100% ACTUAIRES & 100% DATA SCIENCE

INSTITUT DES ACTUAIRES

Etude des traités de réassurance par NLP: méthodes et leviers d'innovation

CCR RE

Reacfin

16 Novembre 2018

Hôtel Marriott Rive Gauche - Paris 14ème

A propos des intervenants

Loris Chiapparo,
Data Scientist
chez Reacfin

Aurelien Couloumy,
Head of Data Science
chez Reacfin

Jérôme Isenbart,
CRO
chez CCR Re

Sommaire

1.1 Data Science et actuariat

1.Performance

Améliorer les processus, réduire le temps et les efforts de travail.

2.Evaluation des risques

Améliorer l'analyse, et la prediction des risques.

3.Compréhension marché

Faciliter la veille réglementaire, compétitive et technique. Mieux comprendre ses clients et leurs besoins.

Souscription

Tarification

Risk management

Comment collecter et exploiter des données non-structurées à des fins actuarielles ?

Provisionnement

Modélisation

ALM

1.2 Contexte de l'étude

- Analyse des données issues des **traités de réassurance et des facultatives**.

Problématiques	Solutions
• *Une **charge de travail lourde** et répétitive pour des équipes métier très occupées.*	• ***Automatiser l'analyse** pour gagner du temps au cours des différents processus*
• *Une **analyse complexe** due aux structures et formes variées des documents.*	• ***Faciliter la compréhension** des documents pour mieux collecter et exploiter l'information*
• *Une **exposition importante face au risque opérationnel** (humain) lié à l'hétérogénéité des contrôles.*	• ***Améliorer les contrôles** afin de réduire les risques et mettre en place des bonnes pratiques de conformité*

1.3 Objectifs de l'étude

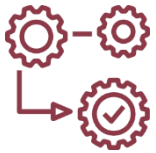
Collecter des critères pour comprendre le document et aider les équipes techniques

Reconnaitre l'architecture des traités et le sujet des différentes clauses

Créer un outil automatisé, simple et évolutif

qui puisse aider les souscripteurs, gestionnaires, et actuaires à…

Gagner du temps et se consacrer à d'autres tâches à plus haute valeur ajoutée



Définir des mesures de pertinence et des contrôles permettant d'assurer la qualité de l'information proposée

Pouvoir exporter et exploiter ces données à des fins métier

1.4 Périmètre de l'étude (1/2)

- Etude réalisée conjointement avec les équipes **CCR Re et Reacfin.**
- Approximativement **450 documents exploités** pour cette étude.
- Des **documents en anglais** dans un premier temps pour simplifier l'approche.
- Des **document à la fois « images » et « digitaux »** qui nécessitaient une attention toute particulière (qualité des données).
- **Différents formats** induits par différentes sources (courtiers, partenaires, etc.) afin de représenter l'activité quotidienne des souscripteurs.

1.4 Périmètre de l'étude (2/2)

- Analyse des **traités non proportionnels** exclusivement.
- Définition explicite **de critères et de clauses** (environ 30) à collecter/analyser :

Nature du traité

Clause sanction

Juridiction

Part cédée

Clause terrorisme

Dépôt de prime minimum

Date de fin de couverture

Priorité

Frais de courtage

Taxes additionnelles

Exclusion

Clause RGPD

Devise de règlement

Périmètre territorial

2.1 Environnement informatique

- Choix de technologies **open sources et évolutives:**

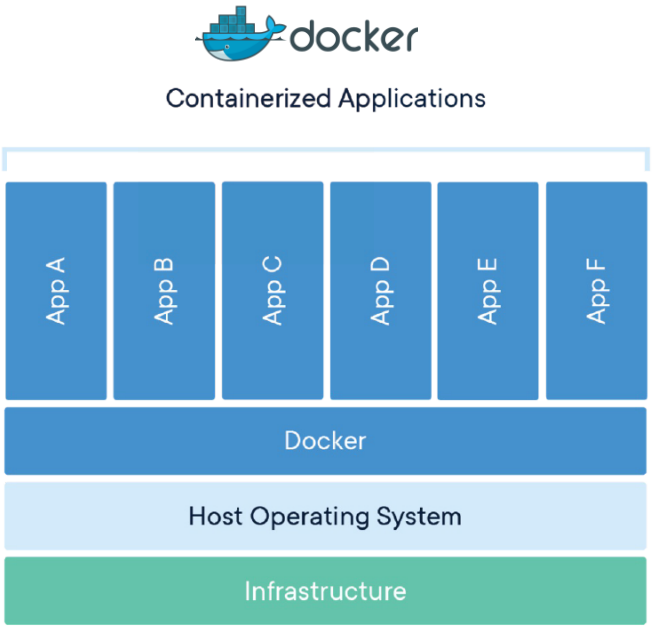
- Démarche progressive basée sur **une organisation agile** (SCRUM) et **une intégration continue** (déploiement de nouvelles versions tous les jours) facilitant l'intégration des demandes des opérationnels.

2.2 Utilisation de conteneurs

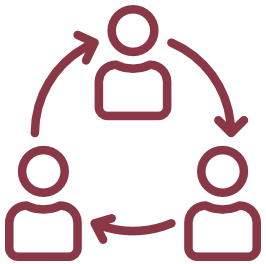
- Un **conteneur est une unité de logiciel standard** qui regroupe à la fois le code d'un programme et ses dépendances de telle sorte à ce que l'outil puisse fonctionner dans n'importe quel environnement.

- C'est un **procédé essentiel au sein du projet** qui permet à la fois:
  - De travailler entre développeurs sur un environnement similaire;
  - De déployer sur un serveur tiers un outil exactement identique à celui développé.

- A l'instar d'être une référence absolue en IT, c'est aussi une **pratique trop souvent ignorée** en actuariat (en développement comme en production).



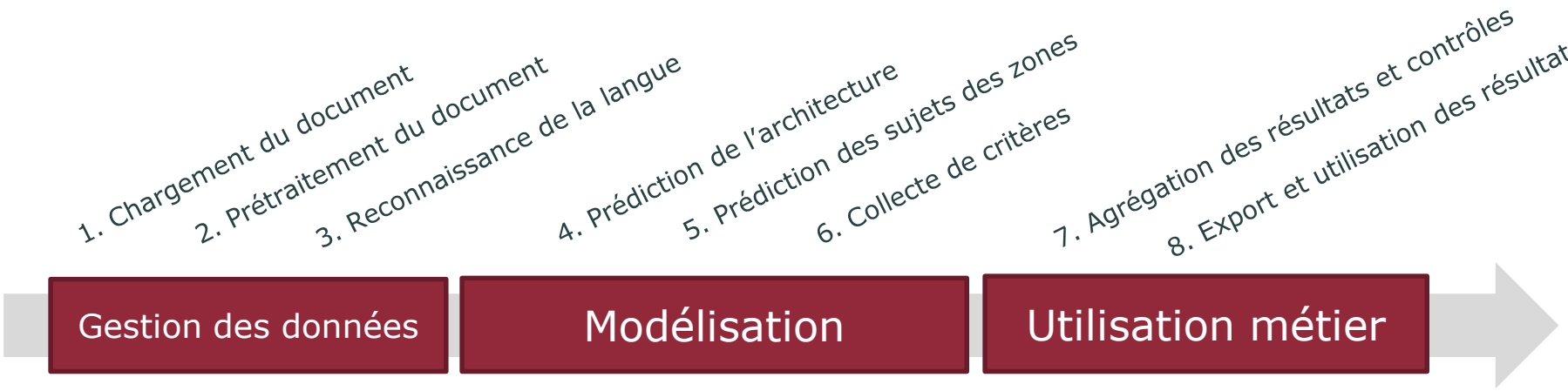
3.1 Retour sur le déroulement de l'étude



- **Organisation**:
  - 28 jours d'étude avec 2 ETP Reacfin et autant côté CCR Re
- **Déroulé**:
  - Définition du cahier des charges et liste des critères à récupérer
  - Création du Minimum viable product (MVP)
  - CoProj pour affiner l'approche et discuter des remarques
  - Evolutions et développements
  - Workshop de pré-restitution
  - Tests de performance pour restitutions
  - Restitutions
  - Documentation, formation et livraison du code

Itérations

3.2 Fonctionnement de l'outil



Représentation des mots + **Modèles deep learning et expressions régulières** + **Visualisation des résultats et KPIs**

3.3 Focus sur la représentation de mots (1/2)

- La **représentation des mots est une étape cruciale** dans le prétraitement des données textuelles.

- L'objectif est de **représenter la signification d'un document** à travers une forme plus exploitable pour les modèles de machine learning (une matrice d'éléments numériques par exemple)

Term document matrix

- Fréquence de mots
- Dim. de 20K à 50K
- Capture des différences générales et basiques mais ne fait pas de lien entre les mots

	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

- Définit l'importance d'un mot selon sa proportion de survenance dans tout le jeu de données
- Dim. de 20K à 50K
- Capture mieux certaines spécificités mais toujours pas de liens

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

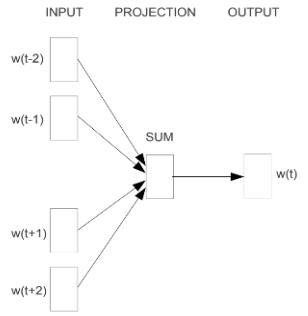
tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Word embedding

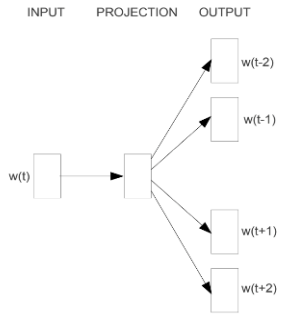
- Utilise un espace vectoriel pour définir la signification d'un mot par rapport à un contexte
- ANN donne une dim. de 250 à 500
- Capture très bien la relation entre les mots
- Word2vec, GloVe, FastText, etc.

3.3 Focus sur la représentation de mots (2/2)

- **Vector space word representation (Word2vec)**:
  - Compréhension des mots selon un contexte local
  - 2 approches complémentaires:
    - Continuous bag-of-words (CBOW)
    - Continuous skip-gram model (Skip-gram)
  - Référence: https://arxiv.org/pdf/1310.4546.pdf

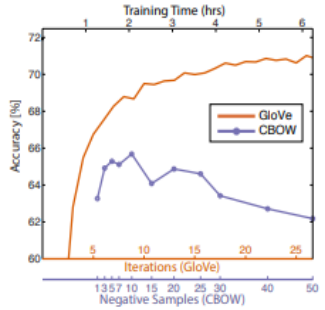


CBOW

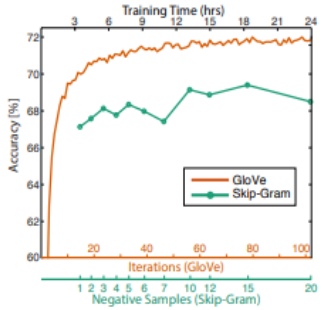


Skip-gram

- **Global vector for Words representation (GloVe):**
  - Utilisation de word2vec et de techniques de factorisation de matrices (analyse sémantique latente, LSA) pour améliorer l'approche initiale
  - Référence: https://nlp.stanford.edu/pubs/glove.pdf



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

- Dans le reste de l'étude la méthode sera **GloVe**

3.4 Focus sur les modèles d'apprentissage et le text mining (1/2)

- Pour comprendre et collecter l'information des traités il est nécessaire de définir **une série de modèles** que l'on peut résumer en **2 catégories:**

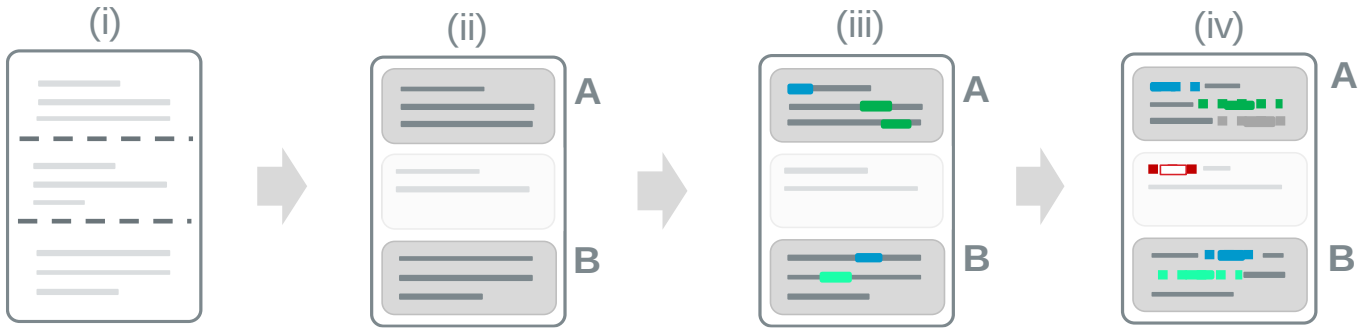
Modèles deep learning

- **Prédiction de la structure du traité** par modèle supervisé. Techniques de classification permettant de scinder le texte en zones. (i)
- **Prédiction des sujets des différentes zones**. Mesure de similarité pour comprendre les signification de chaque zone sur base de seuil de pertinence. (ii)

Modèles text mining

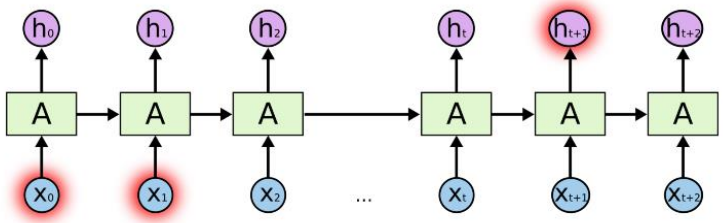
- **Expressions régulières et règles complémentaires** permettant de collecter des éléments candidats à analyser et préciser. (iii)
- **Analyse des contextes** pour définir les candidats les plus pertinents pour un critère à étudier. (iv)

3.4 Focus sur les modèles d'apprentissage et le text mining (2/2)



Focus sur la **classification par RNN:**

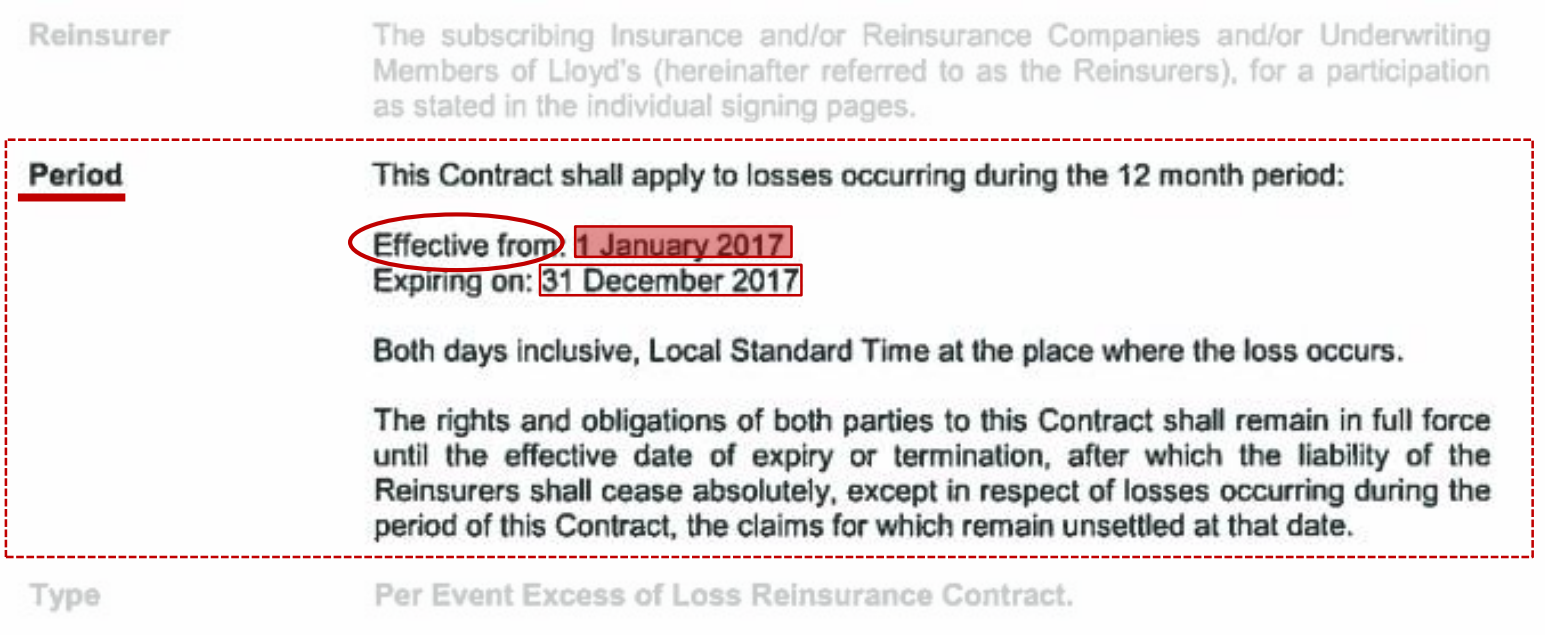
- Tests sur des modèles **SVM, MLP et RNN;**
- Le RNN est le modèle le plus efficace principalement parce qu'il tient compte de la **séquence des éléments** qui lui sont donnés.



3.5 Exemple d'analyse

- Pour résumer, un critère s'obtient donc grâce à **une zone, un sujet, une expression régulière et un contexte**.

Exemple avec la variable *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.

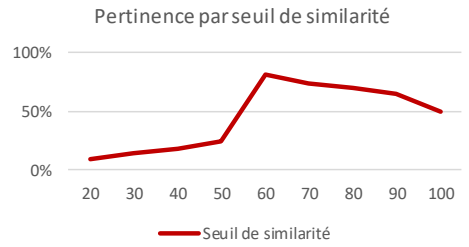
4.1 Résultats de l'étude (1/2)

- (i): Le modèle **RNN prédit correctement l'architecture**. Les 2% restant pourraient être diminués en agrandissant le jeu de données.
- (ii): **La reconnaissance des sujets des zones fonctionne aussi bien** : les erreurs proviennent principalement du seuil d'acceptation de similarités qui est élevé.
- (iii) et (iv): le résultat de collecte des données est satisfaisant puisque l'outil récupère (correctement) **en moyenne 80% des critères pouvant être extraits** dans un traité.

98% de pertinence

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

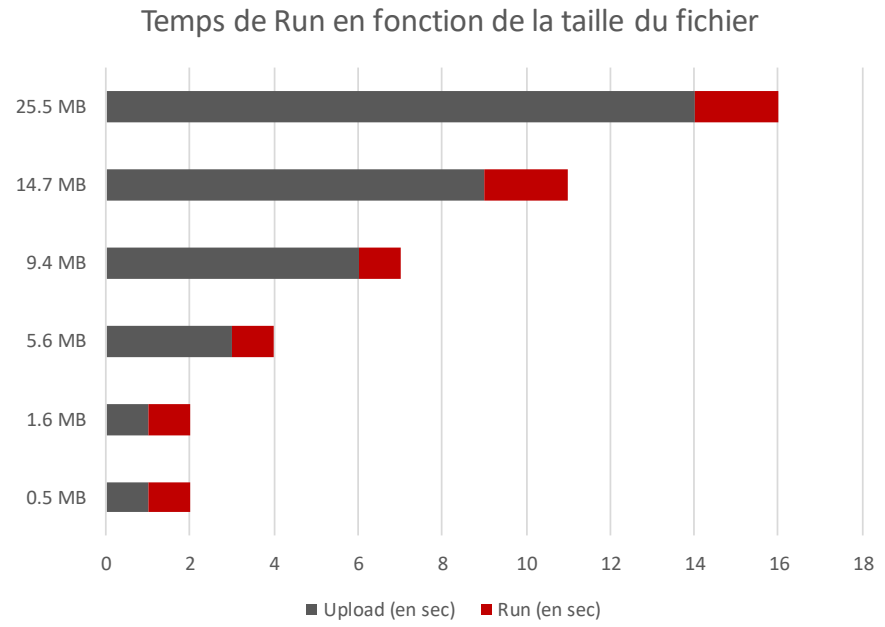
78% de pertinence



4.1 Résultats de l'étude (2/2)

- **L'outil analyse un traité en 2sec à 16sec** ce qui représente un gain considérable par rapport à une analyse manuelle.

- En réalité **le temps de calcul des algorithmes est même inférieur à 2sec:**
  - Le temps de chargement du PDF occupe une partie importante du temps;
  - L'utilisation des RNN permet de gagner du temps par rapport à des analyses de texte traditionnelles;
  - Enfin aucun OCR n'est nécessaire.



4.2 Utilisation en tarification

- Les avantages pour les équipes de tarification d'utiliser de telles techniques sont nombreuses :

Feature engineering
pour créer de nouvelles variables explicatives et améliorer le pouvoir prédictif des modèles

Feature selection
pour évaluer les variables qui ont le plus d'influence afin d'affiner le modèle (via modèle supervisé)

Accélération du processus de quotation
afin de transmettre aux équipes commerciales des informations tarifaires quasi instantanées

Meilleur segmentation de l'offre
en utilisant de nouveaux critères (via modèle non-supervisé)

4.3 Utilisation en Risk management

- Les travaux en risk management peuvent aussi bénéficier de telles pratiques comme par exemple pour :

Améliorer les mesures d'impact de sinistres
pour définir des KPIs compte tenu d'une zone, d'un risque, d'une industrie, etc.

Créer des data visualisation utiles
pour donner une vision commune en interne du risque et du business mais aussi en externe.

Réduire le risque opérationnel
dû aux fautes de frappes, aux informations incomplètes, aux mauvais checks, etc.

Définir des règles de conformité strictes
pour s'aligner au mieux avec la stratégie de risk management de la société.

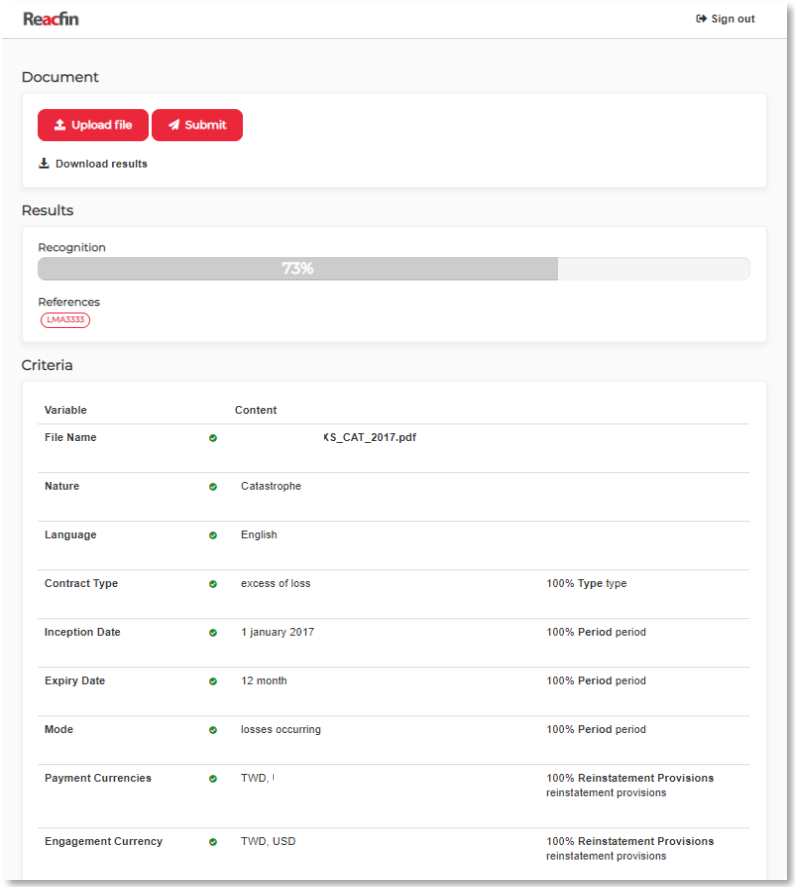
4.4 Conclusion et retour utilisateur

- La **collecte et l'exploitation des données non structurées** est un excellent exemple d'utilisation de la data science pour les actuaires.
- L'usage de techniques telles que le **word embedding** ou le **deep learning** sont des moyens efficaces pour comprendre des documents comme **les traités de réassurance**.
- Au delà du **gain de temps considérable**, l'apport pour les souscripteurs, de tarification et de risk management sont nombreux: création de nouvelles variables d'études, mise en place de contrôles automatiques de conformités, etc.
- Les premiers retours utilisateurs sont excellents et vont dans le sens d'une profonde modernisation des pratiques sans pour autant laisser l'humain de côté:

« Vers une **I**ntelligence **A**ugmentée plutôt qu'une **I**ntelligence **A**rtificielle »

5. Démonstration

- Exemple d'exploitation de l'outil avec un traité de réassurance anonymisé:



Merci

Avez-vous des questions ?

Loris Chiapparo,
Data Scientist
Loris.Chiapparo@reacfin.com

Aurélien Couloumy,
Head of Data Science
Aurelien.Couloumy@reacfin.com

Jérôme Isenbart,
CRO
jisenbart@ccr-re.fr

Reacfin

https://www.reacfin.com/

CCR RE

https://www.ccr.fr/en/ccr-re/