Text Mining and Ruin Theory: A case study on Risk Models with Dependence
Symposium on Big Data in Finance, Retail and Commerce: Statistical and Computational Challenges
Lisbon, 2-3 November 2017, Portugal

Text mining and ruin theory: A case study on risk models with dependence

Renata G. Alcoforado
ISEG & CEMAPRE, Universidade de Lisboa & Universidade Federal de Pernambuco, alcoforado.renata@gmail.com

Alfredo D. Egídio dos Reis
ISEG & CEMAPRE, Universidade de Lisboa, alfredo@iseg.ulisboa.pt

Abstract

This work aims to analyse unstructured data using a text mining approach. In our study, the subject is composed by 27 published papers of the risk and ruin theory topic, area of actuarial science, that were coded in 32 different categories. For the purpose, all data was analysed by using the software *NVivo 11 plus*. Software *NVivo* is a specialized tool in analysing unstructured data.

Keywords: Big data; Unstructured data; Text mining; Risk theory; Ruin probability; Dependence.

Introduction

Big Data is an area of great development in statistics. We can define Big Data as "a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data. Big Data is often defined along three dimensions – volume, velocity, and variety" (TechAmerica Foundation's Federal Big Data Commission, 2012).
According to Han *et al.* (2012) data mining is the process of mining through large amount of data to extract meaningful information, knowledge. It's also treated by many people as a synonym for knowledge discovery from data, or KDD.
Text mining in an analogous manner as data mining, aims to extract information from data, but in this case the data comprehend to texts and does it

through identification and exploration of interesting patterns (Feldman and Sanger, 2006).
The manuscript is organized as follows: In Section A we speak about the collected data to be analysed. Section B is about the coding of the data, the coding matrix, the relationship between the nodes, that is, we plotted the nodes hierarchically. Then, we present the cluster analysis for the nodes and the sources (papers), the comparison diagrams and to finalize, a structural matrix. To finalize we exhibit some conclusions.

Section A - The Data

In our particular work we uploaded 27 papers in the platform, codified and then analised all data. These papers are references for a particular research project in development in risk theory. These papers are: Afonso *et al.* (2017), Ammeter (1948); Asmussen and Albrecher (2010); Bergel and Egídio dos Reis (2016); Constantinescu *et al.* (2011); Constantinescu *et al.* (2012); Constantinescu *et al.* (2016); Czado *et al.* (2011); Frees and Wang (2006); Frees *et al.* (2011); Frees *et al.* (2016); Garrido *et al.* (2016); Gschlößl and Czado (2007); Jasiulewicz (2001); Jørgensen and Paes De Souza (1994); Krämer *et al.* (2013); Kreer *et al.* (2015); Li *et al.* (2015); Maume-Deschamps *et al.* (2017); Ni *et al.* (2014a); Ni *et al.* (2014b); Quijano Xacur and Garrido (2015); Renshaw (1994); Rolski *et al.* (1999); Schulz (2013); Shi *et al.* (2015); Song *et al.* (2009).
The software builds a *word cloud* composed by the the most pertinents words in our entire data base to use in our study. After removing all the verbs, articles and non-meaningful wording, the words are then gathered according to their *stem*, making possible to obtain the *cloud* in Figure 1.

Figure 1: Word Cloud Source: The authors

Section B - The Coding and Analysis

Data mining assumes that the data is already stored in an structured way, whereas text mining assumes that the data is unstructured and still needs coding. (Feldman and Sanger, 2006)
In sequence, all the unstructured data was codified, that is, we put in each one of the categories the respective parts from text to be able to analyse it in a mathematical way. In other words, after coding we get a structure to be able to analyse with clusters and matrices. Then, plotting the data which was not possible previously. The categories were selected after extensive reading and observing the *word cloud*.
In our particular exercise the *codes* are: Actuarial; Aggregate Claims Model; Claim Frequency; Claim Severity; Compound Poisson; Conditional; Copulas; Covariates; Dependence; Exponential; Formula; Function; Gamma; Independence; Insurance; Joint Distribution; Loss; Markov; Martingale; Mixed-Poisson; Parameters; Prediction; Premium; Randomness; Regression; Renewal; Risk Theory; Ruin Probability; Simulation; Spatial; Stationary and Stochastic Process.
After the code and data organizing, we constructed the *coding matrix* presented on the left graph of Figure 2, which shows both in numbers and in graph what is the relationship between the coded categories. A *word tree* in each category is plotted to see the connection from that word (or expression) in the sentence where it belongs.

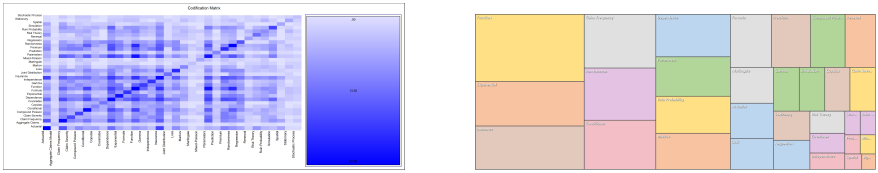

Figure 2: Coding Matrix - Heat and Nodes Hierarchically respectively

We plotted the *nodes* hierarchically presented on the right graph of the Figure 2 to observe which categories are most frequent, the most important among the data available. The *cluster analysis* was performed in cluster by word similarity using Pearson's correlation coefficient as the similarity metric. We made it for both the categories and the sources to see how they relate. The cluster analysis for the nodes is presented in Figure 3, a diagram on the left and a circle graph on the right.

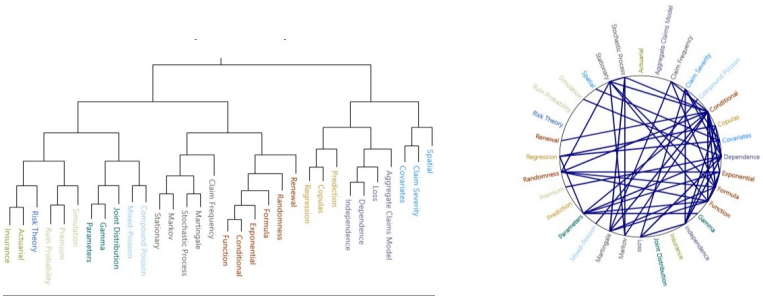

Figure 3: Cluster Analysis of the Nodes

Conclusions

Our source intended to talk about the calculation of premiums and ruin probabilities for insurance application, also how to associate the claim frequency with their severity. Some authors use copulas, others use covariates in a regression model, and others try to find a distribution that can capture that dependence.
The result showed to be interesting to compare respective categories and plot comparison diagrams, for instance, comparing *Dependence* with *Independence*; *Simulation* with *Formula*; *Copulas* with *Covariates*; *Regression* with *Copulas*; *Claim Severity* with *Claim Frequency* among others.
To finalize, we built the *structural matrix* where each row shows each paper and each columns the category mentioned above, in order to identify subtle connections which can allow a thorough and rigorous study.

References

Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook*. Cambridge University Press, New York, USA.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining. Concepts and Techniques*. Elsevier, Waltham, USA, third edition.

TechAmerica Foundation's Federal Big Data Commission (2012). Demystifying Big Data: A Practical Guide To Transforming The Business of Government. Technical report, TechAmerica Foundation's. Retrieved July 10, 2017, from https://www.attain.com/sites/default/files/take-aways-pdf/Solutions_Demystifying Big Data - A Practical Guide To Transforming The Business Of Government.pdf