Probabilistic Programming for Voucher Information Extraction
Preliminary Practical Experiences
Al-Sibahi, Ahmad Salim; Hamelryck, Thomas Wim; Henglein, Fritz

*Publication date:* 2018

*Document version*
Early version, also known as pre-print

*Citation for published version (APA):*
**Probabilistic Programming for Voucher Information Extraction**

**University of Copenhagen**

**Ahmad Salim Al-Sibahi**

**Thomas W. Hamelryck**

**Fritz Henglein**

**University of Copenhagen**

**Preliminary Practical Experiences**

**Introduction to Skanned.com**

Skanned.com provides a Voucher Scanning service for extracting information from vouchers like product lines, total amounts, payment date, sender, and recipient.

Vouchers vary heavily in size, layout, purpose, and content; the scan quality is occasionally suboptimal. Probabilistic programming provides an opportunity to:

- Combine domain knowledge and machine learning to effectively extract features in a systematic fashion.
- Quantify confidence in results, which is important for manual validation.

**Skanned.com’s Pipeline**

- **OCR** Optical Character Recognition extracts textboxes from PDFs.
- **Feature Extractors** extract information from the text boxes.

**Finding Features w/Keywords**

Features are usually located around identifying keywords. Keywords can be positive or negative depending on the feature to be found.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Amount Excl. VAT</td>
<td>23613.00</td>
<td>DKK</td>
</tr>
<tr>
<td>Total VAT</td>
<td>5903.25</td>
<td>DKK</td>
</tr>
<tr>
<td>Total Amount</td>
<td>29516.25</td>
<td>DKK</td>
</tr>
</tbody>
</table>

Probabilistic model below tries to infer a latent score \( r \) from the vector of observed angles \( \overrightarrow{\theta} \) and distances \( \overrightarrow{d} \) from positive keywords to potential target features.

\[
\begin{align*}
    r & \sim \mathcal{B}(0.5,0.5) \quad \bar{r} = (r, 1 - r) \\
    w^+_1 & = (0.7, 0.3) \quad \mu^+_1 = (0, \frac{\pi}{2}) \\
    w^+_2 & = (0.5, 0.2, 0.3) \quad \mu^+_2 = (-\frac{\pi}{4}, \frac{\pi}{4}, 3\pi) \\
    \overrightarrow{\theta}^+ & \sim \sum_{j=1}^{2} \sum_{i=1}^{w^+_j} \frac{w^+_j}{\mu^+_j + \pi (\frac{4}{\pi})} \\
    \overrightarrow{d}^+ & \sim \bar{r}_1 \mathcal{N}(500) + \bar{r}_2 \mathcal{N}(1500, 1000)
\end{align*}
\]

Evaluating extended version on 1000 vouchers:
- **80%** of the time the expected score found the target feature
- **99%** of the time it was within confidence interval

**Voucher Grouping**

To provide more accurate models, to partition the voucher into groups of similar layout and style. We rely on probabilistic Latent Dirichlet Allocation (LDA) to perform the grouping, using visual (colors, lines) and textual cues (keywords).

**Practical Experiences**

- **Sampling**
  - Ease of use
  - Precision
  - Scalability

- **Variational Inference**
  - Scalability
  - Set-up
  - Precision

- **GPU Support**
  - Discrete Latents
  - Ease of use
  - Precision

The first author would like to thank Dan Rose and the Scanner Team at Skanned.com (Bjørn Kaae, Toke Reines, Danni Dromi) for discussions and feedback on the work!