Dependencies Report

This report outlines the dependencies used in the project, their purposes, and their roles in the overall system. The project relies on a combination of \*\*Python libraries\*\* for data processing, graph construction, natural language processing (NLP), and evaluation. Below is a detailed breakdown of the dependencies:

1. Core Libraries

1.1 Pandas

- Purpose: Data manipulation and analysis.

- Role: Used for loading the dataset (`final\_labels.csv`), cleaning missing data, and creating new columns for reconstructed threads and summaries.

- Key Functions:

- `pd.read\_csv()`: Load the dataset.

- `df.dropna()`: Remove rows with missing values.

- `df.apply()`: Apply functions to rows or columns.

- `df.to\_csv()`: Save processed data to a CSV file.

1.2 NetworkX

- Purpose: Graph construction and manipulation.

-Role: Used to create a directed graph (`DiGraph`) to represent the hierarchical structure of discussion threads.

- Key Functions:

- `nx.DiGraph()`: Create a directed graph.

- `G.add\_node()`: Add nodes (comments) to the graph.

- `G.add\_edge()`: Add edges (parent-child relationships) to the graph.

- `nx.dfs\_preorder\_nodes()`: Perform depth-first search to reconstruct threads.

2. Natural Language Processing (NLP) Libraries

2.1 Transformers (Hugging Face)

- Purpose: Pre-trained NLP models for summarization and text generation.

- Role: Used for generating summaries of discussion threads using the T5-base model.

- Key Components:

- `pipeline("summarization")`: Load the summarization pipeline.

- `GPT2LMHeadModel` and `GPT2Tokenizer`: Used for calculating perplexity of summaries.

2.2 Sentence-Transformers

- Purpose: Compute semantic similarity between text pairs.

- Role: Used for detecting context mismatches between comments and their parents, as well as evaluating semantic similarity between summaries and original threads.

- Key Components:

- `SentenceTransformer("all-MiniLM-L6-v2")`: Load the pre-trained sentence embedding model.

- `util.pytorch\_cos\_sim()`: Compute cosine similarity between embeddings.

2.3 NLTK

- Purpose: Text processing and evaluation.

- Role: Used for computing the BLEU score to evaluate summary quality.

- Key Functions:

- `sentence\_bleu()`: Compute BLEU score between a candidate summary and reference text.

2.4 Rouge-Score

- Purpose: Evaluation of summarization quality.

- Role: Used for computing ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) to measure overlap between summaries and original threads.

- Key Components:

- `RougeScorer()`: Initialize the ROUGE scorer.

- `scorer.score()`: Compute ROUGE scores for a pair of texts.

3. Machine Learning and Deep Learning Libraries

3.1 PyTorch

- Purpose: Deep learning framework.

- Role: Used for loading and running the GPT-2 model to calculate perplexity.

- Key Components:

- `torch.device()`: Manage GPU/CPU device selection.

- `torch.tensor()`: Convert data to tensors for model input.

- `torch.exp()`: Compute perplexity from model loss.

3.2 NumPy

- Purpose: Numerical computations.

- Role: Used for handling numerical operations, such as computing averages and filtering invalid perplexity values.

- Key Functions:

- `np.nan`: Represent missing or invalid values.

- `np.isfinite()`: Check for finite values.

4. Utility Libraries

4.1 OS and Sys

- Purpose: System-level operations.

- Role: Used for file handling and system configuration (e.g., GPU/CPU selection).

- Key Functions:

- `torch.cuda.is\_available()`: Check for GPU availability.

5. Dependency Installation

To install all dependencies, use the following commands:

Command:

pip install pandas networkx transformers sentence-transformers nltk rouge-score torch numpy

6. Dependency Management

6.1 Virtual Environment

- It is recommended to use a virtual environment to manage dependencies and avoid conflicts.

- Create and activate a virtual environment:

Command:

python -m venv myenv

source myenv/bin/activate # On Windows: myenv\Scripts\activate

6.2 Requirements File

- Save the dependencies in a `requirements.txt` file:

plaintext

pandas

networkx

transformers

sentence-transformers

nltk

rouge-score

torch

numpy

- Install dependencies from the file:

Command:

pip install -r requirements.txt

7. Key Dependencies and Their Versions

| Library | Purpose | Version (Recommended) |

|--------------------|----------------------------------|-----------------------|

| Pandas | Data manipulation | 2.0.0+ |

| NetworkX | Graph construction | 2.8.0+ |

| Transformers | NLP models (T5, GPT-2) | 4.30.0+ |

| Sentence-Transformers | Semantic similarity | 2.2.0+ |

| NLTK | BLEU score calculation | 3.8.0+ |

| Rouge-Score | ROUGE score calculation | 0.1.1+ |

| PyTorch | Deep learning framework | 2.0.0+ |

| NumPy | Numerical computations | 1.23.0+ |