Requirements for Bug Report Classification with RoBERTa

This document lists all dependencies and requirements needed to run the Bug Report Classification tool with RoBERTa, as described in the accompanying research paper.

Hardware Requirements

For Google Colab Implementation

- **GPU**: NVIDIA L4 (recommended) or other CUDA-compatible GPU
- **RAM**: At least 16 GB (High-memory runtime recommended)
- **Storage**: Sufficient Google Drive space (minimum 10GB) for model artifacts, datasets, and results

For Local Implementation

- **GPU**: NVIDIA GPU with at least 8GB VRAM (strongly recommended)
- **RAM**: Minimum 16GB, recommended 32GB for larger datasets
- **Storage**: At least 10GB free space for model artifacts, datasets, and results
- **CPU**: Multi-core processor (8+ cores recommended for faster preprocessing)

Software Requirements

Python Environment

- Python 3.8 or higher

Required Python Packages

The following packages with their versions are required to ensure reproducibility:

```
transformers==4.36.2
datasets==2.15.0
scikit-learn==1.2.2
pandas==1.5.3
numpy==1.24.0
matplotlib==3.7.2
seaborn==0.12.2
torch==2.1.2
nltk==3.8.1
tqdm==4.66.1
```

Google Colab Installation

In Google Colab, use the following code to install the required packages:

```
!pip install transformers==4.36.2 datasets==2.15.0 scikit-learn==1.2.2 pandas==1.5.3 numpy==1.24.0 matplotlib==3.7.2 seaborn==0.12.2
```

Local Installation

For local installation, it's recommended to create a virtual environment:

```
# Create virtual environment
python -m venv ise_env

# Activate environment (Windows)
ise_env\Scripts\activate

# Activate environment (Linux/Mac)
source ise_env/bin/activate

# Install packages
pip install transformers==4.36.2 datasets==2.15.0 scikit-learn==1.2.2 pandas==1.5.3 numpy==1.24.0 matplotlib==3.7.2 seaborn==0.12.2 t
```

Pre-trained Models

The tool uses the pre-trained RoBERTa base model from Hugging Face Transformers:

- `roberta-base` (approximately 500MB download)

This model will be automatically downloaded when running the code for the first time. Ensure you have an internet connection for the initial setup.

NLTK Resources

The following NLTK resources are required:

- stopwords

These resources will be automatically downloaded when running the code using:

```
import nltk
nltk.download('stopwords')
```

Dataset Requirements

Dataset Format

The tool uses CSV files containing bug reports with the following columns:

- `Title`: The title of the bug report
- `Body`: The body/description of the bug report (can be NaN)
- `class`: Binary classification label (0 for non-performance bug, 1 for performance bug)

Included Datasets

The following datasets are included in the repository's `data/` directory:

- caffe.csv
- tensorflow.csv
- keras.csv
- pytorch.csv
- incubator-mxnet.csv

Dataset Characteristics

As referenced in the research paper:

- Combined dataset comprises 3,712 GitHub issue reports from five deep learning frameworks
- Class imbalance exists with approximately 16.4% of reports labeled as performance-related (class 1)

Storage Structure

Google Colab Structure

```
/content/drive/MyDrive/BugReportClassification/
├── data/             # For storing CSV datasets
├── models/           # For storing trained models
└── results/          # For storing evaluation results and metrics
```

Local Structure

```
BugReportClassification/
├── data/             # For storing CSV datasets
├── models/           # For storing trained models
└── results/          # For storing evaluation results and metrics
```

This structure will be automatically created by the code, but you may need to create it manually if running a modified version of the code.

Important Note About Models

The trained models are not included in the GitHub repository due to their large file size. When you run the code, the `models` directory will be created automatically, and trained models will be saved there. You will need to train the models yourself by following the instructions in the manual and replication guide.

For more information, please refer to the project repository: https://github.com/Alva1103/ISE-Coursework.git (https://github.com/Alva1103/ISE-Coursework.git)