Formation of Knowledge - Data and Methods

Bhargav Srinivasa Desikan

May 2019

1 Introduction

The project involves understanding the research output and teaching/course content of universities and cities by vectorizing the documents which can provide this information. To be able to do this, I must first collect and organise the data, set it up to be vectorized, and then run further analyses on the data using the vectors created. I will be describing the data organising and cleaning, the methods used to vectorize, and preliminary results of the analysis.

2 Data

I will be using two primary data sources - one for modelling the research output of cities and organisations, and the other for modelling the teaching output of cities and organisations. Here, an organisation can refer to a University or an independent research organisation which has scientists or researchers publishing papers, or government organisations (such as NASA).

2.1 Research - WoS Dataset

The Web of Science Reuters [2012] dataset includes metadata of research articles published across the world. Some quick facts about the dataset:

- The database contains publications from 1900 to 2015.

- The publications table contains 57M records.

- The references table contains 1.08B records.

While there are a couple of ways to access the dataset online, the massive size makes it very difficult to easily mine. I will be accessing the dataset through the Knowledge Lab at the University of Chicago, where the dataset, in both raw XML and in a SQL database form, is stored on Cloud Kotta, an online platform for managing data.

Accessing the data through Cloud Kotta requires me to

2.2 Teaching - Syllabus Dataset

Open Syllabus Project Syllabus research

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 Results

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.

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.