UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

TASK AUTOMATION THROUGH EMAIL DATA ANALYSIS

MARIE MAYOL

Thesis supervisor: LLUIS PADRO CIRERA (Department of Computer Science)

Degree: Master Degree in Innovation and Research in Informatics ( )

Thesis report

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor, Professor Lluis Padro, for his guidance throughout my research. His insightful feedback and valuable suggestions have been instrumental in shaping this work.

I am also grateful to the Universitat Politècnica de Catalunya for giving me the opportunity to do this master thesis.

I would also like to express my gratitude to the members of the jury who devoted their time and expertise to the evaluation of this work.

Abstract

Currently, many companies do not use the information contained in their emails, yet it is a data set that is full of information and could be very useful.

This thesis report focuses on **email data analysis** and task automation, particularly in the area of email-based process mining. The state of the art section reviews existing research on extracting information from email content using techniques such as lexical analysis, language detection, semantic analysis and **machine learning** methods [12]. It explores different areas of process mining, including process pattern discovery, anomaly discovery, and process extraction from texts.

The objectives of this research are to assess the feasibility of extracting candidate processes from emails, to develop human-understandable metrics to classify processes, to propose a system to identify automation opportunities in email templates and explore possibilities for automation in email interactions.

To do this, we carried out different steps such as data preparation, chains detection, text representation, distance matrix calculation and **grouping methods**.

Key words: email data analysis, machine learning, grouping methods

Contents

List of Figures

List of Algorithms

Introduction

The extraction of valuable knowledge from email data has the potential to improve internal processes within businesses. Although comprehensive email logs exist in most organizations, a significant portion of this data remains untapped.

The main objective of this research is to uncover repetitive procedures carried out by humans via emails that could be automated. To achieve this goal, we will implement clustering algorithms on the Enron dataset.

This research aims to contribute to the enhancement of internal processes within organizations, ultimately leading to increased productivity and performance.

This thesis begins with a review of the existing literature on the subject to provide a context for previous work. Relevant theoretical concepts are then discussed to establish a solid foundation for this study. The methodology employed to achieve the various objectives is explained in detail. Subsequently, the obtained results are presented, highlighting the discoveries made during this study.

Finally, a discussion is introduced to analyze and interpret the results, and perspectives are proposed to guide future work in this field.

1 State of the art

In this section, we conduct a review of existing research on the analysis of email data and task automation. We explore different approaches and techniques utilized to extract information from the textual content of emails, including lexical analysis, language detection, and semantic analysis. Additionally, we highlight the application of machine learning methods, such as neural networks, classification algorithms, and clustering techniques, to this type of data. This literature review enables us to comprehend the advancements achieved in the field and identify the challenges and opportunities for our project.

1.1 Email-Based Process Mining

Process mining is a specialized field dedicated to the discovery, monitoring, and enhancement of real-world processes by extracting valuable insights from event logs in information systems. Our research focuses specifically on email-based process mining. It involves directly analyzing emails rather than relying on event logs.

Within the realm of process mining, there exist various domains, each with its own distinct objectives. One such domain is process model discovery. It involves generating a process model based on event logs to reconstruct the underlying processes. However, due to the presence of hidden structures, the generated model may not faithfully represent the true underlying model. To address this challenge, several methods have been developed to effectively handle these hidden structures.

Another domain is anomaly discovery, which aims to identify anomalies within the process model or conduct compliance checks. Although this domain has received less attention compared to process model discovery, it benefits from utilizing more general data mining tools.

A less explored domain is process extraction from texts. It presents a more intricate task as it involves extracting complex entities from highly diverse and heterogeneous textual data. Unlike process model discovery, which solely deals with event logs, process extraction from texts requires approaches capable of handling the complexities associated with natural language. Consequently, it is not uncommon to find approaches that combine process model discovery with process extraction from texts by generating an intermediate representation resembling an event log.

Although not a distinct domain on its own, email-based process mining holds significant relevance to our research. It encompasses aspects of both process model discovery and process extraction from texts, both of which contribute to our investigation.

1.1.1 Process Extraction from Texts

Extracting processes from textual data is a challenging task due to the inherent properties of natural language and the potential for multiple valid interpretations to arise from it. [4]

To address this problem, several approaches have been developed and categorized based on the methodology employed to extract processes or activities from the text. These approaches include rule-based systems, template-based systems, and neural network-based systems.

These approaches exhibit varying levels of user involvement, which can be classified as semi-guided or fully automatic. In semi-guided approaches, the system primarily operates automatically but may require user intervention in certain aspects of the process extraction. User intervention can take the form of providing domain-specific knowledge, defining rules or templates, or manually reviewing and correcting the extracted processes.

Rule-based systems rely on predefined patterns or linguistic rules to extract processes from text. These rules are typically created by domain experts and customized for specific domains or applications. While rule-based systems can be effective in certain scenarios, they often necessitate manual effort for rule definition and maintenance, which limits their flexibility and scalability.

Template-based systems utilize pre-defined templates or structures to extract processes from text. These templates capture common patterns or structures found in the text and map them to specific activities or process elements. Template-based approaches offer a more flexible and scalable solution compared to rule-based systems, as they can be easily adapted to different domains by modifying or adding templates.

Neural network-based systems leverage machine learning techniques, particularly deep learning, to automatically learn and extract processes from text. These approaches involve training models on annotated data, where the input is the text and the output is the extracted processes or activities. Neural network-based systems have demonstrated promising results in various natural language processing tasks, including process extraction from texts.

They have the ability to capture intricate patterns and dependencies within the text, enabling more accurate and robust process extraction.

1.1.2 Process Model Discovery

Process model discovery aims to generate a process model that represents the underlying process based on observed event logs. In the context of email-based process mining, the event logs are derived from email data, capturing various actions taken on emails, such as sending, receiving, forwarding, and replying.

Several techniques have been proposed for process model discovery, including the alpha-algorithm, heuristics miner, genetic algorithms, and Petri net-based methods. [17] These techniques analyze the event logs to identify common patterns, dependencies, and sequences of actions, which are then used to construct a process model.

The alpha-algorithm is a popular process discovery algorithm that constructs a process model based on the observed order of activities in the event logs. It employs a systematic approach to determine the relationships between activities and represents them in a directed graph format.

Heuristics miner is another widely used process discovery technique that uses heuristics to identify the most likely process model based on the event logs. It leverages different heuristics, such as activity frequency, activity order, and activity concurrency, to construct a process model that best fits the observed behavior.

Genetic algorithms apply evolutionary principles to discover process models. They utilize a population-based search approach, where multiple candidate models are iteratively evolved and evaluated based on their fitness to the event logs. The fittest models are selected and combined to produce offspring models, which undergo further evolution.

Petri net-based methods represent processes as Petri nets, a mathematical modeling technique that captures concurrency, synchronization, and causality between activities. Petri net-based process discovery algorithms analyze the event logs to construct a Petri net model that accurately represents the observed behavior.

1.1.3 Integration of Email Mining with Process Mining Techniques

The integration of email mining with process mining techniques has emerged as a promising direction in research. By combining the insights extracted from email data with the analysis of process behavior, researchers aim to gain a comprehensive understanding of organizational processes and their underlying dynamics.

One approach is to utilize email data to enrich existing process models or event logs. For example, Li et al. proposed a method that incorporates email communication patterns into the process discovery process. By integrating email data, they were able to identify hidden relationships and dependencies between activities, leading to more accurate process models.

Another line of research focuses on using email data to discover new process perspectives or alternative process models. Chen et al. presented a technique that leverages email content and email communication networks to identify informal or ad hoc processes that are not explicitly captured in existing process models. This approach provides valuable insights into the informal collaboration and decision-making processes within organizations.

Furthermore, researchers have explored the use of email mining for process conformance checking and anomaly detection. By analyzing email communication patterns and comparing them to expected process behavior, deviations or anomalies can be identified. [10] Hu et al. developed a method that combines email data with process event logs to detect non-compliant activities or unusual process instances. This approach enhances the ability to monitor and ensure process compliance within organizations.

In summary, the integration of email mining with process mining techniques offers a promising avenue for advancing the field. It enables researchers to leverage the rich source of information present in email data to enhance process discovery, conformance checking, and anomaly detection. By considering email communication patterns and content, a more holistic view of organizational processes can be obtained, leading to improved process understanding and management.

2 Theoretical background

2.1 Text representation

Text representation is a crucial step in email mining, as it allows transforming unstructured textual data into a format that can be effectively processed by natural language processing (NLP) algorithms and clustering techniques. In this section, we will explore two of the most popular text representations: TF-IDF (term frequency-inverse document frequency) and ParagraphVector (also known as doc2vec). These methods have been widely used in various domains, including text mining and data analysis, to efficiently represent and analyze textual documents.

2.1.1 TF-IDF

TF-IDF (term frequency-inverse document frequency) is a classic method for text representation that quantifies the relative importance of a term in a document compared to a collection of documents. It consists of two main components: term frequency (TF) and inverse document frequency (IDF).

Term frequency (TF) measures the number of occurrences of a given term in a document. It is calculated using the equation:

$$\mathrm{TF}(t, d) = f(t, d)$$

where:

- TF(t, d) represents the term frequency of term t in document d, and f(t, d) is the frequency of term t in document d.

Inverse document frequency (IDF) measures the inverse of the term's frequency in the entire document collection. It is calculated using the equation:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{DF}(t)}\right)$$

where:

- IDF(t) represents the inverse document frequency of term t, N is the total number of documents in the collection, and DF(t) is the number of documents that contain term t.

The TF-IDF representation of a document consists of a numerical vector where each dimension corresponds to a term, and the value represents the importance of that term in the document, calculated by multiplying the term frequency (TF) by the inverse document frequency (IDF):

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \cdot \mathrm{IDF}(t)$$

The advantage of TF-IDF lies in its simplicity and interpretability. It allows representing documents using term weights that are informative and discriminative. However, it does not capture semantic relationships between words and can be sensitive to document length.

2.1.2 ParagraphVector (doc2vec)

ParagraphVector, also known as doc2vec, is a more recent method based on neural networks for document representation. It was introduced as an extension of the Word2Vec method, which aims to capture semantic relationships between words.

The ParagraphVector model represents each document as a dense and continuous feature vector. It uses a neural network to learn vector representations of documents that capture contextual and semantic information.

The model is trained using unsupervised learning algorithms, such as the distributed memory model of ParagraphVector (PV-DM) or the distributed bag of words model of ParagraphVector (PV-DBOW).

The PV-DM model predicts the target word in a context window given the document vector, while the PV-DBOW model predicts the words in the context window given the document vector. The document vector is learned alongside the word vectors during the training process.

The advantage of ParagraphVector is its ability to capture semantic relationships and similarities between documents, even for documents of different lengths. It also allows performing vector operations on document representations, such as searching for similar documents.

2.1.3 Comparison between TF-IDF and ParagraphVector

TF-IDF and ParagraphVector are two popular approaches for text representation, but they differ in their calculation methods and the information they capture.

TF-IDF is simple and easy to interpret, but it does not consider semantic relationships and relies on term frequencies in the documents. It is more suitable for tasks where word frequencies are informative, such as keyword-based document classification.

On the other hand, ParagraphVector uses neural networks to capture semantic and contextual information of documents. It may be more appropriate for tasks where semantic relationships between documents are important, such as information retrieval or document recommendation.

The selection of the text representation method will depend on the specific goals of the email mining task and the characteristics of the available data. It is also possible to use a combination of both approaches or explore other text representation methods based on the project's requirements.

In conclusion, text representation is an essential step in email mining, and the TF-IDF and ParagraphVector methods offer distinct approaches to capture the characteristics of textual documents. The appropriate method selection will depend on the specific objectives and requirements of the email mining task.

2.2 Distances

Similarity measures play a crucial role in clustering algorithms as they quantify the proximity or similarity between data points. Different similarity measures are used depending on the nature of the data and clustering objectives. In this section, we will examine four commonly used similarity measures: Euclidean distance, L2 norm, Word Mover's distance, and cosine similarity.

2.2.1 Euclidean Distance

Euclidean distance is one of the fundamental similarity measures used in clustering. It calculates the geometric distance between two data points in Euclidean space. The Euclidean distance formula between two points (x_1, y_1) and (x_2, y_2) is given by:

$$\mathrm{d}(x_1, y_1, x_2, y_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Euclidean distance is often used in cases where the data dimensions are numeric and continuous. However, it can be sensitive to variable scales, which may require data normalization beforehand.

2.2.2 L2 Norm

The L2 norm, also known as squared Euclidean distance, is a variant of Euclidean distance. It is defined as the sum of squares of differences between the coordinates of the data points. The L2 norm formula between two points (x_1, y_1) and (x_2, y_2) is given by:

$$\|\mathbf{x}_1 - \mathbf{x}_2\|^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

The L2 norm is used in many clustering algorithms, such as K-Means, to compute distances between data points. It is similar to Euclidean distance but does not involve the square root operation, which can make the computations more efficient.

2.2.3 Word Mover's Distance

Word Mover's Distance (WMD) is a similarity measure specifically designed for textual data. It is based on the notion of distance between words, represented as pre-trained word vectors. WMD quantifies the amount of energy (or cost) required to move words from one document to another, taking into account their semantic meanings.

WMD is calculated by minimizing the sum of distances between words in two documents, weighted by the distances between the corresponding word vectors. This measure is useful for clustering similar textual documents as it captures semantic similarities rather than just word co-occurrences.

2.2.4 Cosine Similarity

Cosine similarity is a measure used to evaluate the similarity between two vectors in a vector space. It is widely used in natural language processing to compare word vectors and textual representations.

Cosine similarity is calculated by measuring the angle between two vectors. The smaller the angle, the more similar the vectors are. The formula for cosine similarity between two vectors A and B is given by:

$$\text{sim\_cosine}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where:

- $(A \cdot B)$ represents the dot product between vectors A and B, and $\|A\|$ and $\|B\|$ represent the norms (lengths) of vectors A and B, respectively..

Cosine similarity is a robust measure that can be used to compare word vectors, document representations, and even user profiles. It is scale-invariant and can capture similarities even when vector magnitudes differ.

In conclusion, similarity measures such as Euclidean distance, L2 norm, Word Mover's distance, and cosine similarity are essential for evaluating the proximity or similarity between data points in clustering algorithms. The choice of appropriate similarity measure depends on the nature of the data and clustering objectives. These measures are widely used in various domains, including natural language processing, text mining, and numerical data analysis, to cluster and identify structures in datasets.

2.3 Clustering methods

2.3.1 K-Means

In the context of email mining, data clustering is an important task to discover patterns and similarities among emails. The K-Means algorithm is one of the most commonly used clustering methods. It divides a dataset into K clusters, where each data point is assigned to the nearest cluster.

Description of the K-Means Algorithm:

The K-Means algorithm [1] works as follows:

- Randomly select K initial cluster centers.
- Assign each data point to the nearest cluster center using a distance measure, typically the Euclidean distance.
- Update the cluster centers by computing the means of the data points assigned to each cluster.
- Repeat steps 2 and 3 until the cluster centers converge to a stable position.

Algorithm 1 K-Means

procedure K-MEANS(data, K)
  randomly select K initial cluster centers
  repeat
    assign each data point to the nearest cluster center
    update cluster centers by calculating the mean of the data points assigned to each cluster
  until cluster centers converge to a stable position
  return the final cluster assignments
end procedure

By accomplishing these objectives, this research aims to contribute to the refinement of internal processes within organizations by leveraging the knowledge embedded in emails, leading to improved efficiency, effectiveness, and automation of email interactions.
The K-Means algorithm aims to minimize the sum of distances between data points and their respective cluster centers. It seeks to create compact and well-separated clusters.

Step 2 of the K-Means algorithm can be mathematically formulated as follows:

Let C = C_1, C_2, ..., C_K be the set of cluster centers, and x be the data point to be assigned to a cluster. The distance between x and a cluster center C_i is calculated using a distance measure, typically the Euclidean distance:

$$\text{dist}(x, c_i) = \sqrt{\sum (x_j - c_{ij})^2}$$

where x_j and c_{ij} represent the coordinates of the data points and cluster centers, respectively.

Step 3 of the K-Means algorithm involves updating the cluster centers by computing the means of the data points assigned to each cluster. The mathematical formula for updating the cluster center C_i is as follows:

$$C_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

where $|C_i|$ represents the number of data points assigned to cluster C_i, and $\sum_{x \in C_i} x$ represents the sum of the coordinates of these data points.

The K-Means++ Variant:

K-Means++ is an enhancement of the K-Means algorithm that aims to select initial cluster centers more judiciously. Randomly selecting cluster centers can lead to suboptimal clustering results that are dependent on the initialization.

The K-Means++ algorithm works as follows:

- Randomly select the first cluster center from the data points.
- For each remaining data point, calculate the distance to the nearest already selected cluster center.
- Select the next cluster center with a probability proportional to the squared distance.
- Repeat steps 2 and 3 until K cluster centers are selected.

Selecting cluster centers using the K-Means++ method promotes better spatial distribution of the initial cluster centers, which can lead to improved overall convergence of the K-Means algorithm.

By using the K-Means++ variant, researchers can obtain more stable and consistent clustering results, regardless of random initialization. This helps improve the quality of obtained clusters and facilitates interpretation of email mining results.

2.3.2 DBSCAN

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a widely used non-parametric clustering method in email mining. Unlike the K-Means algorithm, DBSCAN is capable of detecting clusters of arbitrary shapes and can also identify noise points that are not included in specific clusters. [7]

The DBSCAN algorithm is based on two key concepts:

- Density: In DBSCAN, density is defined by the number of data points present in a given neighborhood. Data points with a minimum number of neighbors within a specific radius are considered as core points, and data points that are close to a core point but do not have enough neighbors themselves are considered as boundary points.
- Accessibility: The accessibility of a data point is determined by its proximity to other core data points. If a data point can be reached from another core data point by passing through a series of neighboring data points, then they are considered connected.

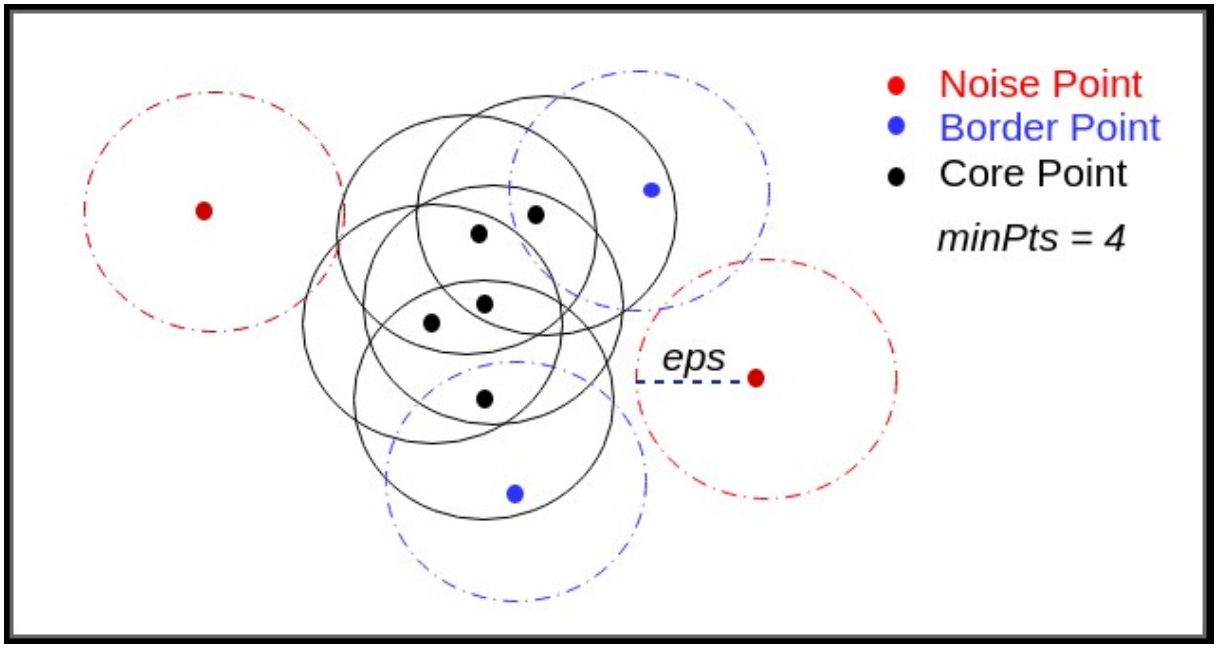

Figure 1: Principle of DBSCAN

The DBSCAN algorithm proceeds by following these steps:

- Select an unvisited data point from the dataset.
- Check if the data point has enough neighbors in its neighborhood, using a specific radius and minimum number of neighbors.
- If the data point is a core point, create a new cluster and add the data point and its accessible neighbors to this cluster. Consequently, all these data points are marked as visited.

- If the data point is a boundary point, mark it as visited.
- Repeat steps 1 to 4 for all unvisited data points until all points are visited.

As the algorithm progresses, it forms clusters by connecting accessible data points to each other. Data points that are not included in any clusters are considered noise.

The DBSCAN algorithm has several advantages:

- Ability to detect clusters of arbitrary shapes: DBSCAN can find clusters that are not necessarily spherical or linear. It can detect complex-shaped clusters in email datasets where patterns can be more diverse.
- Robustness to noise: DBSCAN can identify and isolate data points that are not included in specific clusters, distinguishing them as noise points. This is particularly useful in email mining where unwanted or atypical emails can be considered as noise.

However, DBSCAN also has some limitations:

- Sensitivity to parameters: DBSCAN requires the specification of two key parameters, namely the radius and minimum number of neighbors. The algorithm's performance can vary depending on these parameters, and it can be challenging to adjust them optimally.
- Sensitivity to density: The algorithm may struggle with detecting clusters of varying densities, where data points can be unevenly distributed. In such cases, data preprocessing or exploring algorithm variants may be necessary.

In conclusion, the DBSCAN algorithm is a powerful method for cluster creation and noise detection in email mining. Its ability to identify clusters of arbitrary shapes and isolate noise points makes it a valuable tool for analyzing complex email datasets. However, careful attention should be given to parameter selection to achieve optimal results.

2.3.3 HBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a robust clustering algorithm that combines hierarchical approaches, minimum spanning trees, and density regions. It aims to identify clusters of arbitrary shapes while also detecting noise points in a dataset. Compared to other clustering algorithms like k-means, HDBSCAN is more resilient to outliers.

The HDBSCAN algorithm follows several steps to perform clustering. Firstly, it calculates the density of each data point using a density measure based on the distance between points and their neighbors. [3] This step helps identify regions of high density, which are potential clusters, and regions of lower density, which may contain noise points.

To address points with low density, HDBSCAN introduces the concept of mutual reachability. This technique involves defining new metrics between points, which are calculated using the mutual reachability distance. The mutual reachability distance, denoted as $d_{mreach-k}$(a, b), is the maximum value among the core distances of points a and b, as well as the distance between a and b:

$$\mathrm{d}_{mreach-k}(a,b) = \max\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}$$

This distance transformation helps improve the algorithm's robustness to noise and outliers.

Next, a connectivity graph is constructed by connecting data points that exceed a defined density threshold. This graph captures the neighborhood relationships between the points. From this graph, a minimum spanning tree is created by selecting edges with the lowest weights. The minimum spanning tree highlights the most significant connections between data points.

Using the minimum spanning tree, HDBSCAN determines clusters by traversing the tree and applying a stability measure to decide if a branch corresponds to a valid cluster. Regions of high density are considered clusters, while branches with lower density are identified as noise points.

For cluster extraction, HDBSCAN introduces a measure of cluster persistence represented by the lambda (λ) value. The lambda value is calculated as the inverse of the distance. Each cluster is assigned a lambda birth and lambda death value, indicating the lambda value when the cluster was formed and when it split into smaller clusters, respectively. Additionally, for each point within a cluster, a lambda value (λ_p) is assigned, representing the lambda value at which the point "fell out" of the cluster. The stability of a cluster is then determined by summing the differences between λ_p and λ_{birth} for all points in the cluster.

To select the most stable clusters, HDBSCAN iterates through the tree hierarchy. At each leaf node, the algorithm compares the sum of stabilities of the child clusters with the stability of the current cluster. If the sum of stabilities is greater, the cluster stability becomes the sum of the child stabilities. However, if the stability of the cluster is greater than the sum of its children's stabilities, the cluster is considered a candidate, and its children are excluded from further consideration. This process continues until the root of the tree is reached, and the remaining candidates are identified as the stable clusters.

Overall, the HDBSCAN algorithm can be summarized in five stages:

- Transform the space based on density to enhance robustness.
- Generate a minimum spanning tree from the distance-weighted undirected graph.
- Apply hierarchical clustering on the connected components of the minimum spanning tree.
- Condense the resulting hierarchy based on the minimum cluster size.
- Return the stable clusters obtained from the condensed tree.

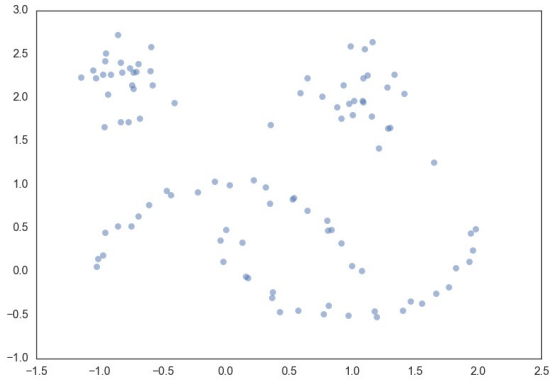

Figure 2: Initial Data Set

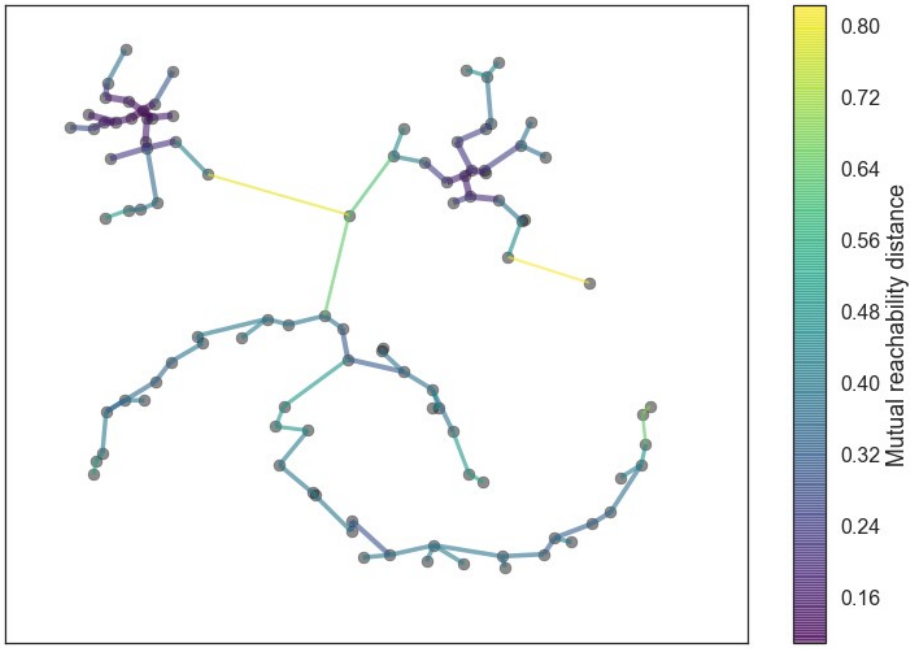

Figure 3: Minimum Spanning Tree

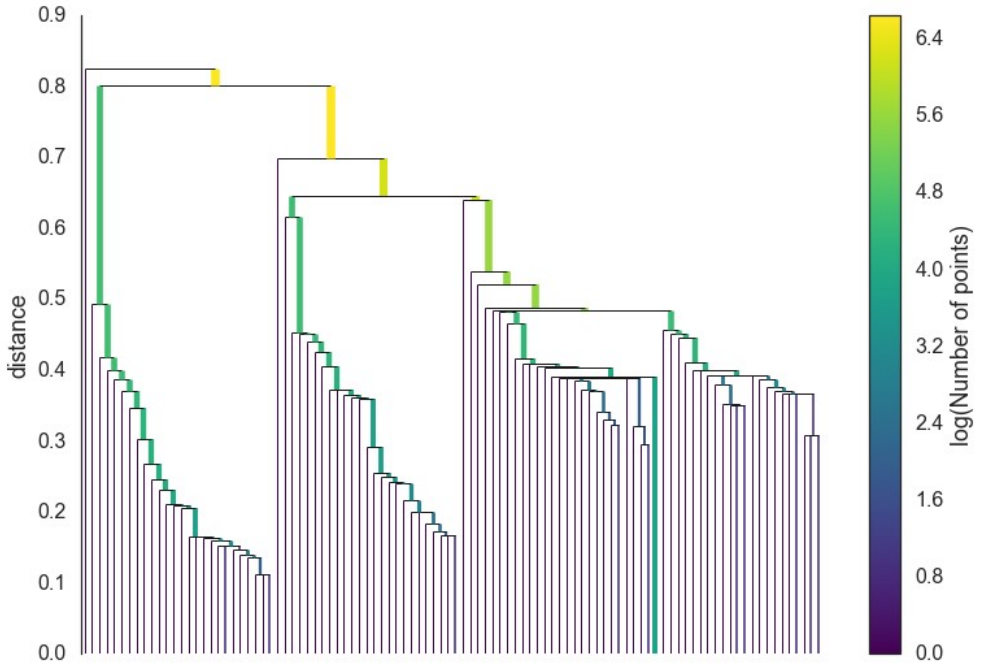

Figure 4: Hierarchy Clustered Tree

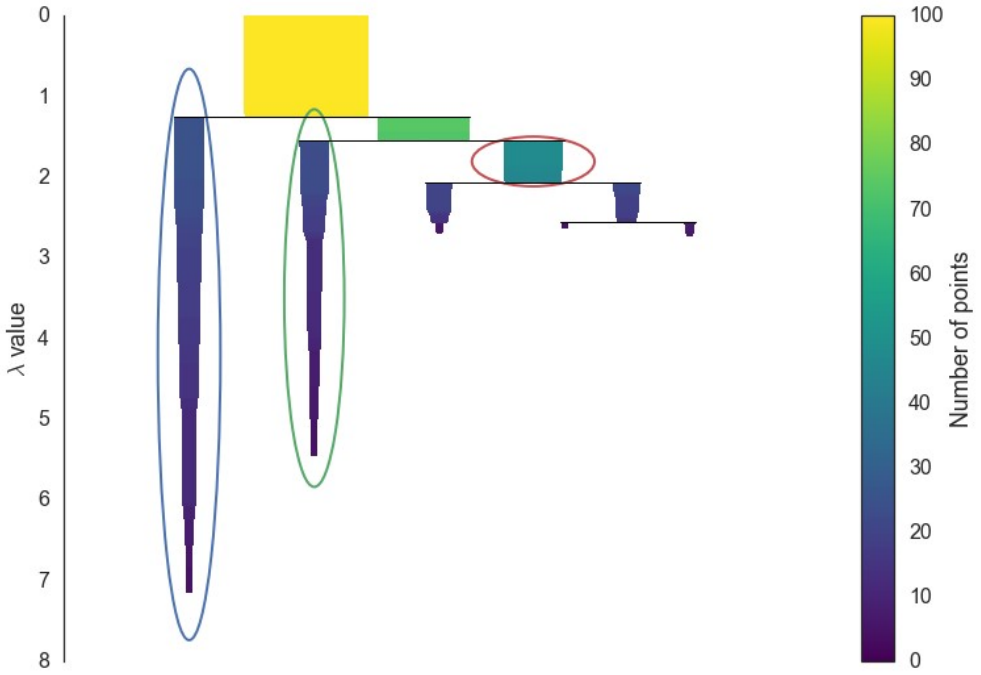

Figure 5: Condensed Tree

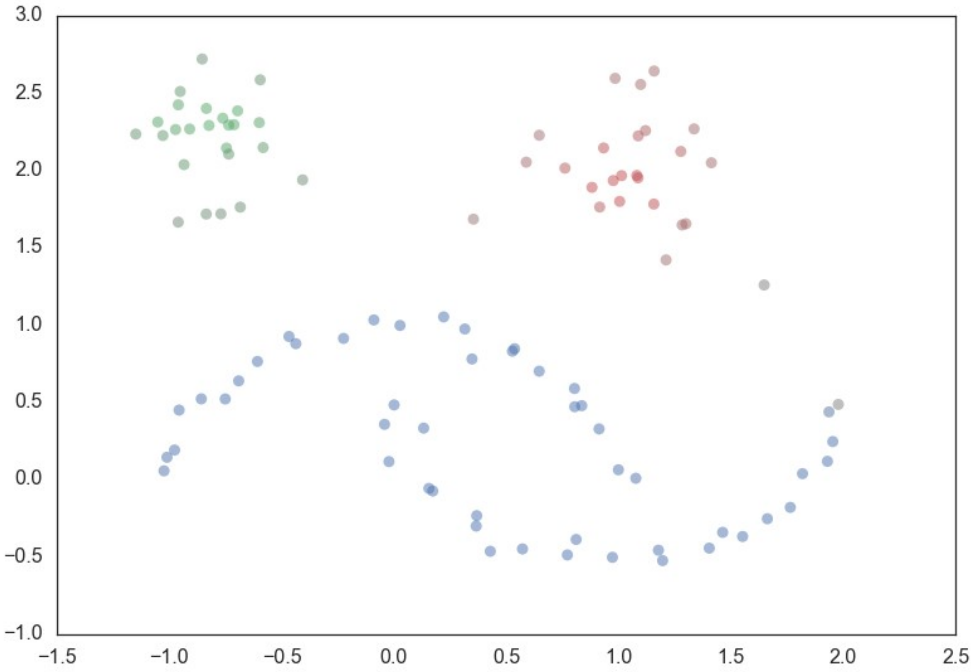

Figure 6: Final Data Set

By following these steps, HDBSCAN effectively identifies clusters of arbitrary shapes and detects noise points, making it a valuable tool for clustering analysis in various applications. However, it is important to note that HDBSCAN is sensitive to parameter settings, such as the density threshold and stability measure, which may require careful consideration. Additionally, the hierarchical nature of the algorithm and the construction of minimum spanning trees may demand more computational resources, especially for large datasets. Despite these limitations, HDBSCAN offers significant advantages in its ability to capture complex cluster structures and handle noise effectively. [15]

2.4 Cluster evaluation measures without ground truth information

Evaluating the quality of clusters is crucial for assessing the performance of clustering algorithms. However, in many cases, the data lacks ground truth information (actual labels) to compare the clustering results. In this section, we will discuss several cluster quality measures that can be used in such situations, namely the silhouette score, Calinski-Harabasz index, Davies-Bouldin index, and entropy.

2.4.1 Silhouette Score

The silhouette score is a commonly used measure to assess the cohesion and separation of clusters without requiring ground truth information. [14] It provides an estimation of cluster quality by calculating the average similarity of points within their own cluster compared to the average similarity with other neighboring clusters.

The silhouette score of a point i is calculated as:

$$\text{silhouette score}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:

- a(i) is the average dissimilarity between point i and all other points within the same cluster.
- b(i) is the average dissimilarity between point i and all points in the nearest neighboring cluster.

The overall silhouette score is computed as the average of silhouette scores for all data points. It ranges from -1 to 1, where a value closer to 1 indicates well-clustered points, a value close to 0 indicates overlap between clusters, and a value close to -1 indicates poor cluster separation.

2.4.2 Calinski-Harabasz Index

The Calinski-Harabasz index is a cluster quality measure that evaluates the separation between clusters and intra-cluster compactness. [2] This measure is calculated using Analysis of Variance (ANOVA) to compare the variance between clusters and the variance within clusters.

The Calinski-Harabasz index is defined as:

$$\text{Calinski-Harabasz Index} = \frac{\text{SSB}/(k-1)}{\text{SSW}/(n-k)}$$

where:

- SSB is the between-cluster sum of squares, which measures the variance between clusters.
- SSW is the within-cluster sum of squares, which measures the variance within clusters.
- k is the number of clusters.
- n is the total number of data points.

A higher Calinski-Harabasz index score indicates better separation and compactness of clusters.

2.4.3 Davies-Bouldin Index

The Davies-Bouldin index is a cluster quality measure that evaluates the separation between clusters while considering intra-cluster compactness. [6] This measure is calculated by comparing the average distances between points in a cluster and the centroid of the cluster with the average distances between centroids of neighboring clusters.

The Davies-Bouldin index for cluster i is defined as:

$$\text{Davies-Bouldin Index}(i) = \max\left\{\frac{(R(i)+R(j))}{d(c(i),c(j))}\right\}$$

where:

- R(i) is the intra-cluster scatter of cluster i, which represents the average distance between points in cluster i and its centroid.
- R(j) is the intra-cluster scatter of cluster j, which represents the average distance between points in cluster j and its centroid.
- d(c(i), c(j)) is the distance between the centroids of clusters i and j.

The Davies-Bouldin index is calculated as the average of the indices for all clusters. A lower Davies-Bouldin index score indicates better separation and compactness of clusters.

2.4.4 Entropy

Entropy is a measure of uncertainty that can also be used to evaluate the quality of clusters without having ground truth information. Entropy measures the dispersion of class labels within each cluster. A cluster with more homogeneous class labels will have a lower entropy, indicating better cluster quality.

The entropy of a cluster C is calculated using the frequencies of different class labels present in the cluster:

$$\text{Entropy}(C) = -\sum(p \log(p))$$

where:

- p is the frequency of a specific class label within the cluster.

The weighted average of cluster entropies is used to calculate the overall entropy of the clustering. A lower entropy score indicates better cluster separation and higher homogeneity of class labels.

In conclusion, cluster quality measures such as silhouette score, Calinski-Harabasz index, Davies-Bouldin index, and entropy are important tools for evaluating the performance of clustering algorithms in the absence of ground truth information. These measures provide estimations of cluster cohesion, separation, and dispersion of class labels within clusters. It is recommended to use multiple cluster quality measures to obtain a comprehensive evaluation of clustering quality.

2.5 Comparison with ground truth-based measures

In addition to cluster quality measures that do not require ground truth information, there are also methods that utilize ground truth data to evaluate clustering performance. Two commonly used methods are the homogeneity score and completeness score.

2.5.1 Homogeneity score

The homogeneity score is a measure that assesses the extent to which each cluster predominantly consists of samples from a single class. It is calculated by comparing the actual class labels of samples with the predicted class labels by the clustering.

The homogeneity score ranges from 0 to 1, where a value closer to 1 indicates better homogeneity of clusters, i.e., samples within the same cluster belong to a single class. However, the homogeneity score does not consider the separation between clusters, which can lead to biased results if the classes are inherently mixed.

2.5.2 Completeness Score

The completeness score is a measure that evaluates the extent to which all samples of a given class are assigned to the same cluster. Similar to the homogeneity score, it uses ground truth information to assess clustering performance.

The completeness score also ranges from 0 to 1, where a value closer to 1 indicates better completeness of clusters, i.e., all samples of the same class are grouped together in the same cluster. However, the completeness score does not consider the separation between clusters and can also be biased if the classes are inherently mixed.

2.5.3 Comparison with Ground Truth-independent Measures

Compared to cluster quality measures that do not require ground truth information, such as silhouette score, Calinski-Harabasz index, Davies-Bouldin index, and entropy, the homogeneity score and completeness score provide a more direct and explicit evaluation of the correspondence between clusters and actual classes.

However, it is important to note that the homogeneity and completeness scores are based on ground truth data, which means they require actual labels for comparison. Therefore, they are not applicable in situations where actual labels are unavailable or in cases where data is unlabeled.

In conclusion, cluster quality measures that require ground truth data, such as the homogeneity score and completeness score, offer a direct evaluation of the correspondence between clusters and actual classes. However, these measures are not applicable in all situations. On the other hand, ground truth-independent measures such as silhouette score, Calinski-Harabasz index, Davies-Bouldin index, and entropy provide a more general evaluation of cluster cohesion, separation, and dispersion. The choice of the appropriate measure will depend on the specific context and clustering objectives.

3 Methodology

3.1 Objectives

Emails, as highly heterogeneous data, hold a wealth of valuable knowledge that can be extracted to refine internal processes in businesses. Despite the presence of comprehensive email logs in most organizations, only a fraction of this data is currently being utilized. Therefore, the primary focus of this research is to extract additional dimensions of knowledge from email data. The objectives of this study can be outlined as follows:

- Assess the feasibility of extracting candidate processes from emails, considering computational efficiency and data suitability. This objective involves determining if the data supports the existence of true underlying processes and evaluating the computational viability of the extraction process for businesses.
- Develop human-comprehensible metrics that assist analysts in classifying candidate processes as true processes or non-processes. These metrics aim to provide interpretable insights and support the evaluation and understanding of the extracted processes.
- Explore opportunities for automating various aspects of email interactions, including tasks such as enhancing automatic categorization systems to mitigate negative effects on personal and organizational performance.

By accomplishing these objectives, this research aims to contribute to the refinement of internal processes within organizations by leveraging the knowledge embedded in emails, leading to improved efficiency, effectiveness, and automation of email interactions.

3.2 Work plan

In the context of this project, we consider that a successful model deployment should adhere to the CRISP-DM methodology [5]. This methodology consists of five phases: business understanding and data understanding, data preparation, modeling, evaluation, and evaluation phase.

In the business understanding phase, we seek to understand the company's needs regarding emails, such as folder classification, ticket generation, or automated responses. The data understanding phase involves collecting email data, exploring their properties and metadata, and evaluating data quality.

Next, in the data preparation phase, we create data subsets, perform pre-processing to obtain suitable text representations, remove erroneous values, and integrate external data.

The modeling phase entails selecting appropriate models, creating corresponding distance matrices, building the model, and evaluating different models to determine the best one.

In the evaluation phase, we utilize scores such as silhouette coefficient, Calinski-Harabasz score, Davies-Bouldin score, and entropy to assess the quality of the obtained clusters. We also conduct visual analysis of a sample of emails to verify the significance of the results.

Finally, the deployment phase involves developing a deployment plan, monitoring the model in production, producing reports, and reviewing the project.

It is important to emphasize that this project primarily focuses on the data understanding, data preparation, and modeling phases, with less emphasis on the other phases of the CRISP-DM methodology.

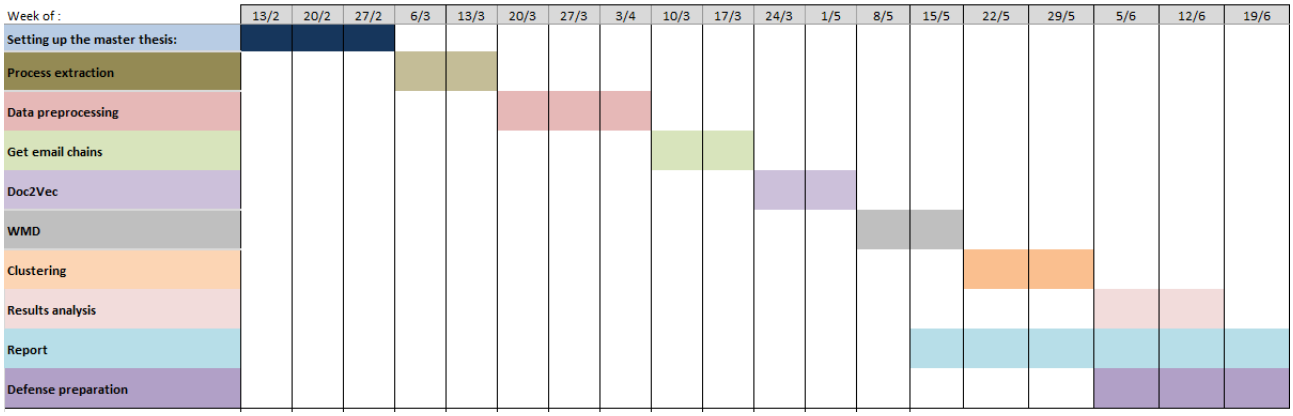

Figure 7: Forecast Gantt chart

3.3 Data description

To provide some context, in December 2000, Skilling replaced Lay as CEO, while Lay became Deputy Chief Executive Officer. However, in August 2001, Skilling abruptly and unexpectedly resigned from his position as CEO. This decision led to another change in leadership, with Lay taking back control of the company as CEO.

Unfortunately, Enron's difficulties did not end there. In December 2001, the company was forced to file for bankruptcy. Following Enron's bankruptcy, the company was dismantled and divided into several separate entities. One of these entities was Accenture.

Due to the magnitude of the events surrounding Enron's downfall, the emails from that period have generated significant interest in research and analysis. Studying them helps to better understand the internal dynamics of the company, the interactions between employees, and potential signals of fraudulent behavior. Thus, the Enron dataset has become one of the most widely used datasets in the field of text analysis, fraud detection, and email exploration.

It's important to note that the Enron dataset was initially made public by the Federal Energy Regulatory Commission (FERC) as part of its investigation into Enron's practices. The data was anonymized to protect the privacy of individuals mentioned in the emails. Although the data is from that period, it continues to be a valuable resource for researchers.

Different versions of the email dataset can be found. They haven't all been preprocessed in the same way. Personally, I have worked with the Enron Email Dataset from CALO (2015 version). It contains 517 401 emails from 150 users. It has undergone the following modifications:

- Some messages have been deleted.
- Attachments are not included.
- Invalid email addresses have been converted to user@enron.com when the employee's name is known; otherwise, they are converted to no_address@enron.com.

Here is an example of email:

Message-ID: 16688729.1075854479282.JavaMail.evans@thyme
Date: Thu, 7 Dec 2000 07:00:00 -0800 (PST)
From: kay.chapman@enron.com
To: michael.guerriero@enron.com
Subject: Organizational Changes
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Kay Chapman
X-To: Michael Guerriero
X-cc:
X-bcc:
X-Folder: David_Delainey_Dec2000 Notes Folders 'sent mail
X-Origin: Delainey-D
X-FileName: ddelain.nsf

Here you go. Sorry, I have Enron Messaging working on it...But here you go for now.....

I focused on the body of the message, the sender, recipients, carbon copy (cc) recipients, date, and the subject of the email. I did not take into account the other information.

3.4 Detection of chains

I constructed the chains in the following way by traversing all the emails:

If the subject does not start with "Re:", it means it is the first email of a new chain. The function checks if this subject is already present in another chain. If not, the email is recorded as the first entry in the list of messages associated with this subject.
If the subject already exists, it means there are already emails associated with that chain. Then, based on the date, if the chain is older than one month, it is assigned to a new chain. If the subject of the message starts with "Re:", it means it belongs to an existing chain.

3.5 Preprocessing

I first realized that there were many duplicate emails. For that, I created a new column called "Participants." This step aims to store the list of participants in each email. To do this, I used the sender, recipients, and carbon copy recipients.
I created a unique key for each email by combining the email subject with the list of participants. This allows differentiating emails based on their content and participants.
Duplicate emails are removed based on the key created in the previous step, as well as the timestamp and message content. This ensures that each email in the DataFrame is unique.

Before performing doc2vec, I tokenized each email and removed noisy data using regular expressions.

3.6 Text representation

I built and trained Doc2Vec models. Once the Doc2Vec models were trained, I was able to extract the corresponding document vectors. I chose the following parameters:

- Vector Size: This is the dimensionality of the document vectors and word vectors in the Doc2Vec model. It determines the size of the vector representations of the texts. I used a vector size of 300, meaning each document is represented by a 300-dimensional vector.
- Window Size: It is the size of the window, i.e., the number of preceding and succeeding words considered when predicting the current word in the Doc2Vec model. I chose a window size of 15, which means I considered 15 words before and 15 words after the current word during prediction.
- Min Count: It is the minimum number of occurrences a word must have to be included in the vocabulary construction and Doc2Vec model training. I chose a minimum count of 1, meaning all words present in the corpus are included, even if they appear only once.
- Train Epoch: It is the number of training epochs, i.e., the number of times the entire training data is traversed during Doc2Vec model training. I chose 20 training epochs.
- Alpha: It is the initial learning rate of the Doc2Vec model. It controls the magnitude of vector updates during training. I chose an initial learning rate of 0.25.
- Min Alpha: It is the minimum learning rate of the Doc2Vec model. As training progresses, the learning rate is gradually reduced until it reaches this minimum. I chose a minimum learning rate of 1e-5.

3.7 Clustering

I chose min_cluster_size=5. This parameter is used to specify the minimum size of a cluster. It means that only points with at least 5 neighbors will be considered as members of a cluster.
Then, I applied the clustering algorithm to the document vectors. This assigns a cluster label to each document.

I displayed the PCA representation of the document vectors based on the clusters. This allows visualizing the obtained groupings.
I created a dictionary to group the documents by cluster. Each cluster is represented by a key in the dictionary, and the corresponding documents are added to the list associated with that key.

4 Results

4.1 Chains

In this section, I will present my findings regarding the chains. We can already observe that the dataset has significantly decreased as there were many duplicates and spam in the initial dataset.
Here is an example of chains of length 2:

From: Kate Symes 03/20/2001 05:22 PM

To: Sharen Cason
cc:

Subject: Re: #555410

I've changed the deal - we're running into some problems rolling out the new Full Service Power application. Apparently it's messed with some details of Deal Entry, including the deletion of strips. So you'll see three strips in this deal, two of which have no price and no volume. I hope that's okay! Will Smith in Houston IT is working on it now. Let me know if you see the right term and volume now.
Thanks,
Kate

From: Sharen Cason 03/20/2001 02:56 PM

To: Kate Symes
cc: Kerri Thompson/Corp/Enron@Enron

Subject: #555410

I believe this deal was changed, bu tnot changed correctly. The hours are doubled up and it is running 32 hours a day.
Thanks!

Chain Length	Number of Chains
1	15140
2	1478
3	5791
4	1847
5	732
6	341
7	159
8	96
9	65
10	41
11	17
12	12
13	10
14	9
15	7
16	6
17	7
18	3
19	1
20	2
22	1
23	3
24	1
27	1
29	1
30	1
31	1
32	1
35	1
38	1
50	1
74	1

Figure 8: Chain Length and Number of Chains

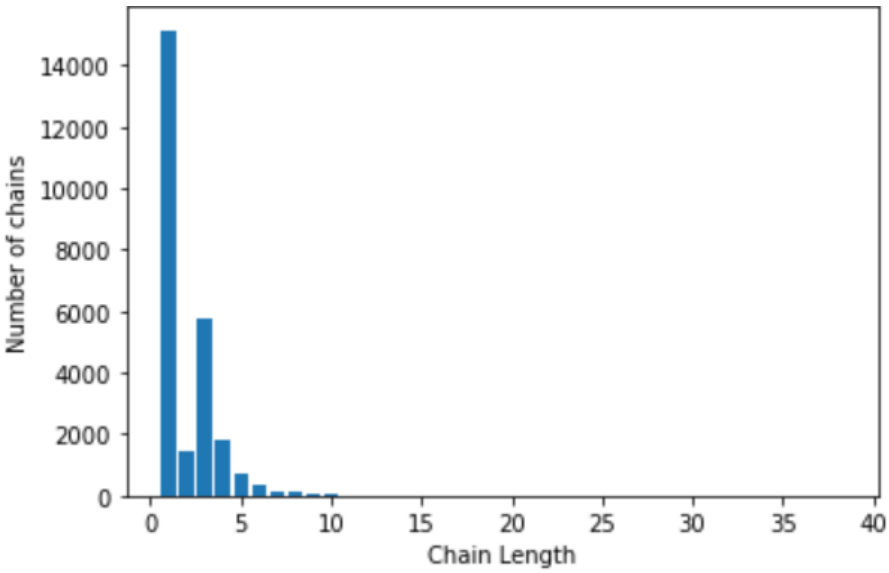

Figure 9: Number of chains by chain length

4.2 Clustering

4.2.1 Only chains

Initially, I decided to focus only on email chains and remove all chains of size 1. Here are the results I obtained:

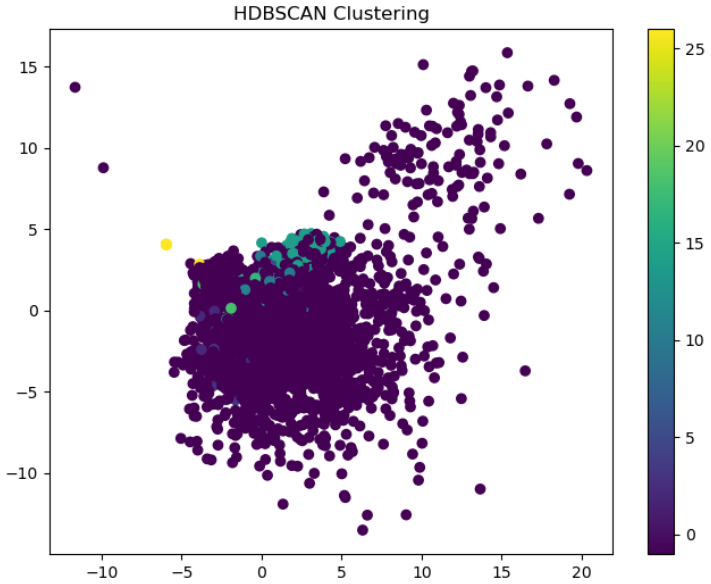

Figure 10: PCA visualization of clustering performed

During the analysis of my cluster, I evaluated its performance using three commonly used validation measures: Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. [18] These measures allowed me to quantify the quality of my clustering and understand the structure of the grouped data.

- The Silhouette Score is a measure that assesses how similar each data point is to other points in its own cluster compared to points in other clusters. A negative score, like the one I obtained (-0.29), indicates that most points are misclassified because they are more similar to points in other clusters than to those in their own cluster. This result suggests that my clustering may not be suitable for the analyzed data, as there is no clear and significant structure in the formed clusters.

- The Calinski-Harabasz Index [9, 14] is a measure of cluster compactness and separation. A higher index value indicates better clustering quality. In my case, I obtained an index of 7.65. Although this value is relatively low, it may suggest some separation between clusters. However, it is important to note that the Calinski-Harabasz Index should be interpreted in conjunction with other measures to obtain a complete picture of clustering quality.

- The Davies-Bouldin Index measures the average similarity between each cluster and its nearest cluster. A lower index indicates better clustering, with an ideal value of zero. In our case, the Davies-Bouldin Index is 2.23, indicating that the clusters have some similarity but are not distinct enough from each other.

Combining these results, it is clear that our current clustering exhibits significant limitations. The negative values of the Silhouette Score indicate poor point classification, while the relatively low Calinski-Harabasz and Davies-Bouldin indices suggest low separation and similarity between clusters. These results indicate that our current clustering method fails to capture the underlying structures of the data adequately.

4.2.2 Individual emails and chains

In this section, I studied both single email chains and longer chains, and I obtained these results.

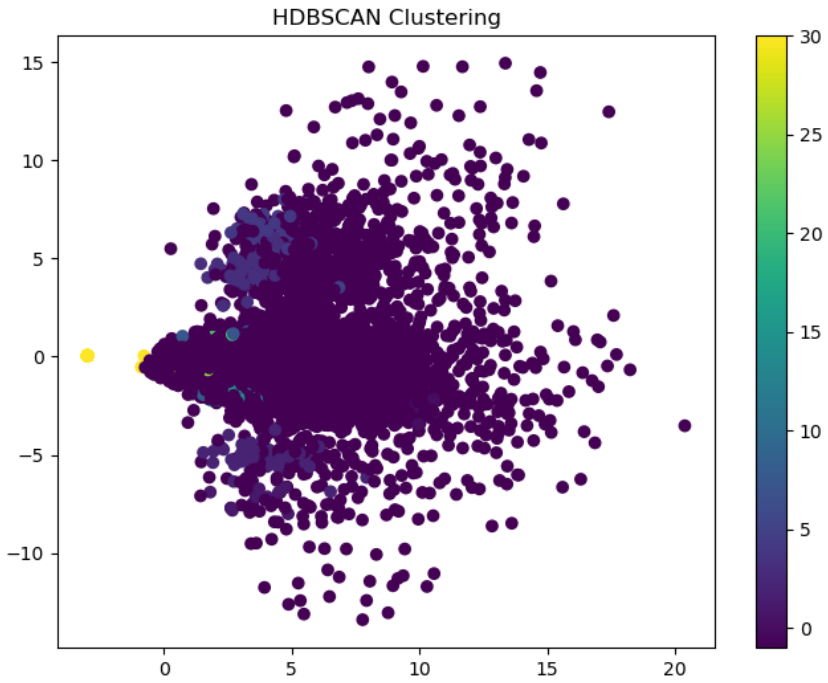

Figure 11: PCA visualization of clustering performed

Cluster 20 appears to contain discussions about a benchmarking study. It seems that there is some uncertainty and hesitation about participating in the benchmarking study, possibly due to past issues and challenges faced by Enron. The final decision on whether to participate or not is not explicitly mentioned.

It appears that the cluster 21 is related to a subject concerning a contract or agreement between Enron and FGU companies. The term "gisb" is mentioned, which could refer to a specific contract or agreement. There is also a reference to documents to be implemented in the contract and discussions with the legal departments of both companies.

To improve the quality of our clustering, we need to explore alternative approaches such as adjusting model parameters, using different clustering methods, or extracting new features from the data. [16] A more in-depth analysis of data features and a thorough exploration of clustering techniques can help us achieve more meaningful and useful results. I have calculated the silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, which are commonly used to evaluate the quality of clusters.

- The silhouette score measures how similar samples are within their own cluster compared to other clusters. It ranges from -1 to 1, where a value close to 1 indicates that samples are well-grouped and separated from other clusters. In your case, the obtained silhouette score is 0.46, suggesting some structure within the clusters, although there is some overlap between samples from neighboring clusters. This may indicate some ambiguity or difficulty in clearly distinguishing the clusters.

- The Calinski-Harabasz index is a measure of cluster compactness and separation between clusters. A higher score indicates compact and well-separated clusters. In your case, the Calinski-Harabasz index is 303.24, suggesting some compactness and separation between clusters.

- The Davies-Bouldin index measures the quality of separation between clusters. A lower score indicates better separation between clusters. In your case, the Davies-Bouldin index is 2.16, indicating some overlap or confusion between neighboring clusters.

In summary, the results of your cluster analysis show a relatively coherent structure but with some overlap or ambiguity between neighboring clusters. There appears to be reasonable compactness and separation between the clusters, although there may be some confusion between certain samples.

5 Discussion and future work

5.1 Evaluation of objectives

5.1.1 Extraction of candidate processes from emails

The primary objective of our study was to assess the feasibility of extracting candidate processes from emails, considering both computational efficiency and data relevance. Through our research, we have successfully demonstrated the extraction of additional dimensions of knowledge from email data. By analyzing email threads and applying clustering techniques, we identified patterns and structures within the data that represent potential processes. This finding confirms the existence of genuine underlying processes in organizational email communication. However, we were unable to extract clear processes that can be automated.

5.1.2 Human-understandable metrics for process classification

The second objective was to develop human-understandable metrics that assist analysts in classifying candidate processes as either real processes or non-processes. We aimed to provide interpretable information and support the evaluation and understanding of the extracted processes. Based on our analysis, metrics such as chain length and the number of chains of different lengths did not help us distinguish authentic processes from random email interactions.

5.1.3 Opportunities for automation of email interactions

The third objective explored opportunities for automation in various aspects of email interactions, including improving automatic categorization systems and mitigating negative effects on personal and organizational performance. Although our research primarily focused on data understanding, data preparation, and modeling phases, we identified potential areas for automation. For example, by accurately categorizing emails into folders or generating automatic responses, organizations can streamline their email workflows and enhance overall efficiency.

Unfortunately, in terms of feasibility, our results suggest that the proposed system is not capable of detecting a clear process from the analyzed emails. These results can be attributed to the poor quality and lack of uniformity in the emails. Specific methods for email data cleaning could be applied to address the poor homogeneity. However, in the case of Enron emails, we believe that standard preprocessing is not sufficient. In fact, until another comprehensive email dataset with cutting-edge formatting standards is published, it is imperative to develop a methodology to properly preprocess this dataset.

5.2 Future work

5.2.1 Exploration of different clustering methods, distance measures, and text representations

One idea to explore in future work is to test different clustering methods, distance measures, and text representations in order to compare the performance of various approaches. [13] Although our study employed specific techniques, it would be interesting to explore alternative approaches and algorithms to assess their effectiveness in extracting processes from email data. For instance, clustering methods such as DBSCAN and K-Means could be evaluated. Similarly, the use of alternative distance measures or more advanced text representations, such as pre-trained language models, could provide new insights and improve results.

By conducting comprehensive comparisons among different methods, we could identify the most effective approaches for extracting and classifying processes from email data. This would enhance the robustness and generalizability of our methodology, providing valuable insights for future practical applications across various domains.

5.2.2 Evaluation and validation of metrics

The human-comprehensible metrics[8, 19, 20] proposed for process classification require further evaluation and validation. Future research can focus on conducting user studies and expert evaluations to assess the effectiveness and reliability of these metrics in distinguishing real processes from non-processes. This evaluation process may involve collaboration with domain experts and stakeholders to ensure that the metrics align with their expectations and requirements.

5.2.3 Ethical and privacy considerations

Given that email data contains sensitive information, it is crucial to consider ethical and privacy considerations in future work. [11] Research should prioritize anonymization and protection of personal data in accordance with relevant regulations and guidelines. Additionally, organizations should strive for transparency and obtain informed consent from individuals regarding the use and analysis of their email data.

Conclusion

In conclusion, this thesis aimed to uncover repetitive procedures carried out by humans via emails that could be automated, with the goal of enhancing internal processes within organizations and increasing productivity and performance. The research utilized clustering algorithms on the Enron dataset to achieve this objective.

The literature review provided a comprehensive understanding of the existing research on the analysis of email data and task automation. Various approaches and techniques, such as lexical analysis, language detection, semantic analysis, neural networks, classification algorithms, and clustering techniques, were explored in the context of extracting information from email content.

The thesis focused on email-based process mining, which involves directly analyzing emails to discover, monitor, and enhance real-world processes. Process extraction from texts, process model discovery, and the integration of email mining with process mining techniques were identified as relevant domains within email-based process mining.

Process extraction from texts was recognized as a challenging task due to the complexity of natural language and the potential for multiple valid interpretations. Rule-based systems, template-based systems, and neural network-based systems were discussed as approaches to extract processes from text, each with varying levels of user involvement.

Future work can explore further methods of text representation and similarity measures for a more comprehensive analysis of email data.

References

[1] David Arthur and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[2] Tadeusz Calinski and Jaroslaw Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics - Theory and Methods* 3.1 (1974), pp. 1–27.

[3] Ricardo JG Campello, Davoud Moulavi, and Jörg Sander. "Density-based clustering based on hierarchical density estimates". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics. 2013, pp. 160–168.

[4] H. Chen and W. K. Ng. "Text mining techniques for process analysis and optimization". In: *Journal of Industrial Engineering and Management* 5.2 (2012), pp. 239–262.

[5] *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium. 2000.

[6] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1979), pp. 224–227.

[7] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press. 1996, pp. 226–231.

[8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3 (1996), pp. 37–54.

[9] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan Kaufmann, 2011.

[10] S. Huang et al. "Combining email and event log data for process anomaly detection". In: *Decision Support Systems* 121 (2019), pp. 1–10.

[11] Richard Johnson and Tong Zhang. "Scalable clustering of categorical data". In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 9–17.

[12] J. Li et al. "Email communication pattern mining for process discovery". In: *International Journal of Information Management* 40 (2018), pp. 111–120.

[13] Xiaodong Li and Raymond T Ng. "An efficient algorithm for large-scale sequence clustering". In: *Proceedings of the 17th International Conference on Data Engineering*. IEEE. 2001, pp. 685–694.

[14] Peter J Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[15] Michael Steinbach, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques". In: *KDD Workshop on Text Mining*. 2000.

[16] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2019.

[17] W. M. Van der Aalst, A. J. Weijters, and L. Maruster. "Workflow mining: Discovering process models from event logs". In: *IEEE Transactions on Knowledge and Data Engineering* 16.9 (2004), pp. 1128–1142.

[18] Xindong Wu and Vipin Kumar. *The Top Ten Algorithms in Data Mining*. Chapman and Hall/CRC, 2009.

[19] Xindong Wu et al. "Top 10 algorithms in data mining". In: *Knowledge and information systems* 14.1 (2008), pp. 1–37.

[20] Yiming Yang and Xin Liu. "A re-examination of text categorization methods". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, pp. 42–49.

A Appendices

In this section, I will display the content of certain clusters.

1.1 Cluster 20:

"I gave you some time ago a brochure on this benchmarking study. They renewed their request for what is your view on? Do you think the benefits of knowing going on offset the loss due information released and time spent on? The my recommendation is to forget."

"Vince forwarded by Vince J. Benchmarking Study. Dear Peter Nance and I wanted to our discussion of participation in our benchmarking. Have you discussed participation in the project with your? Here is a small sample set of metrics that were taken from a comprehensive list of over the study group. Will decide which metrics to with credibility and technical. We intend to become the industry standard in. We would very much like to have Enron participate in the. I would be happy to set up another meeting with you and your colleagues with Peter Nance if you think that would be best. Theresa Sanders, Director of Business Development, Alliance of Energy Suppliers, Edison Electric Institute (EEI)."

"I passed this information along to some of my direct reports for their and truthfully hear back from anyone and follow. Two years we agreed to be part of a benchmarking study that was done by Houston office. Wes Colwell was the AA partner who led this and we had mixed. The study was designed to cover gas and despite our involvement in the steering. The study itself was very such that it was troublesome to Jim Fallon at the time refused to submit information on the majority of the questions surrounding. So our survey data for power was not very. I would agree that at this time we will pass on this. Thanks for passing it. I would be happy to convey this directly to the person that has contacted. Just let me know name and number and I will follow through on that if you would hope that your holidays was enjoyable and that the new year will be a good one for Vince J. Kaminski."

"Sally Vince J. Benchmarking Study. I gave you some time ago a brochure on this benchmarking study. They renewed their request for what is your view on? Do you think the benefits of knowing going on offset the loss due information released and time spent on? The my recommendation is to forget. Vince forwarded by Vince J. Benchmarking Study. Dear Peter Nance and I wanted to our discussion of participation in our benchmarking. Have you discussed participation in the project with your? Here is a small sample set of metrics that were taken from a comprehensive list of over the study group. Will decide which metrics to with credibility and technical. We intend to become the industry standard in. We would very much like to have Enron participate in the. I would be happy to set up another meeting with you and your colleagues with Peter Nance if you think that would be best. Theresa Sanders, Director of Business Development, Alliance of Energy Suppliers, Edison Electric Institute (EEI)."

1.2 Cluster 21:

"Sue Mara, Enron, forwarded by Susan J. on 9 PM, 12 PM. Please respond to Carol Escalante Pigott, Larrea, Castillo, Nguyen, Pelote, Sadler, McMorrow, Sugaoka, Woods, Ellery, Gates, A.

Baker, Linthicum, Carter, Hatton, Kebler, Gosselin, Kerner, Nelsen, Tomeo, Koch, Eisenman, DeRosa, Blue, Boyd, Jack Willey, Greco, Ronan, Stout, Weisgall, Lloyd, Hoffman, Fickett, Lednicky, Fillinger, McFadden, Desrochers, Soos, Hickok, Ain, Newell, Iliff, Ponder, J. Mara, Wetzel, Carlson, Hall, Hudson, Kaplan, Moseley, Kelly, IEP, Legislative Report:
Dear attached is this legislative report for. Despite the summer, leaders in the Assembly planned to call back their side of the legislature to debate and vote on a plan for Southern California. The new effort was outlined in our Tuesday fax to each of in the new plan called for amending SB and AB, among other, provide for a dedicated rate component to support a billion bond financing by but would have prevented Edison from using the funds to pay PX and ISO directly or the billion could be used to pay the dedicated rate component for repayment of the bonds would apply to all consumers for two and then apply solely to under. The account concept was the decision on a transmission line purchase was deferred to settlement discussions between the state and the RPS targeted a growth in new generating capacity through and direct access would be permitted under certain circumstances effective no earlier than January. The Assembly leadership planned on holding an informational meeting on the new plan on July and committee and floor votes on the plan July. The leadership could not garner enough votes for this new plan, this so no informational hearing was nor will any votes be held even if this new plan had been approved by the Assembly this Senator office made it very clear that Senator Burton had no intention of bringing his house back in to vote on the barring the Governor calling the Legislature back into we do not expect any other activity on this new plan before the Legislature reconvenes on August and perhaps not then.
Other events this week in Los Angeles: David Freeman, Governor primary negotiator on the contract deals, authored an editorial arguing that the energy contracts were a primary factor in the recent reduction in energy prices on the spot. You can find the article at It ran on the editorial page on July.
It is our understanding that on Thursday, Senator Byron Sher had draft amendments prepared by his staff to amend SB to address renewable portfolio standards. The amendments are not yet, but we are hopeful that we will be able to obtain them when his staff returns next. We will keep you.
FERC ordered hearings on Wednesday to determine how much money electric generators owe California for wholesale power. With California still insisting that it is owed billion by the expected FERC decision on the refund amount is certain to fall short of the lawsuits are expected to follow issuance of the. In a press statement earlier this, the Governor stated his hopes that California will receive all billion for the energy profiteers, and let me make clear that I will not rest until every dollar gouged from California businesses and residents returns to if the FERC does not make California, we will see you.
Energy Secretary Spencer Abraham unveiled a plan on Monday that would expand power grid along path by attracting private investment to cover the million he hopes by increasing the transmission capacity by this plan would alleviate the that has proved for power to flow freely throughout the state during times of peak. According to Senator, Morgan Stanley destroyed key documents relating to the purchase of California power plants, despite the financial repeated promises to the Senate Select Committee to investigate market manipulation that they would turn them as reported on the committee continued to move forward with their contempt proceedings against Enron and Reliant for refusing to hand over key documents to the. Reliant Energy sued California on claiming for the failure to adequately pay for power contracts that the Governor seized from the power exchange in the state claims those contracts equal about

million while Reliant calculates that the same contracts are worth more than a judge is expected to oversee this.
Best, Tom McMorrow and Mike Martinez, Legislative IEPa Legislative Report. Since I get everything that you, you may remove me from your distribution. Hope all is Jeff."
"ENA and FGU currently operate under an interruptible agreement inherited through the CES. FGU has interest in getting a GISB in place with the attached. Could you please review and let me know? You, I would like to complete this process. Jared forwarded by Jared on PM. On PM, GISB contract info as we discussed earlier. FGU is interested in purchasing firm supply at a fixed price for summer. Enron and FGU only have an interruptible contract in place for supply. In order to get a firm contract in place in a short amount of time, I think we should execute a GISB contract if Enron is acceptable to. I have attached two Word documents to be implemented into the GISB. One is general information, and the other is our special. Please have your contract department implement the attached information into a GISB contract and have it sent to me for review. And if your contract department has any, please have them call me at. Thank, Michelle. GISB Special. Dan and I are working on Debra Perlingiere Enron North America Legal Department. Smith EB Texas Phone Fax. The CES GISB was never entered into between CES and. I have reviewed their requested special provisions, and they do not make any. Why should we pay the days before they pay us if we are each selling to the? The imbalance provisions work the way it is set up in the GISB, and their language does theirs attempts to roll defaults and imbalances together, and those are two separate. As such, I have instructed Debra to prepare our standard GISB with our standard special provisions and forward the same to Michelle. And if you have any questions, please call me, and we can. Jared Kaiser PM. Dan J. GISB contract info. ENA and FGU currently operate under an interruptible agreement inherited through the CES. FGU has interest in getting a GISB in place with the attached. Could you please review and let me know? You, I would like to complete this process. Jared forwarded by Jared on PM. On PM, GISB contract info as we discussed earlier. FGU is interested in purchasing firm supply at a fixed price for summer. Enron and FGU only have an interruptible contract in place for supply. In order to get a firm contract in place in a short amount of time, I think we should execute a GISB contract if Enron is acceptable to. I have attached two Word documents to be implemented into the GISB. One is general information, and the other is our special. Please have your contract department implement the attached information into a GISB contract and have it sent to me for review. And if your contract department has any, please have them call me at. Thank, Michelle. GISB Special. Attached is a draft of GISB for your. Please do not hesitate to give me a call with any quetions and/or comments you may. Debra Perlingiere Enron North America Legal Department. Smith EB Texas Phone Fax."

"Is this a real? Do we need Larry? May to price these spread, Mike?"
"To mention that parking fees are a Nancy? Je mieux je? Did I mention that rental fees to use the truck are? That our net account is. Nancy owes Jeff for abalone. Jeff owes Nancy for parking fees. Nancy owes Jeff. Thanks a million for the delicious, the very competitive game of letting me do my and compelling conversation about Jeff. You back at Nancy Sellers PM Prentice Sellers. Did I forget to mention that parking fees are a Nancy? Je mieux je? Did I mention that rental fees to use the truck are? That our net account is. Nancy owes Jeff for abalone. Jeff owes Nancy for parking fees. Nancy owes Jeff. Thanks a million for the delicious, the very competitive game of letting me do my and compelling conversation about Jeff. You back at Nancy Sellers PM Prentice Sellers. Did I forget to mention that parking fees are a

Nancy? Je mieux je? I think that'll be fine for what you need it for on Thursday with the camper shell on. Trust, we need to take the camper shell. Original message Nancy Sellers Prentice Sellers. Did I forget? I mention that rental fees to use the truck are. Our net account is. Owes Jeff for abalone. Owes Nancy for parking fees. Owes Jeff. Thanks a million for the delicious, the very competitive game letting me do my and compelling conversation about back at Nancy Sellers. We need to take the camper shell. Original message Nancy Sellers Prentice Sellers. Did I forget? I mention that rental fees to use the truck are. Our net account is. Owes Jeff for abalone. Owes Nancy for parking fees. Owes Jeff. Thanks a million for the delicious, the very competitive game letting me do my and compelling conversation about back at Nancy Sellers."

"Congress passes the Commodity Futures Modernization Act of late Friday, December. Congress passed the Commodity Futures Modernization Act of. The purpose of the Act is to eliminate unnecessary regulation of commodity futures exchanges and other entities falling within the coverage of the Commodities Exchange Act and to provide legal certainty with regard to certain futures and derivatives by enacting such. Congress intends to promote product innovation and to enhance the competitive position of US financial. Certain provisions of the Act open the door for further product innovation, and the Act provides important legal certainty for energy and other transactions occurring both on a bilateral basis and on multilateral electronic trading. Enron was a leading advocate of passage of this bilateral transactions under the bilateral transactions in all commodities than agriculture that do not occur on trading facility, not on a are exempt from most provisions of the CEA, as long as the transactions are entered solely between contract persons and certain legal entities satisfying capital thresholds and other requirements under the. This provision essentially codifies and expands existing CFTC exemptions for swaps and forward, thereby eliminating a degree of legal ambiguity that has frustrated product innovation and multilateral transactions in the. The Act creates a broad exemption for any contract or transaction in commodities than agriculture, so long as they the transactions are between commercial are entered into on a and that take place on a. This exemption could facilitate expansion of EnronOnline to allow for. However, certain legal requirements will have to be. Other transactions the Act also creates a broad exclusion for a number of commodities that are of interest to, including risk commodities and weather derivatives. The derivative transaction exclusion will have to meet certain legal. This legislation has been over six years in the making, and the collective support of many Enron employees has been. Thank you to all that have contributed to the passage on this important. We will have a more thorough legal summary of this legislation in the near. Do not hesitate to contact me at of Mark Taylor at with any. Great job for you and everyone else in the office that Chris, Long, am, Jeff, Kenneth, Steven, J., Louise, Greg, David, W., John, J., Mark, E., Mark, Jeffrey, A., Richard, Mike, James, D., Mark, Lisa, Mark, Gary, Linda, Joe, Cynthia, Tom, Stephen, D., Allison, Amy, Carolyn, Jeffrey. Commodity Exchange Act passes Congress. Congress passes the Commodity Futures Modernization Act of late Friday, December. Congress passed the Commodity Futures Modernization Act of. The purpose of the Act is to eliminate unnecessary regulation of commodity futures exchanges and other entities falling within the coverage of the Commodities Exchange Act and to provide legal certainty with regard to certain futures and derivatives by enacting such. Congress intends to promote product innovation and to enhance the competitive position of US financial. Certain provisions of the Act open the door for further product innovation, and the Act provides important legal certainty for energy and other transactions occurring both

on a bilateral basis and on multilateral electronic trading. Enron was a leading advocate of passage of this bilateral transactions under the bilateral transactions in all commodities than agriculture that do not occur on trading facility, not on a are exempt from most provisions of the CEA, as long as the transactions are entered solely between contract persons and certain legal entities satisfying capital thresholds and other requirements under the. This provision essentially codifies and expands existing CFTC exemptions for swaps and forward, thereby eliminating a degree of legal ambiguity that has frustrated product innovation and multilateral transactions in the. The Act creates a broad exemption for any contract or transaction in commodities than agriculture, so long as they the transactions are between commercial are entered into on a and that take place on a. This exemption could facilitate expansion of EnronOnline to allow for. however certain legal requirements will have to be met. Other transactions: The Act also creates a broad exclusion for a number of commodities that are of interest, including risk commodities and weather derivatives. The derivative transaction exclusion will have to meet certain legal requirements. This legislation has been over six years in the making, and the collective support of many Enron employees has been invaluable. Thank you to all that have contributed to the passage of this important legislation. We will have a more thorough legal summary of this legislation in the near future. Do not hesitate to contact me, Mark Taylor, at, with any inquiries."