ICBAA2017-4

JOURNEY FROM TEXT MINING TO NEURAL INFORMATION RETRIEVAL IN AGRICULTURAL DATA SCIENCE

Ayman Salama[1,2], Tomas Maul[1], Ebrahim Jahanshiri[2], Neil Crout[3]

[1]University of Nottingham Malaysia Campus
Tomas.Maul@nottingham.edu.my
hcxas1@nottingham.edu.my
[2]Crops For the Future
ebrahim.jahanshiri@cffresearch.org
ayman.mohamed@cffresearch.org
[3]University of Nottingham, UK
neil.crout@nottingham.ac.uk

Abstract: Agricultural data, like data in other disciplines, tends to suffer from lack of common structure and centralization. Several international and local institutions spend years of research leaving behind large amounts of unstructured and semi-structured data. Data structuring and centralization enables data-wide scale processing and comprehensive decision making. Lately data centralization lost its importance due to the exponential rate of data growth. Current advances in computer hardware and software has provided the capability for online processing of existing and future data. The focus of this paper is to introduce our journey from text mining technology to neural information retrieval in agricultural data science. Text mining techniques that use regular expression to retrieve knowledge out of semi-structured data tend to be deterministic. We have conducted an experiment on 691K words of agronomic datasets and managed to extract 13K local names, 1.7K synonyms and a taxonomy of 2.2K crops. The text mining techniques scored 97% accuracy based on 20% verified samples. However, analysing 691K words is a toy problem compared to analysing 3.4 billion words as with the Wikipedia English corpus. The text mining performed on 691K words was possible partly because the data was recorded in a semi-structured format. However, the format and the structure were not followed in many cases. Text mining was inevitable to transfer the data from a semi-structured information to understandable knowledge. Wikipedia on the other hand is not limited to any category of people for data collection and documentation. Text mining techniques using regular expression will possibly fail to analyse such huge and complex unstructured datasets. Artificial intelligence advances along with advances in computer processing and storage power, empower scientists with extraordinary tools for data analysis and visualization that were not possible a few years ago. Neural information retrieval [Mitara, 2017] was used in our research to process a Wikipedia corpus and extract insights from agricultural data. Word2Vec [Mikolov 2013] and FastText [Joulin 2016] are two recent neural network approaches developed by Google Research and Facebook Research for text analysis. Our experiment used a dynamic version of those tools and connected it to the CropBASE knowledge system to analyse 3.4 billion words from Wikipedia against our knowledge base. The analysis aimed to identify the relationships between common English crop names and different countries. Word2Vec and FastText transfer the corpus of text into a vector space (i.e. word embeddings) and the resulting vectors can be used to measure relationships between words. The results of our experiment show strong relationships between crop names and countries that are aligned with expert knowledge. The results can be used as an intermediate search tool for scientist to explore large text corpora.

Keywords: Neural-Networks, Text-Mining, Agricultural-data, Word2Vec, FastText.

INTRODUCTION

Science advancement resulted into significant amount of research data. Most of the data are scattered in different structured or semi-structured format. Starting from 2355 semi-structured data about crop ecology and taxonomy, the journey started. Hundreds of hours are spent to collect and document those data. Human can easily interpret the data and extract knowledge out of it. We estimated more than 150K pieces of data about 2.3K crops exist in those documents. Manual standardization of this data will take months of work and will inevitably suffer from human errors during data processing. Thus, computer algorithm has to step in to do the work. Scaling up from few thousand documents to millions require change in the computational methodology. Significant amount of reliable data can be extracted from this Corpus of data. Early work of [Sebastiani, 2006] reveal the potential use of machine learning in text classification. The effectiveness of the use of Convolutional Neural Network CNN had shown potential and promising success in text classification [Kim, 2016]. Word2vec as an example had been successfully used to extract ontology concepts from corpus of data [Wohlgenannt, 2016]. In the same context [Sowoboda, 2016] had proven the ability of word2vec to extend seed taxonomy. [Shah, 2016] conducted a successful trial experiment to construct taxonomy using word2vec to automatically build and evolve the taxonomy with high precision with minimal assistance from a domain expert. The explanation of word2vec was confusing for external audience. [Goldberg, 2014] presented an explanation paper to explain word2vec. Imagine a space where all words are floating connected together based on our recorded knowledge. You can navigate this space from a word to another measuring the relationship and understanding the semantic meaning of it. Dimensionality of linguistic data of Wikipedia allow 27 billion words to exist in one location. The advancement of Artificial Neural Networks alongside with the increase of computer machine power, several experiments become possible [Goodfellow, 2016]. Training a neural network model with 51 G of data was never possible 10 years ago. The idea can be simplified in presenting each word in a numerical vector that can be calculated against accompanied words and contained documents. Repeating this process for billions of words creates vector space where words can cross each other. Unlimited amount of analysis can be executed by measuring the distance between words relative to each other. Analysing crowdsourced data have the potential to reveal unpresented conclusions. Research institutions, government and industries spend significant budget to understand the market and human behaviour. Social media, internet and open source projects allowed billions of humans to participate willingly with public datasets. This resulted in what we call today Big Data. Understanding big data and analysing it can save time, effort and money that was spent to study the crowd and their interactions. There are two main obstacles in this kind of research, the first one is obtaining the datasets. The second obstacle is the software and methodology required to process this amount of data. There are several algorithms that are designed to do the task, however the research in this area still growing and several questions are yet to answer. We have conducted previously an experiment to analyse the engine search data where billions of search keywords are executed daily [Salama,2017]. We managed to obtain search data of more than 300K search events and used it to predict cropping pattern for Bambara groundnut by studying people search engine data worldwide. Obtaining this data required significant amount of manual work and currently the extension of that research is not possible as the method for the data extraction become expensive. In this research we manage to find another source of crowd data. And most importantly we manage to develop tools to analyse the big datasets.

METHODS

This research study two methods of understanding corpus of data. The first one is text mining and the second is Neural Information Retrieval. We used text mining to develop an algorithm to extract deterministic data out of small corpus of data “1.2 million words” of taxonomy and ecology documents. The first task was to create simple data collection tools to format all documents and centralized. Given the amount of data and human factor in data collection, several special characters and deformation of data had been discovered. The second task of the algorithm is to clean up the data from special characters and ascii encoding to prevent algorithmic crash. The data can be divided into three categories: Text in a paragraph, data sheet in table and specific text in tables. We develop two algorithms to extract the data, the first was PHP code to work on numerical data tables. The second algorithm was developed in Linux shell scripting for data extraction from text in table format or paragraph format. The output data had been structured and saved in a relational knowledge based system [Jahanshiri, 2015]. Figure1 explains the process on using text mining to extract data from documents. The output of the algorithm was 120K numerical data like optimal maximum temperature for a crop and 30K textual data like crop taxonomical names.

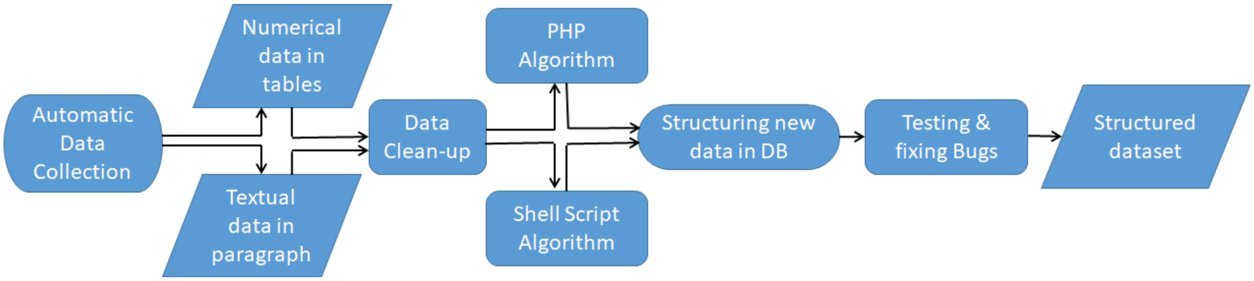


Figure 1: Text Mining process to ecology and taxonomy data.

The second method is to train word2vec on English Wikipedia and then measure the distance between crop common English names versus several concepts. The initiation of the experiment was based on a pre-trained model on English Wikipedia 2015. There were several problems with the retrained model. 80% of the vocabulary that we searched for didn’t exist in the vocabulary dictionary of Wikipedia 2015. The model was case sensitive, so it doesn’t solve the upper and lower cases of the same exact words. Those factors and others affect the accuracy of the measured vocabularies. The results were still semantic and carry respectful amount of data analysis. In order to achieve higher accuracy and enrich the vocabulary dictionary, we trained our own Neural Network model of word2vec on English Wikipedia 2017. The python data processing for training the model took around 10 hours on customized Google Cloud Machine, 7 CPU Cores, 22 G RAM based on Debian OS. The English Wikipedia corpus size as of October 2017 is 14G of compressed XML format. The model can work on compressed file which make it faster and efficient in disk management [Julian FastText zip 2016]. Figure 2 explain the steps used to train and use the model. The output model trained on Wikipedia 2017 resulted in 30% increase of vocabulary dictionary compared to the pre-trained model on Wikipedia 2015.

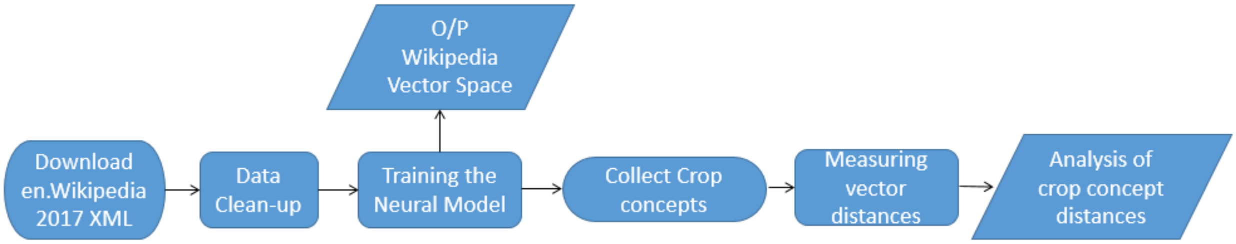


Figure 2: Train Neural Network word2vec on en. wikipedia corpus 2017.

RESULTS AND DISCUSSION

The input data to train the model was English Wikipedia 2017. The model was trained successfully and created a vector space of the vocabulary dictionary of English Wikipedia. Then the model was used to measure the relationship between 91 crops English names and several thousands other concepts like country names, crops local names and others. The output is a number presents the relationship between words ranges from -1 to +1 where +1 means completely related and -1 means complete opposite. We analysed manually couple of the results and decided that the data visualization techniques are the best to present the Big Data analysis. Three method of data visualization are used to present the data.

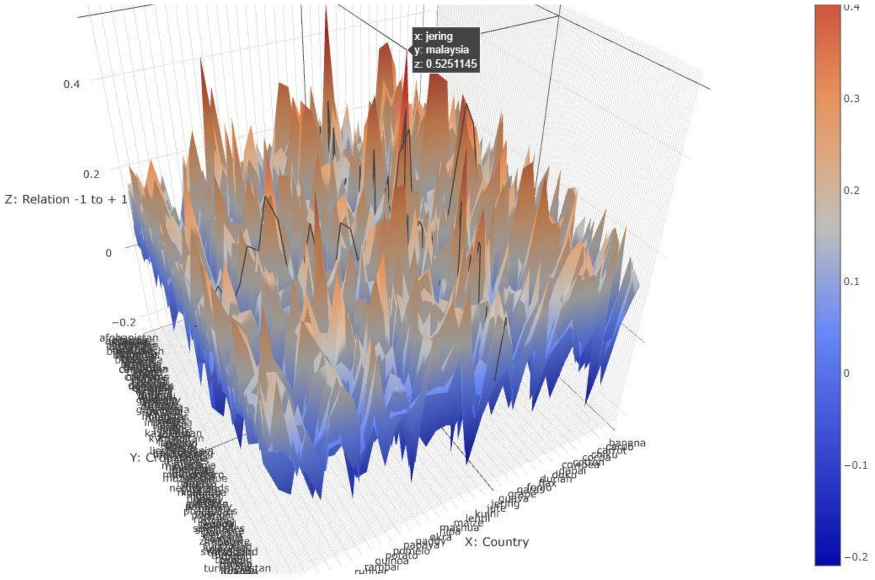
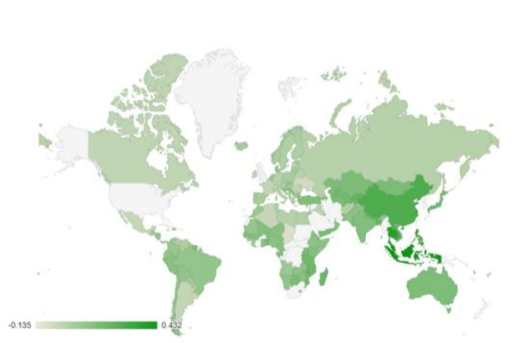


Figure 3: 3D graph presents the relation between country name and crop names.

The first method is 3D graph. It has been created using the open source library "Plotly" [Sievert, C. 2016]. We have created the graph online [3D online graph] to allow the user to study the graphs especially the 3D one. The "Plotly" 3D graph on a website allows the user to rotate the graph, zoom in and out, navigate the graph by the cursor to get specific data point on the graph. The three dimensions are, (1) X Axis: the country names, (2) Y axis: English crop common names, (3) Z axis: the relationship between the country name and the crop English common names. The graph has a colour scale to present the intensity of the relationship. The colour scale range from red to blue where red represents the stronger relationship between concepts and blue represents the opposite relationship between concepts. You can start by checking the peaks in the graph to see the most related concepts. It is noticed that the highest peaks are related to tropical crops. This happened because the focus of our research in the Southeast Asia. So, crops like durian, duku and jering come on the top of the list. Figure 3 shows a snapshot of the online 3D graph. The cursor in the snapshot shows the strong relationship peaks between Malaysia and the crop "jering".

The Second method is the geographical map. The map shows the relationships between countries and crop English common names. We have created online website for the map [Online Map] shown in Figure 4. The online map has been created using google GeoChart library. The map shows the relationship between the crop durian and several countries around the world. The Map has a colour scale of the green to represents the intensity of the relation by the increase of the intensity of the colour. It is noticed that the crop English common name durian has strong relationship with countries around Southeast Asia and equators which match are common knowledge of the crop. Some countries are not included due to technical limitation of the version we are working on. The third graph in figure 5 shows bar chart that represents the relationship between four crops with top 10 countries.



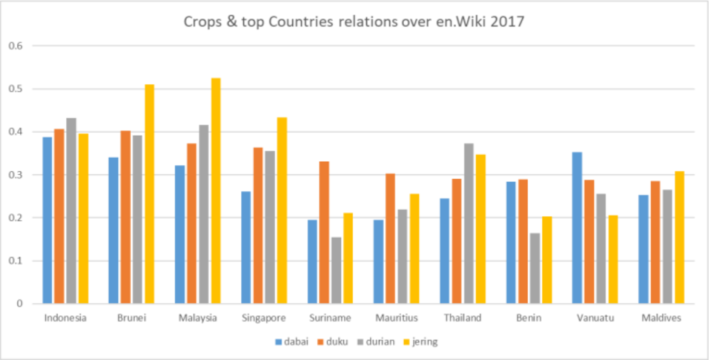


Figure 4. [left] Geographical map presents the relationship between durian and countries.
Figure 5. [right] Bar chart presents the relationship between top crops and top countries.

Can Wikipedia be considered as a reliable source of scientific information? Scientific study were performed to compare the accuracy of Wikipedia against other encyclopaedia and shows that Wikipedia accuracy is over 80% [Holman Rector, L. (2008)]. However as stated in the central limit theorem that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population [Rosenblatt, M. (1956)]. The amount of data presented in Wikipedia are assumed to be large enough to represent the true parameters of population.

Is English Wikipedia is enough to study international datasets? We measured in our experiment the relationship between crops local names from different countries and crops common English name and the output was confusing. For example, studying the closest relationship to rice local name "Padi", the result was a name of a famous character. That happen because his name has the word "rice" and he was mentioned in several Wikipedia articles. We are currently studying how to build a model to accommodate all 198 languages used in Wikipedia. The corpus of scientific literature is assumed to provide more concrete output than Wikipedia. We are planning to incorporate the scientific literature in our future studies. We are hoping that we can achieve deterministic data extraction from the literature to be used in a decision support system.

CONCLUSIONS

Wikipedia corpus analysis through Neural Network model provides significant data insights in crop science. The results of our experiment performed over English language matches the domain expert opinion. Based on our analysis, the current models that we used “word2vec” shows promising potential outcome. The power of data visualization is considered a key in this research. The variety of available libraries starting from bar charts and up to 3D surface graphs provide powerful insights of the output data. The current word2vec and FastText algorithm have limitation that cause inconsistency. For example, the algorithm works best with the analysis of concept that contains one word only. When the analysis is done for a concept composed of two words, it calculates the average of the two vectors that represent each concept. The accuracy will suffer due to this calculation. There were several attempts to solve this problem [Kim, 2014] which we are planning to test. [Bojanowski, 2016] proposed a solution that can include the morphology of the words based on skip-gram model which allow the model to include words that were not even in the vocabulary dictionary. The research is still ongoing to enhance the Neural Information Retrieval to be used in decision support system for agricultural data science. The future work will involve modification of Word2Vec and FastText that were used to enable the integration with decision support systems and its knowledge based system. Given all efforts to collect the data, missing and incomplete datasets is a persistent problem across all disciplines. Gap filling Neural Networks will be addressed in future work to mechanise the prediction of missing and incomplete datasets in agriculture data science.

REFERENCES

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information.
arXiv preprint arXiv:1607.04606.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Holman Rector, L. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. Reference services review, 36(1), 7-22.

http://cropbase.org/cropbase/tools/AI/data_eye.html,.

Jahanshiri, Ebrahim, and Sue Walker. "Agricultural Knowledge-Based Systems at the Age of Semantic Technologies." International Journal of Knowledge Engineering-IACSIT 1 (2015): 64-67.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015, July). Support vector machines and word2vec for text classification with semantic features. In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on (pp. 136-140). IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mitara, B., Craswell, N. (2017). an Introduction to Neural Information retrieval. Foundations and Trends R in Information Retrieval.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. Proceedings of the National Academy of Sciences, 42(1), 43-47.

Salama, Ayman., Maul, T., Jahanshiri, Ebrahim. (2017). “Detecting cropping patterns of underutilized crops using online big data”, 2017 2nd International conference on “Future Technology Conference”. IEEE.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Shah, S. S., Bhattad, S., Lokegaonkar, S., & Ramakrishnan, G. Building Complementary Domain Taxonomies using Query Enrichment.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., & Despouy, P. (2016). plotly: Create interactive web graphics via Plotly’s JavaScript graphing library [Software].

Swoboda, T., Hemmje, M., Dascalu, M., & Trausan-Matu, S. (2016, September). Combining Taxonomies using Word2vec. In Proceedings of the 2016 ACM Symposium on Document Engineering (pp. 131-134). ACM.

Wohlgenannt, G., & Minic, F. (2016). Using word2vec to Build a Simple Ontology Learning System. In International Semantic Web Conference (Posters & Demos).

Zhang, Y., Rahman, M. M., Braylan, A., Dang, B., Chang, H. L., Kim, H., ... & McDonnell, T. (2016). Neural Information Retrieval: A Literature Review. arXiv preprint arXiv:1611.06792.