Google API-Derived Data

Overview

This is an 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 will reference the folder 'haiyan-meta', our first crisis of interest where we learned many data lessons, see history as well as API concept. 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. It should be pointed out that we are including all Google results as well as exemplar media sites. For a listing of the individual sites we are including in our data queries sample crawling activities, refer here. The sites can be further distinguished by media categories *'Traditional'*, *'Blogs-Social'*, *'Independent'* , and 'All'. ***Note: 'All' is a placeholder for Google results not constrained by any site and should not be confused with any sort of artificial summation of results from the sites we track.***

Available Data Resources

1. google-country_stats.csv -- Statistical results for all results relating to the crisis by country, providing the country code used in `cr` query param, e.g. `cr=countryUS`. For a listing of countries and their codes, refer here. *Use(s): choropleth of crisis engagement around the world by country.*

2. google-media-baseline_stats.csv -- Statistical results over a period of 6 months using a period frequency of monthly due to Google API peculiarities. This product establishes the baseline or ***total*** `news OR coverage OR article` result count each site we track as well as all Google, not just the crisis related coverage. *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.*

3. google-media_stats.csv -- Statistical results over 6 months for each site we track as well as all Google, using the actual crisis query information, resulting in a tabular output format having 1 row per period per site tracked. As has been previously stated, the period frequency is monthly due to Google API peculiarities. *Use(s): line chart showing volume at a per site or aggregated into our 3 media categories.*

4. google-media_results_subset.csv -- Subset of total available results, driven by the same data and metadata obtained in the previous output mentioned. However, whereas the previous output produces statistical information, this results in a tabular output format having 10 rows of sample results for each site we track over each period. For example, one site would have 60 result samples available over 6 months. *Use(s): result urls and their contents 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 google-media_results_subset-clean.7z attachment is the output of the original html sample result pages cleaned to plain text files, having filenames corresponding to* `[docId].txt`*. Note: unzipped, the results can be quite large (~10GB) even though compressed they are small (~10MB). Also, any unprocessable files during crawling due to improper mime type or ssl exceptions are provided as crawl-errors.7z.*

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 have mitigated this limitation in our presentation of the data by using monthly periods over anything more granular as monthly is much more predictable.*
- 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.