Data Engineering Task

Introduction

Named-entity recognition (NER) is a task of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, time expressions, quantities, monetary values, etc. One of our key capabilities of Hypatos is the information extraction models (IE models). IE models are trained on labelled entities and predict entities on invoices, e.g., currency, issued date, invoice number, totals amount. Monitoring of models' performance is a crucial part after deploying the model on production system. Important metrics of monitoring include *accuracy* and *straight-through-processing (STP) rate*:

- *Accuracy* is defined on **entity level** and measures if the prediction of an entity is correct.
- *STP rate* is defined on **document level**. If all the key entities on a document are predicted correctly, this document will be marked as STP = 1. Important note: In some documents not all the key entities exist. In this case, if the key entities which are presented on the document are all correct, STP is still counted as 1.

Example

In below we will explain the metrics with 2 documents:

- The table *Document Entity* stores the entity level information. Each row represents a prediction for an entity. This table contains below fields:
    - processed_at: the timestamp when a document is processed.
    - document_id: This is the unique ID of a document.
    - label: The pre-defined categories of entities.
    - prediction_value: This is the value predicted from the models.

	processed_at	document_id	label	prediction_value
1	2022-08-01 22:00:09.558000	9000061420	currency	EUR
2	2022-08-01 22:00:09.558000	9000061420	issued_at	2022-08-01
3	2022-08-01 22:00:09.558000	9000061420	totals_gross	13275340.0
4	2022-08-01 22:00:09.558000	9000061420	vendor_vendor_id	1555111999
5	2022-08-03 16:40:23.030000	6000062535	currency	USD
6	2022-08-03 16:40:23.030000	6000062535	issued_at	2022-08-03
7	2022-08-03 16:40:23.030000	6000062535	totals_gross	24492.16
8	2022-08-03 16:40:23.030000	6000062535	vendor_vendor_id	000033521Z

- The table *Ground Truth* stores the target value for the prediction. Ground truth in machine learning refers to the reality you want to model with your supervised machine learning algorithm. Ground truth is also known as the target for training or validating the model with a labeled dataset. Each row in this table represents a document. This table contains below fields:
    - document_id: This is the unique ID of a document. With this field you can join the *Ground Truth* table with the *Document Entity* table.
    - currency, issued_at, totals_gross, vendor_vendor_id: these fields store the ground truth value.

	document_id	currency	issued_at	totals_gross	vendor_vendor_id
1	6000062535	USD	2022-08-03	24492.16	0000335212
2	9000061420	CLP	2022-08-01	13275340.0	1555111999

- In this example with 2 documents:
  - accuracy = 87,5%.
    - Explanation: prediction for currency for document_id = 9000061420 is wrong. 7 out of 8 entities are predicted correctly.
  - STP rate = 50%.
    - Explanation: Key entities in this case include currency, issued_at, totals_gross and vendor_vendor_id. For document_id = 6000062535, prediction for all key entities is correct and therefore STP = 1. For document_id = 9000061420, one of the key entities is wrong and therefore STP = 0. 1 out of 2 documents are STP documents.

Task

You will receive two tables: *Document Entity* and *Ground Truth*. The key entities are 1) currency, 2) issued_at, 3) totals_gross and 4) vendor_vendor_id.

- Please create an analysis to monitor the model's weekly performance. Some aspects can be for example volume of documents, accuracy per entity and STP rate.
- You can use the technology you prefer. The task should not take longer than 4 hours to prepare. Please send us back the code and a summary of the analysis result.
- Once we review it and want to move further, you will be invited to a presentation with the team. This will be about 40 mins:
  - 10 min introduction round.
  - 10 min presentation. Please guide us through the code as well as the analysis so that we can have an overview about your technical skills.
  - 10 min Q&A on the task.
  - 10 min answering candidate questions.