extracting only the strings that contain "Indonesia." Doing so might shed light on the topics that interested the US in Indonesia, and explain the trends reflected in the exploratory data analaysis above.