GDELT VISUAL GLOBAL KNOWLEDGE GRAPH (VGKG)
POWERED BY GOOGLE CLOUD VISION API
DATA FORMAT CODEBOOK 1.0 ALPHA
12/31/2015
http://gdeltproject.org/

ALPHA RELEASE

INTRODUCTION

This codebook introduces the new GDELT Visual Global Knowledge Graph (VGKG) Version 1.0 Alpha, powered by the Google Cloud Vision API (https://cloud.google.com/vision/). The VGKG extends GDELT's ability to understand global news media, allowing it for the first time to make sense of the vast stream of visual narrative that accompanies the world's news.

Each day GDELT monitors between half a million and a million primary images from across the planet, capturing nearly every event and topic imaginable from almost every corner of the earth. These images offer a vivid and rich visual tapestry of global events and daily life across the planet, reaching far beyond what textual narrative alone can offer. Today, powered by the Google Cloud Vision API, GDELT is able to apply some of the most sophisticated deep learning neural network algorithms in the world to actually make sense of the imagery of the world's news just like a human.

Each image undergoes a range of incredibly sophisticated analyses that catalog the actual topical focus of the image (tagging an image with the kinds of objects, activities, and backgrounds it contains), sentiment mine its emotions (whether people in the image are happy, sad, angry, or surprised), recognize street signs, labels, and other text, identify famous locations, and even flag violent imagery. The topical algorithms in particular we think will entirely transform how we understand the imagery of the world's news media, reaching deeply into how world events are portrayed and understood. We are especially excited about the potential of such algorithms to rapidly triage the earliest imagery emerging from disaster zones or conflict areas, allowing human first responders to assess rapidly changing environments.

The following annotations are produced for each image:

- **Labels**. The API attempts to discern the key topics, objects, and activities depicted in the image and produces a set of category labels that describe its contents. These can be used for macro-level filtering and searching of imagery and their associated articles by category.
- **GeoLandmarks**. The API attempts to identify the location where the photograph was taken, identifying both major landmarks like the Eiffel Tower and individual street corners and public squares around the world, and outputting an estimated latitude and longitude.
- **Logos**. The API attempts to recognize logos of major organizations seen in the image.
- **SafeSearch**. The Google SafeSearch algorithms are used to estimate whether the image likely contains violence, medical imagery, or other characteristics.
- **Faces**. The API attempts to identify the presence of human faces in the image and the general emotional state expressed by that person, such as happiness or sadness, joy or anger. NOTE that by design the API does NOT perform face recognition (identifying who a face belongs to), it ONLY determines that a face exists in the image and its general emotional expression.

- **OCR**. The API attempts to extract any potential textual labels found in the image, ranging from street signs and storefront writing to protest signs and overprinted captions and annotations.

Caveats & Disclaimer: Alpha Release

Unlike the other text-based GDELT feeds, please note that this feed is extremely experimental and is being released as an alpha release. This means that the behavior and supported features of the feed may change at any moment, or it may go away entirely. If ingesting these feeds into an automated workflow, your scripts should perform extensive error checking to ensure that they are able to cope with any changes to the underlying data format. Please check back on the GDELT Blog on a regular basis for any updates or documentation changes to the format and its behavior.

The use of deep learning algorithms for image recognition is still a highly experimental area of active research and the application use case presented by GDELT (attempting to recognize and catalog arbitrary images from almost every corner of the planet on almost every topic imaginable at realtime speed) represents one of the most difficult and wide-ranging applications of such technology today.

What this means is that you will almost certainly encounter a certain level of error in the categorizations and other information computed about each image. Remember that all tags are applied 100% automatically with NO HUMAN INTERVENTION and mistaken categorizations or other tags represent computer algorithm errors, NOT editorial statements. Remember that computer image recognition at these scales is still in its relative infancy and the underlying algorithms are encountering large amounts of imagery utterly unlike anything they've ever seen before, so this data stream is really pushing the boundaries of current deep learning recognition and will make mistakes. Please email cloudvision-feedback@google.com with the image URL and mistaken metadata fields if you find any particularly significant errors so that the system can be constantly improved and refined.

Note also that since this is an alpha release, the total volume of imagery processed may vary considerably throughout the course of a given day, may decrease substantially towards the end of each day, or may stop entirely for long stretches. At present we are only processing the SocialSharingImage field from the GKG. If an image from the GKG is not found in the VGKG that simple means that we ran out of time in a given 15 minute interval to submit it for processing – during this alpha period we are only able to process a small fraction of the total volume of global news imagery.

DATA FORMAT DOCUMENTATION

The VGKG stream updates every 15 minutes in keeping with the rest of the GDELT 2.0 GKG infrastructure. Due to the way the VGKG is currently implemented in the overall GDELT pipeline, there is a 15 minute rolling delay in which each VGKG file encodes the computed metadata for the images of articles found in the GKG file released 15 minutes ago. You can currently access the data through two mechanisms:

- **Google BigQuery**. Similar to the main GKG table, we also populate a publically accessible table housed in Google BigQuery to make it possible to interactively query and analyze the computed metadata and join it against the main GKG and EVENT tables. The table is gdelt-bq:gdeltv2.cloudvision.

- **Raw CSV Files**. Similar to the GKG files, you can access the VGKG data stream via simple CSV files updated every 15 minutes. These files are tab-delimited and gzip compressed. The latest CSV files are released somewhere between 0-5 minutes, 15-20 minutes, 30-35 minutes, and 40-45 minutes after the hour (about 5 minutes after the GKG file is released for each 15 minute increment). To determine when the latest VGKG files are available, check the contents http://data.gdeltproject.org/gdeltv2_cloudvision/lastupdate.txt, which is updated when the latest update files become available. You could poll this file every 5 minutes to ensure you always download the latest files.

The data format of the individual fields is as follows:

- **DATE**. This is the full date and time that the article containing this image was monitored by GDELT in 15 minute resolution. It is in YYYYMMDDHHMMSS format.
- **DOCUMENTIDENTIFIER**. This is the URL of the article in which the image was found. It is identical to the URL used in the main GKG and EVENT tables, allowing cross-referencing and joining.
- **IMAGEURL**. This is the URL of the specific image within the article that was analyzed, since an article may contain multiple images. As of this writing, we are currently processing only the SocialSharingImage for each article.
- **LABELS**. ("<RECORD>" delimited blocks, with "<FIELD>" delimited fields) This is the list of label annotations describing the contents of the image. Each label contains three fields delimited by "<FIELD>" and multiple labels are separated with "<RECORD>". Each record provides the following fields:
    - **Description**. This is the human-friendly description of the label.
    - **Score**. This is a numeric score of the confidence the API had in its assignment of that label to the image.
    - **MID**. This is the unique ID key of the label in Google's knowledge graph.
- **GEOLANDMARKS.** ("<RECORD>" delimited blocks, with "<FIELD>" delimited fields) This is the list of estimated geographic locations where the image was taken. Each label contains three fields delimited by "<FIELD>" and multiple labels are separated with "<RECORD>". Each record provides the following fields:
    - **Description.** This is the human-friendly name for the location. Note that this field may be blank for street-level matches and in certain other cases.
    - **Score.** This is a numeric score of the confidence the API had in its assignment of that label to the image.
    - **Latitude**. This is the general area geographic latitude estimated for the image's location.
    - **Longitude**. This is the general area geographic longitude estimated for the image's location.
- **LOGOS**. ("<RECORD>" delimited blocks, with "<FIELD>" delimited fields) This is the list of logos found in the image. Each label contains three fields delimited by "<FIELD>" and multiple labels are separated with "<RECORD>". Each record provides the following fields:
    - **Description**. This is the human-friendly description of the label.
    - **Score**. This is a numeric score of the confidence the API had in its assignment of that label to the image.
    - **MID**. This is the unique ID key of the label in Google's knowledge graph.
- **SAFESEARCH**. ("<FIELD>" delimited fields). This is the output of the Google SafeSearch algorithms on the image. It contains the following fields:

    - **ViolenceLikelihood.** Estimated likelihood that the image contains violence. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **MedicalLikelihood.** Estimated likelihood that the image is medical in nature. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **SpoofLikelihood**. Estimated likelihood that the image is a spoof. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **AdultLikelihood**. Estimated likelihood that the image is adult in nature. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
- **FACES.** ("<RECORD>" delimited blocks, with "<FIELD>" delimited fields) This is the list of faces found in the image. Each record contains a series of fields delimited by "<FIELD>" and multiple faces are separated with "<RECORD>". Each record provides the following fields:
    - **DetectionConfidence**. The algorithm's confidence that this is a human face.
    - **RollAngle**. The estimated roll angle of the face.
    - **PanAngle**. The estimated pan angle of the face.
    - **TiltAngle**. The estimated tilt angle of the face.
    - **LandmarkingConfidence**. The algorithm's confidence in its landmark determination (facial feature annotation) for this face.
    - **BoundingBox**. The estimated bounding box in pixels containing the majority of the face. Contains four X,Y coordinates separated by semicolons.
    - **EmotionSorrowLikelihood**. Estimated likelihood that this face is expressing the emotion "sorrow." Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **EmotionAngerLikelihood**. Estimated likelihood that this face is expressing the emotion "anger." Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **HeadwearLikelihood**. Estimated likelihood that this face is wearing some kind of headwear such as a hat. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **EmotionJoyLikelihood**. Estimated likelihood that this face is expressing the emotion "joy." Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **EmotionSurpriseLikelihood**. Estimated likelihood that this face is expressing the emotion "surprise." Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **UnderExposedLikelihood**. Estimated likelihood that this face is underexposed. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
    - **BlurredLikelihood**. Estimated likelihood that this face is blurred. Value is either -2 (Very Unlikely), -1 (Unlikely), 1 (Likely), or 2 (Very Likely).
- **OCR**. ("<RECORD>" delimited fields). This is the list of detected text regions found in the image. Each region is separated by "<RECORD>" delimiter and contains a block of contiguous text found in the image.