Evan Hendrickson, Carly Presz, Rafaela Hojda, Julian Esparza, Max Motz
City Data Catalog Revamp: Milestone 3
Professor Zheng
Fall 2022

Milestone 3

Our group is in the process of performing data collection and analysis to find the New Orleans city data that is most important, relevant, and useful to community members and organizations. We began the data collection process by creating a survey to ask community members about the data that is important to their organization and about the DataDriven.Nola.gov site. Our team met with Professor Cabalfin from the Tulane Design School to discuss how we could design our survey in order to get the most useful results. The city data survey was sent out in the November CPS Newsletter and our team is awaiting results from the community.

In addition to the data we will receive from our survey results, we have also obtained two important data sets regarding community data interests. The first contains public records requests, and the second contains user feedback from the city's DataDriven.Nola.gov data catalog site. Our team has performed analysis on this data using components of NLP such as topic modeling and sentiment analysis. To do this, we utilized a variety of Python libraries including pandas, BERTopic, and spaCy.

To determine the most highly requested data topics, we performed topic modeling on the public records requests using BERTopic. We trained 4 different models and produced visualizations of each model's results. The topics that appeared in the results of every model were those that we deemed the most relevant. Our first model had all of BERTopic's default settings and parameters, which we modified to create the other three models. We adjusted the number of words per topic entity, range of topic sizes, and the diversity level, which controls the similarity of words included in each topic to reduce redundancies. These modifications did produce different outcomes from each model, but there were some recurring topics that appeared in the results of all 4 models. We could therefore conclude that these topics were the most highly

requested. These topics were fires (investigations, incident reports), traffic camera footage, tax payments, and crime/police body camera footage.

We also used NLP and BERTopic to do similar analysis on website feedback data to use alongside the survey data to determine what users are asking for. Using the same system, the most often occuring word clusters from the dataset were related to: Ticket Payments, Waste Disposal (Trash and Recycling), Obtaining/Renewing permits, Government Officials, and Monument Removal (the recent process of removing confederate monuments).

On top of this, we have also started using other NLP techniques in an effort to further our analysis of the datasets we have access to. We have begun using NER or Named Entity Recognition on the public records requests dataset to see what entities are commonly mentioned in the queries. This knowledge is different than topic modeling as rather than coming up with different broad categories that fit multiple queries it searches for specific entities that are mentioned in the query. This allows us to find out which organizations or people are mentioned of in queries so that we can gain a better understanding of what kind data users want. For instance, if a certain street is mentioned in a query it will identify it and we can see if it is similar among all queries to find out why a user wants information on this. We have also started the process of performing sentiment analysis on a dataset that has user feedback on the overall DataDriven.Nola.gov website. This datasets main attributes are that users give a yes or no response to if the website was helpful and then are able to explain their problems with the website. Many of the complaints about the website had to do with the ease of navigation and accessing what the consumer needed. Using the spacy and textblob packages we have gone through the responses and performed a rudimentary sentiment analysis. We mainly searched for the polarity of the response, or how positive/negative the response was and also the subjectivity of the response. This should allow us to create a spectrum of responses and find which areas of the website users are having the most difficulty operating. Our hope with this information is to get a better idea of what users go to the website for and how the website can be better tailored to make the user experience easier.

Furthermore, we have started researching and preparing a base outline of what we hope and expect our dashboard to look like. The initial process of creating the dashboard will not incorporate any code and will be done completely on Figma. After it has been approved by the city and Professor Mattei, we will then take what was created on Figma and replicate it into a

functional website using code. The programming languages we will most likely use for the Front-End section of the dashboard will be HTML, CSS, and JavaScript as well as some other smaller libraries. So far, our work on the dashboard is only in its incipient stages. We began by researching different data science dashboards and comparing them to each other in order to collect insight on how to create a structured and functional dashboard. From the information we collected, we then started designing a very basic outline of our dashboard. As our outline becomes more specific, it will become the foundation of our dashboard that we will continuously build on as our project continues to develop.

Alongside our dashboard we will create a website page which will serve as the "Homepage" for our dashboard. This "Homepage" will include a description of our capstone project and a brief description of the DataDriven.Nola.gov website. It will also include pictures of each member of our team and the names of other contributors that helped guide us in our project. The page will serve as a navigation tool where users can click to be directed to either a separate page, which will be the dashboard, or the DataDriven.Nola.gov website.

In our initial timeline, we wanted to have our City Data Survey sent to community partners and organizations by the end of October. However, due to delays from administrators, the survey was not sent out until mid-November. However, hopefully results will become available soon so that we may form a concrete, finalized idea that the City approves by the end of this month. We have already begun utilizing NLP tools for our analysis, and found the most highly requested topics, and once we have the survey results, we will be able to begin other analyses on the applicable datasets that fill within the topic(s).

By the end of the semester/mid-December, we hope to have a working draft of a dashboard. We still plan to use Dash to build this interactive dashboard that visualizes the data found to be most desirable and needed based on the survey results. We still plan to have the dashboard completed by the end of February so that we may make any necessary improvements before the SSE Expo in April. We do not anticipate any future issues that would require us to adjust our timeline again; the main barrier to adhering to our timeline was sending out the survey since that was not something within our control. If something were to arise, however, we will pivot and ensure that we are on track to meet our February deadline.

The workload and roles for each team member up until now have been distributed equally and not task-specific. When working on the survey, each of us added questions and graphics,

were engaging with Professor Cabalfin during our meeting, and revising the survey based on his suggestions. We each have also conducted preliminary analyses on the public records requests and user feedback data we already have. Moving forward, the workload will continue to be distributed equally, with individual tasks assigned. These tasks will be delegated based on the following roles: Julian Esparza and Carly Presz will be the team Data Scientists, Evan Hendrickson will be the Data and Dashboard Manager, Max Motz will be the Back End Developer/Manager (conducting data analysis until back-end is more fleshed out), and Rafa Hojda will be the Front End Developer/Manager. Our team will continue to communicate regularly to ensure that we are on track with our timeline, that we are held accountable for our individual progress, as well as set dates and times to meet weekly or more if needed.

References

Bertopic · spacy universe. BERTopic. (n.d.). Retrieved November 16, 2022, from https://spacy.io/universe/project/bertopic

BERTopic. GitHub. (2022, September 11). Retrieved November 16, 2022, from https://github.com/MaartenGr/BERTopic/blob/master/README.md

Explore Open Data. DataDriven. (n.d.). Retrieved November 16, 2022, from https://datadriven.nola.gov/home/

Interactive topic modeling with Bertopic | Towards Data Science. (n.d.). Retrieved November 16, 2022, from https://towardsdatascience.com/interactive-topic-modeling-with-bertopic-1ea55e7d73d8