Dimensionality Reduction and Hypothesis Testing

Jawadul Chowdhury

April 10, 2025

Contents

1 Introduction

In this section, we discuss about what the paper is about as well as the kind of dataset we use, as well as the features and how big the dataset is.

1.1 Abstract

In this paper, we analyze 63,542 emails. We convert the raw text from these emails into a feature matrix using a "bag of words" model. Each column of the feature matrix corresponds to one word, each row corresponds to one email, and the entry stores the number of times that word was found in the email. We then perform dimensionality reduction, cluster the emails in two clusters. Lastly, we perform binomial testing on all words in each cluster and filter out the top 200 words.

1.2 About the Dataset

Using the 63,452 emails, we converted them into a pandas data frame. A closer look at the pandas data frame tells us that it consists of 5 features, which are `category`, `to_address`, `from_address`, `subject` and `body`. The dataset consists of 63542 rows and 5 columns. For the experiments we perform on the dataset, we are primarily concerned with the `category` and `body`, as will be seen in the sections below.

2 Method

In this section, we discuss the methods we used to perform dimensionality reduction and feature extraction. We discuss in detail all the steps that were performed and why they were performed. The purpose of this section is detail the steps and explain the concepts behind what we did.

2.1 Feature Extraction

Throughout the course of this lab, we extract all the emails that we have stored in a directory from a .json format into a pandas data frame. We then perform feature extraction, where we create a feature matrix from the text that makes up the body of the email. We ensure that we only capture words that are mentioned more than 10 times. The reason we do this is because we would like to reduce noise from rare words as well as prevent overfitting in the unsupervised learning model that will be used later on.

Another point worth noting is that when we reduce the number of words, it helps to lower the dimensionality of the data. Lastly, if we were to include a great number of rare words, this will cause inconsistencies and will disrupt the clustering process when using a unsupervised machine learning model.

2.2 Dimensionality Reduction

With the current feature matrix that we've created, we perform principal component analysis by reducing the number of dimensions in the data. By reducing it to 10 columns, we have data that is easier to analyze. Next, we obtain the explained variance ratios, which tell us how much variance is retained by each component after performing dimensionality reduction.

2.3 The Clustering Process

After performing dimensionality reduction, we use unsupervised learning models to perform clustering. The two methods of clustering we use are centroid based clustering and agglomerative hierarchial clustering.

2.3.1 Centroid Based Clustering

When it comes to centroid based clustering, we used k-means. K-means works by partitioning a set of data points into K clusters based on their features so that data points within the same cluster are more similar to each other than to those in other clusters. Using k-means means we need to determine the value of k (number of clusters), and to find the correct value for k, we determine this using the silhouette score.

2.3.2 Agglorometative Hierarchial Clustering

Agglorometative Hierarchial Clustering works by building a hierarchy of clusters. It works by starting each data point as its own individual cluster and then merging the closest cluster at each step, progressively forming larger clusters until all data points belong to one cluster. For the linkage methods, we used single linkage, which finds the distance between two clusters in the shortest distance between any two points from different clusters.

2.4 Document Frequencies of Words

After clustering all the emails, we now analyze the clusters we've created and how the words we've captured play a role in clustering. We achieve this by creating a separate matrix for each cluster containing the rows for the points in that cluster. We convert these matrices into a CSC format due to the benefit that it is optimized for column slicing. Next, we calculate the document frequency of each word in each cluster. We perform document clustering as it allows us to determine the importance of a word within a cluster and analyze what the cluster represents.

2.5 Enriched Words with Statistical Testing

The aim here is to find words that are enriched in each cluster. As a result, we can interpret the themes in the cluster using statistical tests.

2.5.1 Using Binomial Testing

We use binomial testing to know whether a word appears in more emails than expected in one cluster compared to another. The null and alternative hypothesis for binomial testing are states as follows:

- **Null Hypothesis:** the relative document frequencies of the observed cluster are less or equal to those of the tested

- **Alternative Hypothesis:** the document frequency is higher in cluster 0 than in cluster 1

At the end of the day, we use statistical testing to avoid bias.

2.5.2 The Top 200 Words

We perform binomial testing on all words in each cluster, and then filter out the top 200 words based on their p-value, as the p-value tells us how strong the evidence is, and whether to null hypothesis is true or if there isn't enough evidence to back the null hypothesis.

3 Results

In this section, we look at the visualizations that have been produced.