NLP - P8 - Write Me

Stefano Nava 27436A

Abstract

The extraction of meaningful information from unstructured email text is a critical challenge in the field of natural language processing (NLP). This project explores a two-step approach to the email content analysis: structural classification and topic modeling. First, emails were segmented into structural components - greetings, body, and closings - using pattern-based matching and regular expressions. This classification enabled a more granular analysis of the content and improved the consistency of the subsequent processing. After structural classification, the email body was cleaned through a preprocessing pipeline that included tokenization, stopword removal, and lemmatization. The text data was then vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) and processed with Non-Negative Matrix Factorization (NMF) to uncover distinct topics. The identified topics were evaluated based on their coherence and distribution within the dataset. The results demonstrate the ability of NMF to effectively cluster emails into meaningful categories, highlighting key patterns and recurring subjects. This combined approach provides a foundation for improving automated classification and content analysis of large volumes of email data, with potential applications in spam detection, email filtering, and content-based recommendation systems.

1 Introduction

The rapid increase in the volume of email communication has made it essential to develop efficient methods for extracting meaningful information from unstructured text. Emails are inherently complex due to their partially structured nature, which includes both formal elements — such as greetings, signatures, and metadata — and dynamic content that reflects the dialogical nature of communication. The ability to automatically analyze and categorize email content has significant applications in information retrieval, email filtering, spam detection, and business intelligence.

Natural Language Processing (NLP) provides powerful tools for analyzing unstructured text, but emails pose specific challenges due to their mixed structure and informal language. Previous research on email classification has focused on either structural analysis or semantic content extraction. Structural classification aims to segment emails into standardized components such as greetings, body, and closings, which enables better consistency in processing and improves the accuracy of subsequent analysis. On the other hand, topic modeling allows for the discovery of hidden patterns and recurring themes within large datasets of textual content, revealing meaningful insights into the subject matter of the emails.

This project addresses both aspects by combining structural classification and topic modeling into a unified processing pipeline. The structural classification component segments each email into greetings, body, and closing parts using regular expressions and pattern-based matching. This improves the consistency and interpretability of the data for the next stage. The topic modeling component applies Non-Negative Matrix Factorization (NMF) to identify latent themes within the email bodies after a preprocessing phase involving tokenization, stopword removal, and lemmatization. This combined approach leverages the advantages of both structural analysis and semantic modeling to enable more effective and interpretable analysis of large-scale email datasets.

2 Research question and methodology

2.1 Research Question

The primary goal of this project is to develop an automated system for analyzing unstructured email text by addressing two fundamental challenges: (1) identifying the structural components of an email, and (2) discovering latent topics within the email body to uncover patterns and recurring themes. The research aims to answer the following questions:

- Can structural classification improve the consistency and quality of subsequent topic modeling on email content?
- How effectively can Non-Negative Matrix Factorization (NMF) identify meaningful topics from a dataset of email bodies after structural segmentation and preprocessing?

By addressing these questions, the project seeks to explore the potential of combining structural analysis and semantic modeling to improve the understanding and automatic processing of email data.

2.2 Methodology

The proposed approach is structured into two main phases: structural classification and topic modeling. Their combination is designed to improve the interpretability and performance of email analysis, providing a structured and semantically rich understanding of the dataset.

2.2.1 Structural classification

The first phase involves segmenting each email into three distinct structural components:

- **Greeting**: common opening phrases (e.g., "Dear John," "Hi," "Hello") are detected using regular expressions designed to match common patterns in business and personal email communication.
- **Body**: the main content of the email is extracted after removing the greeting and before detecting the closing section.
- **Closing**: common closing phrases (e.g., "Best regards," "Sincerely," "Thank you") are identified using regular expressions.

This segmentation enables more consistent preprocessing and helps reduce noise in the subsequent topic modeling phase.

2.2.2 Body preprocessing

After structural classification, the body of each email is processed through a text-cleaning pipeline that includes:

- **Tokenization** – Splitting the text into individual words.
- **Stopword Removal** – Removing common, non-informative words using a predefined list of stopwords from the nltk library.
- **Lemmatization** – Converting words to their base form using WordNetLemmatizer to reduce variability and improve clustering accuracy.
- **Character Cleaning** – Removing special characters and retaining only alphanumeric content and periods.

2.2.3 Vectorization and Dimensionality Reduction

The processed email bodies are vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) to create a numerical representation of the text. TF-IDF reflects the importance of each term relative to the entire dataset.

To reduce the dimensionality of the data and improve the efficiency of topic modeling, Truncated Singular Value Decomposition (TruncatedSVD) is applied, retaining the most informative components of the vectorized data.

2.2.4 Topic Modeling

After dimensionality reduction, Non-Negative Matrix Factorization (NMF) is used to extract latent themes from the data. NMF is particularly suitable for text analysis because it imposes non-negativity constraints, which leads to a more interpretable and human-readable representation of topics.

The number of topics is defined through empirical tuning, and the most representative terms for each topic are extracted to provide a clear understanding of the discovered themes.

2.2.5 Evaluation

The evaluation of the topic modeling process is based on three key metrics: *Coherence Score*, *Topic Diversity*, and *Reconstruction Error*. These metrics provide a comprehensive assessment of the quality and consistency of the discovered topics, as well as the model's ability to represent the input data accurately. The evaluation is performed during the topic modeling phase to exploit the available data directly within the processing pipeline.

Coherence Score

The coherence score measures the semantic similarity among the terms within each topic. A high coherence score indicates that the top terms in a topic are semantically related, reflecting a well-defined theme. In this project, the coherence score is calculated using *cosine similarity* between the vector representations of the top n terms in each topic. The average similarity across all topics is computed as the final coherence score:

$$C = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i<j} \mathrm{sim}(w_i, w_j)}{\binom{n}{2}}$$

where:

- T = number of topics
- w_i, w_j = vector representations of terms within a topic
- $\mathrm{sim}(w_i, w_j)$ = cosine similarity between terms

A coherence score close to 1 indicates that the terms within a topic are highly similar, improving the interpretability of the results.

Topic Diversity

Topic diversity measures the degree of uniqueness of the terms among different topics. A high topic diversity score indicates that the topics are well-separated and cover different areas of the text. It is computed as the ratio between the number of unique terms in the top n terms across all topics and the total number of terms ($n \cdot T$) in the top n list for all topics:

$$D = \frac{\text{number of unique terms}}{n \cdot T}$$

where:

- T = number of topics
- n = number of top terms per topic

A diversity score close to 1 means that the discovered topics are distinct and non-overlapping.

Reconstruction Error

The reconstruction error measures how well the model approximates the original data. It is computed directly from the NMF model as the Frobenius norm between the original input matrix X and the reconstructed matrix $\hat{X} = W \cdot H$:

$$E = ||X - W \cdot H||_F$$

where:

- W = document-to-topic matrix
- H = topic-to-term matrix

A low reconstruction error indicates that the model accurately represents the input data, while a high error suggests that the model is struggling to approximate the underlying structure of the text.

Evaluation Strategy

The combination of coherence, diversity, and reconstruction error provides a balanced evaluation strategy:

- A **high coherence score** reflects that the topics are internally consistent.
- A **high diversity score** indicates that the topics are well-separated and distinct.
- A **low reconstruction error** demonstrates that the model captures the key patterns in the data effectively.

In this project, the goal was to optimize coherence and diversity while maintaining a reasonably low reconstruction error. Fine-tuning the number of topics and the regularization parameters for NMF was critical to balancing these objectives.

3 Experimental results

4 Concluding remarks