Mohit Rajput <mohitrajput901@gmail.com>

AI/Machine learning-NLP, Deep learning - Zycus

Mohit Rajput <mohitrajput901@gmail.com> Fri, Mar 29, 2019 at 2:22 AM
To: Anuja Ghosalkar <4e6ey_cqlcav+or9x@inbound.workablemail.com>

Hi Anuja,

Questions
1. Extracting important business fields from invoice document/scanned PDFs.
2. For the extracted data item, your proposed solution should also provide percentage confidence calculation per extracted field
3. Share your detailed approach (Logic/ Algorithms to use, custom logic if any).
4. Important business fields - Invoice Number, Invoice Date, Purchase Order Number, Line Item details (quantity, sku, name, price, etc), Invoice Total, Payment terms.

Constraints
- the solution should be applicable for any invoice in any layout/format.

Factors to consider in deciding the kind of Solution to develop/ Import
- Number of PDFs (scale)
- Total Available Time for getting the job done (time)
- Finance available for getting the job done (cost)
- Acceptable accuracy in getting the job done (performance)
- Time required for reading each PDF
- Cost required for reading each PDF
- Achievable Accuracy in reading information from PDF
- System/ Solution up/development time
- Time and cost required for System Maintenance / Updation / Learning / Evolving.
- Ease of understanding of the System (i.e. interpretability and debuggability of the system.)
- Data confidentiality / agreement

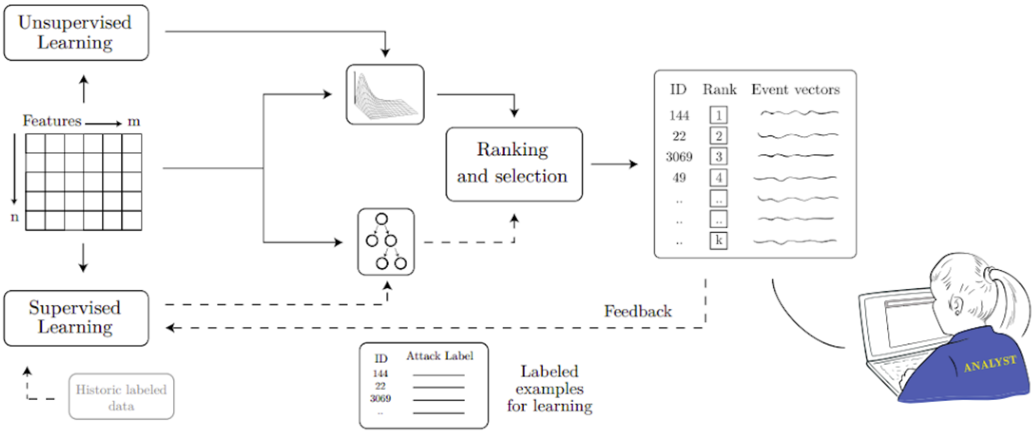
In Summary,
> What's the required/desired balance between Cost, Scalability, Time, Accuracy, Interpretability & confidentiality

Possible Solutions
> Using Services or Tools
This kind of solution can be a cheap alternative in the short run, readily available, possibly a good combination of accuracy and time consumption. Though this route may not suffice if data can't be shared. ...
-- involving Human (https://www.nytimes.com/2010/04/26/technology/26captcha.html)
-- external service (https://www.parascript.com/blog/how-to-extract-reliable-data-from-invoices/)

> Researching & Developing a Solution
...
-- researching on the work which has already been done (https://github.com/m3nu/invoice2data)
-- Developing further based on the desired/required requirement.

Answering the Question
Proposed Solution 1
As highlighted above the definition of the best solution will be based on the business decision. Hence the solution provided below is one way which I feel will be robust enough to handle a large volume of data (i.e. PDFs), will be able to provide confidence about the prediction and will encompass of Natural Intelligence, Artificial Intelligence, Machine/ Deep Learning & Recommendation System. The advantage of such a system will be its ability to evolve itself at a much higher pace by making use of natural intelligence to far exceed in the productivity, performance & scalability. The requirement of an analyst may even cease to exist after some time. Additionally, the incurred cost in the maintenance and the development of such a system will break even in a bit time and the confidentiality of the data will be retained.

High-level of the solution is highlighted below in the image.



The diagram above depicts the use of a structured data which for our case we can assume it to be scanned data in text form (let's just say using some pdf scanner) and this data in the initial runs are directly passed to an analyst. This analyst will highlight/ tag the required field in this. After some iteration, we are gonna train some models using the cleaned data generated by the analyst to identify the required information. For this trained model we are gonna use predict score to identify the less confident prediction from high confident prediction. Another unsupervised model will also in turn run in parallel with this supervised model. This unsupervised model will try to club all the similar document together and also identify the rare documents. The information generated from supervised and unsupervised can then be aggregated together to identify and rank documents in an order i.e. defined by a cost function. This cost function can be a function of club size for the documents, possible aggregated confidence for the club, rare occurrences, cost incurred for the analysis performed by the analyst, Error rectification, etc. The cost function will help in providing a recommendation to the analyst for which it can decrease the weighted cost incurred from using the analyst as well as increasing the overall performance.

Proposed Solution 2

Robustly using Regex. But this will have multiple issues.

Regards,

[Quoted text hidden]

--

Mohit Rajput
Data Scientist and Researcher
Indian Institute of Technology, Roorkee (Graduate)

M: +91 897-957-2630
E: mohitrajput901@gmail.com