תמונה שמכילה טקסט, גופן, לוגו, גרפיקה

התיאור נוצר באופן אוטומטי

Software Engineering Department  
ORT Braude College

Capstone Project Phase B – 61999

**Metalanguage As an Interdisciplinary Classifier for Mathematics and Computer Science Fields**

Ilya Vorotyntsev | Ilya.Vorotyntsev@e.braude.ac.il

Aleksandr Pitkin | Aleksandr.Pitkin@e.braude.ac.il

This Maintenance Guide ensures that future maintainers can **continue to use, update, and enhance** the PDF-to-text and Classification pipeline after the initial project has ended. It provides information on:

1. **Software and hardware environment** required to run the system.
2. **Installation procedures** for the specialized components of this project.
3. **Recommended practices** for updating libraries and modifying the pipeline.

**2. System Environment**

**2.1 Software Infrastructure**

* **Operating System**
  + Primarily tested on Google Colab.
  + Can be adapted for Windows/macOS if Tesseract OCR and other dependencies are properly installed.
* **Python Environment**
  + Python 3.7 or higher (3.9+ recommended).
  + A virtual environment (e.g., venv or conda) is suggested to avoid library conflicts.
* **Key Python Libraries** (installed via pip)
  + **PDF Extraction**
    - PyMuPDF (fitz)
  + **OCR**
    - tesseract-ocr
    - pytesseract
    - pdf2image, poppler-utils
  + **NLP & ML**
    - nltk, gensim, scikit-learn, scikit-learn-extra
    - transformers, torch
  + **Data Handling & Visualization**
    - pandas, numpy, matplotlib, seaborn

**2.2 Hardware Requirements**

* **Memory**:
  + At least 8 GB RAM if dealing with medium-to-large PDF collections or running OCR heavily.
  + 16 GB+ recommended for big clustering/classification tasks or large-scale OCR.
* **CPU/GPU**:
  + A multi-core CPU for parallel operations.
  + A GPU can speed up XLNet or other deep learning tasks (optional, but beneficial).

**3. Installation Instructions (Specialized Components)**

This section focuses on **what is unique to your project**—i.e., how to set up the Python environment, Tesseract, or any other specialized scripts you developed.

**3.1 Project Repository Setup**

1. **Clone or Download** the project folder from your version control system (e.g., Git).
2. **Ensure Folder Structure** is maintained, especially for:
   * The main notebooks/scripts (e.g., final-project.ipynb, pdf\_to\_txt\_conversion.ipynb).
   * Dictionaries (domain-specific .txt files).
   * PDF input folder(s).

**3.2 Data Dictionaries**

* The system references specialized **domain dictionaries** (e.g., cs.txt, math.txt) plus an **English dictionary** (oxford.txt):
  + Place these files under a known directory (e.g., dictionaries/ or directly in BASE\_DIR).
  + If you add new domain terms or remove old ones, keep these .txt files in sync.

**4. Maintaining and Updating the System**

**4.1 Code Structure Overview**

* **pdf\_to\_txt\_conversion.ipynb / .py**
  + Handles PDF extraction using PyMuPDF, and Tesseract OCR.
* **final-project.ipynb / .py**
  + Orchestrates data preprocessing (lemmatization, term replacement), Doc2Vec embedding, clustering, and XLNet classification.
* **dictionaries/**
  + Contains .txt files listing domain terms (CS, Math) and Oxford words.

**4.2 Applying Changes and Updates**

1. **Modifying Preprocessing Logic**
   * Look inside the preprocess\_data() function.
   * Adjust or extend text cleaning routines, dictionary references, or chunk sizes.
2. **Adding a New PDF Extraction Library**
   * Implement a new function (similar to convert\_with\_pymupdf()) for your library of choice.
   * Integrate it into process\_files(folder\_path) to automatically handle the new method.
3. **Updating Doc2Vec Parameters**
   * If embeddings need tuning, edit doc2vec\_params in the main script, adjusting vector\_size, dm, epochs, etc.
4. **Switching the Classifier**
   * The pipeline uses XLNet from transformers.
   * To switch to BERT, GPT, or another model, replace references to XLNetTokenizer/XLNetForSequenceClassification with the corresponding classes (e.g., BertTokenizer, BertForSequenceClassification).