Inferring Google Trend index from GDELT news data

-Ashwin Dhanasamy

Objective

- Inferring the google trend index for the term ‘unemployment’ by analysing news data over the past week.
- Starting axiom: Evolution of *Google Trends keywords* can be inferred from what was written in news media .

Datasets

- GDELT
  - News events containing selected THEMES that might be related to inferring the trend index of the term “unemployment”
- Google Trends
  - Trend Index for the term “unemployment”. This is the target column we are trying to predict.

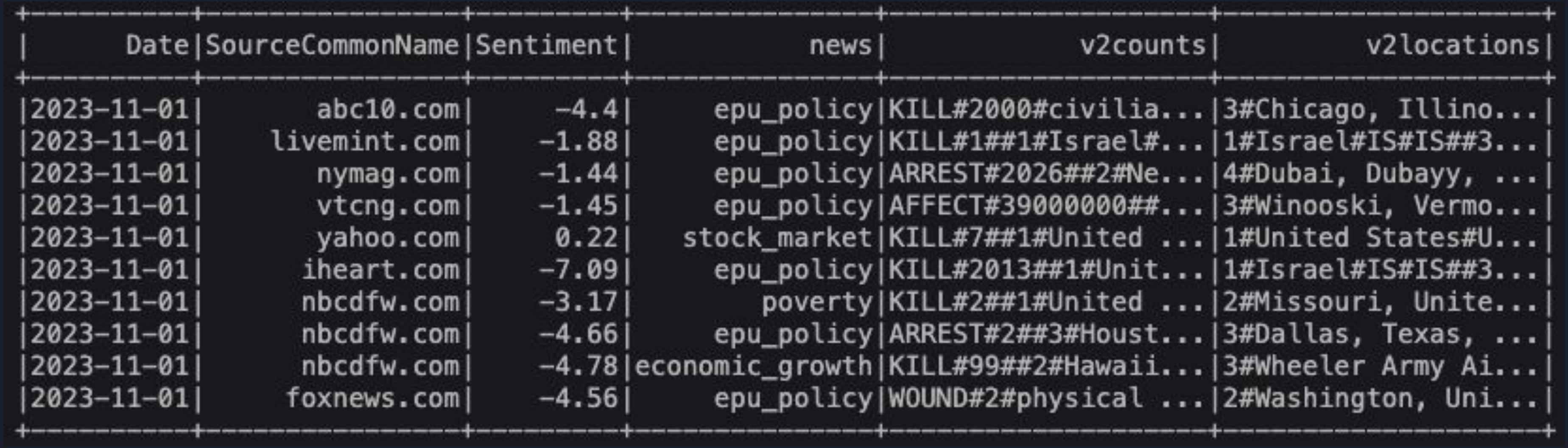
ETL - GDELT

- Used BigQuery to get GDELT records containing the specified themes. Sample of the themes considered on the next slide.
- The returned table is saved directly to GCS.
- Full BigQuery SQL command is available in the file: big_query.txt

SAMPLE THEMES

- Stock Market
- Central Banks
- Inflation
- Bankruptcy
- Debt Vulnerability
- Job Quality and Labor Market Performance
- Poverty
- Economic Growth
- Economic Debt

Raw GDELT Dataframe after ETL

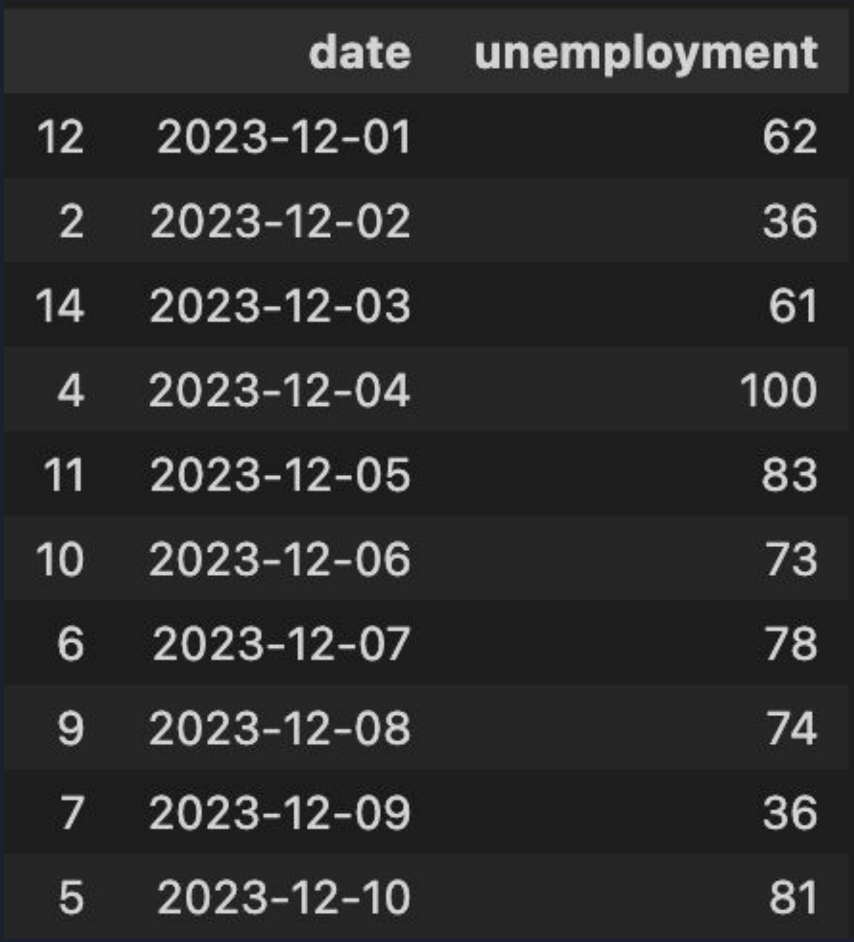


```
+----------+---------------+---------+---------------+--------------------+--------------------+
|      Date|SourceCommonName|Sentiment|           news|            v2counts|         v2locations|
+----------+---------------+---------+---------------+--------------------+--------------------+
|2023-11-01|      abc10.com|     -4.4|     epu_policy|KILL#2000#civilia...|3#Chicago, Illino...|
|2023-11-01|    livemint.com|    -1.88|     epu_policy|KILL#1##1#Israel#...|1#Israel#IS#IS##3...|
|2023-11-01|      nymag.com|    -1.44|     epu_policy|ARREST#2026##2#Ne...|4#Dubai, Dubayy, ...|
|2023-11-01|      vtcng.com|    -1.45|     epu_policy|AFFECT#39000000##...|3#Winooski, Vermo...|
|2023-11-01|      yahoo.com|     0.22|   stock_market|KILL#7##1#United ...|1#United States#U...|
|2023-11-01|     iheart.com|    -7.09|     epu_policy|KILL#2013##1#Unit...|1#Israel#IS#IS##3...|
|2023-11-01|     nbcdfw.com|    -3.17|        poverty|KILL#2##1#United ...|2#Missouri, Unite...|
|2023-11-01|     nbcdfw.com|    -4.66|     epu_policy|ARREST#2##3#Houst...|3#Dallas, Texas, ...|
|2023-11-01|     nbcdfw.com|    -4.78|economic_growth|KILL#99##2#Hawaii...|3#Wheeler Army Ai...|
|2023-11-01|    foxnews.com|    -4.56|     epu_policy|WOUND#2#physical ...|2#Washington, Uni...|
+----------+---------------+---------+---------------+--------------------+--------------------+
```

ETL - Google Trends

- Used the pytrends package to get the trending index for the term “unemployment”, this will be our predicted column.
- Python script is available in the file : trends.py

Raw Trend Index Dataframe after ETL

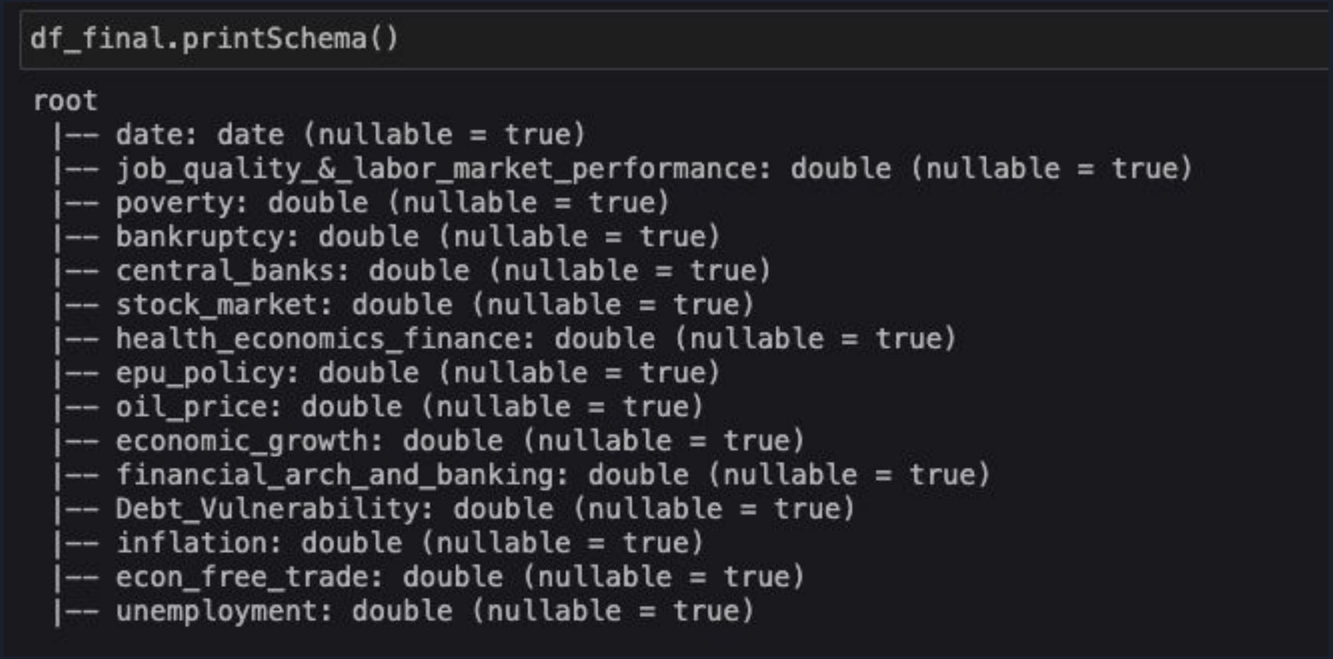


	date	unemployment
12	2023-12-01	62
2	2023-12-02	36
14	2023-12-03	61
4	2023-12-04	100
11	2023-12-05	83
10	2023-12-06	73
6	2023-12-07	78
9	2023-12-08	74
7	2023-12-09	36
5	2023-12-10	81

Data - Preprocessing

- Sentiment scores of each of the themes we have selected will be our features.
- To calculate the sentiment scores for each theme, we consider the sentiment scores associated with each theme over the past week and perform an aggregation to get a final sentiment score for that theme for that week.
- **Our final dataframe will contain the consolidated sentiment scores for each theme and the corresponding trend index that we are trying to predict.**

Schema of Pre-Processed Data Frame

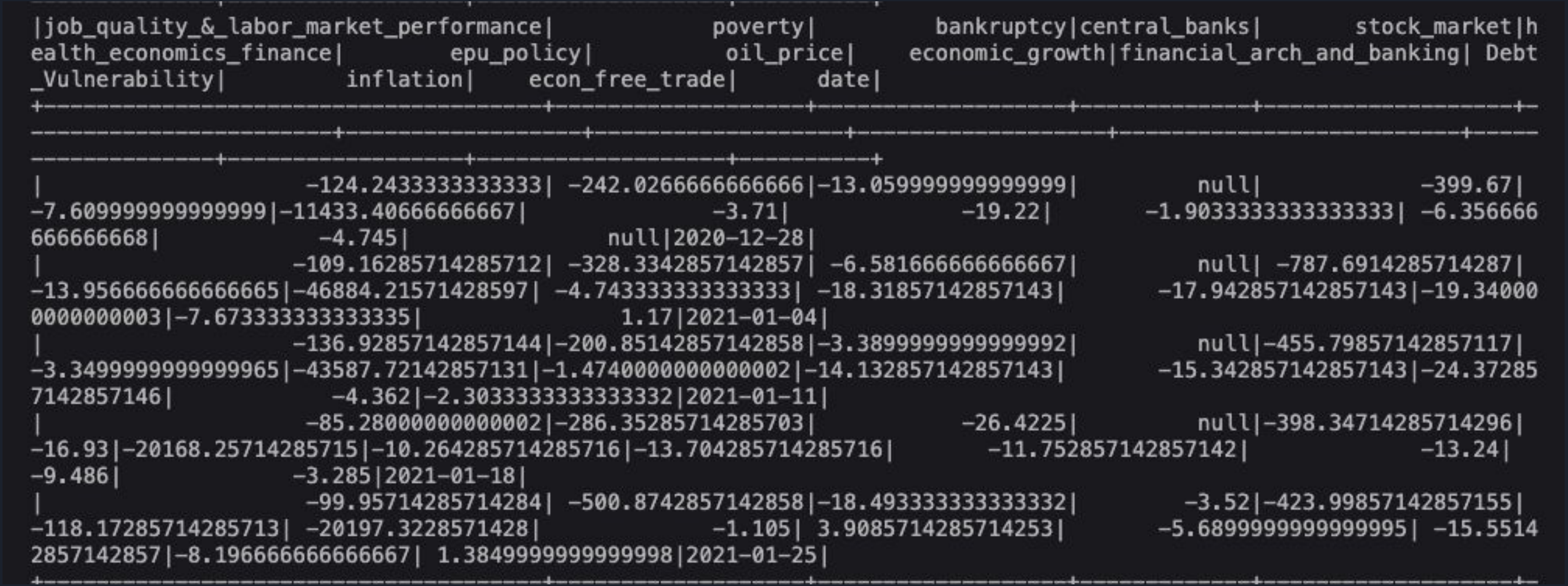


```
df_final.printSchema()
```

```
root
 |-- date: date (nullable = true)
 |-- job_quality_&_labor_market_performance: double (nullable = true)
 |-- poverty: double (nullable = true)
 |-- bankruptcy: double (nullable = true)
 |-- central_banks: double (nullable = true)
 |-- stock_market: double (nullable = true)
 |-- health_economics_finance: double (nullable = true)
 |-- epu_policy: double (nullable = true)
 |-- oil_price: double (nullable = true)
 |-- economic_growth: double (nullable = true)
 |-- financial_arch_and_banking: double (nullable = true)
 |-- Debt_Vulnerability: double (nullable = true)
 |-- inflation: double (nullable = true)
 |-- econ_free_trade: double (nullable = true)
 |-- unemployment: double (nullable = true)
```

Final - Pre-Processed Data Frame

Note: This snapshot contains “null” values, but I’ve removed them in the actual implementation



```
|job_quality_&_labor_market_performance|            poverty|        bankruptcy|central_banks|         stock_market|h
ealth_economics_finance|        epu_policy|          oil_price|    economic_growth|financial_arch_and_banking| Debt
_Vulnerability|         inflation|    econ_free_trade|      date|
+---------------------------------------+-------------------+-------------------+-------------+-------------------+-
-----------------------+------------------+-------------------+-------------------+--------------------------+-----
--------------+------------------+-------------------+----------+
|                    -124.2433333333333| -242.0266666666666|-13.059999999999999|         null|            -399.67|
-7.609999999999999|-11433.40666666667|              -3.71|             -19.22|       -1.9033333333333333| -6.356666
666666668|            -4.745|               null|2020-12-28|
|                   -109.16285714285712| -328.3342857142857| -6.581666666666667|         null| -787.6914285714287|
-13.956666666666665|-46884.21571428597| -4.743333333333333| -18.31857142857143|      -17.942857142857143|-19.34000
0000000003|-7.673333333333335|               1.17|2021-01-04|
|                   -136.92857142857144|-200.85142857142858|-3.3899999999999992|         null|-455.79857142857117|
-3.3499999999999965|-43587.72142857131|-1.4740000000000002|-14.132857142857143|      -15.342857142857143|-24.37285
7142857146|           -4.362|-2.3033333333333332|2021-01-11|
|                    -85.28000000000002|-286.35285714285703|           -26.4225|         null|-398.34714285714296|
-16.93|-20168.25714285715|-10.264285714285716|-13.704285714285716|       -11.752857142857142|             -13.24|
-9.486|             -3.285|2021-01-18|
|                    -99.95714285714284| -500.8742857142858|-18.493333333333332|        -3.52|-423.99857142857155|
-118.17285714285713| -20197.3228571428|             -1.105| 3.9085714285714253|      -5.6899999999999995| -15.5514
2857142857|-8.196666666666667| 1.3849999999999998|2021-01-25|
+---------------------------------------+-------------------+-------------------+-------------+-------------------+-
```

ML Methodologies Implemented

- Linear Regression

Results

- Unfortunately, my implementation of scaling the features is wrong and my scaled features have mean and standard deviation of 0. Thus the model outputs a constant value for every row.
- I am not able to correct this before the submission deadline.

Comments

- Tried to implement an end-to-end ML solution, I could not perform many intricate steps like removing correlated columns that would have made my solution more optimal. Thus I’ve gotten disastrous results.
- I’ve only implemented a simple linear regression model, in practice this problem would warrant the use of better methodologies.
- As a student of Economics, I would have wanted to select more relevant themes than the list I’ve currently selected. Considering GDELT’s massive list of THEMES, a lot of missed potential here in terms of Feature Selection.
- I drew inspiration for the project from: https://lookerstudio.google.com/u/0/reporting/e171bbe8-0db8-49bb-b1d6-86cb4f16acdf/page/DL61B?s=iGEf2faYhUE