Project B - Final Report

Postcard Collection

D’Amico, Christian - Nieto, Andrew - Laycock, Kyle - Sandy Jacob
CSE 470/570 - Machine Learning
Prof. Giabbanelli, Philippe
December 12, 2019

Index

1. Introduction

Technology and the traditional library were originally disconnected. However, as technology grows we are able to utilize all the information stored in these libraries and better both parties from doing so. Resource digitization has helped bridge this gap significantly. By using machine learning, a lot of information can be derived from these digital academic libraries that would not be easy to see from the physical collections.

Attempts have been made to increase the effectiveness of the academic libraries through the use of technology, and we'll be looking specifically at the special collections of these libraries. Analysis of what gets put into these collections has been done in an effort to maximize the use of the collections that the library chooses to collect (Dupont). The main conclusion from that work was that users should be the focus of collections. Another study shows the importance of the user interface for these special collections. Labeling the collections in a way users can find is crucial so the information is usable by those who need it (Burns). User usability is our main focus for assessing these collections.

In this project, we explored Miami's Postcard Special Collections to see if we could aid users' experience when interacting with the database. Our objectives were as follows:

- I. Automatic comprehensive quality assessment of Special Collections data
- II. Data augmentation so that users of Special Collections can efficiently use some of the hidden patterns
- III. Domain-specific research questions

2. Automatic Comprehensive Quality Assessment

2.1 Summary

The first task after exploring the data was to clean the data. Data cleaning is a way to make the data better for the later tasks we perform. We look for parts in the data that have errors or inaccuracies and fix those. This helps keep the data consistent and thorough. For example, some postcards had the same unique identifier. This is important to note because we don't want duplicates in our data. We will now go through every column of the data and describe if it needed cleaning, and if so, what that cleaning entailed.

2.2 Process

1. **Card No. (Identifier) -** In most cases, having the **same Card Number** means that we're referring to the **same postcard**. That's not always true, however, in fact in one case we had 2 identical postcards and one different. Generally, this is due to a double digitation or an update; in one case, the back was uploaded as an extra. We went ahead and removed the repeated row, keeping the most current.
2. **Title -** Some of the postcards had a title that matched the Card No., meaning they really had no title. Every postcard did have a title though, so we didn't need to do anything to this field. We left it as was.
3. **Series -** The series field was not mandatory, so only about 600 cards had a series. When we looked at the data, there were 6 different values for the series field; Flood of 1913, Land of the Cross-Tipped Churches, War and Conflict, Flood of 1913;, Land of the Cross-Tipped Churches;,

and War and Conflict;. As you can see, there’s really only 3 series, with the other 3 being duplicates plus a semicolon. We fixed this in the cleaning process, combining the similar ones, giving us 3 series.

4. **Caption (front) + Caption (back) -** Most postcards are missing a caption but that’s not a big deal. We left this section alone as it’s supposed to be independent and we don’t want to mess with it.
5. **Message -** Cleaning of the messages would be very complex because the message can be almost anything. Some messages are a manual digitization of the card. We left this field alone.
6. **Subject -** The subjects was a very complicated area of focus. There are 3716 unique subjects, albeit, some appear multiple times such as ‘Water’ and ‘Bodies of Water’. We were able to combine some subjects by making them all uppercase and trimming extra whitespace (‘ water ’ became ‘WATER’, ‘trees’ became ‘TREES’). This cut the subjects from 3716 to 1884. The problem with the subjects is that some are very general, while others are extremely specific, such as “Sugar Maple Tapping”. We found that 28.66% of subjects appear only one time. We were not able to generalize or cluster the subjects, but that would be a great step to take.
7. **Location -** We picked apart the Location and created two new columns, State and Location. If the postcard was another country, we left state blank and just put the location.
8. **Coordinates -** There were three types of data in the coordinates field. There were real coordinate (X, y) pairs, there were street address, and there were misspellings of coordinate pairs (41 11.284, -80 58.281). We removed the values that did not match the correct form of a coordinate pair, and fixed the ones that were just missing a period.
9. **Photographer, Printer, and Publisher -** We did not need to change these fields because they were already clean. We also did not use these fields for any of our tasks.
10. **Date Printed -** There were a few errors here. One postcard had the date printed as “1900; 1901; 1902; 1903; 1904; 1905; 1906; 1907; 1908” so we removed that. The other error just required extraction of the date from a datetime object.
11. **Period -** There were 7 unique periods. We did not need to change this field.
12. **Language -** More than 99% of the postcards were written in English. Some had the language listed as English and another language, so we made a column for each extra language.
13. **Date Postmarked -** Some of the dates just have a year, or have an incorrect format. We fixed this by standardizing the format.
14. **Postage -** This field is meant to be a “Yes” or “No”, but we have some errors. Some have misspellings while others have a date and time, possibly the date postmarked. We fixed the misspellings or removed the values that were not yes or no.
15. **Recipient's Name -** Recipient's names are rich of misspellings, but still can give us an idea of the targeted person/family/office. We applied some cleaning that can solve some of the issues. We also added a column for the cleaned name and kept the original.
16. **Recipient's Address -** We were not able to clean the address field because many addresses have different forms.
17. **Notes -** For the sake of preserving the notes on the card, we did not want to alter this field.
18. **Decade -** Some fields had duplicates with semicolons on the end, similar to the series field, so we combined those.
19. **Genre -** There is only one value for the Genre field, and that is “Postcard”, so since they’re all the same, we dropped this attribute.
20. **Digital Publisher -** The only value for this field was “Miami University Libraries” so this is the same situation as Genre, we dropped this field.
21. **Copyright -** This also only had one value, a long sentence about the copyright information of the card. We dropped this as well.

22. **Digital Production -** We did not need to change this field.
23. **Collection Name -** Every postcard is apart of the “Bowden Postcard Collection Online”, so we dropped this column.
24. **Repository -** There are two groupings for this one and both are clean. We left this data alone.
25. **Resource Type -** We did not need to change this field.
26. **Digitized by -** This contained a list of initials, with a few entries being the full name. We converted the full names to the initials, standardizing the column.
27. **Problems -** We left this field alone and did not change it.
28. **Logo -** Only 9 cards had a value for this so we dropped the entire column.
29. **Medium -** Only 1 card had a value for this so we dropped the entire column.
30. **OCLC number -** Only 1 card had a value for this so we dropped the entire column.
31. **Date created -** We did not need to clean this value as there were no errors.
32. **Date modified -** We did not need to clean this value as there were no errors.
33. **Reference URL -** Only 4 postcards don’t have a URL and the ones that do have a URL are all unique. We left this one alone.
34. **CONTENTdm number** + **CONTENTdm file name + CONTENTdm file path -** We kept the file path and dropped the other two columns because the file path contains the name and number.

2.3 Conclusion

As you can see, cleaning data was a very important step in our process. It is important to have data that is all in the same format. If you are comparing dates and one is MM/DD/YYYY and the other is DD/MM/YYYY, there is going to be some issues. Cleaning the data solves these issues before they become a problem. The following table shows the most common errors in the data and how to fix them:

Error	**Recommended Fix**
1. Too much variation in the subjects	Try to group the subjects and narrow it down to maybe 100 or so.
2. Entries like coordinates and dates have many different formats.	Make all the dates and coordinates the same format
3. Some columns are almost completely empty or all have the same value.	Drop columns that are rarely used, like Logo, Medium, and OCLC number. Also drop columns that are all the same, like Collection Name.

3. Data Augmentation

3.1 Summary

The process of gaining insights from the data included the analysis of hidden information that required a preliminary step of *extraction*. Augmenting the data, in fact, is a process of producing new features based on some accessible knowledge that requires manipulation or some sort of computation that will ultimately increase the value of the available resources.

In the postcard collection dataset, it was provided a link to an external source that was intended to be accessible by humans as a summary of each individual postcard. That link was provided in the ‘Reference URL’ column and pointed to a digital filesystem (e.g., https://cdm17240.contentdm.oclc.org/digital/collection/postcards/id/2488). At first glance, it seemed that most of the information given here was already provided into the dataset (Title, Collection, Publisher, …) but one: the postcard’s picture.

The picture is one of the most important aspects of a physical postcard (it is the main reason why somebody decides to buy it, or even before, to print it or, originally, to take a photograph) and holds plenty of intelligible knowledge that a human could immediately understand. One attempt of describing the picture was made in some of the features of our dataset (especially in the ‘Subject’ column) but they didn’t provide a complete depiction of it.

Pictures, for example, could be used in Machine Learning to automatically detect objects (and thus providing a more *standardized* generation of fields like ‘Subject’, that may be influenced by trend topics or the actual person who digitized the information); an attempt was made for that purpose but, not being satisfactory enough, it won’t be described here (other forms of documentation are available if needed). Instead, one of the most successful attempt was made on identifying the colors.

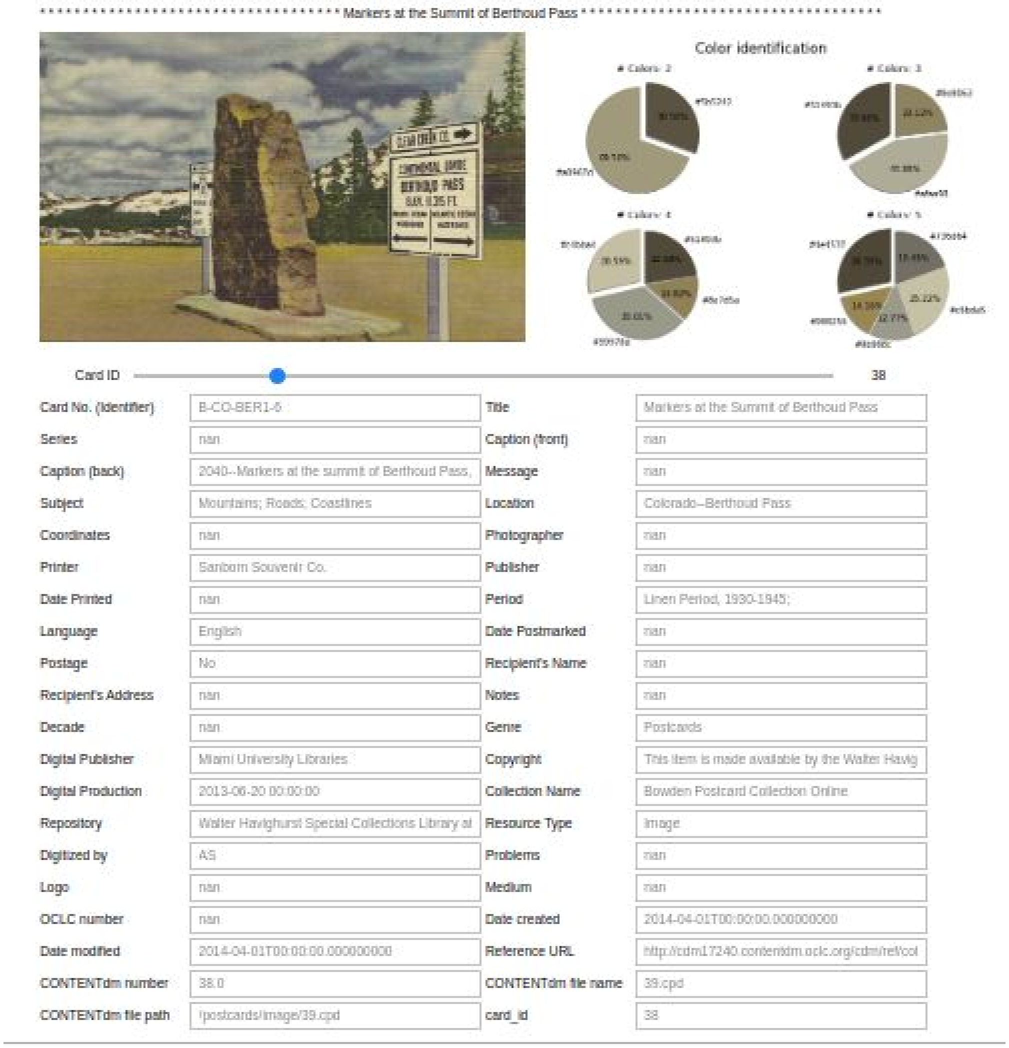
Initially, color identification was thought to serve the goal of making some analysis about feelings, epochs or detecting unwanted data (like the back of the postcards that could’ve been duplicated instances). Surprisingly, the customer had more ambitious ideas that could practically exploit this information on their business. Extracting this information was very valuable for the library’s customers who usually search this kind of collections by colors, but can be also used by artists or in many other scenarios.

3.2 The process

In order to extract the colors from the postcard’s pictures, it was needed to have those pictures available as one of the data sources. The ‘Reference URL’ was the door to access it but an extra effort was required to find a way to actually download the images to be stored in a local filesystem. Luckily, the source code of the referenced webpage was holding the plain URL of the image in one of the embedded widgets and it was possible to find a pattern that allowed to access all the images starting from an identifier. This identifier will now be known as a ‘card_id’ but, in the original dataset, it was provided as the ‘CONTENTdm number’; this column required some wrangling, including removing missing values and casting to integer. The identifier, filled in an appropriate spot of a common URL, will lead to the image file; sometimes this identifier required a little offset to point to the correct resource (e.g. subtracting 2 to the given number). Once the pictures were downloaded, they were placed in a local filesystem; since each picture was about 1MB of size, the total of them (~21500) summed up to a total of ~2GB.

Those pictures were now accessible and it was possible to apply some Machine Learning on them. The tools that we used to extract colors were scikit-image (an implementation of scikit-learn dedicated to images) and OpenCV (an image processing tool). In particular, once the image is read as a sequence of bytes, it is possible to compute the colors composing it. Since that amount can be huge but meaningless (some colors may appear once or maybe a different variety of a color in the same context; one example of it can be the different tonalities of blue in the sky), it was required a stage of clustering that grouped all the similar colors to a unique value and returned a weight for it. The KMeans classifier was also called with a parameter asking for a specific number of clusters (and, thus, a certain number of colors to be extracted).

Deciding the sufficient amount of colors to extract required some experiments and it was possible thanks to some graphical widgets that allowed to compare multiple pie charts. The satisfactory amount may vary but, in the end, it seemed that having 3 colors is a reliable result. Once it was clear what was a reasonable amount of color to extract it was necessary to make this process offline so that a final user should not worry about extracting them on the fly. Extracting colors, in fact, was a time-consuming process and, in our case, took about 2.4 seconds for each postcard. This number may look affordable but, if scaled up to 21500 postcards, it will lead to a total of ~15 hours to be completed. That time can vary according to the resources (if using a single machine, like we did, or a cluster) and the number of requested colors (requesting more colors will take more time to execute). This number is expected to increase considering that the actual collection is not including an enormous amount of postcards that were not digitized yet. In the end, a separate file was provided to describe the colors and their relevance (e.g. color: #000000; relevance:40%).



************************************ Markers at the Summit of Berthoud Pass ************************************

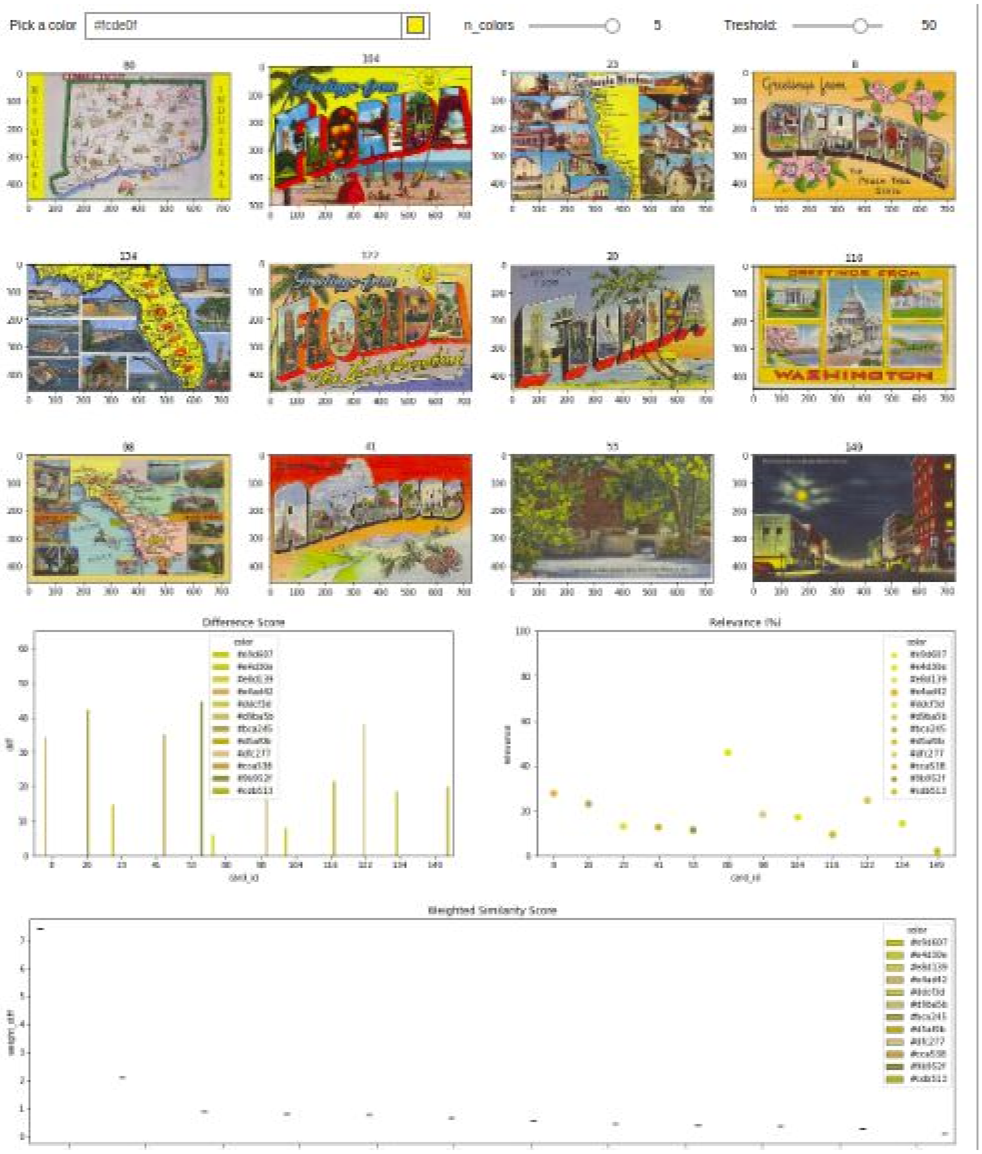
Card ID 38

Card No. (Identifier)	B-CO-BER1-6	Title	Markers at the Summit of Berthoud Pass
Series	nan	Caption (front)	nan
Caption (back)	2040--Markers at the summit of Berthoud Pass,	Message	nan
Subject	Mountains; Roads; Coastlines	Location	Colorado--Berthoud Pass
Coordinates	nan	Photographer	nan
Printer	Sanborn Souvenir Co.	Publisher	nan
Date Printed	nan	Period	Linen Period, 1930-1945;
Language	English	Date Postmarked	nan
Postage	No	Recipient's Name	nan
Recipient's Address	nan	Notes	nan
Decade	nan	Genre	Postcards
Digital Publisher	Miami University Libraries	Copyright	This item is made available by the Walter Havig
Digital Production	2013-06-20 00:00:00	Collection Name	Bowden Postcard Collection Online
Repository	Walter Havighurst Special Collections Library at	Resource Type	Image
Digitized by	AS	Problems	nan
Logo	nan	Medium	nan
OCLC number	nan	Date created	2014-04-01T00:00:00.000000000
Date modified	2014-04-01T00:00:00.000000000	Reference URL	http://cdm17240.contentdm.oclc.org/cdm/ref/col
CONTENTdm number	38.0	CONTENTdm file name	39.cpd
CONTENTdm file path	/postcards/image/39.cpd	card_id	38

3.3 Applications

Having the colors can allow us to perform the aforementioned analysis (on the sentiments, to connect colors and feelings, epochs, to connect to technologies in time, …) but also open a wide range of usages that may not be thought yet. The librarian indeed cited some use cases that can be exploited by artists or colorblind people but there may be some other users interested in having this information.

Filtering is one of the most important practical usages of the extracted colors. Searching from a palette can be useful for anyone exploring such a huge collection and can be made very easily: the only requested parameters are a target color (being a reference) and an acceptable threshold. The reference color will be compared to the extracted ones and, if it is below the acceptable similarity threshold, their difference (similarity score) will be weighted according to their effective relevance (the more a color is relevant, the more likely it will satisfy the search).



Pick a color #fcde0f

n_colors 5

Treshold: 50

Similarly to the color extraction process, filtering can be time-consuming because it is required to compute the difference among all the colors of each postcard before finding the maximum weighted difference score (that indicates what is the likelihood that this postcard is matching the referenced color). In that case, conversely, it is impossible to do that offline because the search is typically made on the moment. Though, this could be improved by determining a subset of frequently asked colors or some standard ones. Also, having a scalable distributed implementation of such a filtering algorithm can help in speeding up the process (maybe using a cluster).

3.4 Conclusions

Gaining insights from the picture of the postcards seemed ambitious but potentially valuable from the beginning. This represented one of the most straightforward ways to augment our data based on a not so explicit feature. Though, the real potential was hidden and was only discovered after the customer had a chance to see the results and was able to propose some possible usages of that knowledge.

An effort was made to document the entire process and a stable version was provided. Anyway, considered the initial goals and the actual situation, it is not excluded that there's something still hidden. One of the mentioned approaches that may deserve a further trial is that of the Object Detection that may require dedicated work.

4. Domain-Specific Research Questions

4.1 Summary

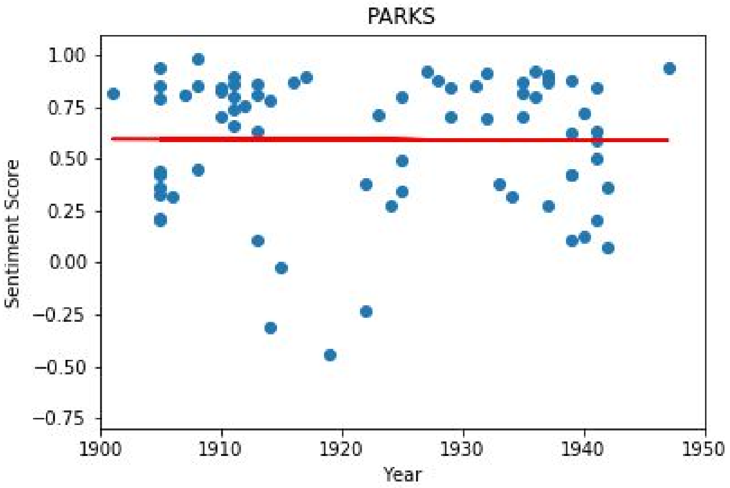
One interesting topic we wanted to explore with the data was whether or not their was an increase in sentiment over time. Meaning, for each subject, did the messages on the postcards change throughout different eras. We found out that sentiment per subject did not change over time. Sentiment for each subject was either sporadic, or consistently high.

4.2 Process

We measured the sentiment for messages using a third-party Python Library called Vader. Sentiment in this case is used to describe the message on the postcard. A sad message would have negative sentiment, a happy message would have high sentiment. Vader analyzed each word in the message, and came up with a score ranging from -1.0 to 1.0. As stated, -1.0 is the lowest sentiment, while 1.0 is the highest sentiment. We then took the postcards from each subject, and plotted the postcards sentiment score with the date printed. We looked for correlation using a line of best fit and R^2 values, finding that in all subjects, there was no correlation between sentiment and date printed. This meant that over time, sentiment stayed the same for all subjects.

4.3 Conclusion

This is an example of the graphs we created. This graph shows all the cards in the subject "PARKS". We plotted the sentiment score for each card, along with the date that card was printed. As you can see, the line of best fit is completely flat, meaning there is no change in sentiment as year increases. Almost all the graphs look identical to this one, proving that there is no change in sentiment over time.



5. Discussion

5.1 Summary

The data cleaning we did on the dataset allowed us to better run our algorithms and models on the data, as well as reach a better insight of what information could be obtained from the dataset. One of the biggest results from our project was the process of using the pictures at the front of postcards to extract colors, features, and other information, as well as the discovery of the utility of this information. Using color extraction was a surprising and informative application of our project that the Miami University Library found very useful and intriguing for their purposes and goals. We also found that there was little change in sentiment regarding nature-related topics in parallel with the growth of the American Environmentalism movement, answering our domain-specific questions. Overall, our group was able to use the project as an opportunity to gain a better understanding of machine learning topics/techniques, and applying those to a real world collection of artifacts that were digitized.

5.2 Other Ideas and Challenges

A particularly difficult challenge we had to overcome during this project was the attempt at using object detection on the postcards. During the pictorial analysis portion of the project (c.f. Section 3: Data Augmentation), we attempted to use object detection (i.e. identifying if something on the picture portion of the postcard is a bottle or a clock) to try and see if we could extract any useful information from the objects/entities that postcards depicted. Unfortunately, object detection is a highly complex operation, and we encountered problems where the objects identified were too specific to be useful (i.e. the model noticed bottles and chairs in a picture of a bar, but not the broader context that what it was seeing was a bar). A major factor of this limitation was the necessity to use a pre-existing model trained with visual data and labels that did not suit our purpose. In the future, trying to utilize object detection on the postcard collection for whatever purpose will most likely require a better-trained model or some other procedure that can better allow one to extract more useful information.

5.3 Future Work

While our group was definitely able to accomplish much during this project in examining the dataset and deriving important (and sometimes mundane) new insights, there can still be work done on this project by another team. One of the biggest things we would have liked to have seen incorporated into this project was the addition of object identification when running algorithms over the front of the postcards. However, due to the complexity of topic modeling and the lack of results from our own, we did not take the opportunity to explore this potentially interesting intersection in more detail.

A rather specific improvement we would have liked to have seen come to fruition is the mapping of a representation of Ohio onto the charts displaying where a postcard was sent to. This is rather just a touch up as opposed to an area to explore, but perhaps it would have added something more to our project if we were able to get it done in the time before the project’s due date. Another potential area of exploration could be for the examination of more than the dataset we were given. The dataset we have is only a fraction of the full collection. If a significant number of postcards are added to the digitized collection, another group could analyze them to find insights or new information that was missing for our group.

Overall, we are pleased that both the library and our team were able to use this collaborative opportunity to gain new insights into a, perhaps neglected, collection of artifacts, and that our team received a new experience applying machine learning techniques to solve challenges outside the realm of purely Computer Science topics.