Mini-Project Report:
Clustering Enron Emails for Process Mining

Aslamhom Sidi Mohamed / C13163

04/06/2025

1 Objectives of the Mini-Project

The aim of this mini-project is to explore how unsupervised clustering techniques can reveal thematic or structural groupings within a real-world enterprise email dataset (Enron). This approach represents a first step toward automated discovery of business process structures from unstructured communications, using a custom "from scratch" implementation of the K-Means clustering algorithm.

2 Libraries Used and Relation to Research Axis

- **pandas**: For loading, manipulating, and cleaning tabular data.
- **numpy**: Used for numerical calculations and matrix operations, particularly in implementing K-Means.
- **scikit-learn**: For text feature extraction (TF-IDF vectorization) and dimensionality reduction (PCA).
- **matplotlib**: For 2D visualization of clusters.

These libraries are fundamental in data science and artificial intelligence, and are commonly used in the early stages of process mining research.

3 Dataset Description

The dataset is a subset of the Enron emails, provided in CSV format.

Data preparation steps:

- **Cleaning**: Empty or very short email bodies were removed, along with special characters, email addresses, and URLs.
- **Feature enrichment**: The columns "subject", "body_clean", "from", and "to" were concatenated to create a richer textual representation for each email.
- **Vectorization**: Text data was converted to numeric vectors using TF-IDF (with max_features=1500).

4 Implementation and Results

K-Means Clustering from Scratch

The K-Means algorithm was implemented manually, without using scikit-learn's KMeans class.

1. Random initialization of k centroids ($k = 10$).
2. Assignment of each email to the nearest centroid (using Euclidean distance).
3. Centroid update as the mean of all assigned vectors.
4. Iteration until centroids converge.

Results

- The emails were divided into 10 distinct clusters.
- Manual inspection shows clusters group together emails of similar intent or content, such as "FYI" notifications, detailed discussions, congratulations, meeting arrangements, etc.
- 2D visualization via PCA demonstrates clear cluster separation in many cases.

Sample Output:

- **Cluster 0:**
  Subject: FYI
  Body: Shirley, fyi vince
- **Cluster 1:**
  Subject: Let's meet
  Body: Bjorn, let's meet 11:30 at the lobby...

(Full results and additional cluster samples can be found in the accompanying Jupyter notebook.)

5 Interpretation

- Clustering reveals distinct "communication roles" or topics within unstructured email corpora, without the need for labels.
- Not all clusters are perfectly pure (short or ambiguous emails may appear in any cluster), but the overall structure is coherent and interpretable.
- This approach demonstrates the potential of unsupervised clustering as a preprocessing step for business process discovery in enterprise communications.
- **Limitations:**
  - Clusters are based on lexical similarity, not semantic meaning.
  - The number of clusters (k) greatly affects granularity: too small merges different types; too large fragments groups.

6 Research Perspectives

- **Improve textual representation**: Apply advanced language models (BERT, LLMs) for more meaningful text features.
- **Automatic activity extraction**: Use LLMs to automatically detect and label business activities in emails.
- **Semi-supervised or guided clustering**: Combine expert input to guide the grouping process.
- **Full process mining**: Link email clusters to real process steps, or combine clustering with sequence extraction for end-to-end process mining.

7 References

- Scikit-learn documentation: `https://scikit-learn.org/`
- Enron Email Dataset (Kaggle): `https://www.kaggle.com/wcukierski/enron-email-dataset`
- van der Aalst, W.M.P. (2016). *Process Mining: Data Science in Action*. Springer.

Document prepared by Aslamhom / C13163