Checkpoint 3: Interactive Visualization and Data Exploration Findings

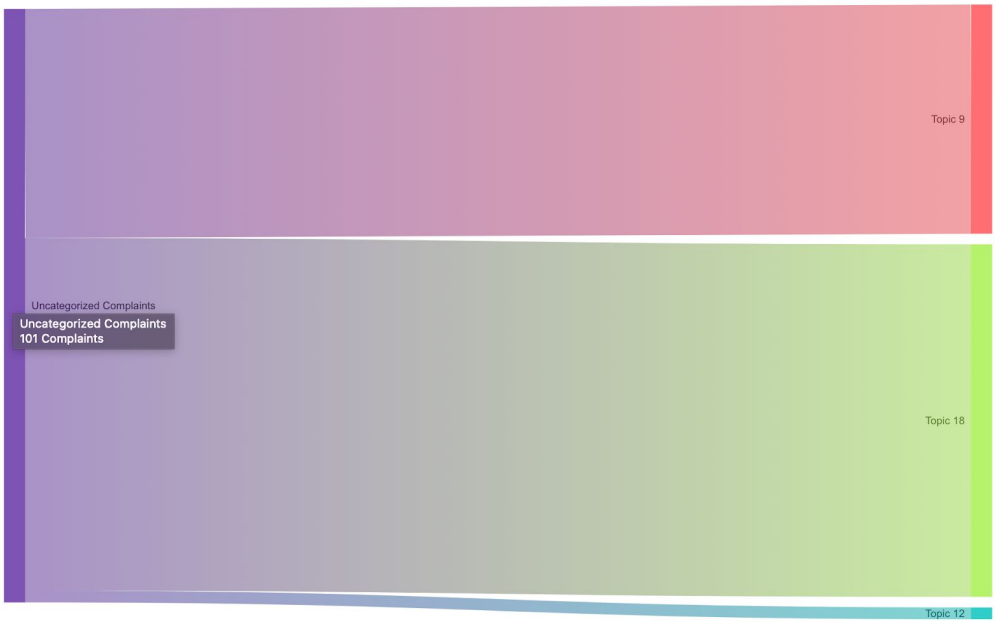
The Enchanted Badgers
Alexander Einarsson, Sergio Servantez, Marko Sterbentz
November 5, 2020

The theme of our project is to identify meaningful categories for uncategorized complaints by analyzing the relationship that exists between the complaint category and the complaint report narrative, and to use this new information to explore various aspects of complaint investigations and outcomes. To this end, we sought to answer the following set of questions by producing interactive visualizations:

1. Sankey diagram to visualize how uncategorized or miscategorized complaints are recategorized by our LDA topic modeling.
2. Filtered Sankey diagram to visualize how 10 different LDA topic models classified preexisting complaint categories into topics.
3. Hierarchical Bar Chart to visualize how preexisting complaint categories are classified by our LDA topic modeling.

The visualization and analysis are provided below. Additionally, we have provided a discussion on our use of Latent Dirichlet Analysis (LDA) to produce new categories for the complaints based on their summaries in the Appendix at the end of this document.

Question 1: Sankey diagram to visualize how uncategorized or miscategorized complaints are recategorized by our LDA topic modeling. *(originally: Chord dependency diagram to visualize how uncategorized or miscategorized complaints are re-categorized by our topic modeling.)*



https://observablehq.com/@servantez/checkpoint-3-interactive-visualization-and-data-explorat/3

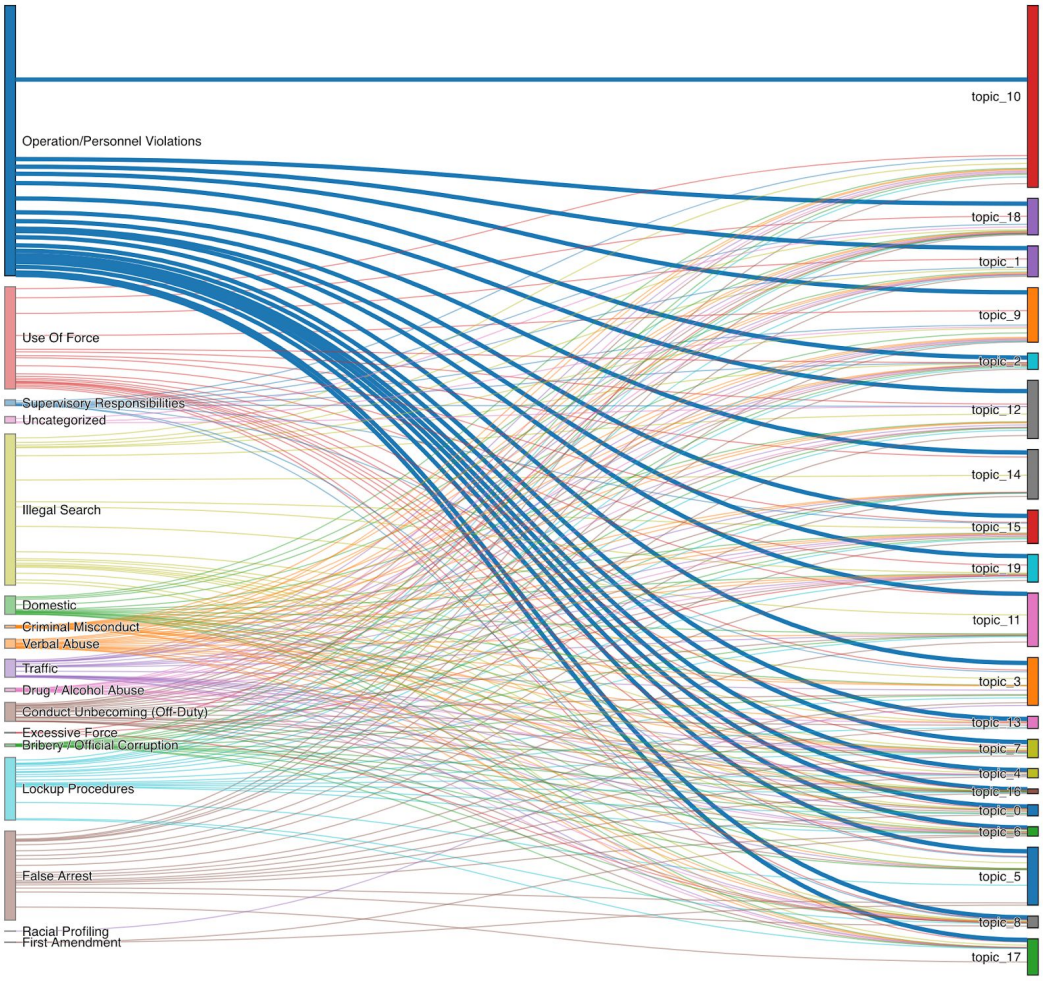
The Sankey diagram for visualizing how previously uncategorized or miscategorized complaints were re-categorized using our topic modeling is shown above, and an interactive version of it is available at the link above as well. We were originally planning to visualize this relationship using a chord dependency diagram, but changed this since chord dependency diagrams are typically used for showing the two-directional relationships between classes of things, rather than the one-directional relationship that exists from classifying complaint categories into LDA topics.

As discussed in our last checkpoint, we found the set of all complaints for which summaries existed and had no category defined for them (the value for this column in the CPDP database was either “Unknown” or NULL). We then used Latent Dirichlet Analysis (LDA), an unsupervised method to produce a new set of categories based on the complaints summaries we have access to. This visualization thus shows which topics the uncategorized complaints were classified as, according to one of the LDA topic sets.

It’s worth noting that we expected there to be a greater variety in the spread of topics the uncategorized complaints would be grouped into. This makes the visualization a little less interesting than we hoped (which is part of the reason we opted to produce a third visualization of the data), but this still provides some valuable insights. As we can easily see in the visualization, the 101 uncategorized complaints were classified into just three topics out of the 20 total topics that were produced by the LDA, with most belonging to just two of those topics. This suggests a high level of similarity in the types of complaints that were originally uncategorized. As a result, it would be interesting to perform a more in-depth qualitative analysis to see if these categorizations make sense to a human observer, as well as determine the appropriate labels for these topics.

Additionally, the small number of topics that the uncategorized complaints were grouped into may indicate that there are particular types of complaints that are being underreported. There are many possible reasons for this. It’s possible that it is simply the result of these complaints being complicated and not neatly fitting into the existing categories, thus resulting in a person being unsure of how to categorize them. However, there is also the potential that these complaints contain extremely serious allegations and the department is intentionally trying to bury these types of complaints or affect the final outcome. As part of future work, it would be interesting to examine the text of these uncategorized complaints more closely and see if either of these possibilities is supported by the data. It would also be very interesting to see how the complaint category and summary text affects the final outcome of the complaint and examine whether a complaint’s summary text and category is predictive of the final outcome.

Question 2: Filtered Sankey diagram to visualize how 10 different LDA topic models classified preexisting complaint categories into topics.



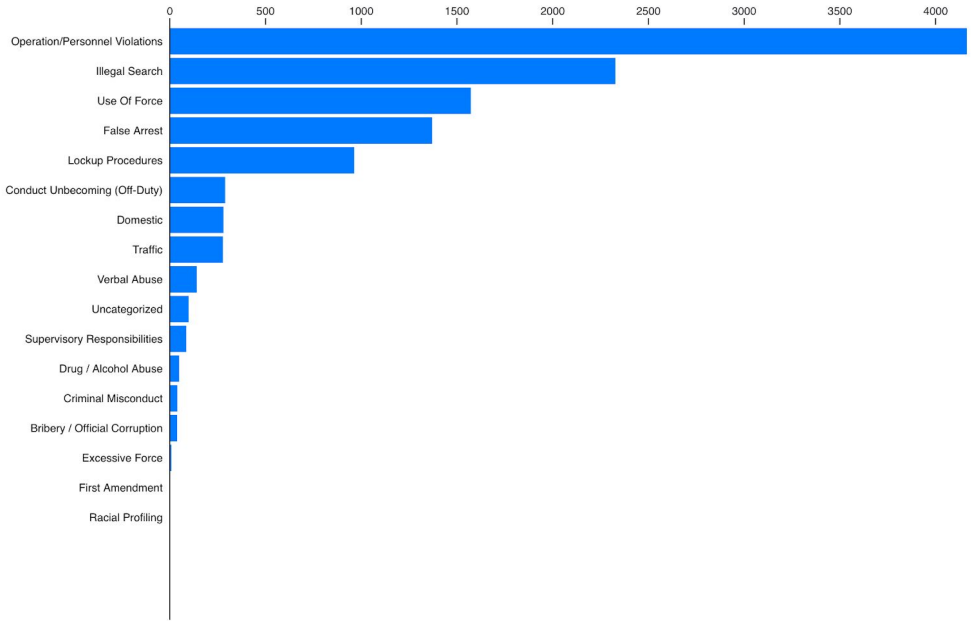
https://observablehq.com/@servantez/checkpoint-3-interactive-visualization-and-data-explorat/2

LDA topic modeling is nondeterministic which means different models are generated even when training on the same dataset. LDA is also an unsupervised machine learning technique so it is difficult to evaluate these models for ground-truth accuracy. We didn't want to assume that our first model would be a good representation of the groupings that exist within the complaints. Thus, we generated 10 different LDA topic models each containing 12 topic keywords in order to explore the potential groupings. To evaluate these models we needed a way to visual the distribution of how each preexisting category was classified by the topic modeling. With these considerations in mind, we chose to use a filtered Sankey diagram where the user could select which LDA model to view. A link to this visualization is provided above.

The above Sankey diagram shows how each of the preexisting complaint categories is classified across the 20 topics. We chose 20 topics to match the current number of complaint

categories. Hovering above a specific complaint category will interactively highlight the flow of that category to the respective topics. The user can easily explore differences in the models using the top dropdown menu to select the desired LDA model. Users can also change the color scheme of the visualization using the second dropdown menu. They may choose to color the diagram links by input or output. This allows the user to visualize both the outflow from complaint categories and the inflow to specific topics. Using this visualization and the keywords for each topic (see instruction for where to find this under question 3), we selected LDA model 7 as our final model based on distribution and topic keywords that readily distinguished the groups. Our third visualization takes a deeper look at this topic model.

Question 3: Hierarchical Bar Chart to visualize how pre-existing complaint categories are classified by our LDA topic modeling.



https://observablehq.com/@servantez/checkpoint-3-interactive-visualization-and-data-explorat

Above is a hierarchical bar chart breaking down the distribution of each category into different topics for a specific LDA model. We chose the LDA model via qualitative analysis, where we found that this model had, to us, the best combination of interpretable topics as defined by their words and a good distribution of complaints to topics (not all visualizations are included in this analysis - some included in the Checkpoint 3 deliverable under the “summaries_and_freq_graphs” directory - see appendix for more information). We found that this interactive visualization was a good help in understanding how the LDA models redistribute the categories into topics, without having to create an abundance of graphs for each category.

By drilling down on each original category we can see how our model redistributed the categories based on their narrative summaries. It appears that most of the complaints classified as Operation/Personnel Violation (the plurality category in the dataset) had summaries that

were similar enough for the LDA to classify them together, onto Topic 10 in this case (for example summaries and the words that define the topic, see the “summaries_and_freq_graphs” and “lda_models_and_test_files” directories in the Checkpoint 3 deliverable).

However, looking at the next highest category, Illegal Search, we find that it’s roughly split down the middle between topics 11 and 14. Both of these topics in our model are search related, but while topic 11 clearly focuses on vehicle search (includes words such as ‘stop’, ‘vehicle’, ‘traffic’, and ‘impound’), topic 14 appears focused on apartment search (includes words like ‘warrant’, ‘apartment’, ‘enter’, and ‘residence’). Other notable categories include Use of Force (four different topics dominate the distribution), and Conduct Unbecoming and Verbal Abuse (well distributed among many topics).

The classification of ‘Uncategorized’ complaints is discussed above, but it’s of note that the vast majority of ‘Uncategorized’ complaints are recategorized similarly to the Supervisory Responsibilities complaints, and largely seem department related based on the words in those topics (such as ‘department’, ‘duty’, and ‘misconduct’ for topic 18, and ‘department’, ‘member’, and ‘incident’ for topic 9). It’s unclear whether it’s obfuscation or lack of clarity about how to file department related complaints about officers that caused this, and without an in-depth qualitative analysis about the department procedure it’s out of scope for this project to come up with any hypotheses to test.

For future work, it should be determined if these recategorizations make the categories more concise, and if so whether the categorization can be kept to 20 classes and automatically classified based on the narrative summarization, rather than categorized by a police officer.

We will go more into detail about the LDA models, how we chose the final model, and what we can find out about the data by using that model in Checkpoint 4.

Appendix: Using Latent Dirichlet Analysis to Create Complaint Categories

For these questions we wanted to evaluate the original categories given to us, so we opted to train a model using Latent Dirichlet Allocation (LDA) to redistribute the complaints into 20 new categories based exclusively on their summaries. For this, we needed a large set of summaries, and as described in the previous checkpoint, we originally only had ~1100 complaints that contained narrative summaries - but because we had already put effort into getting an additional batch of narrative summaries, we simply added the summaries from “data/raw/complaints.json” to the set of original summaries for the training.

While our exact topics can’t be recreated due to the nondeterministic nature of the LDA, a user can create their own 20 topics, bar charts visualizing the distribution of topics for each model, and accompanying CSV files which were used to create our interactive visualizations by following our process. From a terminal, make sure you have a new virtual environment setup and run the following commands.

1. Run “pip install -r requirements_lda.txt”
2. Run “python -m spacy download en_core_web_lg”
3. Run “python lda.py”. This file preprocesses the narrative summaries and feeds it into Scikit-Learn’s LDA. We train 10 different LDA models, each with 20 topics described by 12 words each.
    a. This choice was driven by qualitative analysis. We opted for 20 topics because we wanted roughly as many categories as for the original complaints. The 12 words per topic was largely subjective - we decided on it because we found the different topics were more explainable with that number of words, in addition to not having single words show up repeatedly over many topics - a word helping define a few topics is good, but if it defines many topics it becomes meaningless.
    b. “lda.py” outputs a new directory “lda_models_and_test_files” which contains the models and the words that define the topics for each model
4. Run “python lda_topic_freq_and_summaries.py”
    a. This outputs a new directory “lda_topic_csv_files” with accompanying csv files used for the interactive visualizations
5. Run “python csv_script.py”
    a. This inserts “aggregation.csv” into the “lda_topic_csv_files” directory, which is the final file needed for the interactive visualizations

With this done, you can see the generated topic clusters and example narratives for each topic and lda model in the “lda_models_and_test_files” directory. You can see the distribution over the topics for each model in the “.png” files in the “summaries_and_freq_graphs” directory, and the data visualized via our interactive visualizations in the “.csv” files in the “lda_topic_csv_files” directory.