Truly refreshing document analysis

MINT.extract
Automated Data Extraction from Documents using Machine Learning

19.06.2019
Swisstext 2019
Aaron Richiger, Co-Founder turicode
Agenda

MINT.extract

- What it is
- How it works
- Results
- How to use it
Problem Statement

Input:
- Scans, PDFs, Emails, Excel/CSV, etc.
- Different structures and layouts
- High volumes and/or frequencies

Our Focus:
- Classification
- Extraction
- Semantic analysis
- Validation

Output:
- Machine-readable data
- Web or mobile apps
- Simple integration into existing systems

Software as a Service

Input: Scans, PDFs, Emails, Excel/CSV, etc.
Different structures and layouts
High volumes and/or frequencies

Output: RESTful API
- JSON
- XML

06.08.2019

MINT.extract
How MINT.extract structures your documents

Input

Pre-Processing

Classification

Data - Extraction

Data - Analysis

Output

OCR / Text Recognition

Pension Statements

Tax Statements

Bank Statements

Insurance Policies

Extraction of text, images etc. with Machine Learning

Web-Application for visual inspection

Automated Validation and Enrichment

RESTful API

{ "Pensionfund": "Personalvorsorgestiftung der CSS Versicherung", "Salary": 99585, "Saldo": 496863 }
How it works: Classification

Image Based Classification

Text Based Classification

Combination of Text and Image Based Classification

• Showed the best results

• Importance of text or image input depends on the documents
How it works: Data Extraction

1. Word level classification based on
   - Text value
   - Textual context
   - Layout information

2. Grouping datapoints to structured entities

3. Character level classification
How it works: Data Extraction
How it works: Analytics & Validation

Based on the extracted data, we build project-specific ML systems for further analytics and validation:

Data Analytics with ML
- NLP (e.g. NER, topic detection, etc.)
- Predictive Maintenance
- Trigger a specific workflow

Validation with ML
- Anomaly detection
- Validation classification (critical vs. non-critical validation violations)
Results: Classification

Documents: Financial documents
Case: Incoming mail classification
Data: 17k sample documents for 4 classes
Approach: Combined text- and image-based classification (80/20 split with CV)

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official Documents (Amtliche Dokumente)</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Card Application (Kartenanträge)</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Direct Debiting Application (LSV)</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Payment Order (Zahlungsauftrag)</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Results: Data Extraction

Documents: Purchase Orders
Case: Data extraction and validation
Data: 70 documents from 3 customers with 8 data points to extract
Approach: Word level classification with 80/20 split and CV
How to use MINT.extract

Option 1: Service provided by turicode

• You send us the documents and the spec

• We provide you with:
  • REST API (for ongoing services)
  • Excel, JSON, XML, TXT (for projects)

Option 2: Try and train it yourself

• Request access to MINT.extract on our website turicode.com

• We provide you access to your own MINT.extract web application

• Upload, annotate and process your documents
Excerpt from our References

Supported by:

- Kantonspolizei Zürich
- Migros Bank
- Novartis
- ETH Zürich
- Gesellschaft für Schweizerische Kunstgeschichte
- Mettler Toledo
- Vontobel
- SRF
- Wüest + Partner
- Stadt Zürich
- Universität Zürich

Swiss made software
Contact

turicode AG
Technoparkstrasse 2
CH-8406 Winterthur

Aaron Richiger
aaron.richiger@turicode.ch
+41 79 613 71 78

www.turicode.com