TEXT ANALYTICS USING RAPID MINER

By

Aaron Ward

Supervisor(s):

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
B.SC IN COMPUTING AND INFORMATION TECHNOLOGY
AT
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN
DUBLIN, IRELAND
2017

Declaration

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, except where otherwise stated. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references.

I/We understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I/we engage in plagiarism, collusion or copying. I acknowledge that copying someone elses assignment, or part of it, is wrong, and that submitting identical work to others constitutes a form of plagiarism. I/We have read and understood the colleges plagiarism policy 3AS08.

This material, or any part of it, has not been previously submitted for assessment for an academic purpose at this or any other academic institution. I have not allowed anyone to copy my work with the intention of passing it off as their own work.

Dated: 2017

Author: Aaron Ward

Text Analytics Using Rapid Miner

Business Understanding

This following sections shall describe the business objectives, the mining objectives, a analysis and plan for the proposed project.

Business Objective

The objective of this project is to mine unstructured data using Rapid Miner.

Data Mining objective

- The main objective is to mine articles about Machine Learning, Deep Learning and Robotics.
- Web crawling will be implemented to find related text on the topics.
- Preprocessing steps shall be put in place in order to produce the most predictive words for modeling.
- Multiple clustering and classification models will be used to identify the texts.

Project analysis

In terms of a cost benefit analysis, this project deems relatively efficient due to the benefits out weighing the costs. The proposed assignment poses minimal risk as only a few events may delay the production of the project, such as unavailable texts and misclassification,

which are very unlikely. The resources required for this project include: The RapidMiner software and all it's operators related to the data mining objective, articles from online resources, 13 documents based on 3 categories and a online word cloud service for visualisation of predictive words or phrases.

Project plan

There are a number of steps taken in relation to this project.

- Initially, Three categories are decided on. Two topics should be relatively similar, and the last on will be unrelated. In this case, they are based on Deep Learning, Machine Learning and robitics.
- Thirteen texts are sourced online for each category. Ten shall be used for training datasets and three shall be used for testing. Web crawling will be implemented using RapidMiner for three of those texts. The steps taken will be documented
- These source articles will be downloaded into their respective folders based on their class label.
- A word cloud will be creating to visualise the frequently occurring terms.
- Preprocessing steps shall be experimented with to compare stemmers, determine which words are predominantly more predictive, apply pruning and comparing accuracy based on different vectors and documenting the results.
- Apply two algorithms for clustering and classification and discuss the accuracy of them.
- Lastly, an evaluation of the project will be performed, which will assess the overall result in relation with the business objectives.

Data Understanding

As previously mentioned, thirteen texts have been obtain online according to three categories: machine learning, deep learning and robotics. In order to obtain a better insight to the commonly occurring word in the files of each category, word clouds were made. As seen in Figure 1, Figure 2 and Figure 3 three word images were made to visualise the frequent terms. The larger the word appears in the image, the more often it appears in the documents that have been collected. For machine learning, it is clear that the terms **knn**, **training**, **decision tree**, **naive bayes** and **classification**.

Figure 1: Word cloud for the machine learning category

A more concise description of deep learning can be seen in Figure 2. Although the two categories are related, these terms seem to focus more so on words like **networks**, **neural**, **CNN's**, **model** and **recurrent** to name a few.

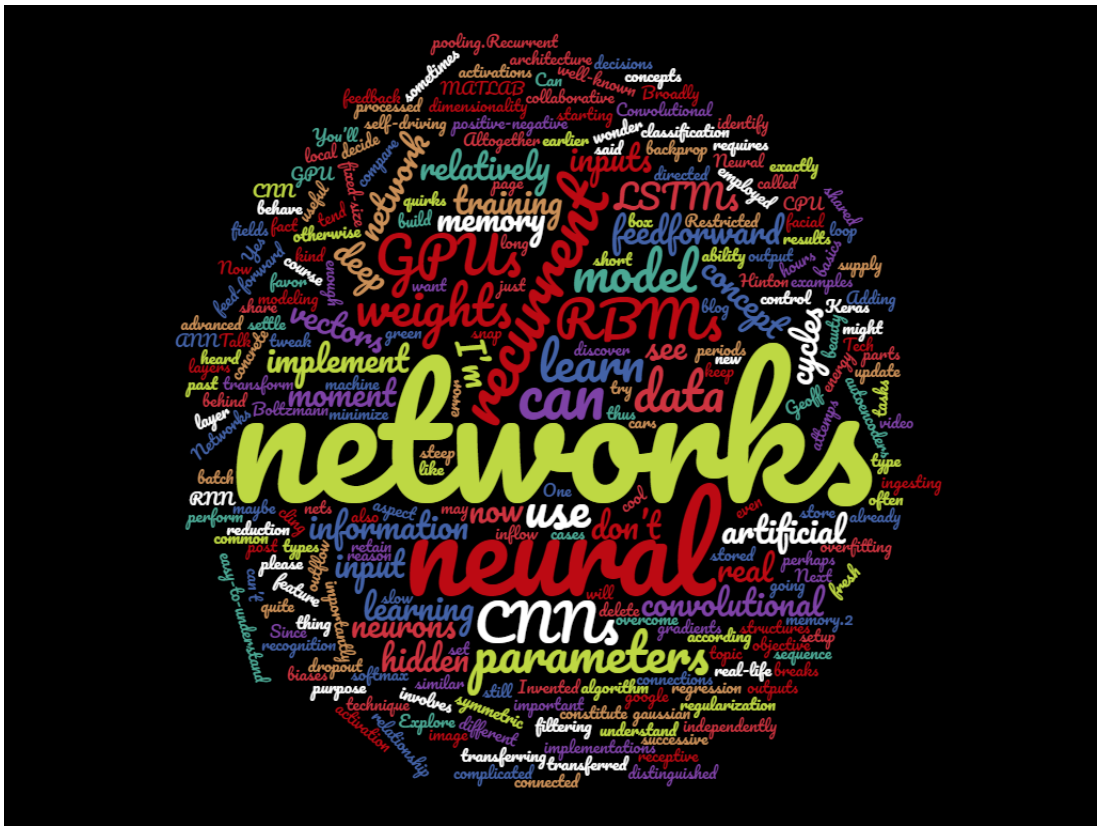

Figure 2: Word cloud for the deep learning category

Lastly, we see the terms generated for the robotics category. Evidently, these words are not related to machine learning or deep learning as they share very few words to the documents in the other two categories. The prominent words found here are **robots**, **humans** or **humanoid** and **robotics**

Figure 3: Word cloud for the robotics category

Not only do these word clouds help determine predictive words, but also provide certain words that are not useful for the mining objects,also known as stop words. Certain stop words in these figures are ”I’m”, ”thing”, ”makes” for example. These words can be ambiguous depending on the sentiment of the article being read or just common English phrases, therefore they should not be included in the working model.

Furthermore, potential synonyms can be found in these word clouds. To reduce dimensionality, it is beneficial to specify words that may hold the same meaning. For instance, it can be seen that in the category of robotics, the terms ”robot” and ”robotics” are in interchangeable, therefore they can be classes as synonyms of one another. Another example of this could be the phase ”neural network” and ”ANN”, which the latter is a acronym for ”neural network” therefore that can also be classed as a synonym.

Data Preparation

Modeling

Modeling techniques

Clustering

Classification

Test Design

Build and Assess the Model

Evaluation

Project review

Project deployment

Appendices

Appendix A

Rapid Miner Processes in XML