Cortex
Loss Runs

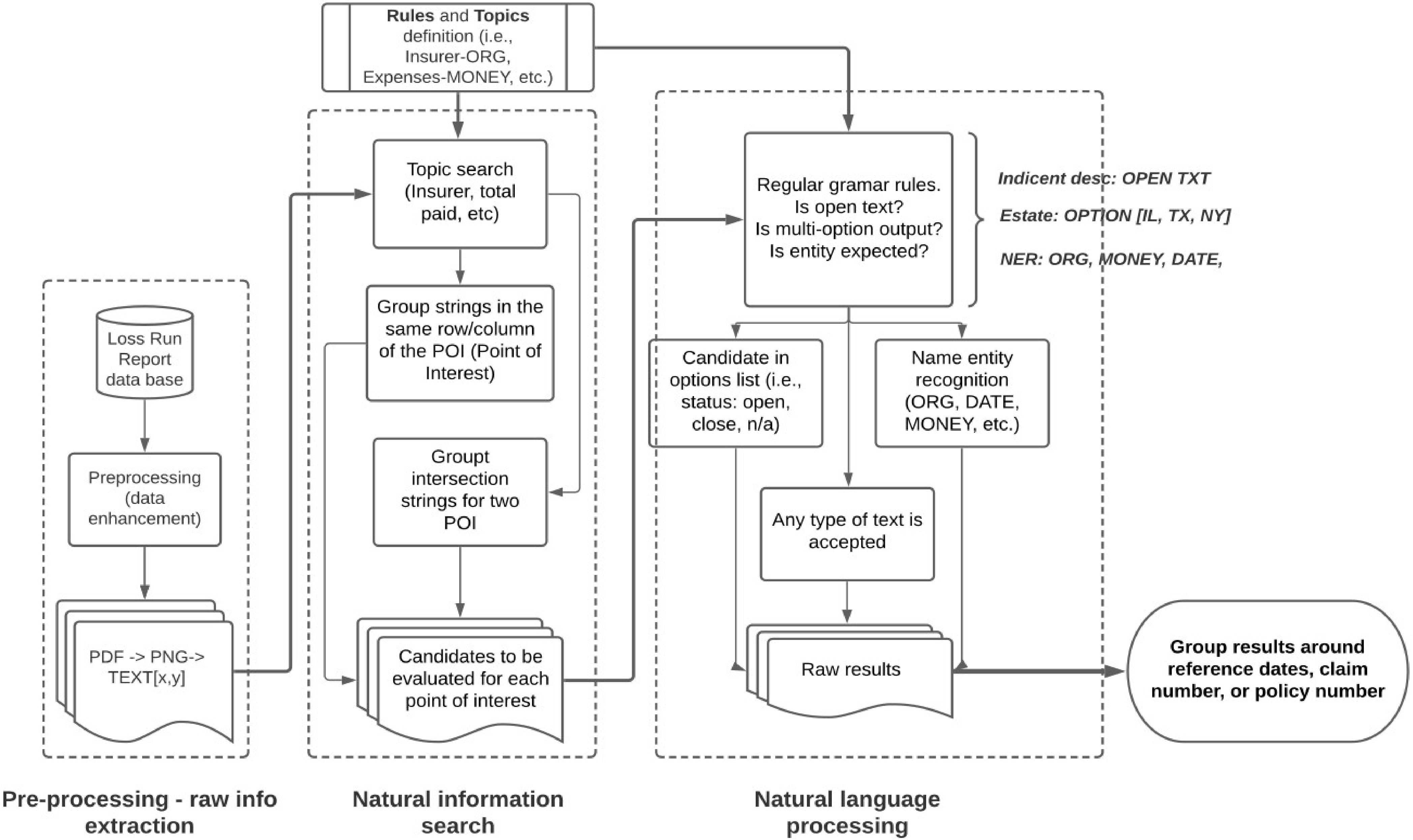
Relevant information extraction in Loss Run reports based on Natural Language Processing and Regular Grammar

A brief description of the core modules.

Assym Developers, Now Insurance. March 2021

Eric C., Ari L. Phill C., Eduardo S. Alberto de O. Romulo T. Raul V., Marvin H., Hugo G.

Approach overview



Rules and **Topics** definition (i.e., Insurer-ORG, Expenses-MONEY, etc.)

Loss Run Report data base

Preprocessing (data enhancement)

PDF -> PNG-> TEXT[x,y]

Topic search (Insurer, total paid, etc)

Group strings in the same row/column of the POI (Point of Interest)

Groupt intersection strings for two POI

Candidates to be evaluated for each point of interest

Regular gramar rules.
Is open text?
Is multi-option output?
Is entity expected?

Indicent desc: OPEN TXT

Estate: OPTION [IL, TX, NY]

NER: ORG, MONEY, DATE,

Candidate in options list (i.e., status: open, close, n/a)

Any type of text is accepted

Name entity recognition (ORG, DATE, MONEY, etc.)

Raw results

Group results around reference dates, claim number, or policy number

Pre-processing - raw info extraction

Natural information search

Natural language processing

Optical Character Recognition

Optical Character Recognition (OCR) is a Machine Learning (ML) technique that extracts the text and its respective spatial distribution in a loss report.

- The OCR is based on an English dictionary.
- The results are stored in a list of words with *x*, and *y* coords.
- OCR is an open source engine (**Tesseract, apache 2.0 [1]**).

Natural information search

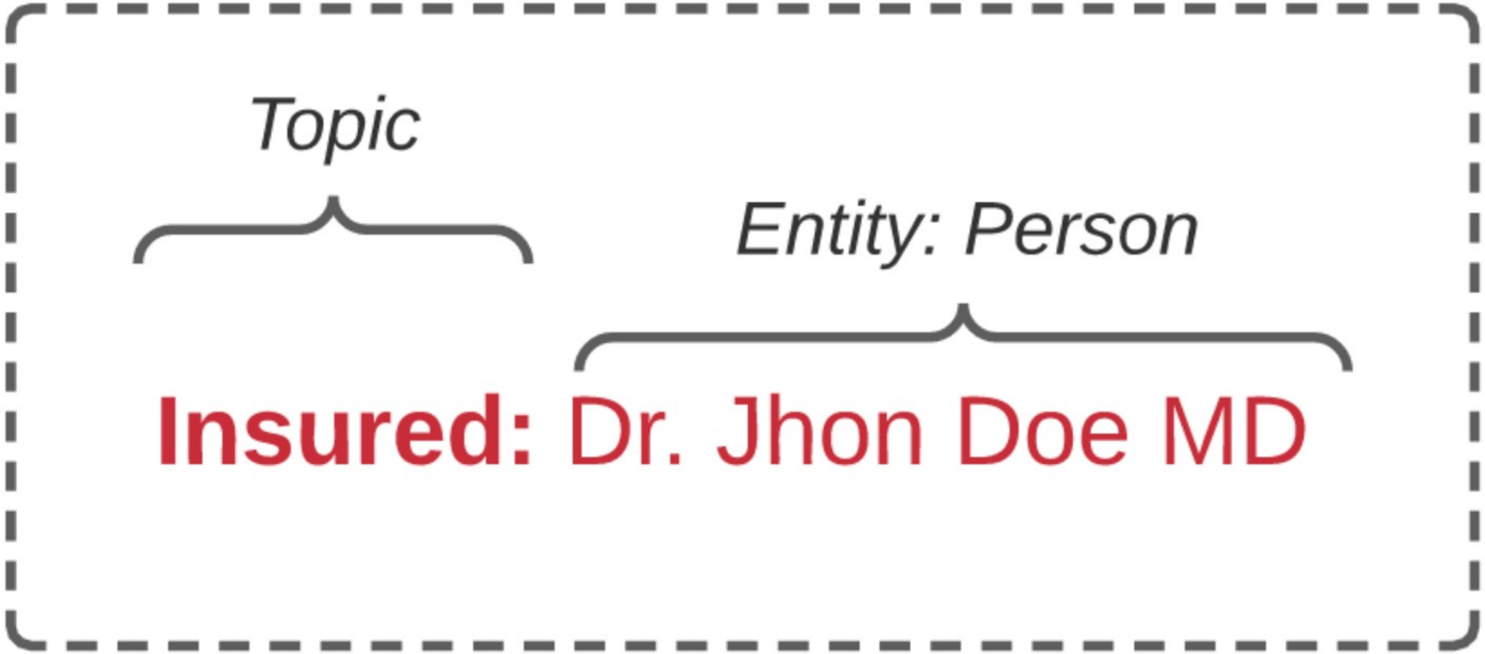
At this stage of the algorithm, the aim is to filter the information in the reports through two processes [3]:

- **Mathematical analysis of the spatial distribution**: associates the coherence of the spatial relationship between the points of interest and the rest of the words extracted by the OCR.
- **Search for the natural distribution of the text**: complements the spatial relationship between words with the way the text is naturally associated in language (forward and downward).

Forward search, same row

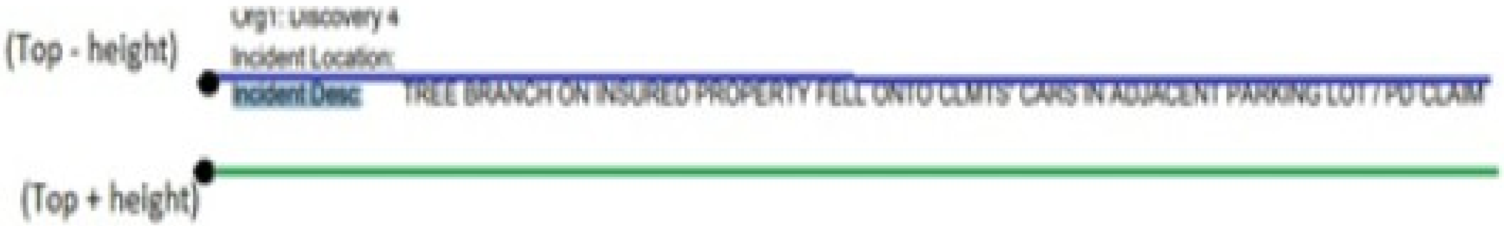
Since each word (string) extracted by the OCR has four coordinates (beginning and end in *x* and *y*) it is possible to group the words that are in the same row. So the following should apply.

- Get *x1* and *x2* coords for each entity.
- Keep all the words which coords *wx1*, *wx2* that fits with the same row (*x1* < *wx2* and *x2* > *wx1*).
- Limit the search range based on the size of the report.
- Check if the obtained value match with the expected value *(TOPIC*, *OPTIONAL*, *OPEN TEXT*).



Insured: Dr. Jhon Doe MD

Forward search



Org1: Discovery 4

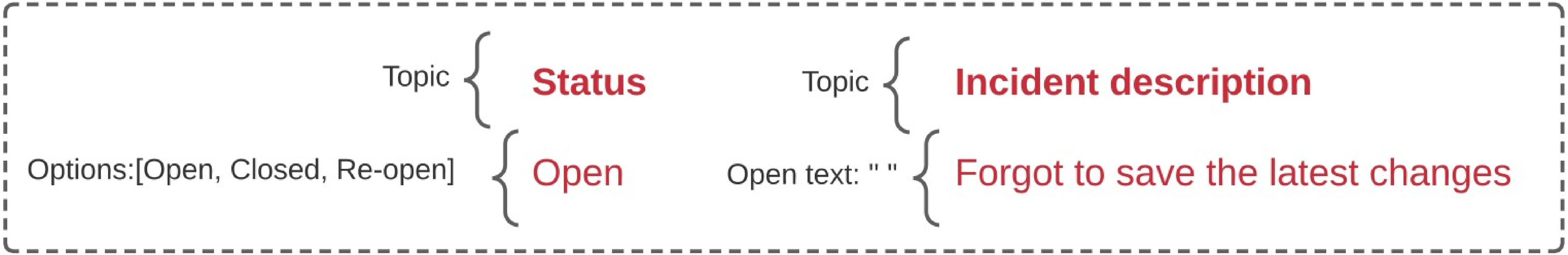
Incident Location:

Incident Desc TREE BRANCH ON INSURED PROPERTY FELL ONTO CLMTS' CARS IN ADJACENT PARKING LOT / PD CLAIM

Search under, same column

Since each word (string) extracted by the OCR has four coordinates (beginning and end in x and y) it is possible to group the words that are in the same column. So the following should apply.

- Get $y1$ and $y2$ coords for each entity **if any match in forward search**.
- Keep all the words which coords $wy1$, $wy2$ that fits with the same column ($y1 < wy2$ and $y2 > wx1$).
- Check if the obtained value match with the expected value *(TOPIC, OPTIONAL, OPEN TEXT*).



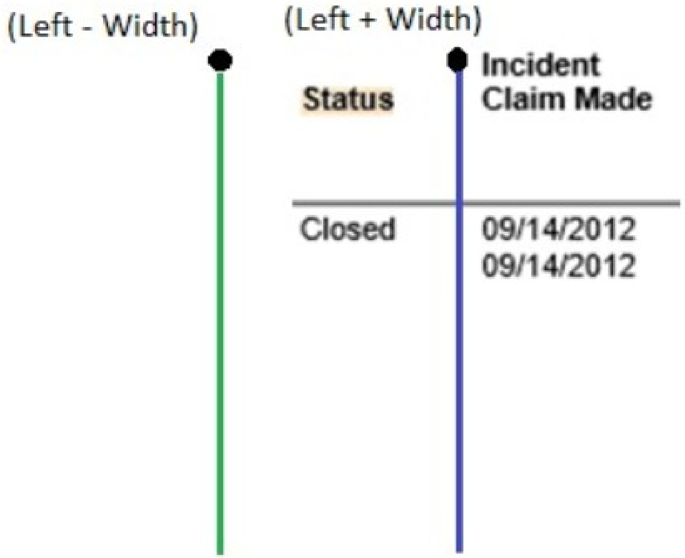
Topic **Status**

Options:[Open, Closed, Re-open] Open

Topic **Incident description**

Open text: " " Forgot to save the latest changes

Search Below

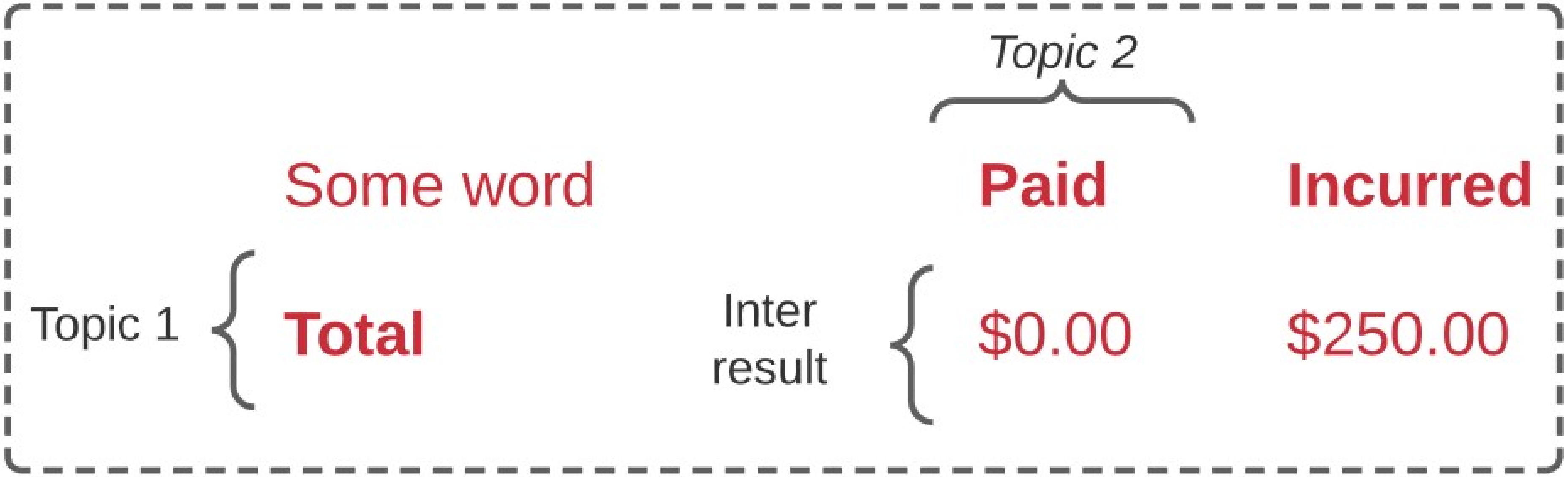


Status	Incident Claim Made
Closed	09/14/2012 09/14/2012

Intesection seacrch

The intersection search will be carried out in the event that the value to be found is composed of two entities [*Total* (entity) - *Paid* (entity)]. The steps to follow are:

- Find the coordinates of the two entities.
- Find the upper entity.
    - Then, elements in the same column.
- Find the entity with the lowest height.
    - Then, elements in the same row
- Get the value where both searches intersect (same row for *Total*, same column for *Paid*).
- Verify that the value obtained is the expected one.



Topic 2

Some word

Paid

Incurred

Topic 1

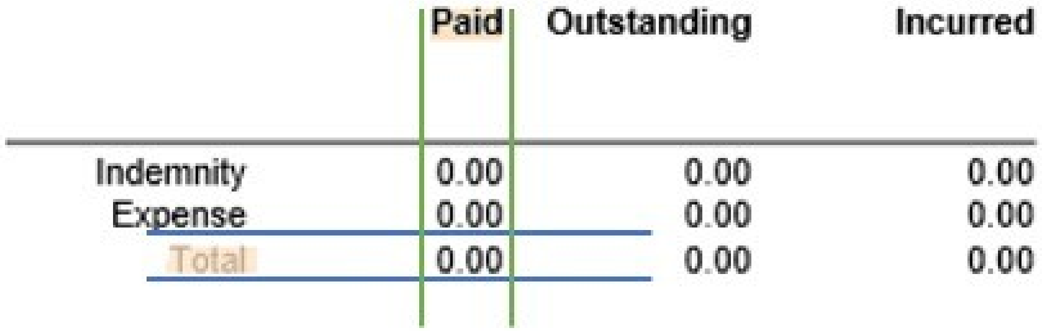
Total

Inter result

$0.00

$250.00

Intersection search



	Paid	Outstanding	Incurred
Indemnity	0.00	0.00	0.00
Expense	0.00	0.00	0.00
Total	0.00	0.00	0.00

Natural Language processing

The Natural Language Processing is a sub-field of computational linguistics and machine learning that is applied in this case, to identify entities (*DATES, ORGANIZATIONS, PERSONS*, etc.) given a specialized linguistic context in the lexicon of Loss Run reports. For this module it is necessary to configure two sections:

- Natural Language Model[2] whit the Name Entity Recognition module in the pipeline (English-Loss Run lexicon based).
- Entities associated with each topic if applicable (*INSURED [TOPIC] – ORGANIZATION[ENTITY].*

References

[1] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE. *https://ieeexplore.ieee.org/abstract/document/4376991*

[2] Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

https://books.google.es/books

[3] Cortex-NLP, Assym Developers, Now Insurance (2021). Cortex NLP applied to relevant Information Extraction in Loss Run reports.

https://github.com/Asymm-Developers/cortex-npdb