# DATA301 Project

In this series of artifacts, students will complete a short, application focused project using the GDELT 2.0 data set.

<https://www.gdeltproject.org/>

## Background

Please read this description of the GDELT project itself:

（GDELT的文件类型）

<https://blog.gdeltproject.org/the-datasets-of-gdelt-as-of-february-2016/>

The event data from this set has been used for analysis using text and images and geolocation data. Please read this article to get a basic idea:

一个运用新闻（反映情绪）与人脸图像（反映情绪）的分析，把全球的人脸图像识别出情绪，分别绘制在地图上，并结合当地新闻所反映情绪，推断冲突或者幸福事件发生概率

<https://www.forbes.com/sites/kalevleetaru/2016/01/13/mapping-world-happiness-and-conflict-through-global-news-and-image-mining/#66f3263ce224>

You should familiarize yourself with the Python package, gdelt:

GDELT包的信息 ----好像没什么卵用

<https://linwoodc3.github.io/gdeltPyR/>

Review the tables and columns of data available in the event, mentions, and GKG sets and their documentation: -----标准查询教学

<https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>

<https://www.gdeltproject.org/data.html>

You might be interested in full text searches of the articles:

----全文搜索

<https://blog.gdeltproject.org/announcing-the-gdelt-full-text-search-api/>

<https://blog.gdeltproject.org/gdelt-doc-2-0-api-debuts/>

You can also craft a query using the GDELT summary, extract the URL from the article list, and then use that URL to query the web API directly:

<https://api.gdeltproject.org/api/v2/summary/summary>

(see my example here: <https://colab.research.google.com/drive/1hXAeG6yheFUQiHfc9Z5ISfNBqAQw47Dq>)

You can also use Google’s BigQuery to fetch filtered data from the GDELT summary:

<https://colab.research.google.com/drive/11mF5gwgsoOPmwd5rDwDmDWMOUvoHoYsv?usp=sharing>

## Sample project code

I am providing a starting point for projects to help. See this minimal complexity project code that answers the question “How similar was the response of the US to the US election of 2016 to the responses of other countries to the US election of 2016?”:

<https://colab.research.google.com/drive/1sTsl_-f2ipgzqM6htsVdKjf4MZ3Ds2CW#scrollTo=fVf4e7R5gM-n>

additional sample (not a full solution, just some starter code to process raw text from an article):  
<https://colab.research.google.com/drive/1hXAeG6yheFUQiHfc9Z5ISfNBqAQw47Dq>

## Requirements (and when to meet them!)

1. **Students will complete their own individual project.** Students may share code and get help from other students in the course with planning, algorithms, etc as long as the individual student writes their own proposal, research question and design, and reports.  
     
   Please let us know in your final report if you have shared code or worked with anyone during the project (similar to listing code / papers you have used from online sources).
2. Artifacts (things you need to submit), project total = 40% of final course grade
   1. **[6%: May 6 at 11:59PM]** Project Proposal including Motivation, Research Question, and Design
   2. **[4%: May 20 at 11:59PM]** Progress Reports (1)
   3. **[15% June 3 at 11:59PM]** Software Implementation
   4. **[15% June 3 at 11:59PM]** Written report including Test Results, Critique of Design and Project Reflection
3. Methods

You will use python and Apache Spark to analyze the GDELT 2.0 data set. You need to formulate a research question that can be answered by the data set, choose an algorithm from the ones we have studied in the MMDS textbook, and implement this algorithm to perform a data analysis that answers your research question.

Subjective requirements: your algorithm must process a sufficiently large portion of the data set and involve inter-relationships between data records. Complexity: algorithms should involve multiple steps or combine multiple algorithms and heuristics / metrics to guide the analysis. Looking at the article listed at the beginning about sentiment analysis and other mapping algorithms might be a good start.

**Note: this is NOT a machine learning project** so certain algorithms not already covered in DATA301 are discouraged although we have done some predictive analysis and clustering.

REGION=australia-southeast1

ZONE=australia-southeast1-a

PROJECT=**superb-flag-345723**

CLUSTER=data301-lab4-**zhe27**-cluster

BUCKET=data301-lab4-**zhe27**-bucket