Formation of Knowledge and Invention of Skills

Bhargav Srinivasa Desikan

April 2019

1 Introduction

How are modern skills and knowledge forged for the 21st century data economy? Is the primary crucible the academic arena of skill discovery, the industrial domain of skill invention, or the educational space of skill translation? Academic reflection and strategic or accidental investigation yield insights and the most successful (e.g., topic modeling Blei and Lafferty [2009], Western blot analysis Mahmood and Yang [2012]) harden into theorized protocols and skills deemed valuable in the marketplace. Alternatively, industry experimentation and refinement yield patterns that generate value and demand greater theoretical insight (e.g., neural word embeddings, plant genetic transformation). We also see that the act of translating and transmitting skills through educational instruction helps to encapsulate and crystallize them. Moreover, how do academic, industrial, and educational skills and knowledge respond to each other over time and space?

With access to large-scale geo-tagged *textual* data on science and scholarship, course syllabi, and job advertisements, it is possible to begin to answer these questions.

2 Data

The Web of Science Reuters [2012] dataset which includes metadata of research articles published from the year 1900, a syllabus data set which includes details of syllabi of courses being taught at universities worldwide, and

details of job postings across all major American cities. I will be receiving all three of these datasets through the Knowledge Lab at University of Chicago.

The data provides a way to ask large questions about how different kinds of institutions conduct research and create knowledge and how they are influenced by market demands and large changes in trends of research nationwide or worldwide. Because the data also includes locational information, it means that we have two simultaneous spaces to explore - a complex (vector) space that will explain the nature of the research, and the physical space in which those skills and knowledge reside. An additional layer of complexity is added when considering the temporal nature of the dataset, and allows us to also see changes in these trends over time, and look for plausible and ultimately causal explanations for the changes in these trends over time.

3 Methods

Previous research on linking invention done in cities to city size and growth (Bettencourt et al. [2007a], Bettencourt et al. [2007b]) has shown one way to measure knowledge growth - by using patent data. Evans [2010] has also discussed how academia and industry interact with a case study (of the Arabidopsis thaliana plant), but did not employ a large scale data analysis. While previous research has used counts of patents and networks of patents, the contents of the patents (i.e, the text itself) was not used to identify the *nature* of the invention or innovation. A combination of both the actual text and the metaknowledge contained in these articles will be crucial

I plan to use machine learning techniques applied to text to understand how different universities, cities and companies produce knowledge, how this relates to what is currently being taught at universities, and how both of these may be related to the demand for skilled workers by companies. In particular, I will use Topic Modelling (Blei and Lafferty [2009]), and Word2Vec Mikolov et al. [2013]) to represent the words as vectors to compare between the various cities and organisations.

4 Progress and Results

I have already started the data cleaning process, and have consolidated all the 3 data sources and cleaned them and set it up for vectorisation.

I have piloted the study for the state of Texas. By creating vectors for cities (e.g Austin, Houston), and for organisations (e.g Texas AM University, UT Austin), I can now find the angle between these vectors and see how different kinds of cities and organisations relate to each other. The vectors for cities or organisations are created by finding the centroid of all the article kind (paper, syllabus) of that particular city or organisation.

References

Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007a.

Luis MA Bettencourt, Jose Lobo, and Deborah Strumsky. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research policy*, 36(1):107–120, 2007b.

David M Blei and John D Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

James A Evans. Industry induces academic science to know less about more. *American Journal of Sociology*, 116(2):389–452, 2010.

Tahrin Mahmood and Ping-Chang Yang. Western blot: technique, theory, and trouble shooting. *North American journal of medical sciences*, 4(9): 429, 2012.

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

Thomson Reuters. Web of science. 2012.