

Defining Data Science

A Case Study in Australia

Xinrui WANG and Tsai-Chun TSOU

r Sys.Date()

Table of contents

Preface	4
1 Abstract	5
2 Introduction	6
3 Data Collection	7
3.1 Web Scraping	7
3.1.1 Inconsistent Information	8
3.1.2 Difficult to Manage Websites	8
3.1.3 Collected Data	9
 I Text Analysis	 11
4 Unit Text Analysis	13
4.1 Unit Code Analysis	13
4.2 Unit Overview and Learning Outcome - Bigram	16
5 Job Text Analysis	18
 II text2vec	 19
Algorithm and Model Fitting	20
Model Adjustments	21
6 Apply the Selected Model to Collected University Data	24
7 Apply the Selected Model to Employer Data	27
 III Conclusion	 28
8 Limitations	29
9 Conclusion	30

Preface

What is Data Science? What do you learn in Data Science? What kind of jobs do Data Scientists do? These are the type of questions we've encountered during social occasions and despite spending almost two years studying Business Analytics, we still struggle to give a definite answer. Data Science, Business Analysis, and Data Analysis are new degrees that are created only in recent years thanks to the advance in technology and increasing demand for talent to work with data. The sudden rise of these degrees gave authorities little time to regulate them. Even the Australian and New Zealand Standard Classification of Occupations (ANZSCO), the Australian Bureau of Statistics (ABS) and the Department of Home Affairs are slow to catch up with their [description](#) of this occupation.

After four months of gathering data, conducting analysis, and running models, we have attempted to define Data Science in Australia. Special thanks to Professor Di Cook and Professor Rob Hyndman for their support. A huge thank you to Dr. Emi Tanaka for the opportunity to work on such a special project and for her guidance throughout the whole thing¹.

¹This report is written jointly by Tsai-Chun Tsou and Xinrui Wang as part of the ETC5543 research project supervised by Dr. Emi Tanaka.

1 Abstract

With the increase in demand for data scientist and the rising popularity of the degrees, how does one define Data Science in Australia. The Australian Mathematical Sciences Institute (AMSI) and the Statistical Society of Australia (SSA) are conducting a review of Australian Data Science Degrees with [surveys](#) and focus groups. The survey method is prone to biases due to the subjective nature. Our research attempts to tackle similar topic but with a more objective approach by collecting data directly from public resources. We collected Data Science related unit information from Group of Eight Universities and [Data Scientist Job Listings](#) from kaggle. We used the data to decompose the core disciplines involved in the degree as well the type of skill sets that may be required. To expand on the initial exploratory data analysis, we also build Latent Dirichlet Allocation models to construct our own text corpus.

From the exploratory data analysis, we observe a lack of homogeneity within the Universities' analysis. The inconsistent data metrics made it difficult to draw direct comparison between the employer data and university data. Nonetheless we were able to conclude that computational disciplines are more prominent on both sides.

2 Introduction

Data Science has ranked as one of the most in-demand jobs in Australia in recent consecutive years. As demands steadily grows, students are also increasingly interested in Data Science degrees, yet recruiters still seem to struggle to fill up data science positions. This leads to our main question: what is Data Science? Is there a shared structure or skill set of Data Science courses offered at Australian universities? Are students and employers' perception of data science similar? To answer these questions, we looked a data from both University and Employer perspectives.

There is no readily available data from Australian universities, so we had to build our own data set through web scraping. The initial target was to collect data from all universities in Australia including both undergraduate and postgraduate courses, however, due to time constrain, the data collected for this project only contains Master of Data Science courses from the Group of Eight (Go8) universities. The employer data was retrieved from [Data Scientist Job Listings](#) on kaggle.

By exploring the current situation and potentially a definition of Data Science in Australia from both university and employer perspectives, the findings would help students and recruiters have a clearer picture of what to expect, as well as raising attentions and awareness to potential gaps between employer demands and university offerings.

3 Data Collection

3.1 Web Scraping

In order to explore the Data Science degrees around Australia Universities, we compiled a list of universities in Australia and the Data Science or related degrees they offered, then web scraped required information from each university's website using R. In total, we collected 298 units from eight postgraduate courses in Data Science across all Group of Eight (Go8) universities.

To start off the project, Professor Tanaka provided sample code for data scraping using Monash Handbook as an example. Libraries `rvest` and `rSelenium` are two of the main tools. Initially, we studied her code and then tried to replicate her code to be applied to other university's websites.

The flow of the data scraping is as follow:

1. Identify the main page (url) where the degree information is contained, which usually is the most updated version of the handbook.
2. Use functions from `rvest` to retrieve all the course unit code (or course unit url). Retrieve the degree code and formal degree name and save it for later.
3. Use `rSelenium` functions and course unit information, to direct R to the unit information page.
4. Retrieve the following information from the page using `rvest` functions:
 - Unit Code
 - Unit Name
 - Unit Overview
 - Unit Learning Outcome
 - Unit Prohibition/ Pre-requisite/ Co-requisite
5. Repeat step 3 & 4 until all the course information has been retrieved.
6. Compile all the retrieved data from the University into a dataframe and export it as csv file.

Despite the process being similar for each University, we soon realize the process was going to be more challenging than expect.

3.1.1 Inconsistent Information

Monash University's student handbook on Degrees and Courses is a spectacular website for data scraping. Its html code is clearly labeled and anything you need to know about the degree/course can be found on the website. The same cannot be said about other Universities.

The course descriptions on the handbook and universities' website page are usually structured in a different manner, since the majority of the data is collected from universities' handbooks, course descriptions are also extracted from handbooks for consistency purposes.

In addition, the required unit information listed above is not all available at the targeted universities. The handbook from University of New South Wales contains extremely limited information: unit overview is brief, unit prohibition/ pre-requisite/ co-requisite is only available for a few units, and unit outcome is not provided at all.

3.1.2 Difficult to Manage Websites

Each university website is unique. Sometimes the information is not so straightforward. An example of this is University of Adelaide. The main website for the degree does contain the list of units that go into the degree.

However, instead of having just one page with all the unit information, the link takes you to a page with different unit information depending on when the unit is offered and on what campus.

I tried bypassing the pages by directly looking at the url of the final unit information page I want to be on. Unfortunately, the url is not designed or structured in a way which I am able to predict the url based on my unit code. With that said, my only option was to code the function to jump from pages to pages before landing on the right unit information page.

It is also often found that the unit overview and learning outcomes for each unit within the same university could vary slightly in format. For example, unit overview may appears before or after campus location at University of Western Australia, empty spaces could be found after section title at the University of Melbourne, which would break the chain of extracting corresponding information.

3.1.3 Collected Data

The collected data contains **298 units** from 8 universities, and **8 variables** including School, Course, Course_code, Unit, Unit_code, Outcomes, Overview and Description. The full data set is made available below for exploration.

Show 10 of 28 entries					Search		Overview
Subject	Course	Course code	Unit	Unit code	Overview		
1	Research	Master of Data Science	CSDA	ETD102 - Introduction to databases	ETD102	ETD102	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
2	Research	Master of Data Science	CSDA	ETD106 - Researching databases in Python	ETD106	ETD106	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
3	Research	Master of Data Science	CSDA	ETD107 - Researching databases and networks	ETD107	ETD107	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
4	Research	Master of Data Science	CSDA	BAT1004 - Mathematical modelling for data science and AI	BAT1004	BAT1004	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
5	Research	Master of Data Science	CSDA	ETD105 - IT research methods	ETD105	ETD105	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
6	Research	Master of Data Science	CSDA	ETD106 - Researching databases	ETD106	ETD106	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
7	Research	Master of Data Science	CSDA	ETD107 - Data mining and networks	ETD107	ETD107	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
8	Research	Master of Data Science	CSDA	ETD106 - Data mining	ETD106	ETD106	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
9	Research	Master of Data Science	CSDA	ETD107 - Data mining	ETD107	ETD107	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>
10	Research	Master of Data Science	CSDA	ETD108 - Researching data science	ETD108	ETD108	<p>This unit will introduce the student to the concepts of databases and data management. It will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p> <p>The unit will cover the basic concepts of databases, data management, and data security. The unit will also cover the basic concepts of data mining and data analysis.</p>

Showing 1 to 10 of 28 entries

Part I

Text Analysis

To find out what is data science from universities' and employers' perspectives, whether there is a shared structure or common skill set of Master of Data Science degrees offered at Go8, what employers expect from a data scientist in the workplace, an exploratory data analysis, in particular text analysis has been conducted.

For universities, the main purpose is to identify shared skills or concepts offered by Master of Data Science degrees through exploring faculty of the units, detailed teaching contents etc. Whereas for employer data, the focus is to extract information regarding skills and programming languages in demand.

4 Unit Text Analysis

4.1 Unit Code Analysis

To explore the teaching contents of Master of Data Science at Go8, an analysis based on faculty of units offered is conducted to see what components are included in this degree.

Unfortunately faculty information is not directly available on the unit handbooks, in this case, unit code is taken as a surrogate identification. As shown in the sample data below, unit code is a combination of letters and numbers, the first few characters such as FIT, MAT, usually represents the faculty this unit belongs to, we could then make relatively educated assumptions on the content of the unit.

The grouping was performed manually using the code listed below. It is a subjective choice made under careful considerations, we are aware that the grouping is not 100% accurate. For example, the code 'DATA' from University of Sydney is classified under IT, however, some of the units start with DATA are actually more related with statistics, which means 'DATA' belongs to multiple departments. Although there would be misclassified units, the results could still provide a meaningful guidance regarding the teaching components of Master of Data Science at Go8 universities.

```
math <- c("STAT", "MATH", "MATHS", "STATS", "MAT", "MAST", "ACTL", "QBUS")
it <- c("COMP", "FIT", "CITS", "INFS", "COSC", "CSSE", "CSYS", "EDPC", "INMT", "PHIL", "PHI")
commerce <- c("ECON", "FINS", "MARK", "ACCT", "FINM", "MGMT", "MKTG")
spatial <- c("GEOM", "ITLS")
science <- c("EDUC", "SCIE", "SOCR")
health <- c("BINF", "BMS", "HTIN", "PUBH")
```

It is clear from Figure 4.1 that IT and Stat/Math are the two dominating components in the Master of Data Science degrees at Go8. Most units (165 out of 298) fall under the IT faculty, followed by Math and Stats, which has 79 units.

Similar findings could be observed from some but not all Go8 universities. Figure 4.2 shows the faculty breakdown by university. Since the total number of units offered by each university is different, instead of showing the actual number, proportions are plotted to make better comparisons across universities.

School	Course	Unit
monash	Master of Data Science	FIT9132 - Introduction to databases
monash	Master of Data Science	FIT9136 - Algorithms and programming
monash	Master of Data Science	FIT9137 - Introduction to computer arch
monash	Master of Data Science	MAT9004 - Mathematical foundations fo
monash	Master of Data Science	FIT5125 - IT research methods
monash	Master of Data Science	FIT5145 - Introduction to data science
monash	Master of Data Science	FIT5147 - Data exploration and visualis
monash	Master of Data Science	FIT5196 - Data wrangling
monash	Master of Data Science	FIT5197 - Statistical data modelling
monash	Master of Data Science	FIT5149 - Applied data analysis
monash	Master of Data Science	FIT5201 - Machine learning
monash	Master of Data Science	FIT5202 - Data processing for big data
monash	Master of Data Science	FIT5205 - Data in society
monash	Master of Data Science	FIT5212 - Data analysis for semi-structu
monash	Master of Data Science	FIT5230 - Malicious AI
monash	Master of Data Science	BMS5021 - Introduction to Bioinformati
monash	Master of Data Science	BMS5022 - Advanced bioinformatics: eff
monash	Master of Data Science	FIT5126 - Masters thesis part 1
monash	Master of Data Science	FIT5127 - Masters thesis part 2
monash	Master of Data Science	FIT5228 - Masters thesis part 3
monash	Master of Data Science	FIT5229 - Masters thesis final
monash	Master of Data Science	FIT5120 - Industry experience studio pr
monash	Master of Data Science	FIT5122 - Professional practice
unimelb	Master of Data Science	Methods of Mathematical Statistics
unimelb	Master of Data Science	A First Course In Statistical Learning
unimelb	Master of Data Science	Programming and Software Developmen
unimelb	Master of Data Science	Algorithms and Complexity
unimelb	Master of Data Science	Elements of Data Processing
unimelb	Master of Data Science	Database Systems & Information Model
unimelb	Master of Data Science	Statistical Modelling for Data Science
unimelb	Master of Data Science	Multivariate Statistics for Data Science
unimelb	Master of Data Science	Computational Statistics & Data Science
unimelb	Master of Data Science	Cluster and Cloud Computing
unimelb	Master of Data Science	Advanced Database Systems
unimelb	Master of Data Science	Statistical Machine Learning
unimelb	Master of Data Science	Data Science Project Pt1
unimelb	Master of Data Science	Data Science Project Pt2
unimelb	Master of Data Science	Data Science Research Project Pt1
unimelb	Master of Data Science	Data Science Research Project Pt2
unimelb	Master of Data Science	Foundations of Spatial Information
unimelb	Master of Data Science	Spatial Databases
unimelb	Master of Data Science	Spatial Analysis
unimelb	Master of Data Science	Information Visualisation
unimelb	Master of Data Science	Analysis of High-Dimensional Data
unimelb	Master of Data Science	Advanced Statistical Modelling
unimelb	Master of Data Science	Mathematics of Risk
unimelb	Master of Data Science	Optimisation for Industry
unimelb	Master of Data Science	Practice of Statistics & Data Science
unimelb	Master of Data Science	Stochastic Calculus with Applications
unimelb	Master of Data Science	Advanced Probability
unimelb	Master of Data Science	Random Processes

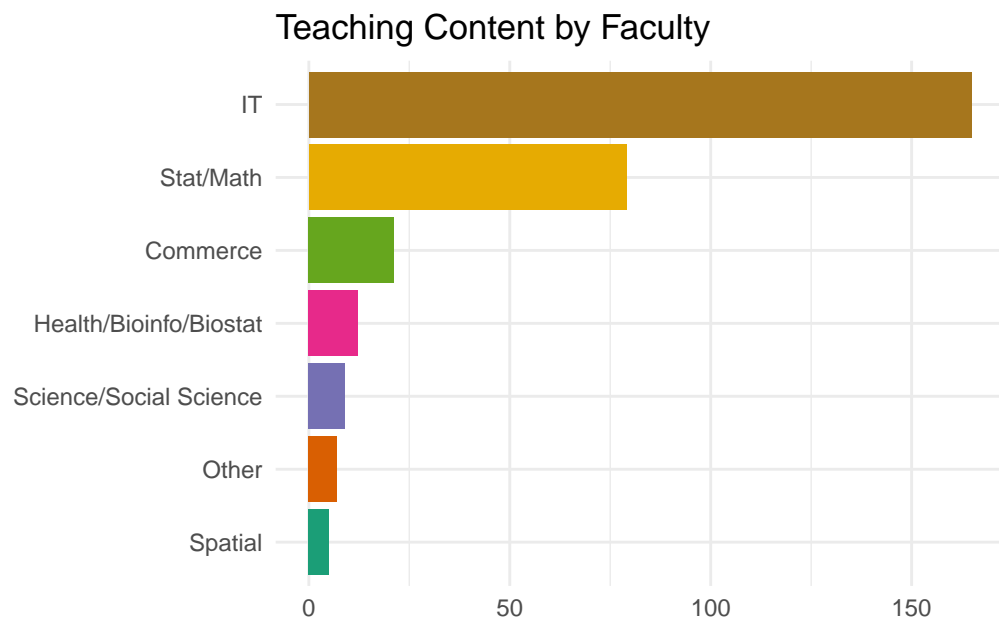


Figure 4.1: Teaching Content by Faculty

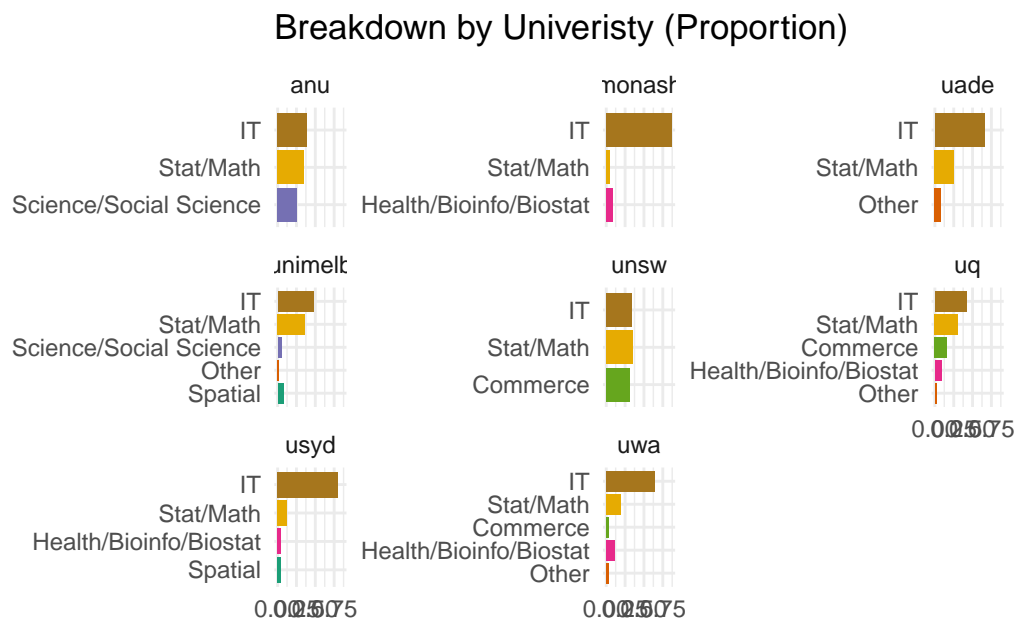


Figure 4.2: Faculty Breakdown by University (Proportion)

At Monash University (monash), University of Adelaide (Uuade), University of Sydney (usyd) and University of Western Australia (uwa), IT is apparently dominating the Master of Data Science degree, especially at Monash University. However, University of Melbourne (unimelb) and University of Queensland (uq) offers relatively higher proportion of statistical and mathematical (Stat/Math) units, whereas units offered at the Australian National University (anu) and UNSW Sydney (unsw) are more evenly distributed across IT, Stat/Math, Science/Social Science and Commerce respectively.

Based on the findings above, it seems that there is a shared structure across Go8 that Master of Data Science is a IT based, computational degree, but the proportion it occupies varies by universities. Monash University tends to be heavily focused on IT and computational aspects, whereas the Master of Data Science degree at UNSW Sydney and ANU are more balanced across IT, statistics and math, as well as science and commerce.

4.2 Unit Overview and Learning Outcome - Bigram

After having a rough idea of the bigger picture, we then moved to explore what exactly are the teaching contents. Single word analysis, bigram and trigram are all produced in order to identify the frequently mentioned words and/or terms. `distinct` function is applied to unit and term, so that same terms are only counted once in a unit to avoid duplicated counting. Words and terms such as ‘student’, ‘successful completion’ add more noises than values to the results, are removed in the pre-processing step.

The bigram, which is Figure 4.3 below provides the most informative results among all. Machine learning appears quite often, as well as software development, linear models, statistical analysis, spatial data. It seems that these frequently mentioned terms are associated with both computational and statistical concepts and skills, which aligns with the findings from the unit code analysis in previous section.

Unfortunately, due to the limited number of observations in the collected data set, the count for each term is too low to make meaningful interpretations or justifications. In addition, similar terms such as research findings, research designs and research literature are supposed to be grouped and counted together, but are not in the bigram. This issue is later solved by introducing the text2vec technique for natural language processing, which will be discussed in the next chapter.

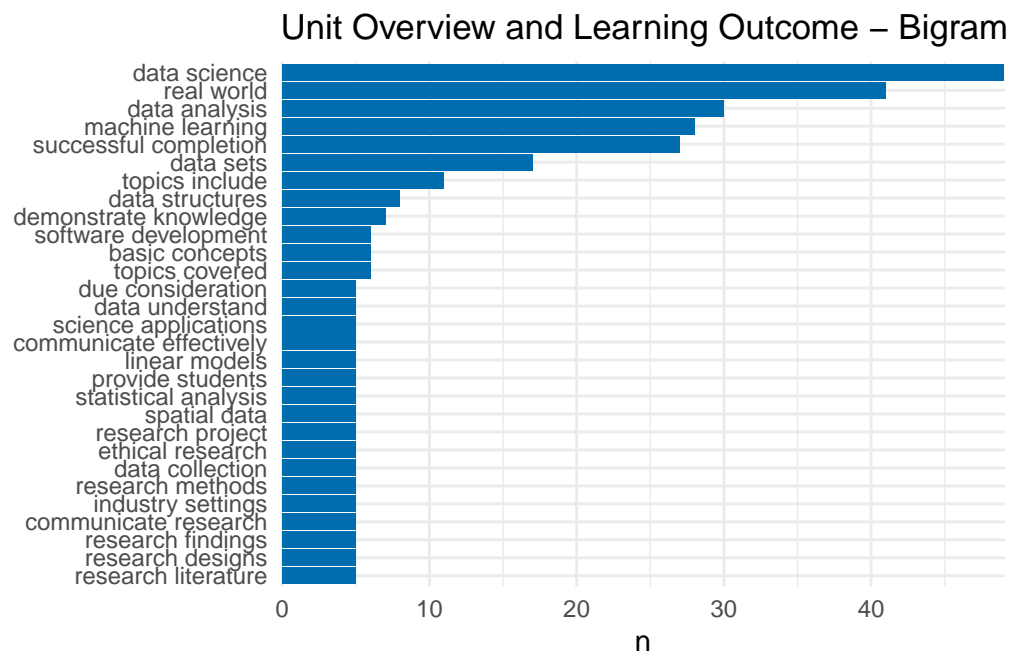


Figure 4.3: Unit Overview and Learning Outcome - Bigram ($n > 4$)

5 Job Text Analysis

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

Part II

text2vec

As mentioned in Section Section 4.2, since the collected university data is relatively small, to make more educated and meaningful interpretations, similar words shall be grouped together and counted by groups. This is usually computed using text corpus, which is a language resource consisting of a large and structured set of texts, since data science is a new term waiting to be defined, there is no available text corpus on this topic. Therefore, we adopted the concept of word2vec and tried to build our own text corpus.

There are multiple publicly available models and packages to conduct similar computations, however, each model takes hours to fit. Due to time constrains, we have only fitted the Dirichlet Allocation (LDA) model with a few parameter adjustments using the `text2vec` package with the concepts illustrated by Das (2016).

Algorithm and Model Fitting

According to Das (2016), the algorithm behind the LDA model is to convert words to document-term matrix (DTM), where the rows, columns and entries correspond to documents, terms and counts respectively. LDA then fits a probabilistic model that assumes a mixture of latent topics, where each topic has a multinomial distribution for all words. The number of topics is a parameter that could be adjusted by needs.

The initial code to build the LDA model was provided by Professor Tanaka, the major part of the code to build the first version of LDA model is also provided below.

```
list(
  tar_target(wiki_stats, get_wiki_articles("https://en.wikipedia.org/wiki/List_of_statisti
  tar_target(wiki_sociology, get_wiki_articles("https://en.wikipedia.org/wiki/Index_of_soc
  tar_target(wiki_computing, get_wiki_articles("https://en.wikipedia.org/wiki/Index_of_com
  tar_target(clean_wiki_stats, map(wiki_stats, clean_wiki_article), format = "rds", reposi
  tar_target(clean_wiki_sociology, map(wiki_sociology, clean_wiki_article), format = "rds"
  tar_target(clean_wiki_computing, map(wiki_computing, clean_wiki_article), format = "rds"
  tar_target(clean_stats, preprocess_text(clean_wiki_stats)),

  tar_target(clean_ssc, preprocess_text(c(clean_wiki_stats, clean_wiki_sociology, clean_wi

  tar_target(itoken_ssc, itoken(clean_ssc, tokenizer = stem_tokenizer),
    cue = tar_cue(mode = "thorough")),

  tar_target(vocab_ssc, create_vocabulary(itoken_ssc, ngram = c(1, 3), stopwords = stopwor
    cue = tar_cue(mode = "thorough")),
  tar_target(vocab_ssc_prune, prune_vocab(vocab_ssc, n_min = 40),
    cue = tar_cue(mode = "thorough")),
```

```

tar_target(dtm_ssc, create_dtm(itoken_ssc, vocab_vectorizer(vocab_ssc_prune)),
           cue = tar_cue(mode = "thorough")),
tar_target(tcm_ssc, create_tcm(itoken_ssc, vocab_vectorizer(vocab_ssc_prune),
                    skip_grams_window = 5L),
           cue = tar_cue(mode = "thorough")),
tar_target(word2vec_model_ssc, model_glove(vocab_ssc_prune, tcm_ssc),
           cue = tar_cue(mode = "thorough")),
tar_target(word2vec_dist_ssc, dist2(t(word2vec_model_ssc$components), method = "cosine"),
           cue = tar_cue(mode = "thorough", format = "rds", repository = "local")),
tar_target(word2vec_res, find_close_words("statistics", word2vec_dist_ssc, 10),
           cue = tar_cue(mode = "thorough")),
tar_target(lda_model03_ssc, model_lda(dtm_ssc, ntopics = 3),
           format = "rds", repository = "local"),
tar_target(lda_model20_ssc, model_lda(dtm_ssc, ntopics = 20),
           format = "rds", repository = "local")
)

```

The model must be trained before it could be used, we web scraped over 4448 Wikipedia articles as training data, including 2816 articles in statistics, 1005 articles in sociology and 627 in computing. The functions used in the codes above such as `get_wiki_articles`, `clean_wiki_article`, `get_clean_combined_wikis`, `model_lda`, `preprocess_text`, `stem_tokenizer`, `prune_vocab`, `model_glove` and `find_close_words` are constructed by Professor Tanaka for pre-processing purposes, the original scripts could be found from the [project repository](#).

Model Adjustments

We have tested using different values for parameter `ntopics` and tried out training the LDA models with different combinations of data.

The results provided differs from models, Figure 5.1 compares the results produced by the full model and model without sociology data on ten topics.

From the results computed by the full model, Topics 9, 10, 2 and 8 occupies relatively higher proportion compare with the others, but the order varies across universities, and their proportions are not significantly larger than the rest of other topics, makes it hard to draw meaningful interpretations. On the right hand side, results from the model without sociology data demonstrates a better picture: Topics 6, 8 and 9 in this case are the top 3 topics across all Go8 universities, however, proportions of Topic 10, 5, 7, 3 and 4 are also obvious higher in some of the universities, brings in difficulties to make justifications.

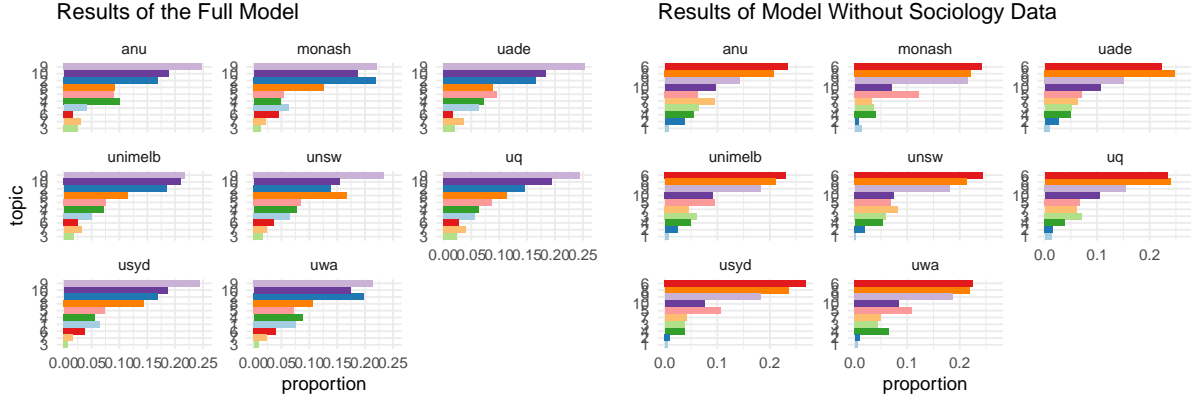


Figure 5.1: Compare Results Between Full Model and Without Sociology Data

As sociology data tends to bring in noises to the model, and is not closely relevant to the data science topic compared with statistics and computing, another two models are fitted using only statistics data and computing data respectively, the results of both models are shown in Figure 5.2.

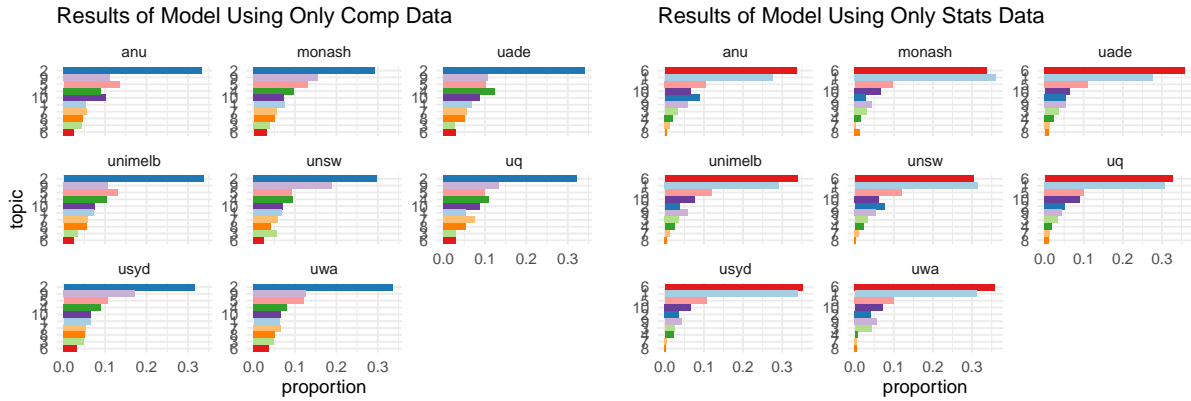


Figure 5.2: Compare Results Between Full Model and Without Sociology Data

Topic 2 is the only dominating topic based on the results provided by the model using only computing data, which provides a clearer picture than the previous models. The table below listed the top ten words for each topic, it turns out Topic 2 contains words like compute (computation etc.), system, program, softwar (software), which are associated with computational aspects, especially software. This model provides a more meaningful result than the prior ones, however, there is not much interpretation could be made for the other topics, the information it offers is still not very satisfying.

In terms of the model trained by only data in statistics, there are also dominating topics across all eight universities: Topics 6 and 1, besides, topics 5, 10, 2, 9 together took a relatively higher proportion compared with the rest of other topics. Both models using only computing

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
window	comput	ibm	algorithm	network	bit	format	intel	softwar
system	system	comput	can	use	instruct	use	chip	compani
version	program	system	number	can	memori	imag	design	appl
releas	use	disk	function	web	use	digit	processor	free
support	machin	drive	set	data	address	can	bit	use
user	design	use	state	internet	regist	video	core	also
oper_system	process	control	languag	secur	processor	standard	mhz	develop
oper	develop	machin	use	access	oper	disc	use	open
os	inform	card	symbol	protocol	can	data	introduc	sourc
file	softwar	unit	problem	link	data	file	microprocessor	user

or statistical data delivers better results, model trained by only statistical data provides more information than the other, hence is selected to use for further analysis on our university data.

Note that it requires highly skilled linguists and huge efforts to establish a proper text corpus, the model we built is still fairly basic and could be further optimised by adjustments.

6 Apply the Selected Model to Collected University Data

Before applying the fitted LDA model to our university data set, words from unit overview and learning outcomes are stemmed using the `SnowballC` package, so that noises like plurals and part of speech are removed. The stemmed words are then assigned to the corresponding topic with the highest probability, instead of counting the appearance of words, the new counts generated are based on topics.

Similar with the university breakdown in Section 4.2, to make more objective comparisons, counts are converted to proportions due to different number of units scraped for the eight universities. Figure 6.1 suggests that Topics 1 and 6 are obviously the dominating topics in Master of data science at all Go8 universities, whereas Topics 2, 5, 9, 10 together also occupies a relatively large proportion.

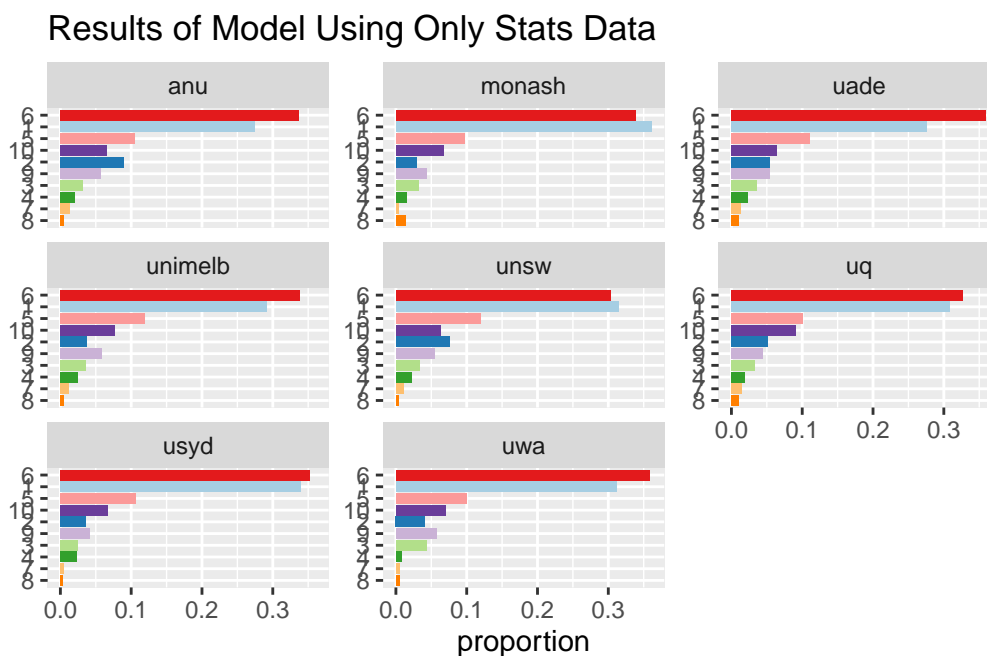


Figure 6.1: Topics Proportion by University

The top ten words based on probabilities for each of the ten topics are provided below, colours

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
studi	model	sampl	theta	probabl	data	distribut	x__	test
use	variabl	estim	function	one	use	frac	frac	statist
statist	regress	mean	probabl	number	algorithm	alpha	left	hypot
research	estim	valu	x_x	theori	analysi	mu	right	valu
data	beta	distribut	distribut	bayesian	can	beta	sum	two
design	linear	varianc	variabl	event	method	function	operatornam	numb
can	use	statist	random	can	model	gamma	sigma	use
effect	y__	use	x__	prior	set	right	y__	correl
popul	squar	standard	random_variabl	infer	cluster	left	cdot	rank
may	can	popul	mathcal	use	comput	normal	operatornam_e	measu

of the columns are aligned with Figure 6.1. Topic 1 contains words like statist (statistics), data, popul (population), and we can see data, algorithm, analysi (analysis), model, cluster, comput (computation etc.) in topic 6, it is a reasonable interpretation that these two topics are both associated with computational aspects.

In addition, words under Topics 2,5,9 and 10 are model, regression, estim (estimate), linear, probabl (probability), bayesian, prior, infer, test, statist (statistics), hypothesi (hypothesis), correl (correlation), most of them are related to math and statistics, and also more on the computational side of them.

The results above further proves the earlier findings discussed in Section Section 4.1 and Section Section 4.2: Master of Data Science degrees offered at Go8 universities tend to be mainly IT based, the major components are computational as well as statistical/mathematical aspects.

Figure 6.2 above demonstrates a breakdown by topics instead of universities, it is clear that compares with the results based on only faculty in Section 4.1, the differences between Go8 are not as much here. The proportions occupied by the eight universities under each topic are fairly similar to each other, indicating the subjective choice made regarding the grouping method in Section Section 4.1 might have provided a slightly misleading information, but it would require further explorations to confirm whether it is truly the case.

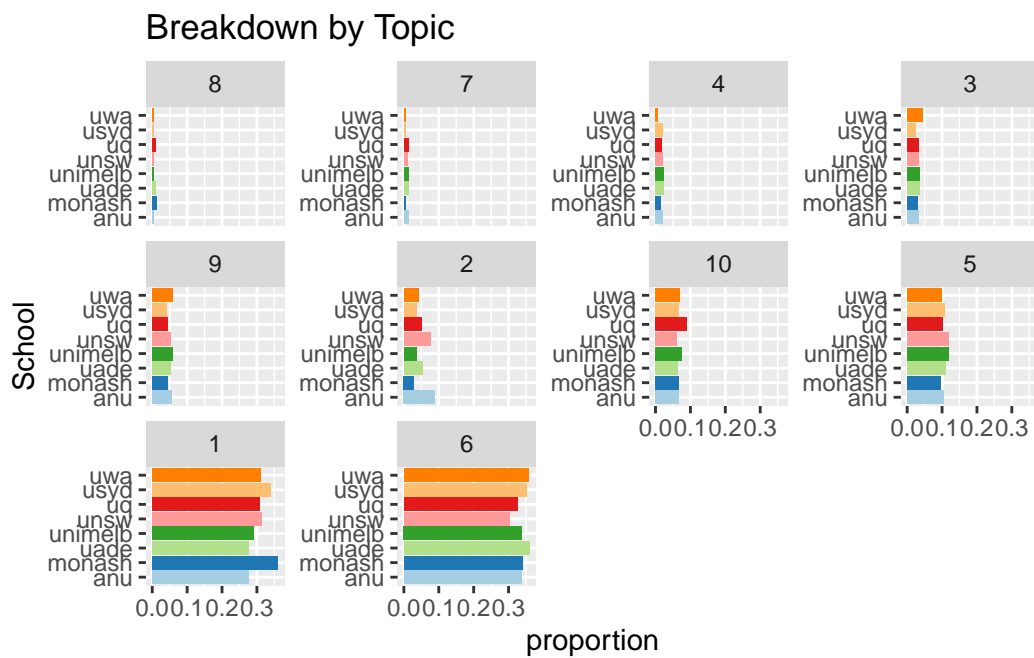


Figure 6.2: Proportion Breakdown by Topic

7 Apply the Selected Model to Employer Data

Part III

Conclusion

8 Limitations

9 Conclusion

References

- Das, Sanjiv Ranjan. 2016. “Data Science: Theories, Models, Algorithms, and Analytics.” *Learning* 143: 145.
- Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. <https://doi.org/10.1093/comjnl/27.2.97>.