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

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

Loris Chiapparo, Data Scientist
loris.chiapparo@reactfin.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.
Agenda

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

1.1 Data Science and actuaries

1.2 Business case context

1.3 Goals

1.4 Scope
1.1 Data Science and actuaries

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

Everything… **no!**

1. **Performance**
   Improving processes by reducing time and efforts.

2. **Risk assessment**
   Improving the analysis and the understanding of risks.

3. **Market overview**
   Facilitating competitive, regulation, market and customer needs watch.

- 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?**
1.2 Business case context

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

<table>
<thead>
<tr>
<th>Issues</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- An heavy and repetitive workload for already very busy business teams.</td>
<td>- Automate the analysis: to save time during underwriting and pricing process.</td>
</tr>
<tr>
<td>- A complex document analysis with different structures and formats.</td>
<td>- Simplify document understanding: to collect and assess accurate and usable information.</td>
</tr>
<tr>
<td>- An incomplete view of criteria and clauses which have been underwritten.</td>
<td>- Improve controls: to reduce risks and set up compliance rules.</td>
</tr>
<tr>
<td>- Operational risks exposition due to heterogeneous and non exhaustive controls applied by hand.</td>
<td></td>
</tr>
</tbody>
</table>
1.3 Goals

Collect criteria and key elements that explain the documents.

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

Recognize treaties architecture and clauses topics.

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

Create a simple, scalable and effective AI tool that can help underwriters and actuaries to...
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.
1.4 Scope – business aspects (2/2)

- Non proportional treaties analysis.

- Collection of business criteria and clauses, among others:

  - Nature of the treaty
  - Expiry date
  - Taxes information
  - Currency
  - Share part
  - Minimum premium deposit
  - Limits
  - Jurisdiction
  - Territorial scope
  - Lloyds references
  - Sanction amended clause
  - Exclusions
  - Terrorism amended clause
  - Brokerage fees
2. IT Framework

2.1 Environment and technologies

2.2 Containerized applications
2.1 Environment and technologies

- Open and scalable technologies:

  - Python
  - TensorFlow
  - JavaScript
  - Flask
  - Amazon Web Services
  - Docker
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**
3. Process general functioning

3.1 Introduction

3.2 Words representations

3.3 Deep learning and text mining

3.4 Example
3.1 Introduction

Process general functioning

1. Document loading
2. Document pre-treatment
3. Language recognition
4. Architecture prediction
5. Topics prediction
6. Collection of criteria
7. Agregation and controls
8. Exports and use of results

Data management | Modeling | Business use

Word vectorization + Deep learning & regex + Visualization & KPIs
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

\[
\begin{pmatrix}
T_1 & T_2 & \ldots & T_t \\
D_1 & w_{11} & w_{12} & \ldots & w_{1t} \\
D_2 & w_{21} & w_{22} & \ldots & w_{2t} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
D_n & w_{1n} & w_{2n} & \ldots & w_{tn}
\end{pmatrix}
\]

### 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

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

- \( w_{i,j} \) = 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.
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)

• **Global vector for Words representation (GloVe)**:
  - Joint use of word2vec and matrix factorization techniques (Latent semantic analysis, LSA) to improve word embedding

• In the next parts, we will use **GloVe**
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)
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 characteristic 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), phone numbers,
  - IP addresses,
  - Dates
  - Hexadecimal values,
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*

| Period          | Effective from: 1 January 2017  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expiring on: 31 December 2017</td>
</tr>
<tr>
<td>Reinsurer</td>
<td>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.</td>
</tr>
<tr>
<td>Type</td>
<td>Per Event Excess of Loss Reinsurance Contract.</td>
</tr>
</tbody>
</table>
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.
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)
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)
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.
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.
5. Demo version
5. Demo version

- A simple demo version of the project already used by many underwriters and pricing actuaries.
The views expressed in this [publication/presentation] are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this [publication/presentation] and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this [publication/presentation].

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this [publication/presentation] be reproduced without the written permission of the IFoA [or authors, in the case of non-IFoA research].