Maximizing Data Communication for Faster Energy Access

Brooke Erickson, Alejandro Ortega, Jade Wu
Dr. Rebekah Shirley, Dr. T. Robert Fetter, Dr. Jonathan Phillips, Scott Barnard, Wayne de Jager

This project explores how ***machine learning*** and ***natural language processing*** tools can facilitate improvements for Power for All's Platform for Energy Access Knowledge (PEAK), which automatically curates, organizes, and streamlines large, growing bodies of data into digestible and sharable knowledge to better educate policymakers and researchers alike.

In order to optimize the usability of this existing data, we have created three tools to be implemented into the PEAK platform:

1. Document Auto-Categorization
2. PDF Table Extraction
3. Data Auto-Identification

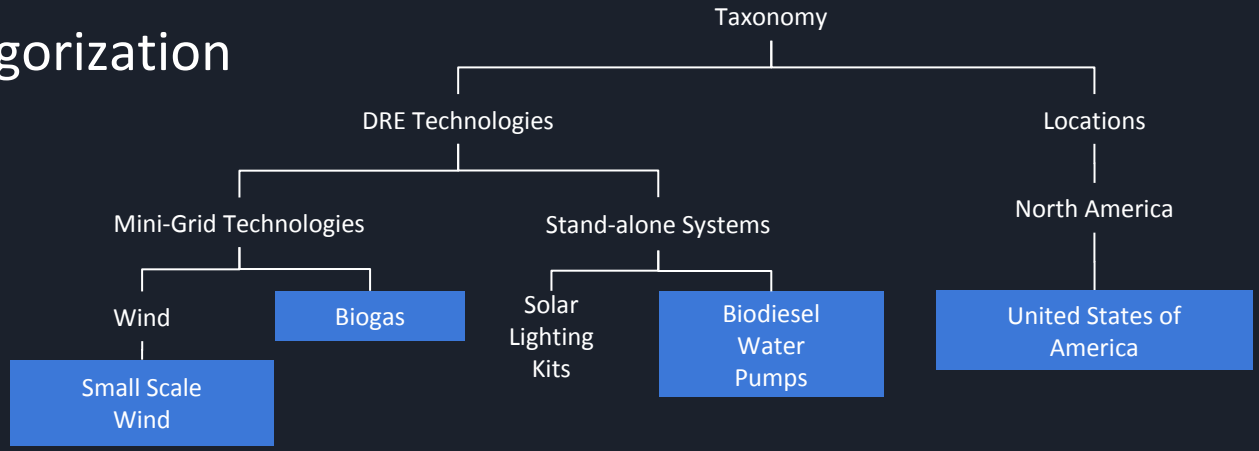


Images: http://www.powerforall.org/newsletters/

Document Auto-Categorization

The document auto-categorization process systematically organizes the high volume of documents present in the PEAK database based on a Decentralized Renewable Energy (DRE) taxonomy. This allows for greater user accessibility without the burden of manual keyword entry and summarization.

Using natural language processing techniques such as stemming and tokenization, we were able to categorize documents hierarchically while accounting for common errors such as synonyms, abbreviations, and regular misspellings. This matching process allows users to quickly find relevant documents based on keywords, and more specifically, find relevant pages within those documents.



Sample Taxonomy Snapshot - matched keywords in blue

Document Analysis

Keyword	Frequency in Document
Small Scale Wind	0.12
United States of America	0.07
Biodiesel Water Pumps	0.03
Biogas	0.01

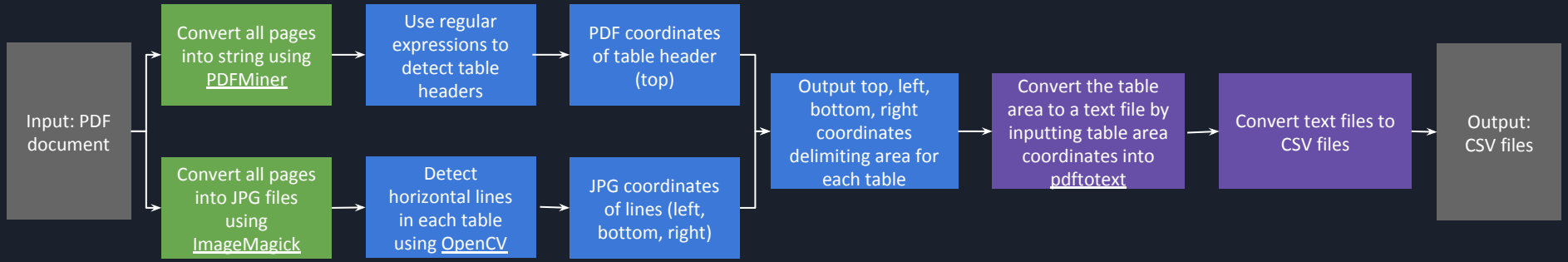
Threshold: 0.05

Page by Page Analysis

Keyword	Relevant Pages
Small Scale Wind	97, 99, 4, 5, 6,...
United States of America	22, 23, 70, 69,...

Page Number	Relevant Keywords
4	Small Scale Wind, Biodiesel Water Pumps
22	United States of America

PDF Table Extraction



Pre-process PDF

Prior to any table detection or data extraction, the PDF document must be converted into two formats: 1) JPG files for line detection and 2) a string for table header detection.

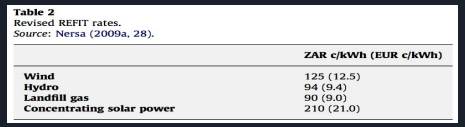


Table 2
Revised REFIT rates.
Source: Nersa (2009a, 28).

	ZAR c/kWh (EUR c/kWh)
Wind	125 (12.5)
Hydro	94 (9.4)
Landfill gas	90 (9.0)
Concentrating solar power	210 (21.0)

Auto-detect table areas

A key component of our auto-detection process is edge detection. Although table formats vary widely, virtually all tables have lines that either delimit table elements or the table from nearby text. We use computer vision to detect these lines and their coordinates on the page.

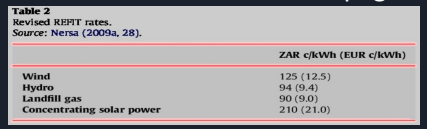


Table 2
Revised REFIT rates.
Source: Nersa (2009a, 28).

	ZAR c/kWh (EUR c/kWh)
Wind	125 (12.5)
Hydro	94 (9.4)
Landfill gas	90 (9.0)
Concentrating solar power	210 (21.0)

Coordinates:
Top: 195
Bottom: 241
Left: 43
Right: 88

81% tables accurately detected

Extract data from tables

Columns are detected in the text files as consecutive vertical white spaces. Commas are inserted to separate columns.

```
Table 2
Revised REFIT rates.
Source: Nersa (2009a, 28).

                                ZAR c/kWh (EUR c/kWh)

Wind                            125 (12.5)
Hydro                           94 (9.4)
Landfill gas                    90 (9.0)
Concentrating solar power       210 (21.0)
```

```
Table 2
Revised REFIT rates.
Source: Nersa (2009a 28).

                                ZAR c/kWh (EUR c/kWh)

Wind                            125 (12.5)
Hydro                           94 (9.4)
LandÕ¨Äll gas                    90 (9.0)
Concentrating solar power       210 (21.0)
```

82% of detected tables accurately extracted

Overall Success Rate: 67%

As compared to 35% success rate by the best open-source table extraction tool

Data Auto-Identification

The auto-identification of relevant documents consists of a two-step pipeline that incorporates web-scraping to gather documents from sites of interests and a binary classifier to determine whether the documents collected are relevant to the current PEAK corpus.

Our classifier consistently achieved an accuracy of 98%, recall of 95% and precision of 95% on the test set. The confusion matrix, right, shows that false negatives are kept to a minimum. However, some tuning must be performed to decrease the rate of false positives.