National University of Computer & Emerging Sciences
Karachi Campus

PROJECT CODE DOCUMENTATION

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:

1. OVERVIEW:

This project implements a Persona-based topic modeling framework for learning latent personas of film characters using dialogue and metadata. It leverages a Dirichlet persona model with regression.

2. ENVIRONMENT SETUP:

- **OS:** Linux/Ubuntu recommended.
- **Java:** Ensure Java JDK 8 or above installed.
- **Shell:** zsh or bash for running scripts.
- **Dependencies:** No external dependencies apart from Java runtime.

3. FILE STRUCTURE:

- project_root/
- |── Learning Latent Personas of Film Characters/
- │ |── java
- │ | | — run.sh
- │ | | — output.properties
- │ | | — output.out/
- │ | | |— 25.100.lda.log.txt
- │ | | |— 25.100.lda.cond.log.txt
- │ | | |— lr.weights.txt
- │ | | |— out.phi.weights
- │ | | |— personaFile
- │ |── preprocess
- │ | | — corenlp_plot_summaries
- │ | | — MovieSummaries
- │ | | — SupersenseTagger
- │ | | — pipeline.sh
- │ |── README.md

4. CONFIGURATION:

run.sh accepts two arguments:

- $1: Project run name
- $2: Input dialogue file path.

It generates a properties file ${name}.properties with parameters:

- K: number of topics.
- A: number of personas.
- V: vocabulary size.
- alpha, gamma: Dirichlet priors.
- L2: regularization for regression.
- maxIterations: maximum iterations.
- Paths for input metadata and outputs.

5. MODEL EXECUTION:

To execute the model, we run the script: **./run.sh output input/dialogues.txt**

- Creates output directory **output.out**.
- Writes **${name}.properties**.
- Executes the PersonaModel Java program.
- Outputs logs and results to output.out directory

6. OUTPUT FILES:

- **25.100.lda.log.txt:** Character posterior probabilities.
- **25.100.lda.cond.log.txt:** Conditional posteriors.
- **out.phi.weights:** Topic-word weights.
- **lr.weights.txt:** Regression model weights.
- **personaFile:** Persona assignments.
- Other supporting files for analysis.

7. MONITORING AND TROUBLESHOOTING:

- Monitor logs for parameter convergence.
- Adjust maxIterations and priors if the model does not converge.
- Check for file paths and permissions.
- Use the commented logging command in run.sh for verbose logs.

8. EXTENDING THE MODEL:

- Modify run.sh to change parameters.
- Integrate additional metadata features by updating input files.
- Explore parallel runs for parameter tuning.
- Implement early stopping in Java code if possible.