Google API-Derived Data

Overview

A quick orientation to the various production files generated for a single crisis and their potential uses. These files will be placed within a crisis specific folder. For this document, we are using the folder 'haiyan-meta'. There are a few considerations and peculiarities with Google API result data for which we have accounted and will call out after describing the available resources. ***Notice: we have too many files and variations as a result of many attempts to work with this date; the files will be reduced pending decision on TODOs listed below -- the files that remain will be included in our final production set in short order.***

Available Data Resources

1. google-all-country_stats_by_query_code.csv and google-all-country_stats_by_query_name.csv -- Two versions of the same data, with the 'query_code' variant providing the country code used in `cr` query param, e.g. `cr=countryUS` and the 'query_name' variant providing the country name instead. *Use(s): choropleth of crisis engagement around the world.* **TODO: PICK ONLY ONE OF THESE GOING FORWARD, MOST LIKELY `*by_query_code.`**

2. google-media-baseline_stats_by_query_distinct.csv and google-media-baseline-monthly_stats_by_query_distinct.csv -- Statistical results over the 10 weeks per media site being tracked for the former and over 6 months for the latter; it is the baseline or total `news OR coverage OR article` results. *Uses(s): show relative percentage of attention given by a site or by aggregation of sites. For example, if BBC had 10K news results the first period of the crisis, of which 5K were directed at the crisis, then 50% of their news coverage was dedicated to the crisis for that period.* **TODO: THE BASELINE STATS SHOULD ONLY INCLUDE THE MONTHLY GOING FORWARD(NOTE: THE MONTHLY BASELINE IS CURRENTLY ONLY A SAMPLE USING ALL GOOGLE RESULTS AND MISSING ALL INDIVIDUAL SITES).**

3. google-media_stats_by_query_distinct.csv and google-media-monthly_stats_by_query_distinct.csv -- -- Statistical results over the 10 weeks per media site being tracked for the former and over 6 months for the latter, using the actual crisis query information. *Use(s): line chart showing volume at a per site or aggregated into our 3 media categories. Note: monthly is the most predictable metric for measuring search results over a period.* **TODO: THE ACTUAL STATS SHOULD ONLY INCLUDE THE MONTHLY GOING FORWARD (NOTE: THE MONTHLY IS CURRENTLY ONLY A SAMPLE USING ALL GOOGLE RESULTS AND MISSING INDIVIDUAL SITES).**

4. google-media_dedup_all_no_text.csv -- 10 weeks of results from the crisis for our media sites, which can be further distinguished by media categories *'Traditional'*, *'Blogs-Social'*, and *'Independent'* , determined from information in the docId. There are 10 results per week per site available in this file, equaling up to 100 results per site over the crisis window captured. *Use(s): result urls can be presented for deeper exploration such as sentiment analysis or word cloud; the urls can also be explored directly in the browser. The 7zip clean-text_w13-to-w22.7z AND clean-text_w23-to-w25.7z attachments are the output of the original pages as text files, having filenames corresponding to `[docId].txt`. Note: unzipped, the results can be quite large (~10GB) even though compressed they are small (~10MB).* **TODO: THIS SHOULD BECOME MONTHLY PERIOD INSTEAD OF WEEKLY (WILL NEED TO BE REDONE).**

5. google-media-year_stats_by_query_distinct.csv -- Statistical results over the past year, not broken down by weeks and extending beyond the 10 week window up to the date of query. This will be a single number per site. *Use(s): May be used in some sort of collapsed temporal window. __TODO: RECOMMEND REMOVING THIS AS WE GET THE NUMBER NEEDED ELSEWHERE_*

6. google-media-all_stats_by_query_distinct.csv -- Statistical results without date restriction. This will be a single number per site. *Use(s): list as part of overview vis, most likely adjacent to the choropleth showing country engagement. Also, and very important, these total results help us temper in-between crawl-periods where Google reports at or nearly the same results as the week prior, see discussion in the following section. Note: these results are favored over the year results as they tend to include more results even if the crisis happened within the past year.* **TODO: RECOMMEND REMOVING THIS AS WE GET THE NUMBER NEEDED ELSEWHERE**

Peculiarities with Google data

- A very big peculiarity is that a manual search of Google data in the browser does not produce the same results as running the same search via API. It is well-known that Google keeps its own special blend of secret sauce for itself; however, exactly what is and is not included in the API is not documented. The best we can piece together is that less-interesting hits determined at some cut-off of Google's page ranking are not available. This includes myriad forum data and individual comments and user posts.
- A continuation of the above point about missing data -- additionally, Twitter data is handled rather specially by Google, where API licensing does not allow Google to provide full results. Therefore, in order to provide the Twitter total (if included), we would insert the results of manual / browser querying (i.e. the total appearing on results page at Google after running a search). It is this same restriction that caused us to keep Twitter out of our specifically tracked media sites.

- Google API does not offer a great deal of clarity when temporal searches are applied to ranges, e.g. w10 (10 weeks) prior to today. The results appear to conform to an opaque crawl-period or even API sync-period window which forces us to interpret weekly results in light of our knowledge of total results and the way the data precipitously adjusts. The pattern we have been able to detect in the data is that over a span of 1+ temporal periods, namely weeks, results appear to stay at or near a level of last crawled by Google and then make a noticeable drop. *We will work to mitigate this in our presentation of the data by interpolating (and visually distinguishing) periods where the data did not adjust.*
- As a follow-on to the above bullet about the opacity and unpredictability of Google API results for specified API periods, we attempted to request weeks prior to our crises and were unable to get the expected minimal results. It appears that Google insists on returning something, even if it is outside the period requested. As has been stated, we will work within the limits of the clarity Google API affords but it does diminish the ability to be too detailed with media responses within smaller windows; however, it remains very capable of showing overall responses and response trends over mid-term to long-term.