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

Preliminary Practical Experiences

Ahmad Salim Al-Sibahi
University of Copenhagen/Skanned.com

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

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

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

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

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!

Thomas W. Hamelryck
University of Copenhagen

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).

Fritz Henglein
University of Copenhagen

Skanned.com's Pipeline

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

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

- ✷ 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!