An Novel RNN Approach to Classification of Complex Textual Scientific Metadata

MACS 30200 Final Paper

Reid McIlroy-Young

June 4, 2017

Abstract

We

Contents

1 Introduction

2 Literature Review

Computer's have been a formal part of scientific work since the 18th Century (Grier, 2013), but the modern day electromechanical machines developed by Turing (Turing, 1937) and many others (Abbate, 2012)(Abbate, 2000) are a much more recent innovation, of the last century (Bauer and Rosenberg, 1972). The introduction of these devices to communities around the world (both metaphorically and literally) has had major impacts on the culture (Lessig, 2007), technology(Abbate, 2000) and rate of development (Bauer and Rosenberg, 1972). Much work has been done to study these effects, but it has been primarily focused on either the macro cultural effects (Pfaffenberger, 1988) or the economic/business usage (Landauer, 1995).

By comparison the usage of computers by scientists has been overlooked by researchers (Lab, 2017). This oversight has many reasons, but one of the most significant is the lack of available data. The primary methods for large scale analysis of the culture or structure of scientific work involve bibliometric techniques (De Bellis, 2009) using large standard datasets(e.g. Boyack et al., 2005; Börner, 2010, 2015; Sugimoto et al., 2013; Shi et al., 2015; Evans and Foster, 2011; Skupin et al., 2013). These dataset are generally lacking information about the computational aspects of the work, e.g. the Clarivate Analytics Web of Science (WOS) does not have any such field (McLevey and McIlroy-Young, 2016) and as such research into this dimension is difficult. Recent developments in natural language processing (NLP) have shown that complex concepts can be extracted reliably from text for a wide variety of tasks (Evans and Aceves, 2016), with some very similar to that done here (Foster et al., 2015).

2.1 Information Extraction

To extract the information about software usage from the available data requires complex NLP techniques and the best methodologies change quickly(Evans and Aceves, 2016). As we are primarily concerned with the classification of meta-data for a record relating it to a new software tool or not, in theory there are a large number of available techniques, as this is a simple binary classification problem (James et al., 2013)(Jurafsky, 2000)(Murphy, 2012). We have considered most of the available techniques:

- Classified based on a simple regular grammar, e.g. regex
- Word collocation frequencies (Manning et al., 1999)

- Term frequency–inverse document frequency vectors with an SVM or other classifier (Collobert et al., 2011)
- Word2Vec vectors with an SVM or other classifier(Mikolov et al., 2013)(Collobert et al., 2011)

The the current state of the art for natural language processing is the usage of deep neural networks for information extraction requiring more than simple word level similarities(Manning et al., 2014). As this is the state of the art there is no simple set of rules to follow, but there are some guidelines (Goodfellow et al., 2016). These have lead us to the use of a recurrent neural network (RNN) (Mikolov et al., 2010) for the classification, although the exact specifics have been determined with cross-validation techniques (James et al., 2013). The main features to consider are the type of regularization (Goodfellow et al., 2016), what representation of words to use (most likely Word2Vec (Mikolov et al., 2013)), what non-textual data will be included as there are in the WOS data set over 60 possible fields for each record (McLevey and McIlroy-Young, 2016) and what values the hyperparameters take(Goodfellow et al., 2016). This tuning is highly specific to the data, framework (in this case TensorFlow (Abadi et al., 2016)) and model and the parameters are provided in the supporting material.

2.2 Data Analysis

Once the records with new software tools have been identified, we can use the existing theory of bibliometrics to look at the network structure. The literature standard approaches are to look at the structure of these nodes in the citation and authorship graphs (de Solla Price, 2002)(Larivière et al., 2006)(Borgatti et al., 2009). This can be a computationally intensive task but tools exists that make it more practical (McLevey and McIlroy-Young, 2017) so once the records have been labelled the analysis techniques are no longer novel.

The literature is silent on basic features of scientific software usage, and even when limited to only new releases there is no existing data. Thus simple measures such as per domain counts/frequencies and basic graph measurements such as the centrality will be new contributions.

The other main question of what causes tools to be successful, has not been answered for scientists. There has been some work in the business domain (Xin and Levina, 2008)(Hsu et al., 2009). The adoption of new tools by businesses is theorized to follow a sigmoid pattern, with successful new entrants having three stages of usage: First they are used by early adopters and have small market penetration. Then they reach a "take off point" and the large majority of users will adopter their tools. Finally there will be slow growth in adoption again as

only the laggards are left as new users (Xin and Levina, 2008). This is based on adopters having a Gaussian distributed chance of adopting the tool and notably this diffusion model does not require that the software have any costs for the users and allows for network effects, thus this signature is considered in our modelling.

There also has been work done examined open source projects (Mockus et al., 2002) which agrees with the theory (Raymond, 1999) of open source that success is derived from openness and collaboration. This would predict that successful tools would come from highly connected groups who are working successfully with the community. This may show up as high connectedness in the co-authorship network correlating with success.

What leads to success has also be been studied in the context of ideas in the scientific literature (Acharya, 2004) (McLevey et al., 2016) or of individuals(Sinatra et al., 2016). In both cases the main measure of success is the cumulative count of citations, which we can also examine on a per paper and a per author basis. We can look for the predictors of success for a new software tool by examining its citations over time and us this as our measurement for the signature. Notably Sinatra et al. (2016) show a that success very unpredictable and can happen years after the paper is published. If the software records have patterns matching this model then the diffusion model may not be a good fit.

3 Data

The source of data used for this analysis is the Web of Science (WOS) database hosted by Knowledge Lab. It has metadata on almost all scientific publications from 1960 to 2015, with new records being more complete. Each publication can be linked to one or more other tables each which contain other metadata than the main table, the number of entries for each table I am concerned with are shown in Table 1 and the complete database schema in Figure 1. Access to the database is controlled by Knowledge Lab so they would need to be contacted to access it, once access rights are obtain the database is found at wos2.cvirc91pe37a.us-east-1.rds.amazonaws.com and the documentation at http://docs.cloudkotta.org/dataguide/wos.html.

The data for WOS were collected by Thompson Reuters until 2016, when it was given to Clarivate Analytics who now maintain it. The contemporary publications are collected from the publishers directly while older and more obscure publications are obtained from scanned copies digitalized with OCR, which is one of the factors that leads to newer publications having much higher quality data.

For this analysis I limited my data to those journals from the top 123 statistics publications between 2005 and 2016, giving me a total of 78 971 articles (publications). From these I derived a training set of classified (training) and unclassified

Table	Number of Entries
publications	57136685
abstracts	26093439
publishers	50668193
keywords	78155603
references	1085738245

Table 1: *Web of Science database number of entries per table*

publications. To do this I found journals that almost entirely publish new scientific software, thus all articles from these can be classified as containing new software, i.e. as positive. There are also some that contain virtually none, all of their publications can then be classified as negative. The classified journals are given in Table 2, note they are all top statistics journals from the data set.

Journal	Classification	Total Citations	Impact Factor
R JOURNAL	Mostly Software	271	1.045
STATA JOURNAL	Mostly Software	2636	1.292
JOURNAL OF STATISTICAL SOFTWARE	Mostly Software	6868	2.379
ECONOMETRICA	Little Software	24957	4.053
TECHNOMETRICS	Little Software	6062	1.435
STATISTICAL METHODS IN MEDICAL RESEARCH	Little Software	2703	4.634
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY	Little Software	2 360	1.702
BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY	Little Software	1 278	3.698
ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION	Little Software	74	3.045
ANNALS OF STATISTICS	Little Software	15 680	2.780
STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT	Little Software	2 297	2.237

Figure 2: *Classified journals, citations and impact factors are both from 2015*

When combined the I have a training set of 1251 positive and 4362 negative

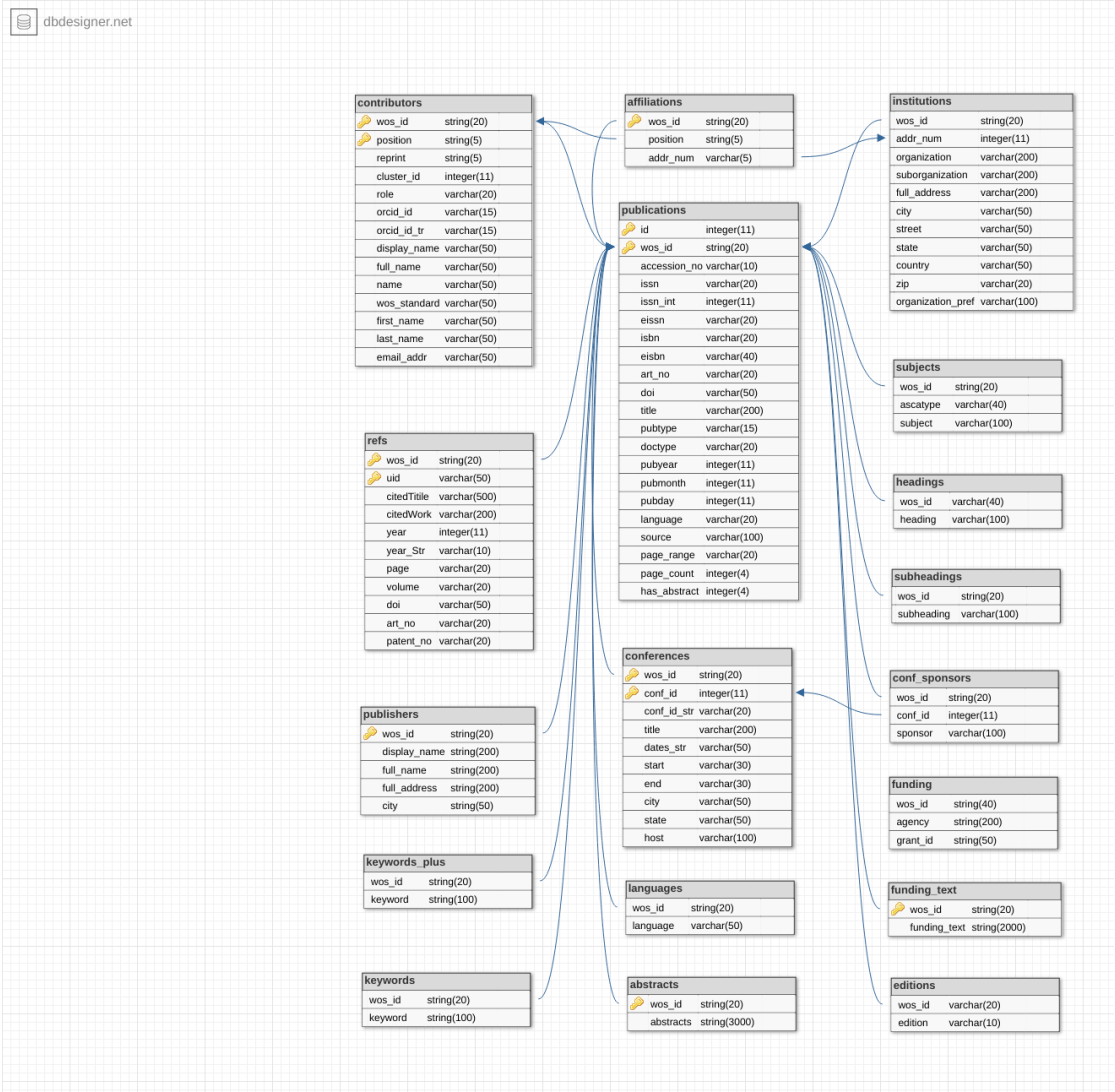

Figure 1: *Knowledge Lab WOS database schema*

examples. This is not a large data set nor is it pure since some articles from the positive set are in fact negative, one example(Wickham et al., 2014) is shown Figure 3. As there is no pre-existing know set of cleaning papers I cannot give an exact count of the incorrectly identified papers, but as I will discuss later the model is capable of identifying them despite their presence in the training set. The paper used here is one of the one identified by the fully trained model as being not software.

Field	Value
ID	WOS:000341806800001
Source	JOURNAL OF STATISTICAL SOFTWARE
Year of Publications	2014
Title	Tidy Data
Abstract	A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Figure 3: *An example of a false positive in the training set*

4 Methods

sdffsdds

5 Results

dfddfs

6 Conclusion

dsdfsdfs

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* .

Abbate, Janet. 2000. *Inventing the internet*. MIT press.

Abbate, Janet. 2012. *Recoding gender: women's changing participation in computing*. MIT Press.

Acharya, Amitav. 2004. "How ideas spread: Whose norms matter? Norm localization and institutional change in Asian regionalism." *International organization* 58:239–275.

Bauer, Walter F and Arthur M Rosenberg. 1972. "Software: historical perspectives and current trends." In *Proceedings of the December 5-7, 1972, fall joint computer conference, part II*, pp. 993–1007. ACM.

Borgatti, Stephen P, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. 2009. "Network analysis in the social sciences." *science* 323:892–895.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge: MIT Press.

Börner, Katy. 2015. *Atlas of Knowledge: Anyone Can Map*. Cambridge: MIT Press.

Boyack, Kevin, Richard Klavans, and Katy Börner. 2005. "Mapping the Backbone of Science." *Scientometrics* 64:351–374.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12:2493–2537.

De Bellis, Nicola. 2009. *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press.

de Solla Price, Derek J. 2002. "The pattern of bibliographic references indicates the nature of the scientific research front." *Social Networks: Critical Concepts in Sociology* 4:328.

Evans, James and Jacob Foster. 2011. "Metaknowledge." *Science* 331:721–725.

Evans, James A and Pedro Aceves. 2016. "Machine translation: mining text for social theory." *Annual Review of Sociology* 42:21–50.

Foster, Jacob G, Andrey Rzhetsky, and James A Evans. 2015. "Tradition and innovation in scientists' research strategies." *American Sociological Review* 80:875–908.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Grier, David Alan. 2013. *When computers were human*. Princeton University Press.

Hsu, Maxwell K, Stephen W Wang, and Kevin K Chiu. 2009. "Computer attitude, statistics anxiety and self-efficacy on statistical software adoption behavior: An empirical study of online MBA learners." *Computers in Human Behavior* 25:412–420.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*, volume 6. Springer.

Jurafsky, Daniel. 2000. "Speech and language processing: An introduction to natural language processing." *Computational linguistics, and speech recognition* .

Lab, Knowledge. 2017. "The Impact of Programming Languages and Datascience Frameworks on Thinking, Software, and Science." Technical report, The University of Chicago. Unpublished.

Landauer, Thomas K. 1995. *The trouble with computers: Usefulness, usability, and productivity*, volume 21. Taylor & Francis.

Larivière, Vincent, Yves Gingras, and Éric Archambault. 2006. "Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities." *Scientometrics* 68:519–533.

Lessig, Lawrence. 2007. *CODE VERSION 2.0*. codev2.cc.

Manning, Christopher D, Hinrich Schütze, et al. 1999. *Foundations of statistical natural language processing*, volume 999. MIT Press.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

McLevey, John, Alexander Graham, Reid McIlroy-Young, Pierson Browne, and Kathryn S. Plaisance. 2016. "Knowledge diffusion and status boundaries: A statistical network analysis of the relationships between philosophy of science and the sciences." Sunbelt XXXVI (Annual meetings of the International Network for Social Network Analysis). Close to publication.

McLevey, John and Reid McIlroy-Young. 2016. "metaknowledge documentation." `http://networkslab.org/metaknowledge/documentation/metaknowledgeFull.html#WOSRecord`.

McLevey, John and Reid McIlroy-Young. 2017. "Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science." *Journal of Informetrics* 11:176–197.

Mikolov, Tomas, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. "Recurrent neural network based language model." In *Interspeech*, volume 2, p. 3.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111–3119.

Mockus, Audris, Roy T Fielding, and James D Herbsleb. 2002. "Two case studies of open source software development: Apache and Mozilla." *ACM Transactions on Software Engineering and Methodology (TOSEM)* 11:309–346.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Pfaffenberger, Bryan. 1988. "The social meaning of the personal computer: Or, why the personal computer revolution was no revolution." *Anthropological Quarterly* pp. 39–47.

Raymond, Eric. 1999. "The cathedral and the bazaar." *Philosophy & Technology* 12:23.

Shi, Feng, Jacob Foster, and James Evans. 2015. "Weaving the fabric of science: Dynamic network models of science's unfolding structure." *Social Networks* 43:73–85.

Sinatra, Roberta, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. "Quantifying the evolution of individual scientific impact." *Science* 354:aaf5239.

Skupin, André, Joseph Biberstine, and Katy Börner. 2013. "Visualizing the topical structure of the medical sciences: a self-organizing map approach." *PloS one* 8:e58779.

Sugimoto, Cassidy, Vincent Lariviere, Chaoqun Ni, Yves Gingras, and Blaise Cronin. 2013. "Global gender disparities in science." *Nature* 504:211–213.

Turing, Alan Mathison. 1937. "On computable numbers, with an application to the Entscheidungsproblem." *Proceedings of the London mathematical society* 2:230–265.

Wickham, Hadley et al. 2014. "Tidy data." *Journal of Statistical Software* 59:1–23.

Xin, Mingdi and Natalia Levina. 2008. "Software-as-a-service model: Elaborating client-side adoption factors." .