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 57 Million records.

- The references table contains 1.08 Billon 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 create an instance of Cloud Kotta, which I do through Turing Compute, also set up through the Knowledge Lab.

Once I have access, I am able to use SQL queries to access the dataset through the Bastion interface. The massive dataset size means that it is wiser to start accessing the data year by year before joining it, or by accessing a smaller sub-set of the dataset.

The most important parts of the WoS dataset for the purpose of my research are the abstract, the city, and the organisation affiliated with the research paper. The abstract allows me to aggregate research papers for each city and for each organisation.

Using the PyMySL package, I use a python script to exract the WoS ID, the abstract, the year the paper was published, the state and city it was published in, and the organisations affiliated with the authors of the paper.

For the purpose of my pilot study and for the results of this assignment, I have extracted papers published between 2008-2016 in the state of Texas in the United States of America. The final study involves the entirety of the United States of America.

2.2 Teaching - Syllabus Dataset

A similar approach is taken for the syllabus data for cities and organisations. The data has been collected and organised by the Open Syllabus Project, undertaken by researchers at Columbia University. It is an ongoing project, and has received a fair amount of attention from the press (Nature, Digital Science), and proves to be a promising way to measure what is being taught at universities across America (and the world). While 3rd party access to the dataset is not publicily available, through the Knowledge Lab I have access to the JSON dump of the open syllabus project, until 2018.

Using a python script I have extracted the contents of the syllabus, the year the course was taught, the state, the city and the organisation affiliated

with the course. The data itself is patchy, and for the pilot study, to compare the research and the syllabus output of cities and organisations, I have extracted syllabi from the state of Texas, from 2008-2016.

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. We have taken this further in two different ways. The work done only measured the count of the papers published and the authors collaborating, but not the *nature* of the patents or the contents of the patents. We move from patents to research publications and courses taught, and also measure the contents of the research and syllabi. To represent the content of the documents we turn to word embedding methods. Word2Vec (Mikolov et al. [2013]) and Doc2Vec (le2014distributed) are two popular methods to represent words and collections of words in a lower dimensional space, and for the primary study I use gensim's (Řehřek and Sojka [2011]) Doc2Vec implementation. Each of the papers and syllabi contained in the datasets were vectorized.

To generate each cities and each organisations research and syllabi vectors, I calculated the mean of each city/organisations vectors.

4 Results

With each dataset having their cities and organisations tagged with their 300 dimensional vectors, it is possible to calculate the similarities between the research and teaching vectors of each city and each organisation.

A way to calculate this relationship can be by measuring the angle between vectors - for our primary analysis we use cosine angle between the vectors. A smaller angle means the vectors are more similar, while a larger angle suggests more dissimilar vectors.

Using this framework, it allows us to do an analysis of each organisations relationship with another organisations research and syllabi vectors. An example is:

Angle Between	UT Austin Research	UT Austin Teaching
UT Dallas Research	17.65	73.07
UT Dallas Teaching	73.44	14.42

Angle Between	UT Austin Research	UT Dallas Teaching
UT Dallas Research	17.65	72.77
UT Austin Teaching	73.30	14.42

For the University of Texas at Austin and the University of Texas at Dallas, we see that teaching/syllabus vectors between them appear to be quite similar (14.42 degrees), research vectors a little less so (17.65 degrees), and that research and teaching vectors are quite dissimilar across universities.

5 Further Research

The bulk of the work so far was in setting up the research and syllabi vectors for universities: now we dive into the analysis. Some questions which I am now in the process of answering are:

- Similarity within universities and between universities for research and teaching vectors
- Change in vector directions over years
- Change in vector directions for different tier universities

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.

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.

Radim Řehřek and Petr Sojka. Gensim—statistical semantics in python. *statistical semantics; gensim; Python; LDA; SVD*, 2011.

Thomson Reuters. Web of science. 2012.