Write Me P8 (2023/2024)

ANALYSIS AND STRUCTURAL AND THEMATIC CLASSIFICATION OF UNSTRUCTURED EMAIL TEXT

STEFANO NAVA

NATURAL LANGUAGE PROCESSING

27436A

INTRODUCTION

- **Challenge**: Extracting information from unstructured email text.
- Emails are **complex**: **formal** elements + **dynamic** content.
- Need for efficient **analysis methods**.
- NLP tools are useful, but emails pose specific challenges.

PROJECT GOAL & APPROACH

- **Goal**: Develop an automated system for email text analysis.

- Address two **challenges**: structural components & latent topics.

- **Approach**: Two main phases combined:
  - Structural Classification
  - Topic Modeling

- **Aim**: Improve interpretability and performance.

PHASE 1: STRUCTURAL CLASSIFICATION

- **Purpose**: Segment email into Greeting, Body, Closing.
- **Method**: Regular expressions and pattern matching.
- Body is the main content, extracted after removing Greeting/Closing.
- **Benefit**: Enables consistent preprocessing, reduces noise, improves interpretability for next phase.

PHASE 2: TOPIC MODELING - PROCESS

- **Analysis** on the extracted Email Body.
- **Preprocessing**: Tokenization, Stopword Removal, Lemmatization, Character Cleaning.
- **Vectorization**: TF-IDF (numerical representation).
- **Dimensionality Reduction**: TruncatedSVD.

PHASE 2: TOPIC MODELING - MODEL

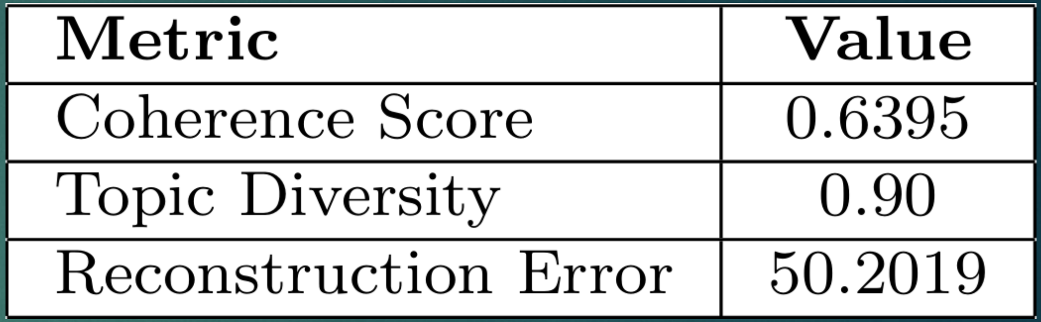
- **Model**: Non-Negative Matrix Factorization (NMF).
- **Reason**: Leads to interpretable topics.
- **Configuration**: 5 Topics, 1000 max iterations, 10 top terms/topic.
- **Data**: 3332 fraudulent emails.

EVALUATION METRICS

Metrics used:

- **Coherence Score**: Measures semantic similarity within topics. (High = good)
- **Topic Diversity**: Measures uniqueness of terms across topics. (High = good)
- **Reconstruction Error**: Measures how well model approximates original data. (Low = good)

EVALUATION RESULTS



Metric	Value
Coherence Score	0.6395
Topic Diversity	0.90
Reconstruction Error	50.2019

Summary:

- **Coherence Score**: good internal consistency
- **Topic Diversity**: high topic separation
- **Reconstruction Error**: relatively high, suggests approximation can improve

Overall: NMF identified coherent and distinct themes despite data approximation challenge.

DISCOVERED TOPICS (EXAMPLES)

Identified 5 **themes** in fraudulent emails:

- Gov/Diplomat Impersonation (e.g., "ministry", "nigeria")
- Inheritance/Financial Scams (e.g., "son", "family", "late")
- Accident-related Unclaimed Funds (e.g., "plane", "crash", "deceased")
- Legal Matters (e.g., "deposit", "document", "attorney")
- Religious Scams (e.g., "lord", "god", "charity")

LIMITATIONS & FUTURE WORK

- **Limitations**: High Reconstruction Error, Bag-of-words ignores context, potential topic overlap.
- **Future Work**: Use contextual embeddings (BERT), Combine NMF with clustering, Expand dataset.
- **Conclusion**: NMF showed promising results for topic separation and coherence on fraudulent emails. Further tuning and hybrid methods could enhance performance.