Quality comparison of python PDF text

extraction libraries

Table of Contents

[Abstract 2](#_Toc190628856)

[Motivation 2](#_Toc190628857)

[Methodology 2](#_Toc190628858)

[PDF Dataset 2](#_Toc190628859)

[Ground Truth 2](#_Toc190628860)

[Libraries under comparison 2](#_Toc190628861)

[Metrics 3](#_Toc190628862)

[Semantic Metrics 3](#_Toc190628863)

[Syntactic Metrics 3](#_Toc190628864)

[Performance Metrics 3](#_Toc190628865)

[Results & Analysis 3](#_Toc190628866)

[Semantic Metrics 3](#_Toc190628867)

[Syntactic Metrics 3](#_Toc190628868)

[Levenshtein distance 3](#_Toc190628869)

[BLEU Score 4](#_Toc190628870)

[Jaccard Similarity 4](#_Toc190628871)

[Performance Metrics 4](#_Toc190628872)

[Execution Time 4](#_Toc190628873)

[Throughput 4](#_Toc190628874)

[Conclusion for Project Chitti 4](#_Toc190628875)

[Code 4](#_Toc190628876)

[References (Citations) 4](#_Toc190628877)

[Future Work 4](#_Toc190628878)

# 

# Abstract

// goal

// key points

# Motivation

// why do this

# Methodology

// analysis setup

## PDF Dataset

## Ground Truth

## Libraries under comparison

// to-do: read about open-source software licenses

// finish the table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Lasy PyPi release | License | Version | Comments |
| Pypdf | [October 28, 2024](https://pypi.org/project/pypdf/) |  | 5.1.0 | PyPdf2 is no longer maintained |
|  |  |  |  |  |
|  |  |  |  |  |

## Metrics

### Semantic Metrics

// ignore for now

### Syntactic Metrics

// describe

1. **Levenshtein distance:**

// define + pros and cons + working example

1. **BLEU Score:**

// define + pros and cons + working example

1. **Jaccard similarity**

// define + pros and cons + working example

### Performance Metrics

// describe briefly: average time taken, throughput

# Results & Analysis

## Semantic Metrics

// ignore for now

## Syntactic Metrics

### Levenshtein distance

// describe the result

// make sure we visualize the results with 1 or more graphs

A comparison of Levenshtein distance indicated the following:

1. pypdf outperformed both pymupdf &amp; pdfminer in terms of edit accuracy with a median of 109 edits per page to match the quality of Pipeline Builder text.

### BLEU Score

### Jaccard Similarity

## Performance Metrics

// describe the result

// make sure we visualize the results with 1 or more graphs

### Execution Time

### Throughput

# Conclusion for Project Chitti

# Code

# References (Citations)

# Future Work