Multilingual Dataset Creation and Cleaning Report

This document describes the methodology and implementation of a Python script that generated cleaned, structured instruction–input–response (IIR) datasets in Finnish and Swedish from PDF files.

1. Purpose

The goal is to extract meaningful sentence-level data from agricultural and environmental PDFs and transform it into high-quality IIR samples suitable for fine-tuning instruction-following language models like Viking LLM.

2. Tools Used

- pdfplumber – for accurate text extraction from PDFs
- langdetect – to verify language of each sentence
- nltk – to split raw text into well-formed sentences
- json – to store output in structured .jsonl format

3. Configuration Parameters

- DESIRED_COUNT = 250: Number of validated samples per language
- OVERGENERATION_FACTOR = 20: Tries 20× more candidates to ensure enough pass
- MIN_INSTRUCTION_WORDS = 2
- MIN_RESPONSE_WORDS = 5
- MAX_RESPONSE_WORDS = 600

4. Supported Languages

The script currently supports:
- Finnish (fi)
- Swedish (sv)

Each language uses its own localized instruction templates and input phrasing.

5. Script Workflow

1. Load all PDF files in a target folder.

2. Extract visible text using pdfplumber.
3. Tokenize text into sentences with nltk.
4. Clean each sentence (remove URLs, ISBNs, institutional tags).
5. Check if sentence is the desired language using langdetect.
6. Wrap sentence in a randomized natural-language instruction template.
7. Attach a corresponding language-specific input phrase.
8. Validate word counts and save structured JSONL.

6. Sample Output Format

Each output entry is structured as:

```
{
  "instruction": "Mitä seuraava väite tarkoittaa käytännössä: Biohiili parantaa maan rakennetta.",
  "input": "Kuvaile tätä tarkemmin:",
  "response": "Biohiili parantaa maan rakennetta."
}
```

7. Output

Two output files are produced:
- dataset_finnish_cleaned.jsonl
- dataset_swedish_cleaned.jsonl

Each dataset is intented to have 250 cleaned and validated samples in instruction–input–response format.

✅ This dataset is now ready for use in multilingual instruction-tuning pipelines such as Axolotl or LoRA fine-tuning.