DS5110 Final Project

Team D.A.D.S. – Tyler Beaulieu, Tim Paylor

December 8, 2025

Abstract

We will be working with the Maine Port Authority to analyze their current data storage and share insights on the types of files they have and how they are organized. Our goal is to use our analysis to make their data more accessible and searchable.

1 Introduction

The Maine Port Authority works in coordination with the Maine Department of Transportation in acquiring, financing, constructing, and operating any marine port terminal facility within the state of Maine.1 [2025] They operate three major ports in Portland, Searsport, and Eastport.2 [2025]

Important documentation for the Port Authority is stored on a SharePoint drive. This drive has become increasingly disorganized over the years. The client is looking for assistance in making their data more organized and accessible.

Our first goal is to analyze the data as it exists today to help formulate a path forward. We'll use text mining and clustering techniques to find similar types of documents. Our secondary goal was to build a standard operating procedure to better name files moving forward.

2 Data Processing

2.1 The Dataset

Instead of working with the contents of their full SharePoint drive, we were given a smaller sample via a Google Drive folder. We will start by analyzing the metadata from these files.

We used the Google Drive API for Python to pull data about the files from Google Drive.Google [2021] We used it to recursively go through folders in the Drive and extract metadata. We retrieve the file ID, filename, file type, file size, file path, and parent folder name. This was then stored in a Pandas DataFrame.

We determined that we were working with 8672 files inside 653 folders. Looking at file types, we found that the files were 78% documents and 22% images (as seen in Figure 1).

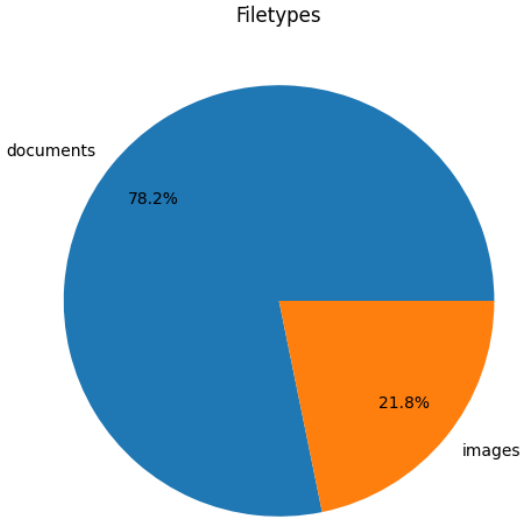

Figure 1: File Types
Proportion of document and image file types.

The most popular document types were Word documents, PDF files, and Excel worksheets (as seen in Figure 2). Word documents made up over half of the non-image files in the drive.

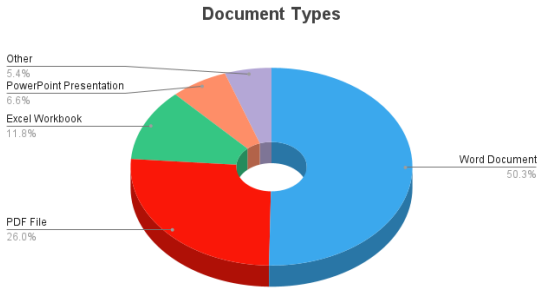

Figure 2: Document Types
Proportion of document types.

2.2 Duplicates

Now that we have our metadata, we can check for duplicate files in the dataset. We made a copy of our DataFrame with just the filename, file size, and file type columns. Then we eliminated duplicate rows. From this we found that there were 3202 unique files of 8672 total files. When we looked at folders, we found 257 unique folder names of 653 total folders.

Next we checked for duplicate folders. We created a DataFrame of folder names, summed up the size of all files insides, and counted the number of documents and images present. Then we eliminated duplicate rows. From this we discover 89 folders with that have a duplicate in the system. Most of them appear to be from the CruiseMaine directories.

Deliverable: From this analysis, we will be able to provide the client a list of duplicate files and folders.

2.3 Data Preparation

Using our new DataFrame of unique files, we have a directory upon which we can begin our analysis using machine learning.

Figure 3: Baked Goods Example
The proportion of ingredients leads to a different category.Rebello [2023]

3 Clustering

3.1 Introduction to Clustering

Our goal is to group documents by finding common words between them. This starts with text mining. Files are opened, and the text of each document is simplified to a recipe of words. In this form, the relative proportions of words are like the proportions of ingredients in baked goods (as seen in Figure 3). More of one word and less of another word could help point us towards the type of document being analysed.

These similarities allow us to categorize and sort the documents without having to read them. We can see this graphically in Figure1 where the three ingredients, flour, sugar, and butter, are used to find clusters of baked goods that fall under a similar category. For the MPA data, the same concept is used on hundreds of words in documents instead of just 3 ingredients in recipes.

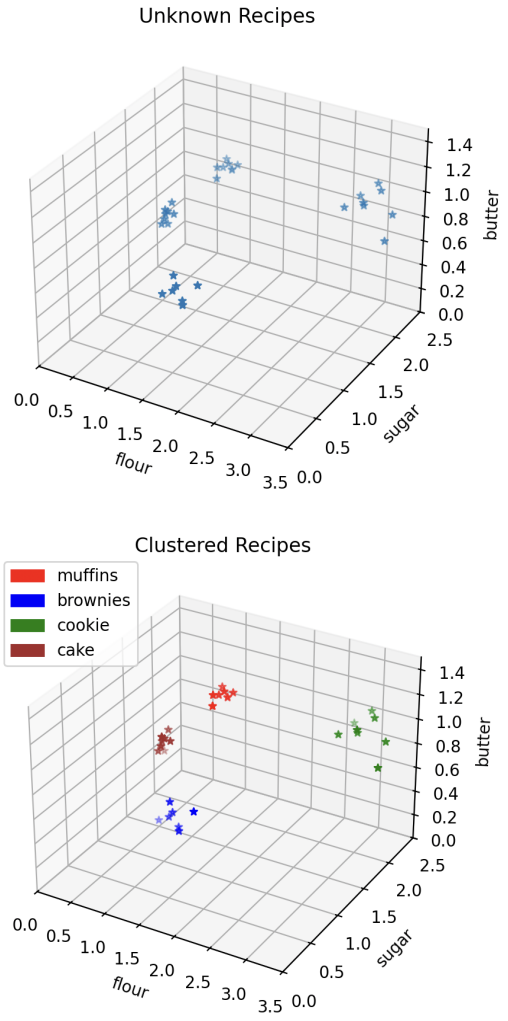

Figure 4: Clustering Example
The algorithm finds the clusters, than the human identifies the category.

3.2 Random Sampling

Given 3,000 unique documents, we can take random samples from across many folders to see that our clustering is working. We perform textual analysis on the samples with the goal of identifying and labelling the types of documents. We put our primary focus on Microsoft Word documents, since they comprise over half of the documents in the dataset.

3.3 Clustering Analysis

Tim was here.

4 Conclusions

4.1 Deliverables and Future Work

For the client we are able to give them Excel files that contain lists of duplicate files and folders.

We would continue our clustering work on other document types (pdfs, excel documents, powerpoints). This would give us the information we would need to produce a Standard Operating Procedure for future naming and storage of files.

We could also build a script to allow us to reorganise and rename their files. We would want to build it to have human validation to avoid major errors in categorization.

References

Title 23: Transportation, 2025. URL https://legislature.maine.gov/statutes/23/title23ch0sec0.html.

Maine port authority, 2025. URL https://www.maineports.com.

Google. Google api client, 10 2021. URL https://github.com/googleapis/google-api-python-client.

Sanchia Rebello. Baking ratios, 12 2023. URL https://thelittleshine.com/baking-ratios/.