VISUALIZING OUR GLOBAL WORLD

COMPREHENSIVE INSIGHTS ON GLOBAL NEWS IMAGERY

Authorship:

Samvat Rastogi
Nirad Ranjan Parhi
Sneha Lakshmi Nyayapati
Sreya Chakrabarti
Niketh Shetty

GDELT AND VGKG

- The Global Database of Events, Language and Tone is an open database collection of news articles from across the globe from the year 1979. GDELT has three datasets :
    - EVENTS DATABASE
    - GKG
    - VGKG
- The Visual Global Knowledge Graph is an enhancement on top of the GKG, which uses the Google Cloud Vision API to extract meaningful information from a stream of visuals that are part of the world's news

DETAILS OF THE DATASET

- We are accessing the data from Google BigQuery . The name of the table accessed is gdelt-bq:gdeltv2.cloudvision
- The table has 248,841,467 rows of data ranging from around 31st Dec 2015 to the present. It has a size of 2.12 TB
- There are 12 columns in the table, out of which 5 are useful to this project
- The envisioned depictions are based on the relevance and meaning that each column has.

SCHEMA OF THE CLOUD VISION DATASET

DATE
DOCUMENTIDENTIFIER
IMAGEURL
LABELS
GEOLANDMARKS
LOGOS
SAFESEARCH
FACES
OCR
LANGHINTS
WIDTHHEIGHT
RAWJSON

COLUMN SPECIFIC INFORMATION

DATE – The date on which that specific image was monitored and recorded by GDELT. This field is used in our timeline based visualizations
LABELS – This column gives information about the contents/objects present in the image. This field can be extensively used to cover various use-cases on a wide variety of topics
GEOLANDMARKS – The attributes 'Latitude' and 'Longitude' of this entity are being used to plot specific information over the world map
SAFESEARCH – This column provides us information about the content being offensive or non-offensive. The four attributes (Violence , Medical , Spoof , Adult) will be used to locate the kind of information
FACES – This field will be used extensively to extract information related to the emotions depicted in an image

EXAMPLE

Sample Query:
select Labels, GeoLandmarks, DocumentIdentifier, SafeSearch, Faces FROM [gdelt-bq:gdeltv2.cloudvision] where ImageURL = "http://www.wafa.ps/http://www.wafa.ps/userfiles/image/multimedia/samah.jpg";

EXAMPLE CONTINUED

- LABELS : GRADUATION<FIELD>0.98690468<FIELD>/M/016C3C<RECORD>ACADEMIC DRESS<FIELD>0.92389059<FIELD>/M/01XQVB<RECORD>EVENT<FIELD>0.55467749<FIELD>/M/081PKJ <RECORD>COSTUME<FIELD>0.55138981<FIELD>/M/0250X
- GEOLANDMARKS : NULL
- DOCUMENTIDENTIFIER : http://www.wafa.ps/ar_page.aspx?id=RcYwVMa695154680682aRcYwVM
- SAFESEARCH : -1<FIELD>0<FIELD>0<FIELD>0
- FACES : 0.99955839<FIELD>12.286129<FIELD>30.525574<FIELD>-7.1071544<FIELD>0.5889504<FIELD>78,87;132,87;132,141;78,141<FIELD>0<FIELD>0<FIELD>0<FIELD>**2** <FIELD>0<FIELD>0<FIELD>0<RECORD>0.50865442<FIELD>6.1033735<FIELD>-5.2280416<FIELD>6.324564<FIELD>0.63740838<FIELD>146,69;196,69;196,119;146,119<FIELD>0<FIELD >0<FIELD>0<FIELD>**2**<FIELD>0<FIELD>0<FIELD>0

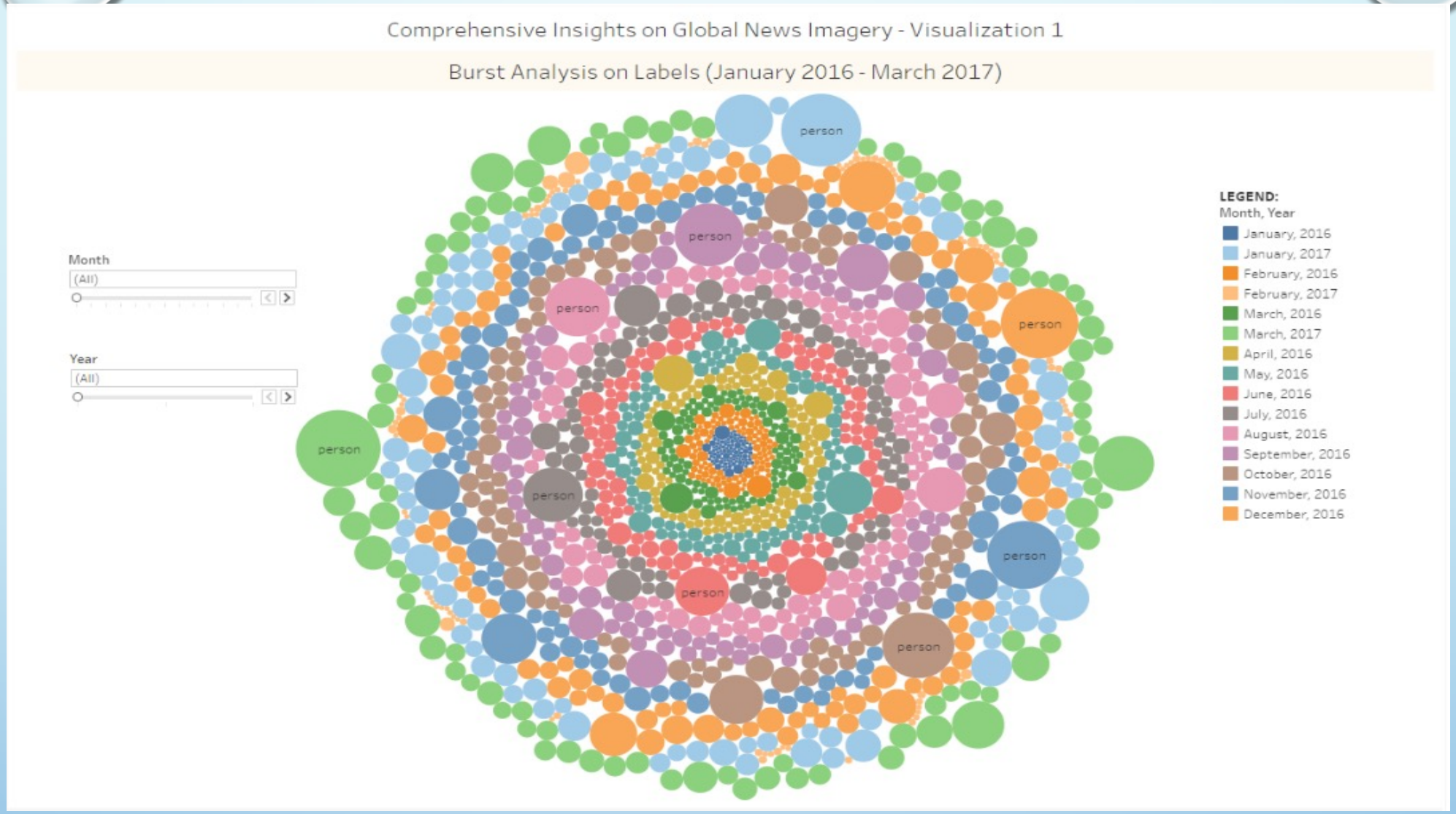
ABOUT THE VISUALIZATIONS

There are four Visualizations in this project:

1. Burst Analysis on Labels
2. Quantified Relationship between Emotions and Associated Labels
3. Geographical distribution of SafeSearch trends
4. Top Labels based on SafeSearch category

BURST ANALYSIS ON LABELS

- This Visualization intends to highlight the variation of interests in major news topics over time.
- The topics of interests from various news articles are shown with partitions of monthly time slices for the time period - January 2016 to March 2017.
- Each label is represented by a bubble.
- The size of the bubble denotes the frequency of appearance of a label number in the news articles for the period in question.
- We captured 15 such subsets of data from BigQuery for the entire time range.



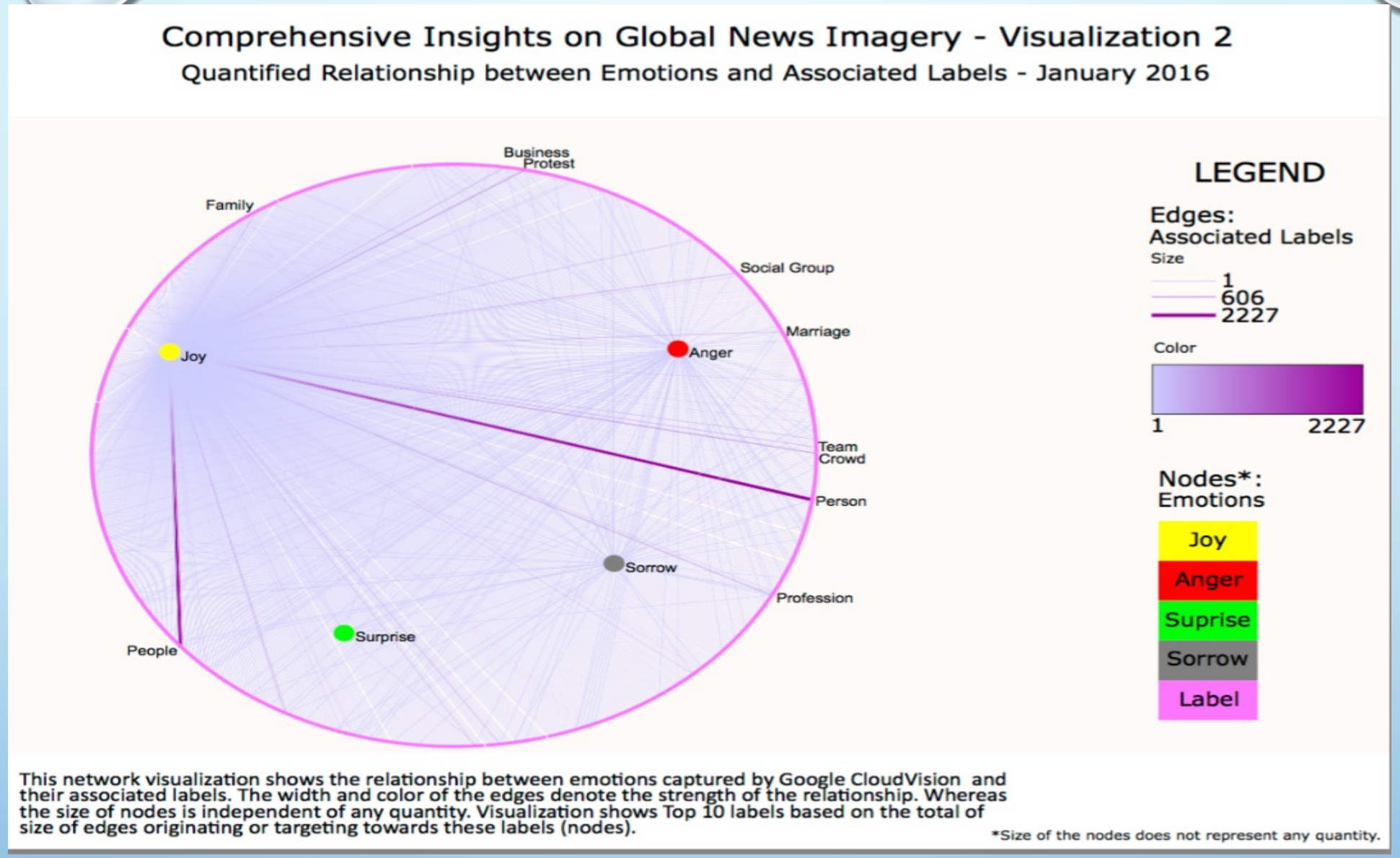
https://public.tableau.com/profile/samvat.rastogi#!/vizhome/ComprehensiveInsightsonGlobalNewsImagery-Visualization1/Dashboard1

INSIGHTS GAINED

- The label **person** is most frequently occurring in all the news articles over the time range
- Other frequently occurring labels are - **people, profession, vehicle, athlete, sports**
- We can conclude that the labels relating people like vehicle, sports are occurring most frequently as the day-to-day news articles are dominated by such content.

QUANTIFIED RELATIONSHIP BETWEEN EMOTIONS AND ASSOCIATED LABELS

- This is a network visualization that helps fetch insights about the co-occurence of emotions like **joy, sorrow, anger** and **surprise** by analyzing the likelihood of each emotion in the images tagged to the corresponding labels in the dataset
- We have determined the top 10 labels associated with all the four emotions
- It gives the quantified relationship between emotions and their associated labels



https://samvat.github.io/ivmooc-gdelt-project/Visualization2.pdf

INSIGHTS GAINED

- Out of the four categories of emotions i.e. "Joy", "Sorrow", "Surprise" and "Anger" , the emotion "Joy" is most frequently occuring.
- The top labels associated with all the 4 emotions are - Person, People, Protest, Crowd, Business,Social Group, Marriage etc.
- The labels occurring most frequently with "Joy“ are Family, Person, People, Profession etc.
- Similarly, we can see other frequently occurring labels with emotions.

GEOGRAPHICAL DISTRIBUTION OF SAFESEARCH TRENDS

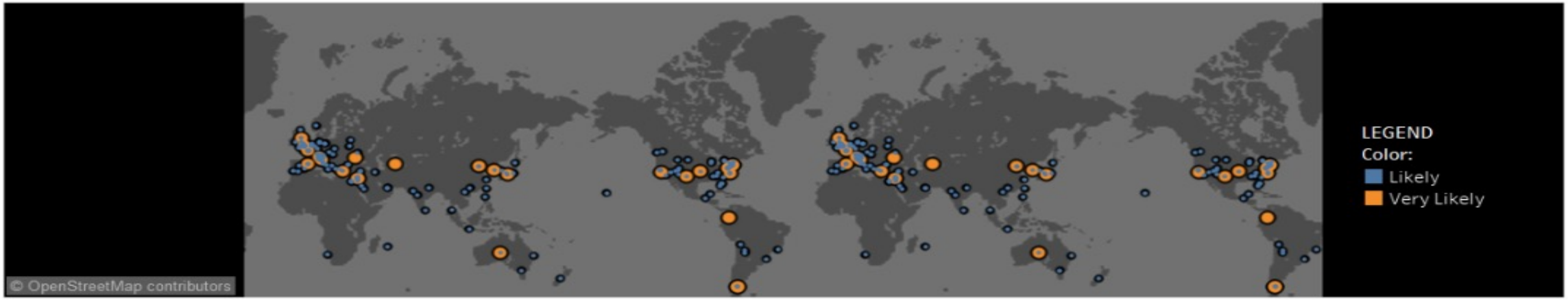
- This visualization depicts the sources of news articles categorized as **Medical, Violent, Adult** and **Spoof** by the Google SafeSearch algorithm on a two dimensional map of the Earth.
- The Geolocations are marked with the number of articles in each category originating from the location in question.
- For this visualization, the two fields in use are GeoLandmarks and SafeSearch.

Comprehensive Insights on Global News Imagery - Visualization 3

Geographical Distribution of SafeSearch Trends

Adult

Medical



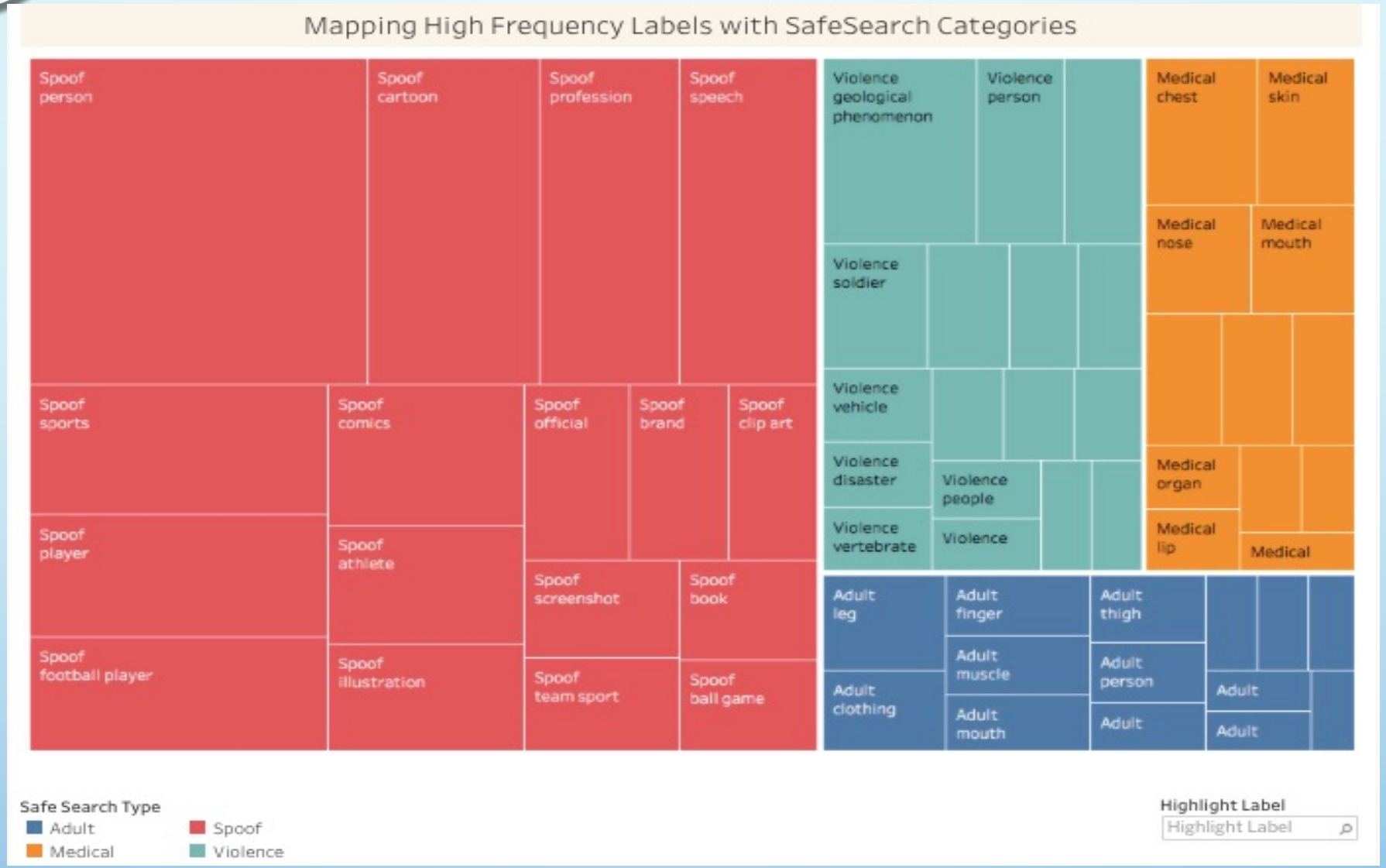
https://public.tableau.com/profile/samvat.rastogi#!/vizhome/ComprehensiveInsightsonGlobalNewsImagery-Visualization3/Dashboard1

INSIGHTS GAINED

- Adult content in the news articles are majorly generated from North America and Europe Regions
- In case of medical content, Europe region dominates the news. The U.S. , South Africa and Australia are other significant contributors.
- News articles coming under spoof category is generated in almost equal proportion from the developed countries
- News articles containing violent images are more predominantly generated from Europe, Russia, India and U.S.

TOP LABELS BASED ON SAFESEARCH CATEGORY

- This visualization is a TreeMap portraying the top 20 labels for each category of the Safe Search field.
- Through this visualization we are trying to find the most common labels or objects present in an image that are linked to the type of search.
- There are four categories of search that can be represented here – **Violent , Medical , Spoof** and **Adult.**



https://public.tableau.com/profile/samvat.rastogi#!/vizhome/ComprehensiveInsightsonGlobalNewsImagery-Visualization4/Dashboard1

INSIGHTS GAINED

- The top labels which generally trigger the Spoof category of Cloud Vision API are Person, Cartoon, profession, speech, sports etc.
- For violence SafeSearch category, the top labels are - geological phenomenon, soldier, vehicle, disaster etc.
- Similarly, for the category **"Medical"**, top labels are **- chest, skin, nose, mouth, human body.**

RELATED WORK

- The blog *New GDELT Daily Trend Reports[3]* discussed daily trends of conflict across the globe by correlating specific news items related to conflicts
- Work done by Kaleev Letaru to find geographical details from news articles and geocodes those articles to specific locations[4]
- Project undertaken by Kaleev Letaru to portray the positive or negative tone of the news articles based on the geo location using GKG dataset[5]

CHALLENGES

- Only around 1.5% of the dataset has geographic information
- Data is sparse in most columns of the dataset
- Huge size of the dataset
- The CloudVision API parses the images and marks images as common nouns and does not give details such as name of a person etc.
- Data ranges only from 2016 – 2017 , posing a limitation on temporal analysis

REFERENCES

1] Börner, K. 2015. Atlas of Knowledge: Anyone Can Map. Cambridge, Massachusetts: The MIT Press.

[2] http://blog.gdeltproject.org/announcing-the-new-gdelt-visual-global-knowledge-graph-vgkg

[3] http://blog.gdeltproject.org/gdelt-daily-trend-reports/

[4] https://www.forbes.com/sites/kalevleetaru/2017/02/21/visual-geocoding-a-quarter-billion-global-news-photographs-using-googles-deep-learning-api/#68bf7a1217fa

[5] https://www.forbes.com/sites/kalevleetaru/2017/02/22/mapping-global-happiness-in-2016-through-a-quarter-billion-news-articles/#4141a7642692

[6] http://analysis.gdeltproject.org/module-gkg-wordcloud.html

QUESTIONS ?