Library Analytics and Access to Data

Loni Hagen, Ph.D.
Assistant Professor of Data Science,
School of Information, University of South Florida
Data-Driven Decision Making

Utilizing library user data to help librarians plan for future programs and predict the needs of their users
User and Library Data

• **Institutional data**
• Collection of data set
Available Data and Possible Outcomes

• Evaluation of library performance
  – Annual statistics of Florida Libraries, Census, Satisfaction Survey
• Textbook affordability
  – Vendor data
• Collection development
  – Institutional data
• Research activities (trending topics in academia)
  – Social Media data and institutional data
• Literacy
  – Job posting + Social Media data
Possible Benefits of Data Analytics

• Efficiency in building collections
• Expenditure and budget justification

=> Evidence-driven decision making
User behavior data from vendors?

<table>
<thead>
<tr>
<th>User ID</th>
<th>Information source</th>
<th>Faculty/student</th>
<th>Faculty productivity</th>
<th>Page view</th>
<th>View time</th>
<th>Time of a day</th>
<th>Day of a year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predict the most valuable information source and better price for textbooks
Search input data can reveal user need/research trend
Library Data and Statistics

• Constantine Gillourakis: Florida Crime and Library Use (Links to an external site.)
• https://csgiallourakis.shinyapps.io/Gi allourakisFinal/ (Links to an external site.)
1. A Sample Student Project

Using Florida Library Statistics
The Goal

• Offense data from the Florida Department of Law Enforcement. This data shows each county, the total number of arrests and the population.

• Annual Statistics the Florida Department of State, Division of Library and Information Science.
  – This is an annual report provided by each library in the state of Florida that provides usage statistics such as circulation, borrowers, funding, and visitors.

• Our goal is to visualize the data and present out ideas to whether there is a correlation between library usage and arrests in an area.
The Datasets (Arrest and Offense data)

- Provided by the FDLE
- Includes 68 observations over 33 variables.
- After cleaning the data we opted to retain the county name, total population, total arrests and total offenses in each county during 2017.
- Several counties had to be modified to match county names for ggplot maps.

```r
dfNew <- merge(arrest, offense, by="county")
dfNew$population <- as.integer(dfNew$population)
dfNew$arrests <- as.integer(dfNew$arrests)
dfNew$offenses <- as.integer(dfNew$offenses)
head(dfNew)
```

<table>
<thead>
<tr>
<th></th>
<th>county</th>
<th>population</th>
<th>arrests</th>
<th>offenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>alachua</td>
<td>260003</td>
<td>9568</td>
<td>1645</td>
</tr>
<tr>
<td>2</td>
<td>baker</td>
<td>27191</td>
<td>1339</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>bay</td>
<td>178820</td>
<td>15202</td>
<td>899</td>
</tr>
<tr>
<td>4</td>
<td>bradford</td>
<td>27642</td>
<td>1334</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>brevard</td>
<td>575211</td>
<td>18945</td>
<td>2703</td>
</tr>
<tr>
<td>6</td>
<td>broward</td>
<td>1873970</td>
<td>53735</td>
<td>7543</td>
</tr>
</tbody>
</table>
The Datasets (Library Usage Data)

• This dataset is provided by a comprehensive annual report all libraries submit to the state. I downloaded several tables and merged them together in Excel to upload for visualization and analysis.

• About 30 libraries are autonomous of their county systems and had to be assigned to their county.

• Again, several counties’ names had to be modified as well.
Modified Scatter Plots

- These scatter plots are much more telling of the dataset than the original.
- We see, in arrests & circulation the higher the arrests the less the circulation.
  - This also doesn’t have a complete correlation, but it is insightful.
- For most variables, we see the data points cluster around the bottom-left corner. With some variation for high library usage and low arrests or high arrests and low library usage.
In this map, I wanted to (1) make sure we could graph the map of Florida; and (2) show data on the graph. Under ggplot, using `geom_map` it doesn’t allow us to map several states including Florida so we used `geom_polygon` instead.

This is the map of Florida colored by region. It is from the University of West Florida, Division of Anthropology and Archeology.

We saw this data fill in the points on the scatter plot.

```r
# Show a map of Florida by Region
flMap.region <- fl_base + geom_polygon(data = flNew1, aes(fill = flNew1$flRegion)) +
  geom_polygon(color = "black", fill = NA) +
  expand_limits(x = fl$long, y = fl$lat) +
  labs(title = "State of Florida by Region") +
  guides(fill = guide_legend(title="region"))
flMap.region
```
Florida Maps

- Seeing these graphs side-by-side is a better way to visualize the data over the map. As can see, several counties that are on the higher end of arrests have fewer library users and circulation. As well, several libraries that have a high usage proportion has much fewer arrests.
The Shiny App

- Being able to visually compare the charts side by side helped with analyzing the data on a county by county basis.

- [https://csgiallourakis.shinyapps.io/allourakisFinal/](https://csgiallourakis.shinyapps.io/allourakisFinal/)

- The above link brings you to the interactive platform.
Social Media Data

Learning trends and influential actors
What can be done?

• Learning trends
• Building multi-media communication strategy during crisis
Network of Actors during Zika Emergency
Authoritative Public Institutions
Twitterers with high PageRank (highly endorsed by influential people) are dispersed around the network. Topologically, these entities are positioned in between communities rather than in the middle of one community. Also, these authoritative entities are predominantly public institutions such as WHO, WHOSEARO, White House, and HHS or news agencies such as TheDailyEdge, NYTheath and NYTScience.

<Visualization of the PageRank>
Note: Larger font size means higher PageRank values.
Multimedia use during the Zika crisis

- Text tweets are sent out more frequently than multimedia tweets, but multimedia tweets have more retweetability.
- **Image and text** tweets were actively sent out by online news media and news aggregates,
- Whereas for **video** tweets, mainstream news media and mainstream journalists were more active agents.
Multimedia type and agent type

Figure 3a: Distribution of types of active agents (%)  
Figure 3b: Distribution of types of influential agents (%)
Topical Distribution by Multimedia Type

Figure 4a: Topical distribution of randomly selected tweets (%)

Figure 4b: Topical distribution of highly retweeted tweets (%)
Open Government Data

311 Citizen Request Data

Census Data
• Analyze citizen demands and living conditions
Phone call is the major methods for 311 requests

<table>
<thead>
<tr>
<th>Method Received</th>
<th>Freq.</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>48,869</td>
<td>76</td>
</tr>
<tr>
<td>Departments Service Request(^a)</td>
<td>7,526</td>
<td>12</td>
</tr>
<tr>
<td>City of Miami Neighborhood-Enhancement-Team</td>
<td>1,689</td>
<td>3</td>
</tr>
<tr>
<td>iPhone</td>
<td>1,450</td>
<td>2</td>
</tr>
<tr>
<td>Web</td>
<td>1,340</td>
<td>2</td>
</tr>
<tr>
<td>In-house</td>
<td>1,011</td>
<td>2</td>
</tr>
<tr>
<td>Email</td>
<td>978</td>
<td>2</td>
</tr>
<tr>
<td>Android</td>
<td>555</td>
<td>1</td>
</tr>
<tr>
<td>Others(^b)</td>
<td>834</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Method 311 Requests Received

\(^a\) Requests made through outside interfaces by public employees.  
\(^b\) Others include, requests made through fax, walk-in and social media.
Distribution of requests
Hurricane Irma and 311 Request
Communications Request by Area

Categorical distribution of 311 reports before Irma (2017)

Categorical distribution of 311 reports after Irma (2017)

(a) 2 days before Irma

(b) 2 days after Irma
311 service request patterns are indicative of underlying socio-demographic factors within the area (K-means clustering)
311 service request patterns are indicative of underlying socio-demographic factors within the area (Merging with Census Data)
Automatically extracting themes

Text Data

We the People e-petition text data
WE THE PEOPLE ASK THE FEDERAL GOVERNMENT TO CHANGE AN EXISTING ADMINISTRATION POLICY:

Divest or put in a blind trust all of the President's business and financial assets

Created by H.B. on January 20, 2017

In keeping with tradition and to avoid the appearance of conflicts of interest, corruption, and violations of the emoluments clause of the US Constitution, President Trump should divest his financial and business holdings or have them administered by a truly blind trust.

Sign This Petition

Needs 0 signatures by February 19, 2017

to get a response from the White House

360,327 SIGNED 100,000 GOAL

First Name *
Last Name *
Email Address *

THE WHITE HOUSE MAY SEND ME EMAILS ABOUT THIS AND OTHER ISSUES
<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Label</th>
<th>Topic words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>People</td>
<td>people time make country american stop government states</td>
</tr>
<tr>
<td>2</td>
<td>President Obama**</td>
<td>president obama congress states united petition act administration</td>
</tr>
<tr>
<td>3</td>
<td>Tax Budget*</td>
<td>tax federal pay government money dollars budget employees</td>
</tr>
<tr>
<td>4</td>
<td>Cancer Disease**</td>
<td>health care cancer disease research medical treatment patients</td>
</tr>
<tr>
<td>5</td>
<td>Election Clinton**</td>
<td>vote investigation election clinton investigate people federal party</td>
</tr>
<tr>
<td>6</td>
<td>Prison Sentence**</td>
<td>justice years prison case life trial court release</td>
</tr>
<tr>
<td>7</td>
<td>Terrorism Syria**</td>
<td>war terrorist people stop government terrorism genocide syria</td>
</tr>
<tr>
<td>8</td>
<td>Guns Firearms**</td>
<td>law amendment gun rights states bans weapons</td>
</tr>
<tr>
<td>9</td>
<td>Children Gender</td>
<td>children child women sex law sexual parents rights</td>
</tr>
<tr>
<td>10</td>
<td>Religion*</td>
<td>rights government religious human freedom god religion church</td>
</tr>
<tr>
<td>11</td>
<td>National Holiday*</td>
<td>day national american house holiday white awareness world</td>
</tr>
<tr>
<td>12</td>
<td>Water Park Energy**</td>
<td>water national energy park land oil areas gas</td>
</tr>
<tr>
<td>13</td>
<td>Police &amp; BLM **</td>
<td>police law officers violence enforcement officer black death</td>
</tr>
<tr>
<td>14</td>
<td>Internet Companies*</td>
<td>internet service information access companies small business government</td>
</tr>
<tr>
<td>15</td>
<td>Students School Education**</td>
<td>students school education schools student public children college</td>
</tr>
<tr>
<td>16</td>
<td>Ukraine Russia*</td>
<td>ukraine russian russia puerto sanctions japan ukrainian rico</td>
</tr>
<tr>
<td>17</td>
<td>Visa Immigration**</td>
<td>visa immigration united states status family green home</td>
</tr>
<tr>
<td>18</td>
<td>Military Veterans**</td>
<td>military service members veterans soldiers war army forces</td>
</tr>
<tr>
<td>19</td>
<td>White Anti Genocide**</td>
<td>white anti genocide countries whites racist word code</td>
</tr>
<tr>
<td>20</td>
<td>Http &amp; China*</td>
<td>http www org chinese people human china world</td>
</tr>
<tr>
<td>21</td>
<td>Animal*</td>
<td>animals animal dogs wild hong dog kong horses</td>
</tr>
<tr>
<td>22</td>
<td>Secession*</td>
<td>states united government state america people powers nature</td>
</tr>
<tr>
<td>23</td>
<td>Vehicle &amp; FAA**</td>
<td>vehicles safety vehicle faa aircraft air cars flight</td>
</tr>
<tr>
<td>24</td>
<td>Medal Award*</td>
<td>medal honor freedom award presidential game team american</td>
</tr>
<tr>
<td>25</td>
<td>Food Labeling**</td>
<td>food fda products foods health safe labeling ban</td>
</tr>
<tr>
<td>26</td>
<td>Marijuana**</td>
<td>marijuana drug cannabis medical schedule hemp states substances</td>
</tr>
<tr>
<td>27</td>
<td>Ebola &amp; TPP*</td>
<td>ebola trans media trump trade partnership people protect</td>
</tr>
<tr>
<td>28</td>
<td>FDA &amp; Blood</td>
<td>fda blood life india drug sri sikhs drugs</td>
</tr>
<tr>
<td>29</td>
<td>Mclellan</td>
<td>mclellan act iran veterans toxic nuclear congress health</td>
</tr>
<tr>
<td>30</td>
<td>Charly Wingate</td>
<td>charly robbery pardon vietnam max retrial wingate circumcision</td>
</tr>
</tbody>
</table>
Fig. 4. LDAvis results using $\lambda = 0.6(b)$ focused on “Police & BLM” topic.
Topic Change Over Time
Fig. 7. Google trends and topic popularity.
How can libraries leverage data analytics?

• Learning trends/demands
• Analyzing use
  – frequency
  – co-occurring use patterns (market basket analysis—display items)
To achieve library goals

• Communication policy
• Library evaluation
• Collection development
• Literacy education
References


- Hagen, L., Neely, S., Scharf, R., Neely., S., & Keller, T., Government Social Media Communications during Zika Health Crisis, ACM Digital Government: Research and Practice (DGOV) (Accepted)


