Institute and Faculty of Actuaries

Reacfin

Reinsurance treaties study using NLP: methods and innovation enablers for actuaries

Aurelien Couloumy, *Head of Data Science*
aurelien.couloumy@reacfin.com

Loris Chiapparo, *Data Scientist*
loris.chiapparo@reacfin.com

08 November 2018

About the speakers

- **Aurelien Couloumy, Head of Data Science at Reacfin**

Aurélien is currently responsible for the Data Science department in the consulting firm Reacfin. He is also associate lecturer at ISFA, Université Lyon 1. Previously, he worked as Head of Models in an international actuarial consulting firm. Aurélien is also fellow of the French Institute of Actuaries and the Institute of Actuaries in Belgium (IA\|BE).

- **Loris Chiapparo, Data Scientist at Reacfin**

Loris joined Reacfin as a Data Scientist. Previously, he worked for a consulting firm as a software engineer on financial applications. Graduated from the Université Libre de Bruxelles (ULB) as an engineer in computer science and computational intelligence, he develops his expertise around machine learning and natural language processing.

Reacfin Data Lab

Institute and Faculty of Actuaries

Agenda

1. Introduction
2. IT framework
3. Process general functioning
4. Results and growth enablers
5. Demo version

Reacfin Data Lab

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

1. Introduction

1.1 Data Science and actuaries

1.2 Business case context

1.3 Goals

1.4 Scope

08 November 2018

1.1 Data Science and actuaries

What does **data science** bring to **actuaries?**

Everything… **no!**

- As complementary approaches for **pricing, underwriting,** reserving, capital modelling, ALM, etc.

- One particularly interesting use case**: how to collect and exploit unstructured data for actuarial purposes?**

Reacfin Data Lab

Institute and Faculty of Actuaries

1.2 Business case context

- Example with **reinsurance treaties and facultatives:**

Issues

- *An **heavy and repetitive workload** for already very busy business teams.*
- *A **complex document analysis** with different structures and formats.*
- *An **incomplete view of criteria and clauses** which have been underwritten.*
- *Operational risks exposition due to **heterogeneous and non exhaustive controls** applied by hand.*

Solutions

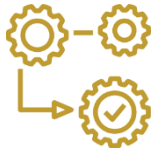
- ***Automate the analysis:***
  to save time during underwriting and pricing process.
- ***Simplify document understanding:***
  to collect and assess accurate and usable information.
- ***Improve controls:***
  to reduce risks and set up compliance rules.

Reacfin Data Lab

Institute and Faculty of Actuaries

1.3 Goals

Collect criteria *and key elements that explain the documents.*



Define accuracy measures *and quality check to assess the information's value.*

Create a simple, scalable and effective AI tool

that can help underwriters and actuaries to…

Recognize treaties architecture *and clauses topics.*

Export and use data *to run actuarial and risk management studies.*

Reacfin Data Lab

Institute and Faculty of Actuaries

1.4 Scope – technical aspects (1/2)

- Business case realized in **partnership with a large French reinsurance company.**
- Around **450 documents** requested for this study.
- Documents **in English** in order to simplify the approach.
- **Image and digital documents** that represent real-life material.
- **Different sources and different formats** to represent day-to-day activities.

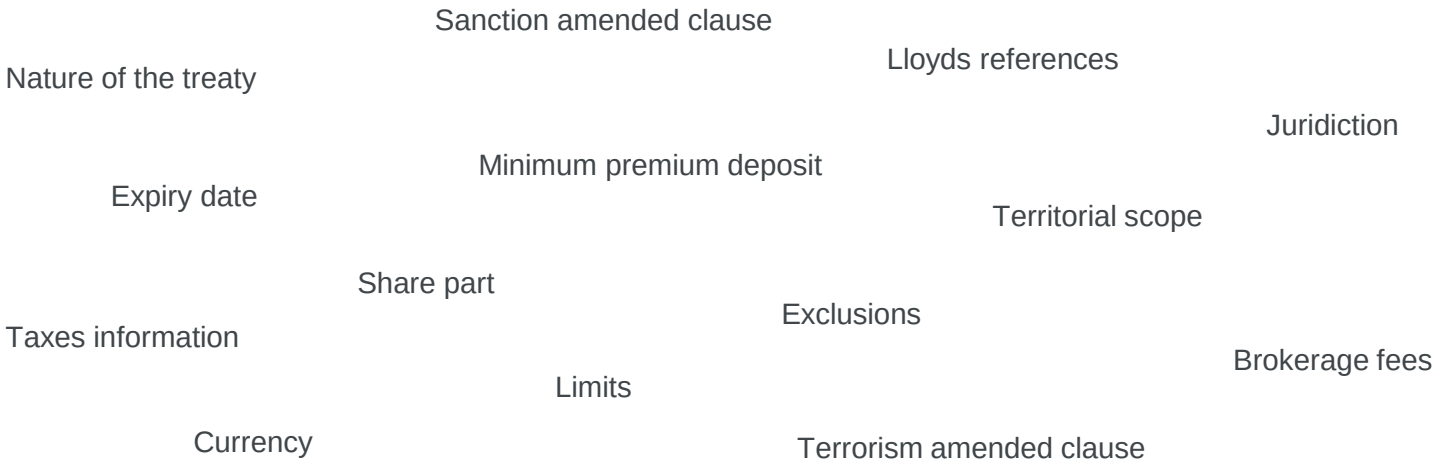
Reacfin Data Lab

Institute and Faculty of Actuaries

1.4 Scope – business aspects (2/2)

- **Non proportional treaties** analysis.

- **Collection of business criteria and clauses**, among others:



Institute and Faculty of Actuaries

2. IT Framework

2.1 Environment and technologies

2.2 Containerized applications

08 November 2018

2.1 Environment and technologies

- **Open and scalable** technologies:

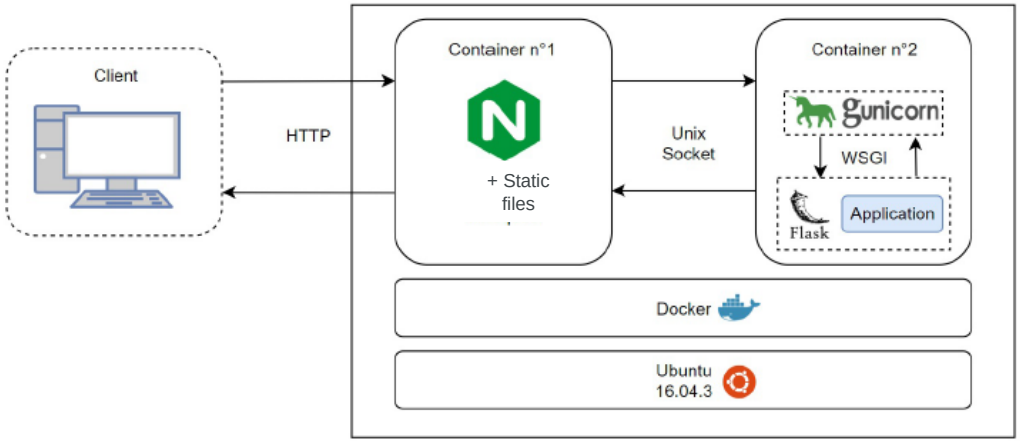
Reacfin Data Lab

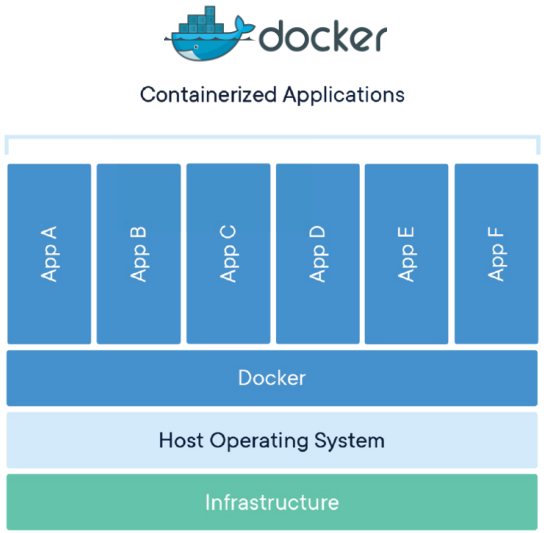
Institute and Faculty of Actuaries

2.2 Containerized applications

- A container is a standard software unit that **regroups both code and dependencies** so that the tool can run quickly from one environment to another.

- Probably the **best way to run uniformly any kind of software**





Reacfin Data Lab

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

3. Process general functioning

3.1 Introduction

3.2 Words representations

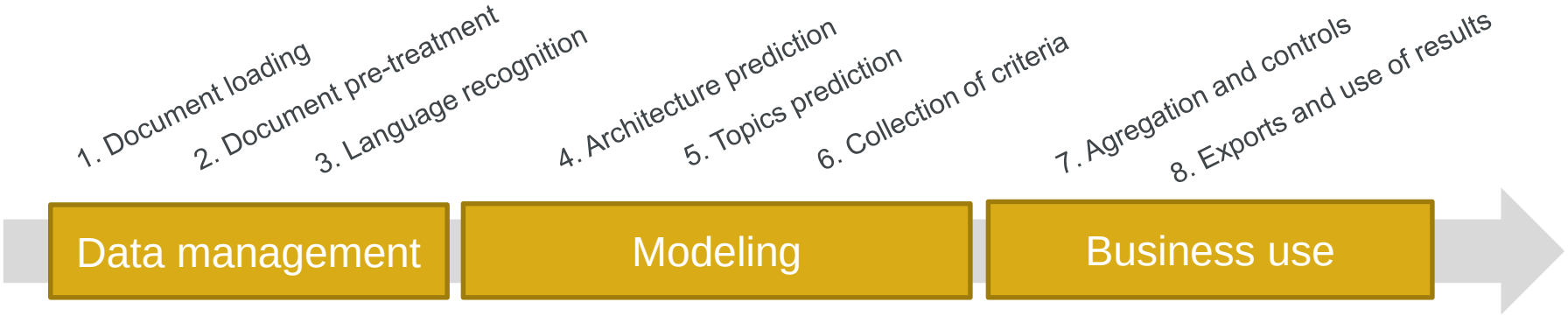
3.3 Deep learning and text mining

3.4 Example

08 November 2018

3.1 Introduction

Process general functioning



Word vectorization + **Deep learning & regex** + **Visualization & KPIs**

Reacfin Data Lab

Institute and Faculty of Actuaries

3.2 Words representation (1/2)

- **Word representation** is one crucial stage of the data pre-treatment part.

- It aims at **representing the meaning of the document** for modeling works.

Term document matrix

- Bag of words frequency
- 20K to 50K dim.
- Capture discrete general differences but not relationship between words

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

TF-IDF

- Score the importance of a word comparing it to the frequency of this word in the whole document dataset.
- 20K to 50K dim.
- Capture specific differences

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Word embedding

- Use of a vector space to predict the meaning according to the context of this word
- ANN that give 250-500 dim.
- Capture regularities and relationship between words
- Many techniques: Word2Vec, GloVe, etc.

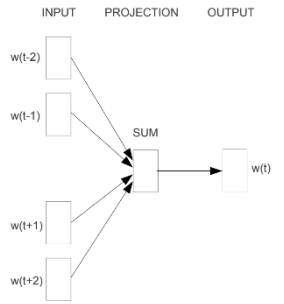
Reacfin Data Lab

Institute and Faculty of Actuaries

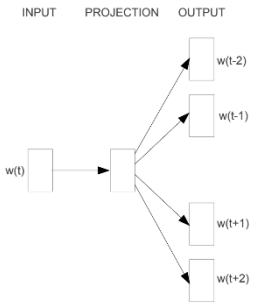
3.2 Words representation (2/2)

- Focus on **Word embedding:** words are represented as vectors in a predefined vector space.

- **Vector space word representation (Word2vec**):
  - Words understanding according to a local context.
  - 2 complementary approach:
    - Continuous bag-of-words (CBOW)
    - Continuous skip-gram model (Skip-gram)
  - Reference: https://arxiv.org/pdf/1310.4546.pdf

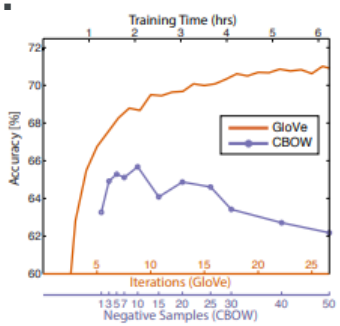


CBOW

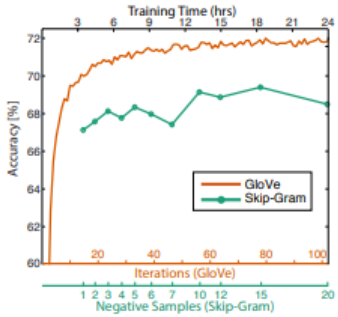


Skip-gram

- **Global vector for Words representation (GloVe):**
  - Joint use of word2vec and matrix factorization techniques (Latent semantic analysis, LSA) to improve word embedding
  - Reference: https://nlp.stanford.edu/pubs/glove.pdf



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

- In the next parts, we will use **GloVe**

Reacfin Data Lab

Institute and Faculty of Actuaries

3.3 Deep learning and text mining (1/2)

- Now data have been prepared, we can **deep dive into the modelling part**.

- To understand and collect information from documents we have to make the split between **2 categories of models:**

Deep learning models

- **Supervised learning model to predict the structure of the document**. Classification technique to split the document into several areas by recognizing titles from common text (i)

- **Similarity measure model to predict the topic of the different areas of the document**. Distance measures to understand the meaning of each area based on an accuracy threshold (ii)

Text mining models

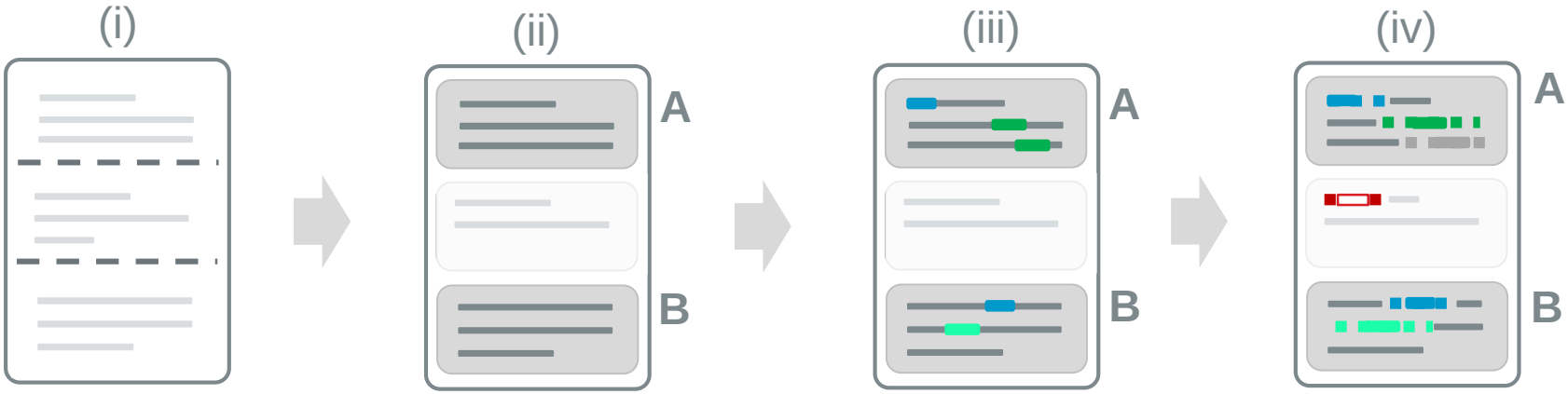
- **Regex lists and rules to collect candidate information** for the different criteria we want to study (iii)

- **Context analysis to assess the most relevant candidates** from all the candidate informations according to reference contexts (iv)

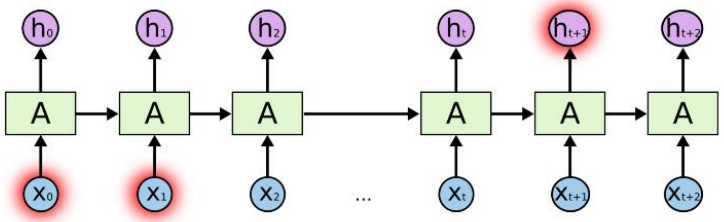
Reacfin Data Lab

Institute and Faculty of Actuaries

3.3 Deep learning and text mining (2/2)



- Deep learning part - Focus on **RNN classification:**
  - Tests on SVM, MLP and RNN models
  - RNN is the most effective model, mainly because it takes into account the sequence characterictic of data



- Text mining part - Focus on **regex and context strategy:**
  - A regex is a string of characters that describes, in a precise syntax, a set of possible strings. Examples:

emails (e.g. from webpages), → `([a-z0-9_\.-]+)@([a-z a-z0-9\.-]+)\.([a-z\.]{2,6})`

phone numbers,

IP addresses, → `(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)`

Dates

hexadecimal values, → `#?([a-f0-9]{6}|[a-f0-9]{3})`

Reacfin Data Lab

Institute and Faculty of Actuaries

3.4 Example

- To sum up, a criteria will be obtained thanks to: **an area, a topic, a regex and context**.

Example with the criteria ***Inception Date***

Reinsurer	The subscribing Insurance and/or Reinsurance Companies and/or Underwriting Members of Lloyd's (hereinafter referred to as the Reinsurers), for a participation as stated in the individual signing pages.
Period	This Contract shall apply to losses occurring during the 12 month period: Effective from: 1 January 2017 Expiring on: 31 December 2017 Both days inclusive, Local Standard Time at the place where the loss occurs. The rights and obligations of both parties to this Contract shall remain in full force until the effective date of expiry or termination, after which the liability of the Reinsurers shall cease absolutely, except in respect of losses occurring during the period of this Contract, the claims for which remain unsettled at that date.
Type	Per Event Excess of Loss Reinsurance Contract.

Institute and Faculty of Actuaries

4. Results and growth enablers

4.1 Results

4.2 Pricing perspectives

4.3 Risk management perspectives

4.4 Conclusion

4.1 Results (1/2)

- (i): **RNN model predicts the architecture well**. 2% remaining could be reduced thanks to a larger training set.

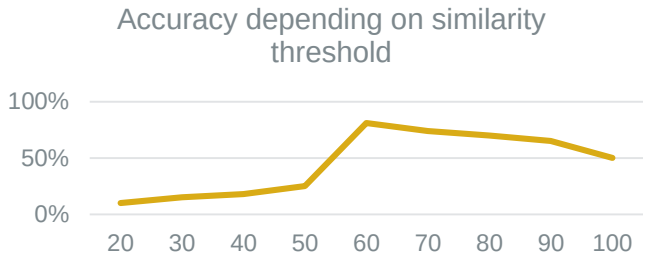
- (ii): **Topic recognition also works well**. Errors are mainly due to an high threshold of similarity acceptation.

- (iii) and (iv): **data collection is very good**. We **collect well almost 80%** of all the relevant criteria we could extract from documents

98% of accuracy

		Predict.	
		Posit.	Negati.
Obs.	Posit.	1852	109
	Negati.	116	18215

78% of accuracy



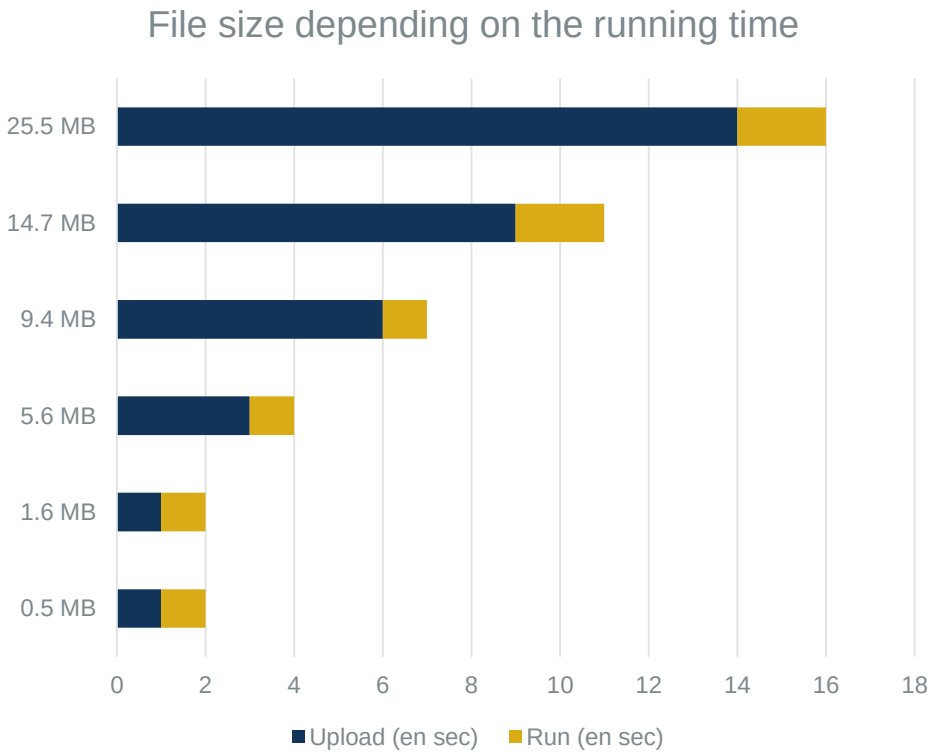
Reacfin Data Lab

Institute and Faculty of Actuaries

4.1 Results (2/2)

- **Running time is between 2sec and 16sec** on average (in comparison – a manual analysis could take more than 4 hours)

- Actually, **algorithm calculation time is not higher than 2sec**. It comes from the use of RNN and from the fact that data are only digital (so we don't need to apply OCR analysis)



Reacfin Data Lab

Institute and Faculty of Actuaries

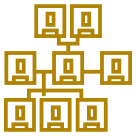
4.2 Pricing perspectives

- Use and benefits of the business case for pricing actuaries are numerous:

Feature engineering to create new explanatory variables to improve the predictive power of the pricing model.

Feature selection to assess the most influencing explanatory variables to precise the pricing model features (using supervised ML)

Accelerate the quotation process to give instantly the information to the business teams or the brokers



Define new product segmentations using these new criteria and the common ones(and using unsupervised ML)

Reacfin Data Lab

Institute and Faculty of Actuaries

4.3 Risk management perspectives

- Same observation for risk management:

Define precisely a combined ratio and other KPIs related to an area, an industry, a risk, in order to assess claims impacts

Get an homogeneous view of the taken risks, clauses specificities, differences from one year to another, etc.

Reduce the operational risks due to typing errors, incomplete information filled, wrong checks, etc.

Define, improve and check strict compliance rules related to risk management strategies

Create useful data visualization to share an internal common vision of the risks and customers.

Reacfin Data Lab

Institute and Faculty of Actuaries

4.4 Conclusion

- **Data science can be used in many different ways** by actuaries.
- One of the most impactful is probably the use of methods that aim at **collecting and enhancing unstructured data**.
- **Pricing and risk management for reinsurance** business is a good example of this.
- Cumulative use of **word embedding, deep learning and text mining** techniques applied on reinsurance treaties or facultatives can be highly effective.
- **Pricing and risk management teams can benefit from such developments** in many different ways: models improvement, quotation process optimization, KPIs definition, compliance rules setting up, etc.

Reacfin Data Lab

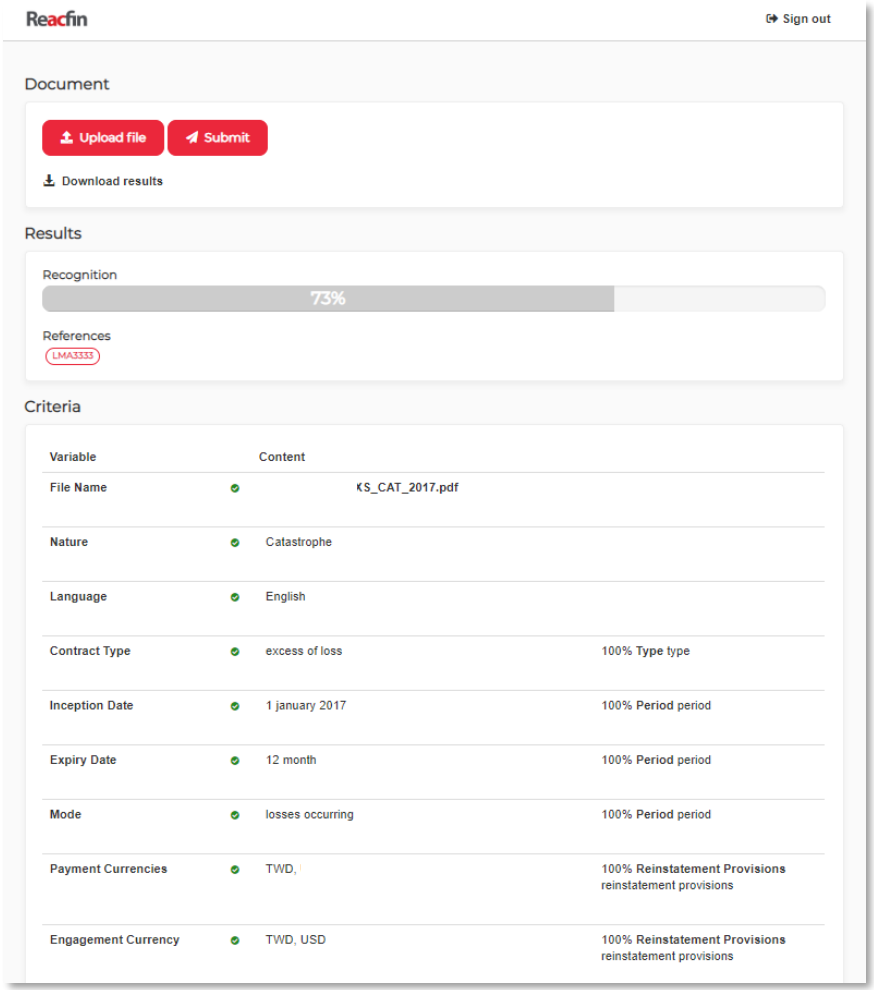
Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

5. Demo version

08 November 2018

5. Demo version



- A simple demo version of the project already used by many underwriters and pricing actuaries.

Reacfin Data Lab

Institute and Faculty of Actuaries

Questions

Comments

Institute and Faculty of Actuaries