Monthly Report on "An Extensible Model for Deduplication of the GDELT Events Database"

Andee Kaplan (andee.kaplan@colostate.edu; 832-526-7947)

October 2019

Progress and Plans

1. Description of progress made against each deliverable during the reporting period

   Casey has written `R` code to compare the two models under consideration. He has run the Steorts model on a test data set and is continuing to improve the speed of the code.

   Additionally, I have pulled the West African GDELT data from the website and compiled into a single csv file. I am currently exploring appropriate blocking variables for the data that will be used for the record linkage model.

2. Brief description of significant results

   Casey's work has been focused on comparing the methods on a benchmark dataset from the Italian Survey on Household and Wealth which includes the 2008 and 2010 surveys. This dataset contains ground truth, and so serves as a method of comparison between methods. Additionally, The Steorts method has been performed on this dataset, so he is able to compare our results to ensure the code is written correctly.

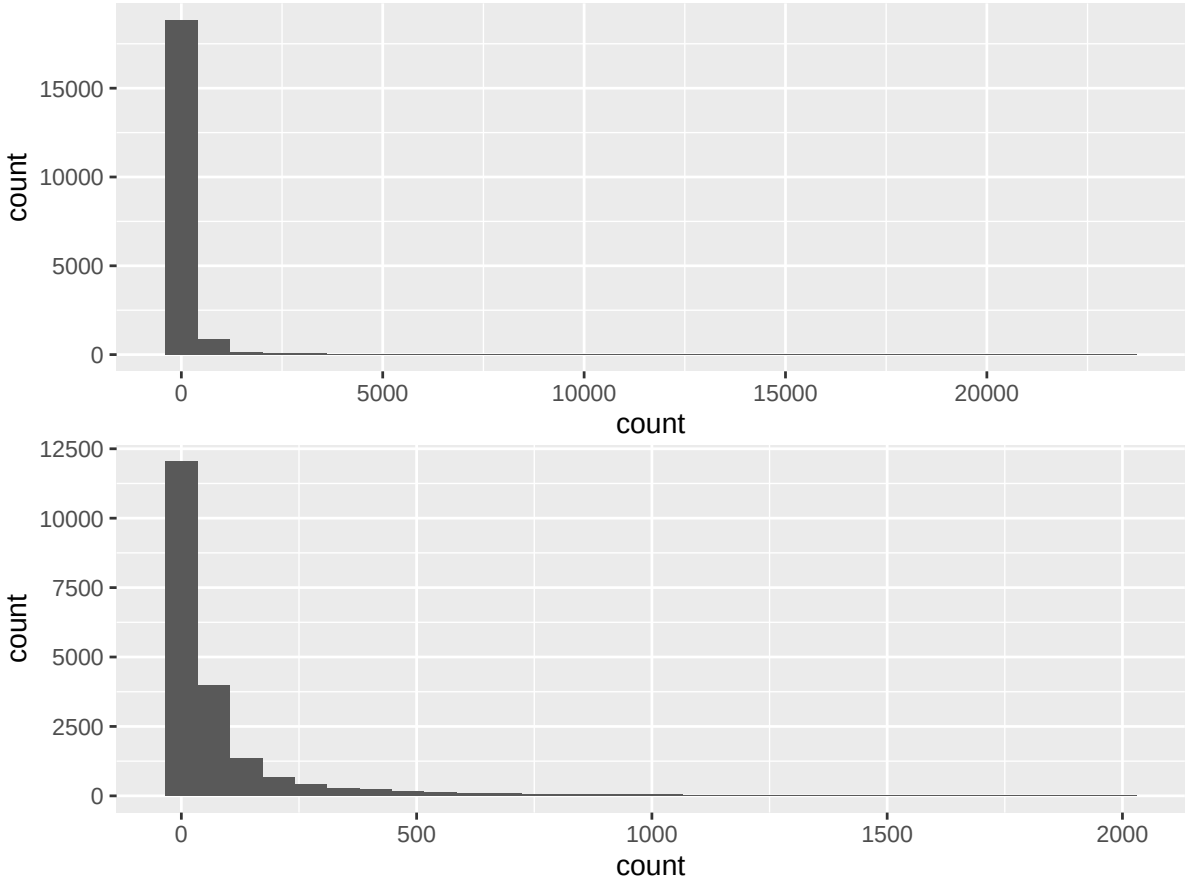
   Additionally, we have explored the possibilities of different blocking variables for the GDELT Data. The data we have downloaded ranges in date from 1979-11-01 to 2019-11-30, and there are 3,242,977 records to be deduplicated. We have 58 columns to work with, and they are

```
names(gdelt)
```

```
##  [1] "GLOBALEVENTID"         "SQLDATE"
##  [3] "MonthYear"             "Year"
##  [5] "FractionDate"          "Actor1Code"
##  [7] "Actor1Name"            "Actor1CountryCode"
##  [9] "Actor1KnownGroupCode"  "Actor1EthnicCode"
## [11] "Actor1Religion1Code"   "Actor1Religion2Code"
## [13] "Actor1Type1Code"       "Actor1Type2Code"
## [15] "Actor1Type3Code"       "Actor2Code"
## [17] "Actor2Name"            "Actor2CountryCode"
## [19] "Actor2KnownGroupCode"  "Actor2EthnicCode"
## [21] "Actor2Religion1Code"   "Actor2Religion2Code"
## [23] "Actor2Type1Code"       "Actor2Type2Code"
## [25] "Actor2Type3Code"       "IsRootEvent"
## [27] "EventCode"             "EventBaseCode"
## [29] "EventRootCode"         "QuadClass"
## [31] "GoldsteinScale"        "NumMentions"
## [33] "NumSources"            "NumArticles"
## [35] "AvgTone"               "Actor1Geo_Type"
## [37] "Actor1Geo_FullName"    "Actor1Geo_CountryCode"
## [39] "Actor1Geo_ADM1Code"    "Actor1Geo_Lat"
## [41] "Actor1Geo_Long"        "Actor1Geo_FeatureID"
## [43] "Actor2Geo_Type"        "Actor2Geo_FullName"
## [45] "Actor2Geo_CountryCode" "Actor2Geo_ADM1Code"
```

```
## [47] "Actor2Geo_Lat"          "Actor2Geo_Long"
## [49] "Actor2Geo_FeatureID"    "ActionGeo_Type"
## [51] "ActionGeo_FullName"     "ActionGeo_CountryCode"
## [53] "ActionGeo_ADM1Code"     "ActionGeo_Lat"
## [55] "ActionGeo_Long"         "ActionGeo_FeatureID"
## [57] "DATEADDED"              "SOURCEURL"
```

From these, we will use `MonthYear`, `ActionGeo_CountryCode` (the country that the action occured in), and `EventRootCode` (the broad event type) as blocking variables, which leads to the following distribution of block sizes.



There are 20163 blocks that result, with 318 over 2,000 records within them, which can lead to computational challenges.

3. Planned activities for the following reporting period

   We plan to start running a record linkage model on the GDELT data in November.

4. List of any LAS-funded trips during reporting period, with description of work presented

   None.

5. Description of any significant meetings/events held (e.g., a focused discovery activity, LAS Symposium) or conducted related to this task

   Casey and I continue to meet weekly for one hour.

Issues

1. Progress on/results obtained related to previously identified problem areas

   None.

2. Proposed significant changes to your methodology, goals, milestones, or deliverables

   None.

3. New challenges affecting technical performance or schedule, with background

   None.

4. Has anything happened to impact your anticipated schedule?

   Casey has broken his collar bone and has just had surgery (11/2) to repair. It is unclear how long he will be impaired, but this may make our progress slower than anticipated.