National University of Computer & Emerging Sciences
Karachi Campus

PROJECT REPORT

Learning Latent Personas of Film Characters

Course Name: EL-2003 INFORMATION RETRIEVAL
Section: 6-H

Group Members:

- 22K-4148 Shayan
- 22K-4159 Rania Ghazanfar
- 22K-4320 Marium Arif

TABLE OF CONTENTS:

ABSTRACT:

This project implements and evaluates a Persona-based topic modeling framework to uncover latent personas and topics from movie character dialogues and metadata. Using a Dirichlet persona model with regression, the system aims to learn meaningful latent structures representing characters' personas and their interactions. The model's iterative optimization process is examined with respect to convergence, accuracy, and computational efficiency. The report covers methodology, experimental setup, analysis, and recommendations for future improvements.

1. INTRODUCTION:

Topic modeling is a powerful technique to extract latent themes from large text corpora. This project focuses on an advanced topic modeling framework — a Persona Model — which jointly models topics and latent personas of film characters using probabilistic graphical models. Unlike traditional Latent Dirichlet Allocation (LDA), the Persona Model integrates persona regression to link character metadata and dialogue for richer semantic understanding.

The project utilizes a dataset consisting of movie scripts, character metadata, and movie metadata, implementing the Persona Model based on the CMU Personas project codebase. The objective is to analyze and improve convergence behavior, assess model outputs, and recommend efficient parameter tuning strategies.

2. FEATURES:

- Latent Persona Discovery: Identifies hidden personas of characters from dialogues.
- Topic Modeling: Extracts meaningful topics associated with personas.
- Persona Regression Model: Incorporates metadata for improved persona estimation.
- Configurable Parameters: Adjustable topic numbers, vocabulary size, iterations.
- Output Analysis: Logs model parameters at each iteration for convergence tracking. Support for Large Datasets: Handles metadata integration for real-world film datasets.

3. METHODOLOGY:

3.1 Data Preparation

- Input Data: Dialogue corpus from movie scripts. Character metadata (demographics, traits). Movie metadata (genre, release year).
- Preprocessing: Tokenization, stop-word removal, and vocabulary limiting to top 1000 tokens. Metadata cleaning and feature extraction.

3.2 Model Setup

Parameters:

- Number of topics (K=50).
- Number of personas (A=50).
- Vocabulary size (V=1000).
- Dirichlet priors (alpha=10, gamma=1).
- L2 regularization (0.01).
- Maximum iterations (originally 50000; tuned down for efficiency).

Algorithm:

- Persona Model with Persona Regression.
- Iterative optimization with convergence monitoring.

3.3 Implementation

- Shell script **(run.sh)** generates a properties file with model configurations.
- Java executable PersonaModel is run with the properties.
- Logs capture parameter updates per iteration (gamma, nuA, nuP, nuM).

3.4 Evaluation

- Monitor convergence of parameters over iterations.
- Compare results with baseline LDA and simpler topic models.
- Analyze output personas and topics qualitatively.

4. CONSTRAINTS AND ASSUMPTIONS:

Constraints:

- High computational cost due to large iteration counts.
- Memory limits in handling large vocabularies or datasets.
- Dependence on quality and completeness of metadata.

Assumptions:

- Dirichlet priors adequately model persona-topic distributions.
- Character metadata is relevant and accurately reflects persona features.
- Stochastic optimization converges given enough iterations.

5. COMPARISONS WITH OTHER MODELLING TECHNIQUES:

Modelling Technique	Description	Strengths	Weaknesses	Suitability for Persona Extraction
Latent Dirichlet Allocation (LDA)	A generative probabilistic model for discovering topics in text.	Simplicity, scalability, interpretable topics.	Ignore metadata and persona structure.	Good for topic modeling but limited for personas.
Author-Topic Model (ATM)	Extends LDA by associating authors with topics.	Captures author influence on topics.	Metadata limited to authors only.	Useful if authorship is key, less fine-grained personas.
Persona Model with Regression	Joint model of topics and latent personas with metadata regression.	Integrates rich metadata; discovers latent personas.	Computationally expensive; convergence can be slow.	Best suited for persona and topic joint extraction.
Neural Topic Models (e.g., Neural Variational LDA)	Neural network based probabilistic topic modeling.	Handles complex dependencies, flexible priors.	Requires more data; less interpretable.	Can be adapted for personas but less straightforward.
Clustering + Feature Engineering	Unsupervised clustering of character features and dialogue.	Easy to implement, scalable.	May miss latent topic-persona structure.	Baseline approach, less principled than generative models.

This project's Persona Model outperforms traditional topic models for persona discovery by leveraging character metadata and a regression framework, at the cost of increased computational demands.

6. OUTPUT AND ANALYSIS:

The model outputs several files containing:

- Character posterior probabilities.
- Conditional posterior distributions.
- Persona assignments and regression weights.

Parameter logs show gradual reduction in change magnitudes (gamma, nuA, etc.), indicating convergence.

Sample topics extracted relate closely to known character traits (e.g., hero, villain, mentor). Persona profiles capture latent character dimensions combining dialogue and metadata.

7. RECOMMENDATIONS:

- Parameter Tuning: Reduce maxIterations to 10,000 or fewer based on convergence monitoring to save time.
- Early Stopping: Implement or use an early stopping mechanism based on parameter change thresholds.
- Data Enhancement: Improve metadata quality for better regression modeling.
- Parallelization: Explore parallel processing to accelerate training.
- Model Extensions: Incorporate additional modalities (e.g., scene context, sentiment).

8. CONCLUSION:

This project successfully implemented a Persona-based topic modeling framework that leverages character and movie metadata to extract latent personas from film dialogue data. While the model produces richer representations than traditional LDA, it requires careful parameter tuning and monitoring to balance computational costs with output quality. The insights gained here can inform future developments in persona and topic modeling in multimedia contexts.

9. REFERENCES:

- Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning Latent Personas of Film Characters. ACL 2013.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. McCallum, A. K. (2002).
- MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu