BSS Implementation

Parent Project	Cyber Vulnerability Prediction Link: https://bitbucket.org/dpatel24/visionbox-services/src/master-dev
Current Branch	feature/VS-6-rohit-compute-brand-sentiment-score
Owner	@ Rohit Krishnan Somasundaram
Reviewer	@ Vishal @ abhisavaliya
Version date	27 Jun 2022

🏛️ Architecture

I have created separate documentation for the architecture of BSS.

Link: https://visionbox-ai.atlassian.net/wiki/spaces/VS/pages/203456513/Brand+Sentiment+Score+BSS+using+GDELT

💻 Current Implementation

Our current implementation of the BSS follows the flow:

Initial account setup for accessing BigQuery from a Python script:

1. Initially, use the Google Cloud console to create or select a project (this project should have access to the gdelt-bq project) and enable billing.
2. We use BigQuery's Storage API to provide fast access to data/projects stored in BigQuery. We then use the BigQuery Storage API to query data stored in BigQuery for our use.
3. Next, we need to create a service account and use that to retrieve data from BigQuery's using a python script. Necessary steps to follow: https://cloud.google.com/bigquery/docs/bigquery-storage-python-pandas (**Note**: I have already created a service account on my Google Cloud account and enabled billing for the same. I have also created a key and downloaded it for authentication purposes.)
4. After implementation of step 3, we can access to store/retrieve data in BigQuery from a python script.

Python Implementation of the current version of BSS:

1. Before starting the implementation, we have to set the `GOOGLE_APPLICATION_CREDENTIALS` (The key you received in step 3 for accessing your service account) as an environment variable. Unfortunately, I was unable to set it using windows. So I have added it as an environment variable inside my python script. Here I have saved my `GOOGLE_APPLICATION_CREDENTIALS` in the project folder under the folder name Key.

```
# setting envoirment variable
os.environ[
    "GOOGLE_APPLICATION_CREDENTIALS"
] = "\\Key\\brand-sentiment-score-1b3f73b30ff6.json"
```

2. Steps to run logger.py (current working python script):

- Install the latest version of Python or anything above Python 3.6.
- Create a virtual environment to isolate dependencies.

    1. cd your-project

    2. py -m venv "**venv_name**" [Replace **venv_name** with your desired name]

- Set your shell/terminal to use the venv paths for Python by activating the virtual environment.
  1. "**venv_name**"\Scripts\activate Replace venv_name with your venv name.
- Install all the necessary packages from requirements.txt.
- Run the following command in your terminal to compute the BSS score.
  1. python -m logger "company_name" "from_date" "to_date"
  Here replace "company_name" with your desired company name and replace "from_date" and "to_date" with just the dates.

```
(env) PS C:\Users\VisionBox\Visionary_Farm\GDELT\BSS\Test> python -m logger google 26 27

Sample of the extracted data:
         Date      tone                                DocumentIdentifier
0  2022-06-26  0.000000                https://www.bhol.co.il/news/1405047
1  2022-06-26  2.625821  https://www.deseret.com/2022/6/26/23181928/goo...
2  2022-06-26 -3.079710  https://www.wicz.com/story/46762528/one-of-the...
3  2022-06-26  2.405498  https://y105fm.com/body-found-in-minnesota-riv...
4  2022-06-26  0.924214  https://www.phonearena.com/news/osom-ov1-has-a...
```

3. Implementation details: Once you have installed all the required packages, you are ready to know what's happening inside the script.

    1. Initial setup involves *importing necessary packages*, *saving the command line arguments as variables* and *setting the environment variables.*
    2. Next is to connect to the BigQuery service account.

```
# connecting to big query client
client = bigquery.Client()
```

    3. Next, I have implemented a simple function to *accept a query, retrieve data from BigQuery and store it as a data frame.*

```
# Function to extact results to dataframe given a query as input
def gcp2df(sql):
    query = client.query(sql)
    results = query.result()
    return results.to_dataframe()
```

    4. For the query we have used is a modified version of the existing query that I have experimented with (Sample Big Query Experiment and result) and added some customizations to the query to use user inputs.

```
# adding customizable query values (from-to date and desired company names are given as command line arguments)
from_date = '  DATE(_PARTITIONTIME) >= "2022-06-' + str(date) + '"'
to_date = '\n  AND DATE(_PARTITIONTIME) < "2022-06-' + str(date1) + '"'
organizations = "\n  AND lower(V2Organizations) LIKE  '%" + company + "%'"
```

    5. The current Brand Sentiment Score for a company just sums up all the tone values obtained on the given date range (usually per day) and divides by the number of documents received for that query.

```
(env) PS C:\Users\VisionBox\Visionary_Farm\GDELT\BSS\Test> python -m logger google 26 27

Sample of the extracted data:
         Date      tone                                DocumentIdentifier
0  2022-06-26  0.000000                https://www.bhol.co.il/news/1405047
1  2022-06-26  2.625821  https://www.deseret.com/2022/6/26/23181928/goo...
2  2022-06-26 -3.079710  https://www.wicz.com/story/46762528/one-of-the...
3  2022-06-26  2.405498  https://y105fm.com/body-found-in-minnesota-riv...
4  2022-06-26  0.924214  https://www.phonearena.com/news/osom-ov1-has-a...

Number of documents mentioning the company google: 869

Sum of all tone values (from 1st March to current date): -68.1619262636062

The Brand Sentiment Score for google is: -0.0784
```

    6. We have stored all the results for a company in a CSV file date-wise for further graphical analysis. Data Storage link: https://1drv.ms/u/s!Aj7ldg8i-QXtavs29oSTrx0IR4U?e=KxBOdm Documentation: BSS Daily Data Tracking

🔥 Known Issues

	Issues	Error date	Jira Ticket

1	Unable to set up `GOOGLE_APPLICATION_CREDENTIALS` as an environment variable.	20 Jun 2022	
2	To find a way to handle config /credentials in BSS. (Packaging)	20 Jun 2022	**VS-16** - Remove the config/credentials file in brand-sentiment-score **DONE**
3	To find a way to add entire date(DD-MM-YYYY) as input instead of just the (DD) format.	27 Jun 2022	