Datasheet for Google Trends data (Search Interest)

Motivation for Dataset Creation

- Created by Google to analyze relative search interest of search queries over custom time ranges.
- The data can also be used to assess search interest in different categories, and regions.

Dataset Composition

- Multiple instances with explicit relations between each other i.e. country with keyword, a category with keyword where the range of popularity is standardized between 0-100 (relative).
- The data points display interest relative to the highest point in the time frame for a given country and category.
- Data consists of dates and numerical data with identified subpopulations sourced directly from and by google itself.

Data Collection Process

- Google records search based on terms and not personally identifiable searches in its large data warehouses.
- We sampled data deterministically on search interest of the term 'startups' from 2004 to 2014.
- The population is the search interest for all terms that were searched for on Google.
- The missing information for any category means that the search has insufficient data. No known errors or sources of noise were identified in the available data.

Data Preprocessing

- For our use case, we collected data on relative search interest for the topic 'startups' in the USA (2004-2014) using the Pytrends API with a specific payload.
- Appropriate renaming and indexing of the interest levels were applied.

Dataset Distribution

- The trends data is distributed via an API and is updated daily.
- The terms set by Google include use case limitations, access methodology, and non - exclusivity. More on: https://developers.google.com/terms

Dataset Maintenance

- Google maintains search activity data that is anonymized (no one is personally identified), categorized (determining the topic for a search query), and aggregated (grouped together) as a display of interest transparency in its community of users.
- There is no mechanism for a third party to augment Google trends data.
- The Pytrends API is not an official or supported API.

Legal & Ethical Considerations

- The graph extracted is owned and protected by Google's Terms of Use.
- Since no personal data is stored, GDPR does not apply.
- Since data is normalized to the time and location of queries, it is likely to be appropriate for projects that analyze 'topical interest' as a feature.
- An ethical constraint must be to account for the search activity that is reflected due to automated searches/spam queries in Google as they do not reflect the public interest and can be deceptive at large.