Vectors of Research and Teaching Across Universities in the USA

Bhargav Srinivasa Desikan

June 2019

Contents

Abstract

Traditional research in analysing relationships between research and teaching conducted in universities largely use qualitative methods and only measure individual outcomes of those conducting the research and teaching. I propose using a big data method which involves aggregating documents associated with research and teaching and vectorising them. This allows us to create research and teaching vectors for universities across the USA, which gives us a flixible framework for further analysis. I use this framework to demonstrate that teaching across universities is more similar than research across universities.

1 Introduction

The relationship between research and teaching in universities has long been questioned, either from a policy perspective (uz Zaman [2004]), or in the

field education research (Hattie and Marsh [1996], Hattie and Marsh [2004]). The consensus has largely been that research and teaching in universities are *not* intimately related, at least in terms of performance of the professors and teachers - success in research does not necessarily indicate success in teaching, or vice versa.

However, most of the questions have largely been focused on the researchers themselves and less about the *kind* of research being done, at the level of universities or cities. A qualitative enquiry into this is incredibly difficult, needing one to closely look at ways research and teaching across departments at universities are being conducted.

Access to large datasets of research and teaching allows us to begin to start asking more complex, larger questions: how similar are research and teaching *within* universities? And how similar research and teaching *between* universities? Apart from the relationship between research and teaching, it is also possible to probe the relationships of research between universities, as well as the relationships of teaching between universities.

To be able to do this, we need to be able to organise these datasets to be *able* to answer these questions, as well as a framework to measure similarities of research or teaching. To do this, I gather large amounts of textual data on research *output* and teaching *intent*, and vectorize them using word embedding techniques.

The data section discusses how I have prepared and organised the data to be able to vectorise them. This was a crucial part of the analysis and project,

and allows us to use word embedding methods which can later be used for comparative analysis, which is described in the methods section. The results section describes the information which these vectors can give us about the research and teaching across universities in the USA.

2 Theory

Previous work has largely been to address the relationship between the *performance* of professors teaching and conducting research. The metrics to measure this has been teaching reviews, and research output, while some studies measure the time spent on preparing for teaching and time spent on research. The papers by Hattie and Marsh (Marsh and Hattie [2002], Hattie and Marsh [2004], Hattie and Marsh [1996]) largely find no correlation between research and teaching, and other studies (Ramsden and Moses [1992], uz Zaman [2004]) conducted across Australian and British universities confirm this. Qualitative studies (Smeby [1998]) suggest that while professors themselves might believe research and teaching are intimately related, performance metrics suggest they are not.

All of this literature suggests that the question of the links between research and teaching are an important question in education research and science of science, but the methods used to probe it, as well as the *intention* of the research are not about patterns of research and teaching across universities. The few which might touch upon the topic use highly qualitative

frameworks - the Schimank and Winnes [2000] paper, for example, uses what they describe as a pre-Humboldtian, Humboldtian and post-Humboldtian distinction, where a Humboldtian approach was to being research into universities without displacing teaching. It is largely policy related, and again, doesn't begin to describe the *kinds* of research or teaching being done. Indeed, one of the first points noted in the paper is that the lack of quality data makes that kind of probe very difficult to do.

Now, however, we are beginning to have access to the kind of data previous researchers could have found invaluable. I use this opportunity to begin to probe a new, quantitative way to measure *general* trends and relationships among research and teaching, instead of measuring personal outcomes of research and teaching performance.

3 Data

I use two separate data sources to conduct my analysis - one to identify research output, and the other to identify teaching intent.

3.1 Research Output

To measure the research outputs of universities, I use the widely popular Web of Science dataset Reuters [2012].

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 extract 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.

The dataset contains articles published all across the world; for the pur-

pose of my study I only extract papers which have been published after the year 2006 and in the United States of America. It should also be noted that because each abstract is linked to both a university and a city, I am able to create both university and city vectors of research. The data I have extracted from the dataset is a total of 29 GB.

3.2 Teaching Intent

A similar approach is taken for teaching intent - I use 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 publicly 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 current study, to compare the research and the syllabus output of cities and organisations, I have extracted syllabi of universities in the United States of America, from the year 2006 onward.

Similar to the way I have aggregated data from the WoS dataset, syllabi are associated to both cities and universities.

3.3 Matching Data

The WoS database and the open syllabi project are quite different in the level of pre-processing, in IDs, and in the way universities and cities are made. To make sure I am correctly matching universities across the two databases, I have written a script which regularises the strings across both the data-sets to a common representation, and use string similarity to match the names. There are a total of 1782 universities which perfectly match across both the databases, and I will be using these universities to conduct my analysis.

4 Methods and Models

Now that I have textual data organised by city and universities, it is possible to start thinking of a framework which will allow us to explore similarities between research and teaching within/between universities.

Textual similarity is a problem with many solutions, ranging from a bag of words, TF-IDF or jaccard similarity match, but these prove to be inefficient when dealing with larger documents. Also, since we are interested in aggregating the texts for each city, merely appending all texts associated with an institution or city would not be the best idea.

Here, vectorization of documents start to become a good idea. By vectorizing documents we can play around them more easily. Aggregating all the documents in an entity (an university, research organisation or city) could become merely adding all the document vectors associated with an entity,

and we can also define the vector associated with such an entity as the mean of all the document vectors of that entity.

We can create these vectors associated with the entities by using word embedding methods.

4.1 Creating Entity Vectors

The module developed allows us to use any kind of text vecrtorisation method, and any kind of aggregation method. Since we are interested not in just the count of words (using bag of words, TF-IDF), or in topics (topic model vectors of documents Blei and Lafferty [2009]), a word embedding method such as Word2Vec (Mikolov et al. [2013]) or Doc2Vec (Le and Mikolov [2014])which use semantic information will be very useful.

Word embedding algorithms use large collections of text and output a n-dimensional vector-space model in which each unique word is represented as a vector in a vector space. By using the location of a word in a document relative to the other words in the document, it is possible for us to capture the context - words are usually given meaning by the company they keep, and this method soon becomes an efficient way to capture meaning in text. Research has shown that words frequently share linguistic contexts, and thus those words which are located nearby in the vector space tend to share similar meanings.

The first step in creating an entity vector would be to vectorise all the documents associated with that entity. In our case, entities are universities,

so we create our university vector by simply finding the mean of all the document vectors associated with that university. I run this algorithm for all the universities associated with the research and teaching datasets.

I have created 2877 university research vectors and 2122 university teaching vectors. 1782 Universities have both research and teaching vectors. Since it would be interesting to see Research-Research relationships, Teaching-Teaching relationships and Research-Teaching relationship, we will be using the 1782 universities with both research and teaching vectors, as we can have all of the above vectors available.

I use cosine similarity, and the subsequent angle between the vectors to understand the similarity between the vectors.

Again, it should be noted that any vectorisation method can be theoretically used, any vector similarity measure can be used.

5 Results

I will be using the angles between the research and teaching vectors to describe the relationship between them. This will form the crux of my current analysis, though it should be noted that the vectors themselves can be used for a variety of other analyses, some of which I will outline in the Future Work section.

Our research and teaching vectors allow us to calculate similarities/differences between the universities associated with those vectors. I use three different

vector relationships:

- Research-Research angles
- Teaching-Teaching angles
- Research-Teaching angles

The research-teaching angles can be further broken into between university angles (for different university pairs), and within university angles (for the same university).

For the purpose of illustrating the utility of this framework, I have used 3 different subsets of the universities dataset. Out of these, one set contains 10 randomly sampled universities, 9 randomly sampled medical or health universities, and 10 community colleges.

The 10 randomly sampled universities are John Hopkins University, Rutgers State University, University of Chicago, University of California at Los Angeles, University of Toledo, University of Notre Dame, Georgetown University, Harvard University, Iowa State University, and San Diego State University.

The 9 randomly sampled medical or health universities are Oregon Health and Science University, Medical College of Wisconsin, University of Arkansas Medical Center, University of Nebraska Medical Center, Eastern Virginia Medical School, New York Medical College, Western University Health Science, SUNY Upstate Medical University.

The 10 randomly sampled community colleges are Flathead Valley Community College, Quinsigamond Community College, Community College Rhode Island, Borough Manhattan Community College, Anne Arundel Community College, Tacoma Community College, Nassau Community College, Florida Community College, Kingsborough Community College, Roane State Community College.

5.1 Research-Research

The average angle between research vectors is **57.62**°and the variance is **263.04**°.

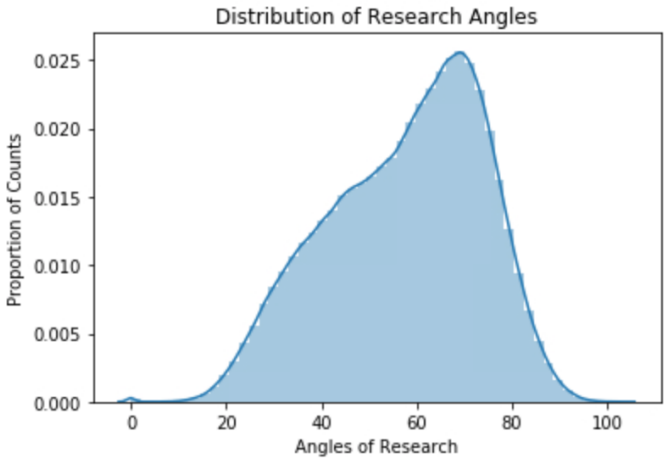

Figure 1: Distribution of research

This is very telling: it means that on average the differences in research between universities is fairly high. This is also expected behaviour, as different tiers and kinds of universities are likely to have very different kinds of

research output.

5.2 Angle Heatmaps of Research

I use heatmaps to illustrate how the distribution of angles between our three different samples of universities.

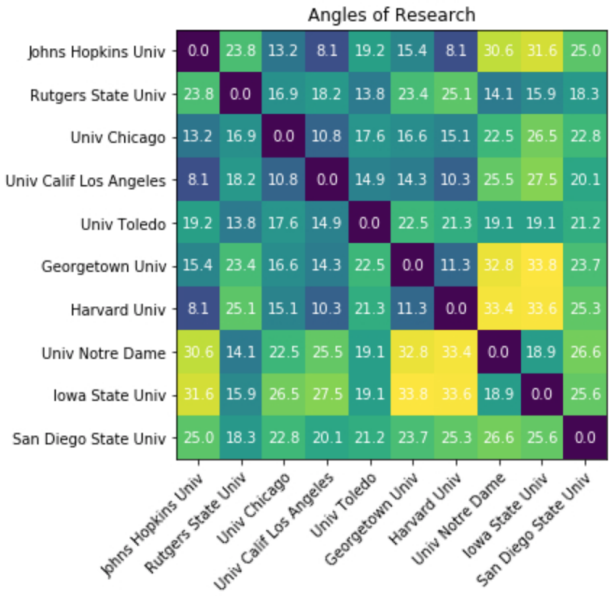

Figure 2: Angles between Universities

These heatmaps are very informative: we see that within universities, research focused universities such as Harvard and Chicago tend to have similar research outputs. In figure 5, where we compare community colleges and universities, we see much larger angles, indicating quite different research outputs, consistent with what we would imagine would be the difference in research between a research and teaching based institution.

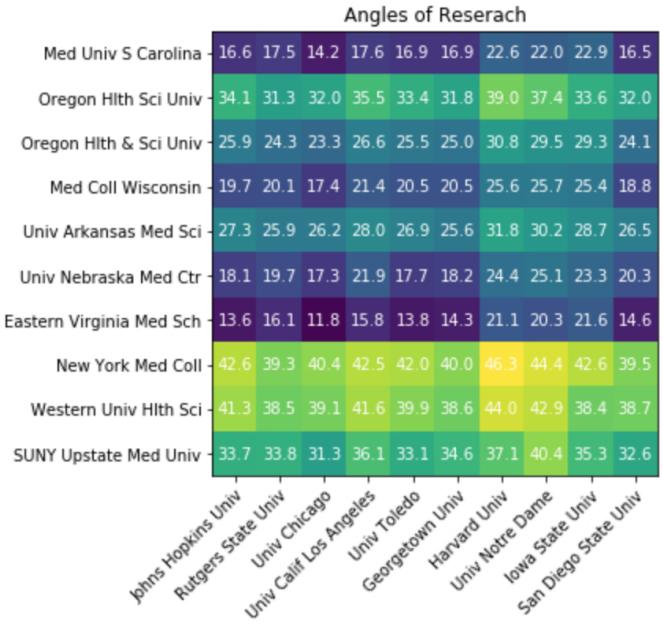

Figure 3: Angles between Universities and Medical Schools

It is interesting to note that there is a lot of variation in the research between different community colleges: it would be worth diving more into the nature of these vectors to identify the reason for this.

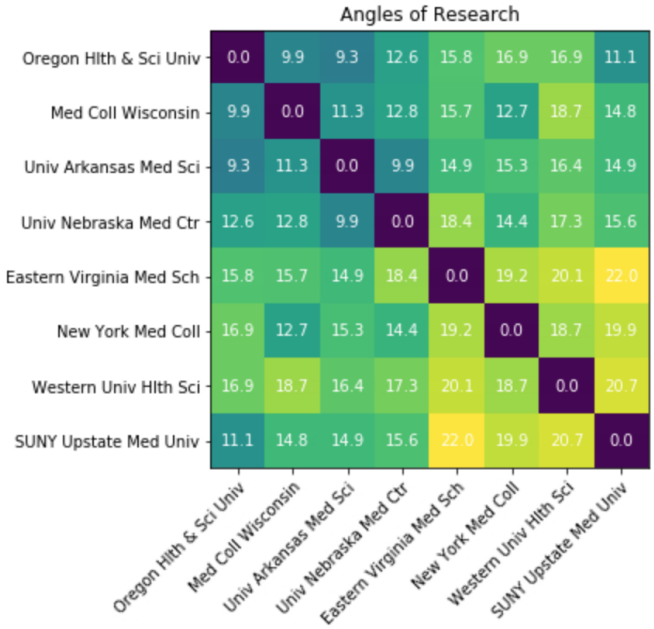

Figure 4: Angles between Medical Schools

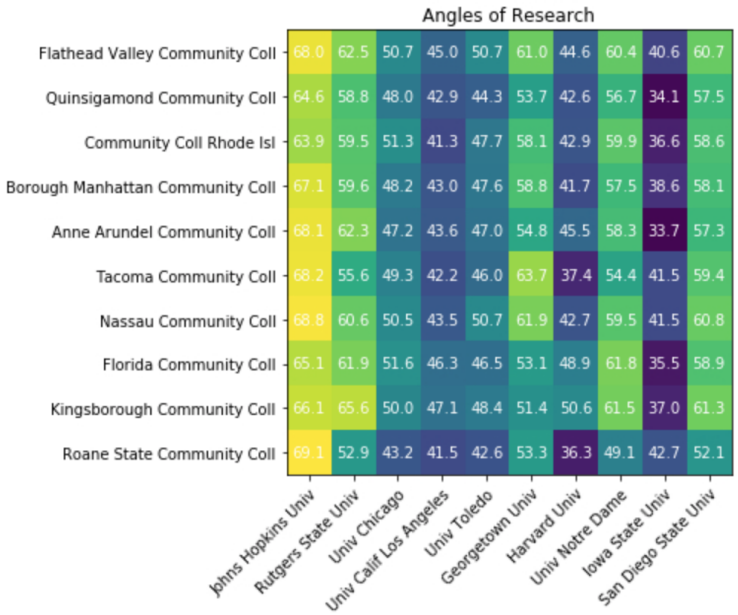

Figure 5: Angles between Universities and Community Colleges

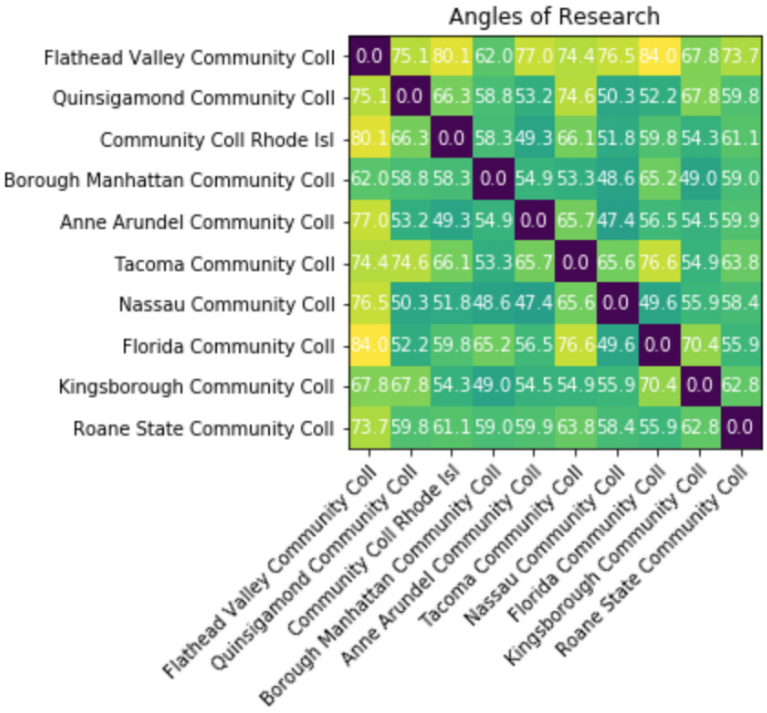

Figure 6: Angles between Community Colleges

5.3 Teaching-Teaching

I conduct a similar analysis for the teaching vectors of universities. The average angle between research vectors is **30.22**°and the variance is **106.31**°.

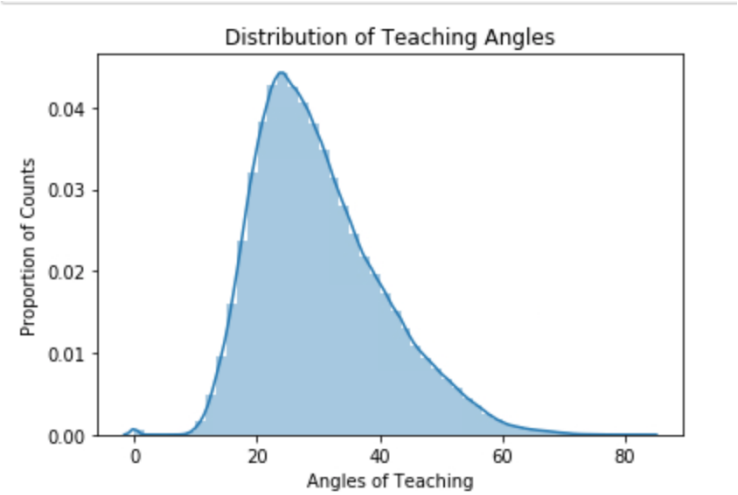

Figure 7: Distribution of teaching

This is again very informative: our distribution of angles has shifted by about 30 degrees to the left, suggesting that teaching across universities is a lot more similar than research. If we assume that most universities have similar introductory material, and would have similar syllabi, then this result is an important demonstration that teaching *intent* across universities in the USA is a lot less differential than research. Let us now look at heatmaps for teaching.

5.4 Angle Heatmaps of Teaching

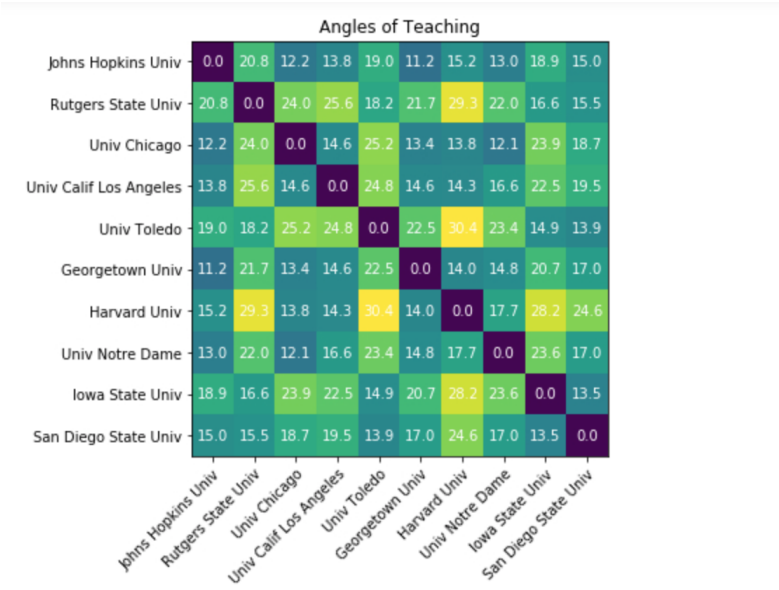

Figure 8: Angles between Universities

We can begin to see patterns of the way teaching may be conducted in universities. Teaching intent at Harvard is more similar to Univerity of Chicago, UCLA, John Hopkins, Georgetown as opposed to Iowa State University or San Diego State University. Angles of teaching at community colleges and medical colleges are less similar, suggesting that they might have different specialisations and may not be as general as universities, though we would need to do a more thorough analysis to confirm this hypothesis.

It should be noted that these heatmaps only represent a small subset of all the universities and kind of universities: there are many more particular questions which may be explored using this framework.

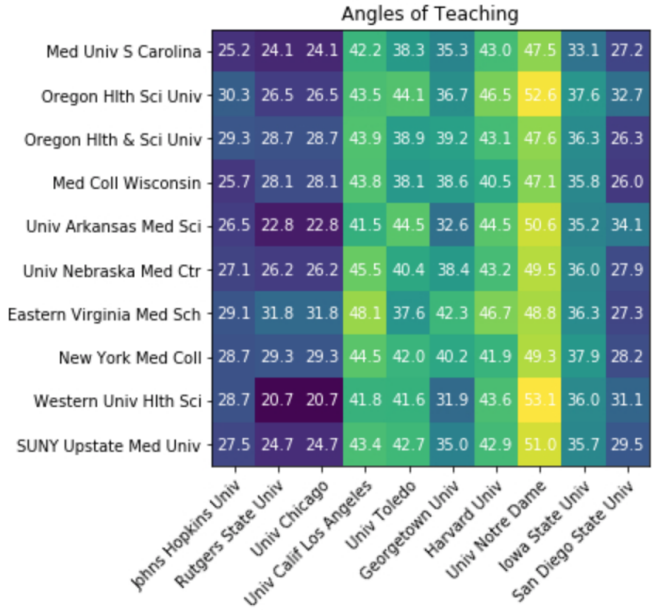

Figure 9: Angles between Universities and Medical Schools

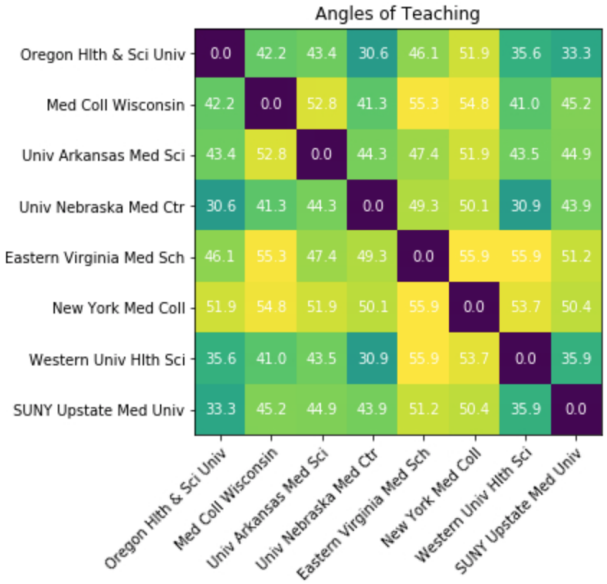

Figure 10: Angles between Medical Schools

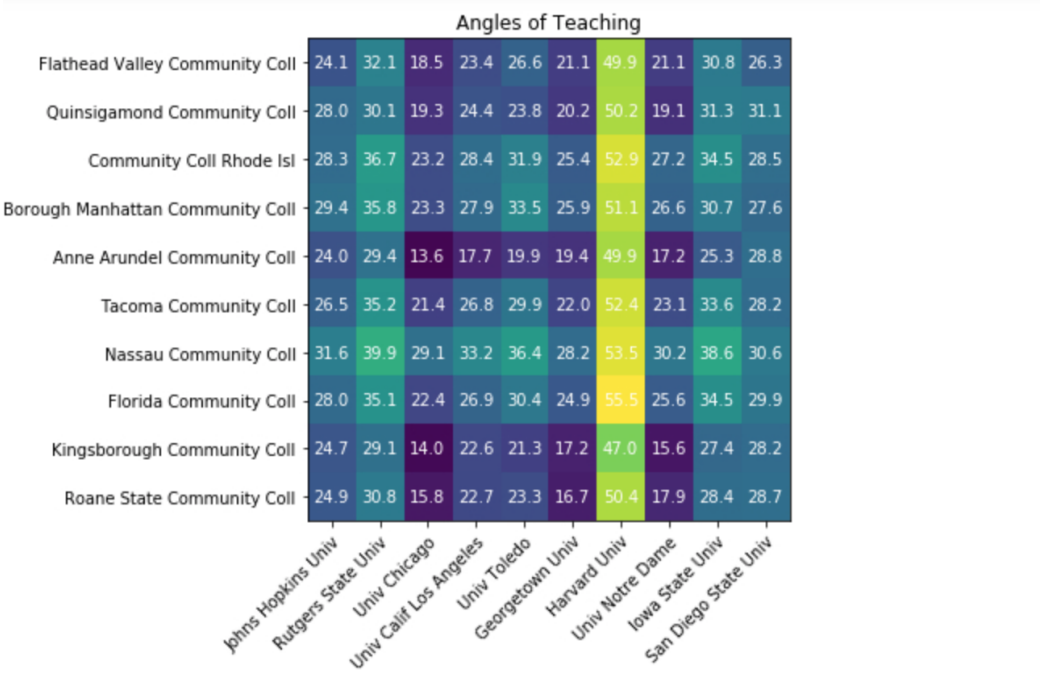

Figure 11: Angles between Universities and Community Colleges

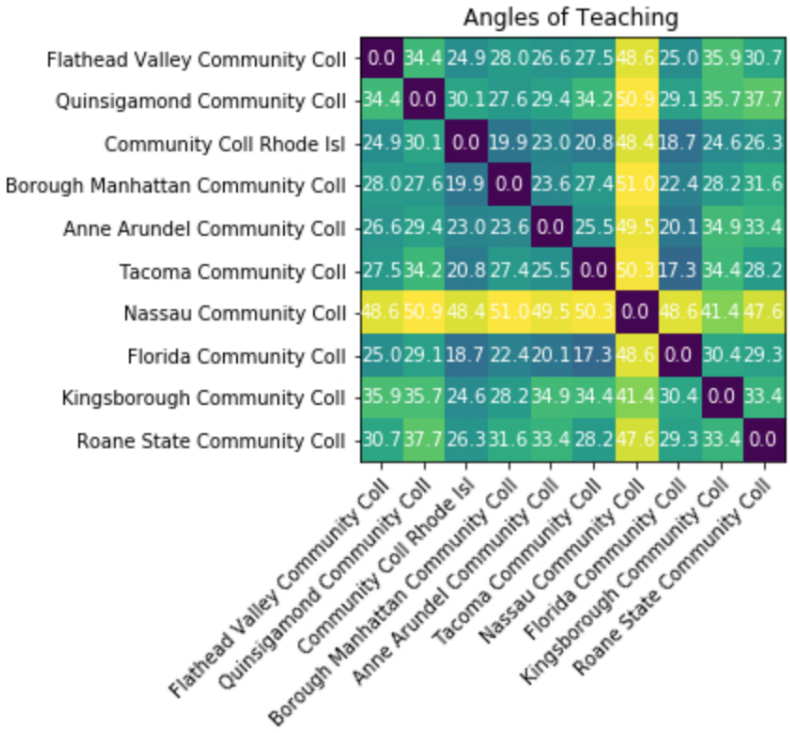

Figure 12: Angles between Community Colleges

5.5 Research-Teaching

We note that research-teaching vectors are quite different: because they are mined from very different sources, some of the angle difference may exist due to inconsistent cleaning and organising. The current average angle *between* universities is **72.27**°and the average angle *within* universities is **73.03**°. There is no consistent signal here, and we would need to dive down deeper into the nature of the vectors before making any assumptions, though at a glance we can conclude that research and teaching vectors across universities are quite different.

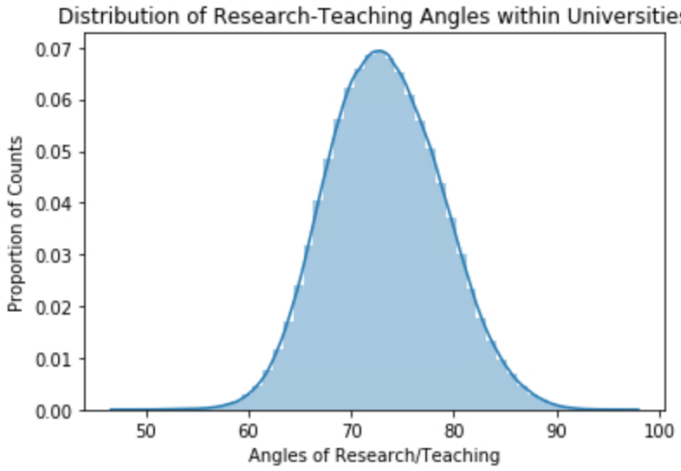

Figure 13: Distribution of research-teaching within universities

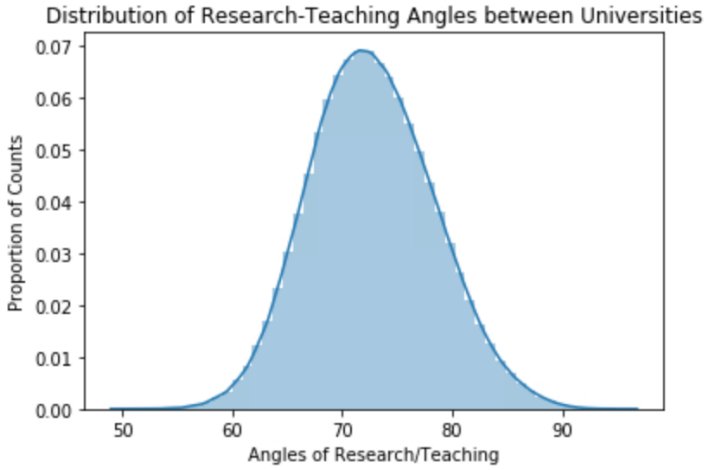

Figure 14: Distribution of research-teaching between universities

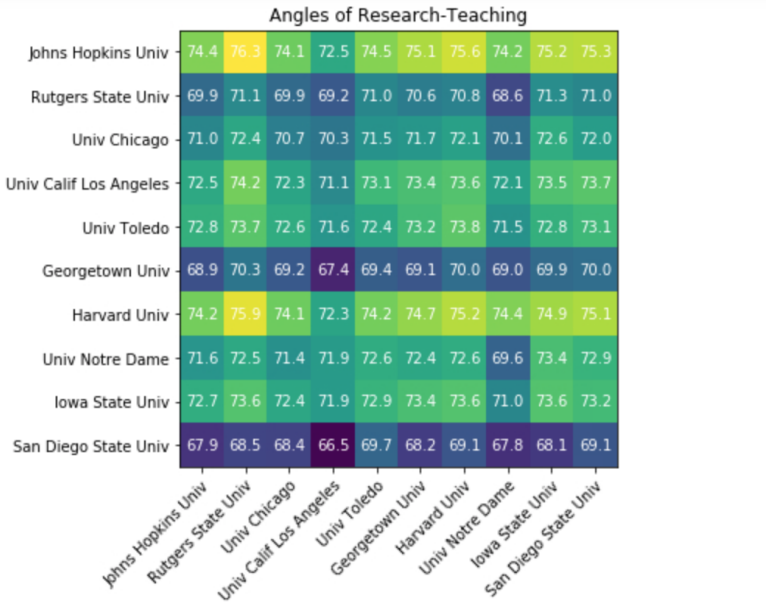

Figure 15: Heatmap of Research-Teaching

6 Future Work

We have seen how vectorising and aggregating research and teaching based documents can be a very useful way to quantitatively compare how universities approach them. The data is set up for us to analyse the same patterns across cities instead of universities, opening up room for interesting geographical analysis. Using patent data and job advertisements can allow us to create similar *invention* vectors (Bettencourt et al. [2007]), as well as job *demand* vectors. We can then measure how teaching intent may distribute itself to research and academia, and in which cases, with none of these happening, the teaching intent does not result in any perceivable output.

Using just the research and teaching vectors we can also identify how different tiers of universities or kinds of universities differ in their research and teaching. The preliminary analysis demonstrated in the heatmaps suggests that a detailed analysis into this would prove to have very interesting results.

We have also conducted a purely static analysis, and not used the rich time based data also available to us. Using the year of the publication or syllabus can indicate to us how these vectors may differ over time, and how increase of certain research trends (such as Machine Learning) influence research and teaching in universities. It would also be possible to compare which individual word vectors are most similar for a particular university or city.

Creating aggregated vectors for entities opens us to a whole world of spa-

tial and temporal analysis and would help us answer many complex questions previously untouched.

7 Conclusion

We have seen how creating a vectorised framework to measure the relationship between research and teaching leads to many useful results. We demonstrate using this framework that teaching across universities is more similar than research across universities. Different kinds of universities have very different vectors, allowing us to also use this framework to cluster universities based on their research and teaching. While this is a purely geographical analysis, it is trivial to add a time based modelling by incorporating vectors for different time slices. Using a big data approach to research and teaching relationships allows us to ask larger questions which we can generalise across states and countries.

References

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, 2007.

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

John Hattie and Herbert W Marsh. The relationship between research and teaching: A meta-analysis. *Review of educational research*, 66(4):507–542, 1996.

John Hattie and Herbert W Marsh. One journey to unravel the relationship between research and teaching. In *Research and teaching: Closing the divide? An International Colloquium*, pages 18–19, 2004.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

Herbert W Marsh and John Hattie. The relation between research productivity and teaching effectiveness: Complementary, antagonistic, or independent constructs? *The journal of higher education*, 73(5):603–641, 2002.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their composition-

ality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Paul Ramsden and Ingrid Moses. Associations between research and teaching in australian higher education. *Higher Education*, 23(3):273–295, 1992.

Thomson Reuters. Web of science. 2012.

Uwe Schimank and Markus Winnes. Beyond humboldt? the relationship between teaching and research in european university systems. *Science and public policy*, 27(6):397–408, 2000.

Jens-Christian Smeby. Knowledge production and knowledge transmission. the interaction between research and teaching at universities. *Teaching in higher education*, 3(1):5–20, 1998.

Mohammad Qamar uz Zaman. *Review of the academic evidence on the relationship between teaching and research in higher education*. Department for Education and Skills London, 2004.