How to generate Data for ProntoTrends

Intro

Data for ProntoTrends is mostly sourced from Google Trends, where we retrieve data on the level of an individual Keyword or alternatively multiple Keywords against each other. The complexity of using Google Trends data comes from the fact that Google scales the data always with respect to the maximum level that is currently viewed by the user. This means that the absolute values for keyword A when looked at individually can be/ are being scaled if you compare it with keyword B. This only happens in relative terms though so that you can definitively say that the keyword with the highest absolute value on GTrends' scale also has the highest search interest and how other keywords compare in relative terms (i.e. a value of 100 versus a value of 50 means that the bigger one is double as high as the lower one). In addition to this, Google does not allow temporal data to be compared across regions. It does offer geographic comparison over time however, which is used by the scripts to adjust the temporal data when comparing between regions

Generally, there are two distinct classes of data used in ProntoTrends:

- Keyword/Tag level data: Individual
  - created by querying individual keywords
  - for tag level keywords get averaged (and rescaled)
  - data is used for:
    - line chart over time (requests-trend.csv/ Chart_Data)
    - table over time (requests-details.csv / Table_Data)
    - Map (interest-map.csv / Map_Data)
- Comparison / detail / questions data: Comparisons
  - created by querying different options against each other (to observe scaling)
    - within each "category" (e.g. Ceremony type) different options exist (e.g. civil wedding, religious wedding, etc.) with different keywords that users may use to search for the option (e.g. religious wedding, wedding in church, etc)
      - of the possible keywords the one with the highest search volume is used to represent the option
  - data is used for:
    - any graph or tool where relative strengths of two concepts matter (e.g. Top5, preferences)

How to scrape

1. Preparing input files:

- For Individual: Use GSheet or Input_Set-Up/prepareKeywordsFile
  - CSV-File
- For Comparisons: Use GSheet or Input_Set-Up/prepareKeywordsFile
  - JSON-File
- Save in Input_Files

2. Scrape using mainProxy.py
   - For Individual: Choose `'Individual Keywords ("Keywords_CC.csv) - All Regions'` or `'Individual Keywords ("Keywords_CC.csv) - Only Country'`
     - will save data in `out` distributed by Keyword - individual files per Region and Dimension (Time & Geo) (different country could appear in same folder)
   - For Comparisons: Choose `'Comparisons (ProntoPro_Trends_Questions_CC.csv) - All Regions'`
     - will save data in `comparisons` distributed by Category - individual files per Region and Dimension (different country could appear in same folder)
3. If Individual:
   - You'll have to merge the Keywords to Tag level ➡ use generateSummaries (needs a Tag_Keyword.csv file) ➡ mergeData
   - You'll have to adjust the raw data to make regions relavive ➡ use generateSummaries ➡ adjustData
4. Create Final CSVs
   1. 'Create Category Overviews' ➡ Uses comparisons to generate summaries of the data within each category
      - normalizes data within each region and year to sum up to 100%
   2. 'Create Top5' ➡ Uses comparisons to determine the highest demand tag
      - for each year it calculates the seasonality, max month, min month, rank
   3. 'create Main Section' ➡ summarizes the CSVs generated in Category overviews
      - for each year, region and csv it selects the most selected option and saves it as representative for the category
   4. 'create Chart Data' ➡ uses individual data to create an overview of how tags develop over time. Requires Adjusted data to scale regions appropriately
   5. 'create Table Data' ➡ uses Chart Data and adjusts data so that monthly values for each year together make 100%
   6. 'create Map Data' ➡ can use Chart Data or Geo data from individual to generate data for Map
      - for every tag it calculates the region values by relative strength on a scale of 0 - 10