Australian Universities Accord

Australian Government

Submissions Analytics

Overview

- Summary Statistics
- Topic Modelling
- Large Language Modelling
  - Document Level
  - Paragraph Level
- Proposed Next Steps
  - Cohort Analysis
  - Subtopic modelling by themes
  - Subtopic modelling by specific questions

Australian Universities Accord

Submission Raw files

Total submissions (n=297*)

- Organisations… 219 submissions
- Individuals… 78 submissions

Includes submissions from

- 38 Universities covering all states and territories

Total submissions (n=297)

Six batches of zip files

- Batch 1
- Batch 2
- Batch 3.1
- Batch 3.2
- Batch 3.3
- Batch 4
- Batch 5
- Batch 6

Plus one extra submission (#295).
Plus Two last minute submissions (#296) & (#297)

*Note there's 299 submissions as two 56 and 59 have two submissions as separate attachments.

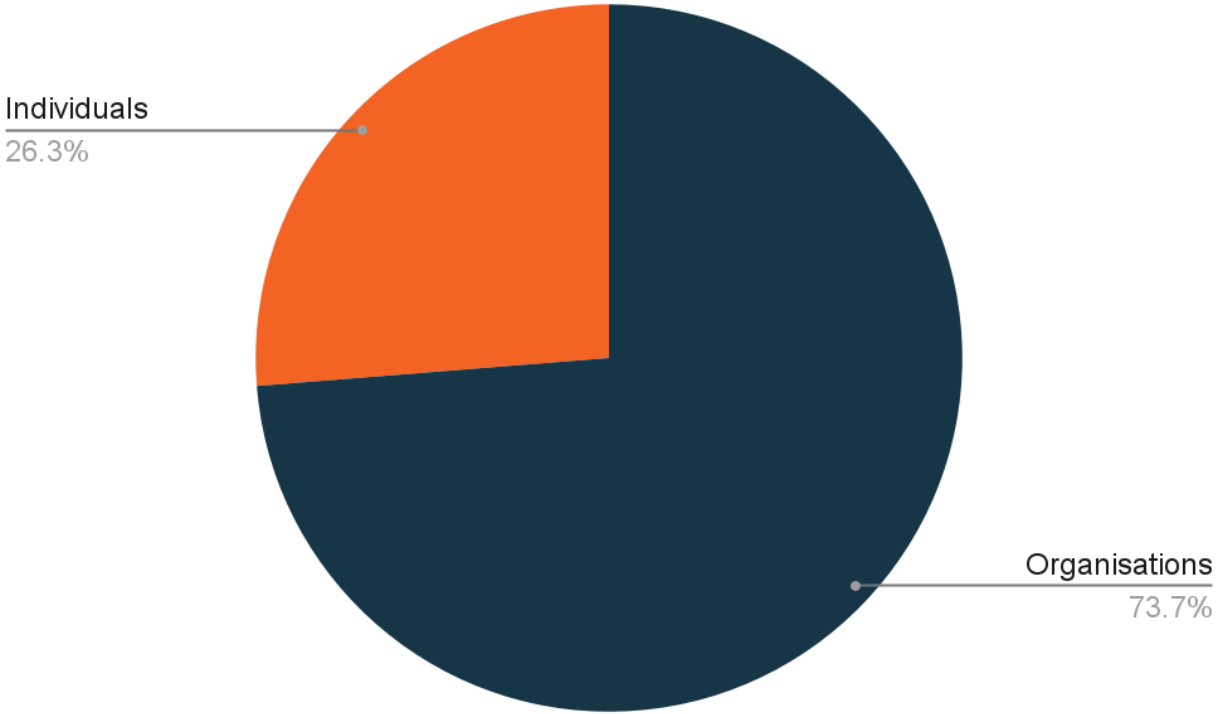
LEAGUE OF SCHOLARS

Summary Statistics

Anatomy of submissions

Number of Submissions

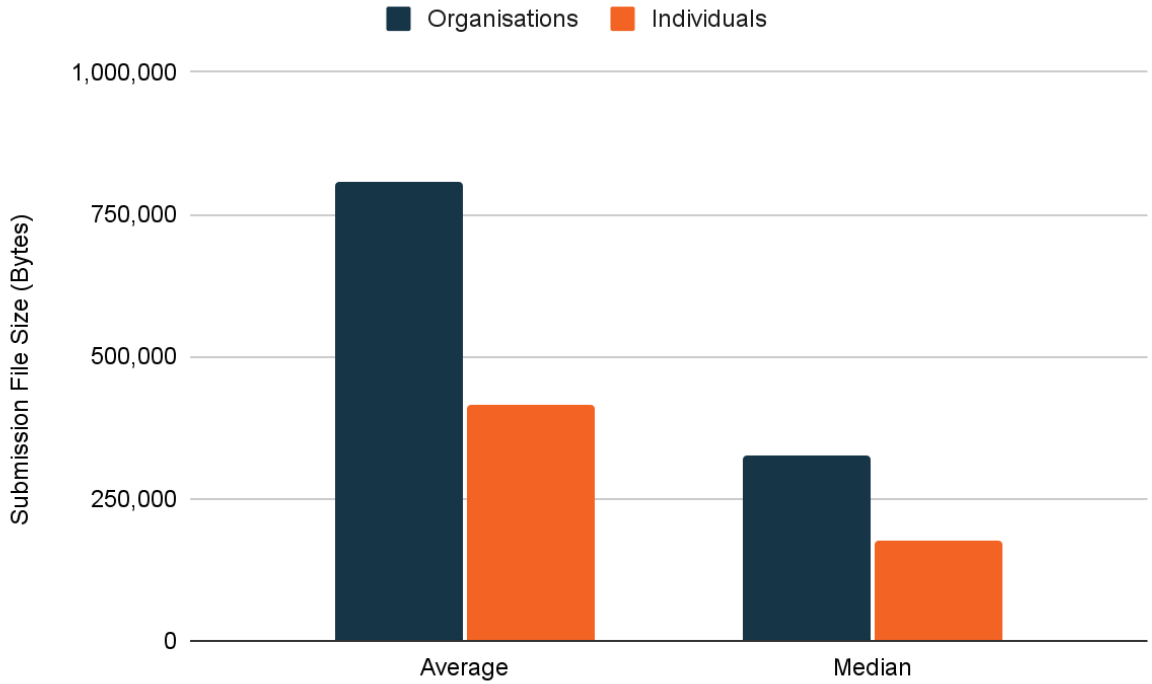
Most submissions are from organisations



LEAGUE OF SCHOLARS

Organisations submissions are ~ 2x as long

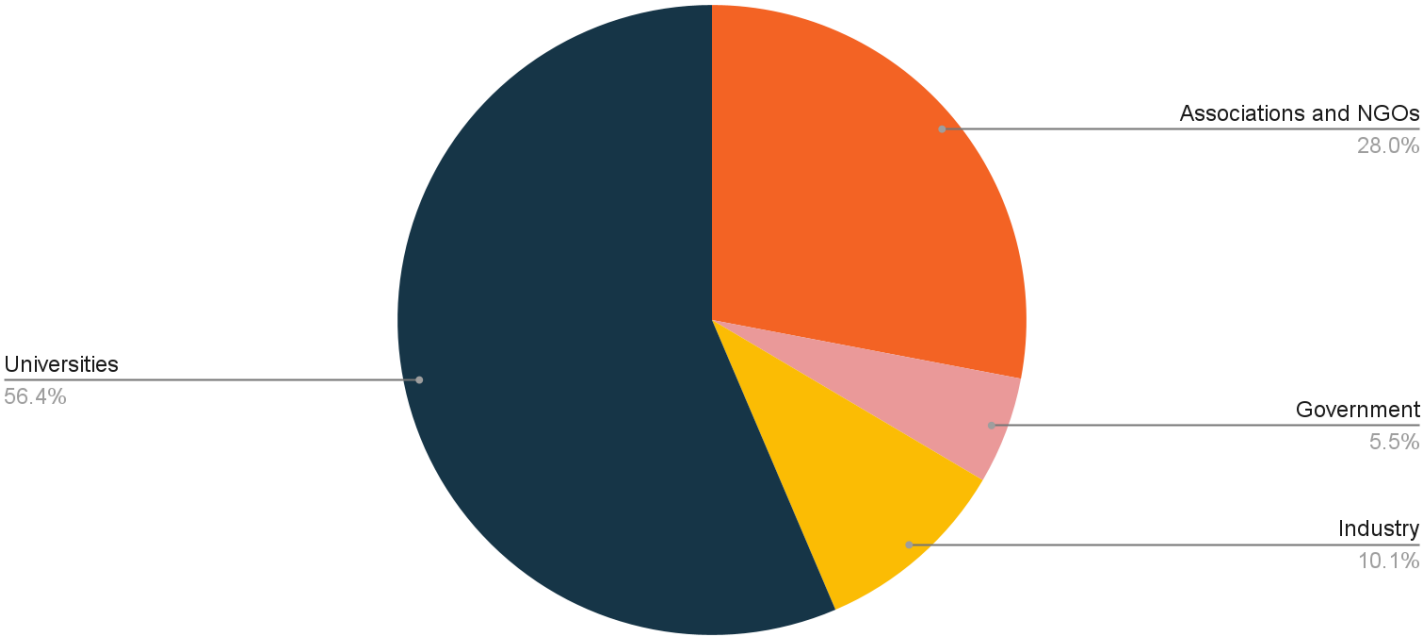
Most organisation submissions are about twice as long



LEAGUE OF SCHOLARS

Universities: majority of submissions

Organisation submissions by source



LEAGUE OF SCHOLARS

Universities with submissions

All states and territories represented

Australian National University	ACT
University of Canberra	ACT
Alphacrucis College	NSW
Charles Sturt University	NSW
Macquarie University	NSW
University of Newcastle Australia	NSW
Southern Cross University	NSW
University of Sydney	NSW
University of New England	NSW
UNSW Sydney	NSW
University of Technology Sydney	NSW
Western Sydney University	NSW
Charles Darwin University	NT
Australian Catholic University	QLD
Bond University	QLD
Central Queensland University	QLD
Griffith University	QLD
James Cook University	QLD
University of Queensland	QLD
University of the Sunshine Coast	QLD
University of Southern Queensland	QLD
University of Adelaide	SA
Flinders University	SA
Torrens University	SA
University of South Australia	SA
University of Tasmania	TAS
Deakin University	VIC
Federation University	VIC
La Trobe University	VIC
Monash University	VIC
Open Universities Australia	VIC
RMIT University	VIC
University of Melbourne	VIC
Victoria University	VIC
Curtin University	WA
Edith Cowan University	WA
Murdoch University	WA
University of Western Australia	WA

LEAGUE OF SCHOLARS

Initial Analysis

Our process

Initial submissions analysis

1. Preprocessing
    a. Convert all PDF & Word submissions into text
    b. Clean text to remove artifacts
    c. Parse and label for paragraph-level segmentation

2. Parse to create submissions corpus
    a. Determine length and frequency of n-grams

3. Topic modelling
    a. Determine optimum number of topics
    b. Determine distinctive phrases and SIPs
    c. Label and characterise topics.

Subsequent submissions analysis

- Identify the stakeholders who have made submissions and what cohorts of stakeholders responded to which discussion questions.

- Identify dominant topics and discussion questions commonly responded to

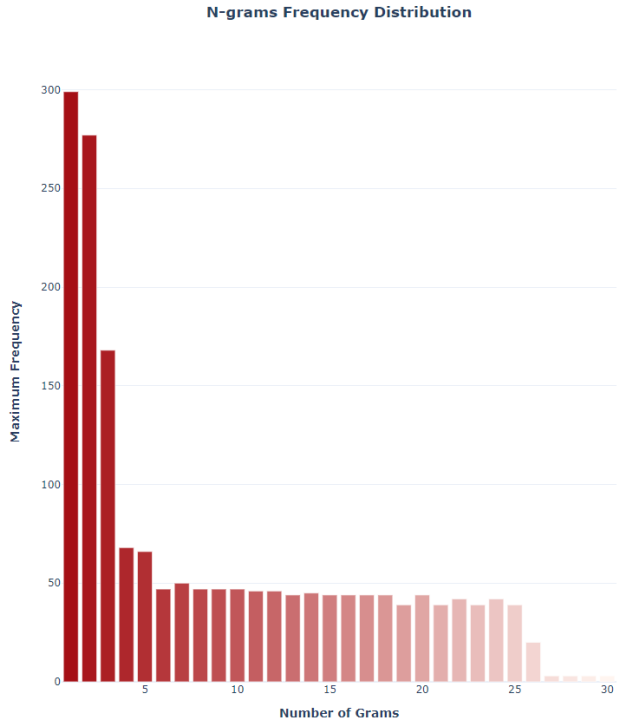
LEAGUE OF SCHOLARS

Topic modelling

Automatic bottom-up approach

Terms and phrases used in all submissions

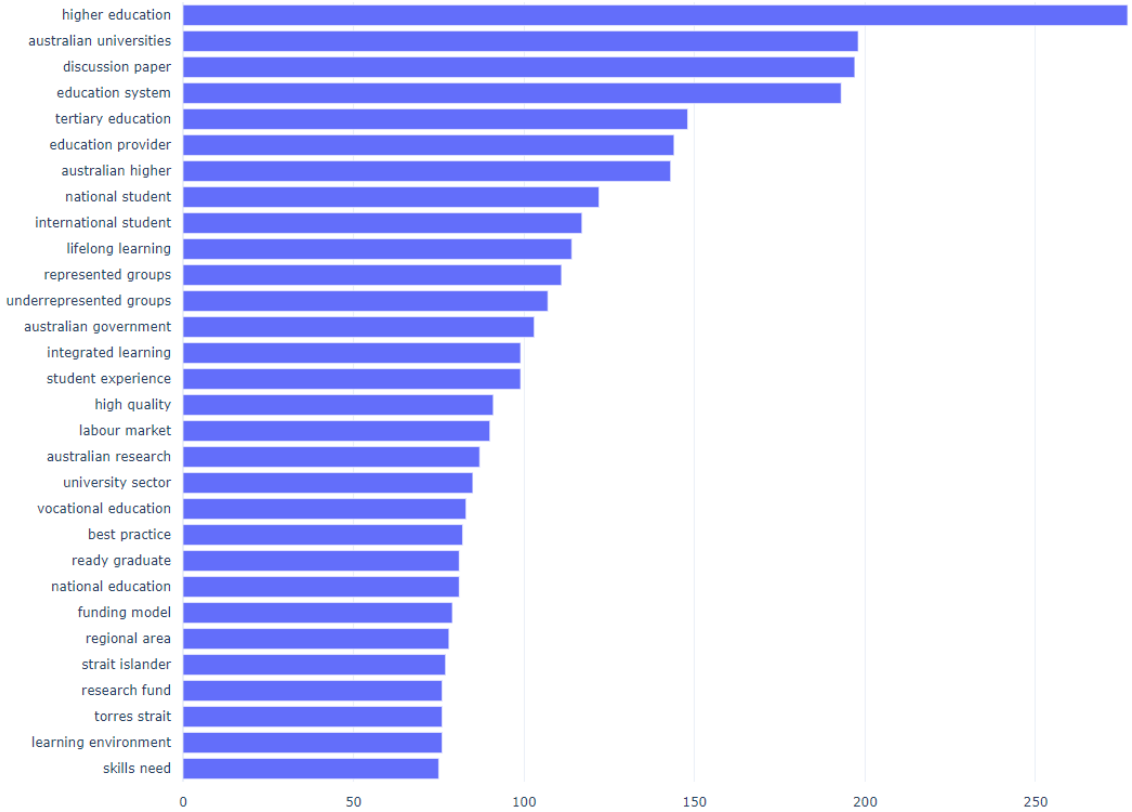
N-Gram Frequency Distribution



LEAGUE OF SCHOLARS

Top 30 Bigrams

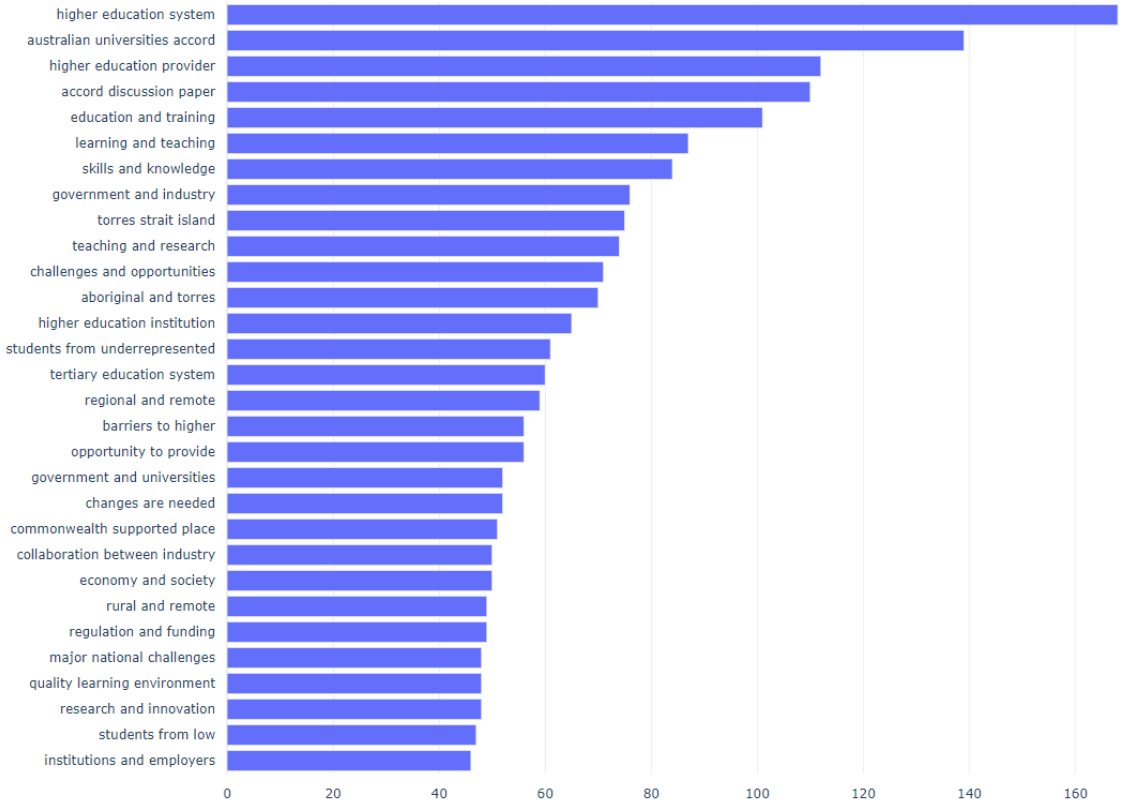
By Frequency Distribution



LEAGUE OF SCHOLARS

Top 30 Trigrams

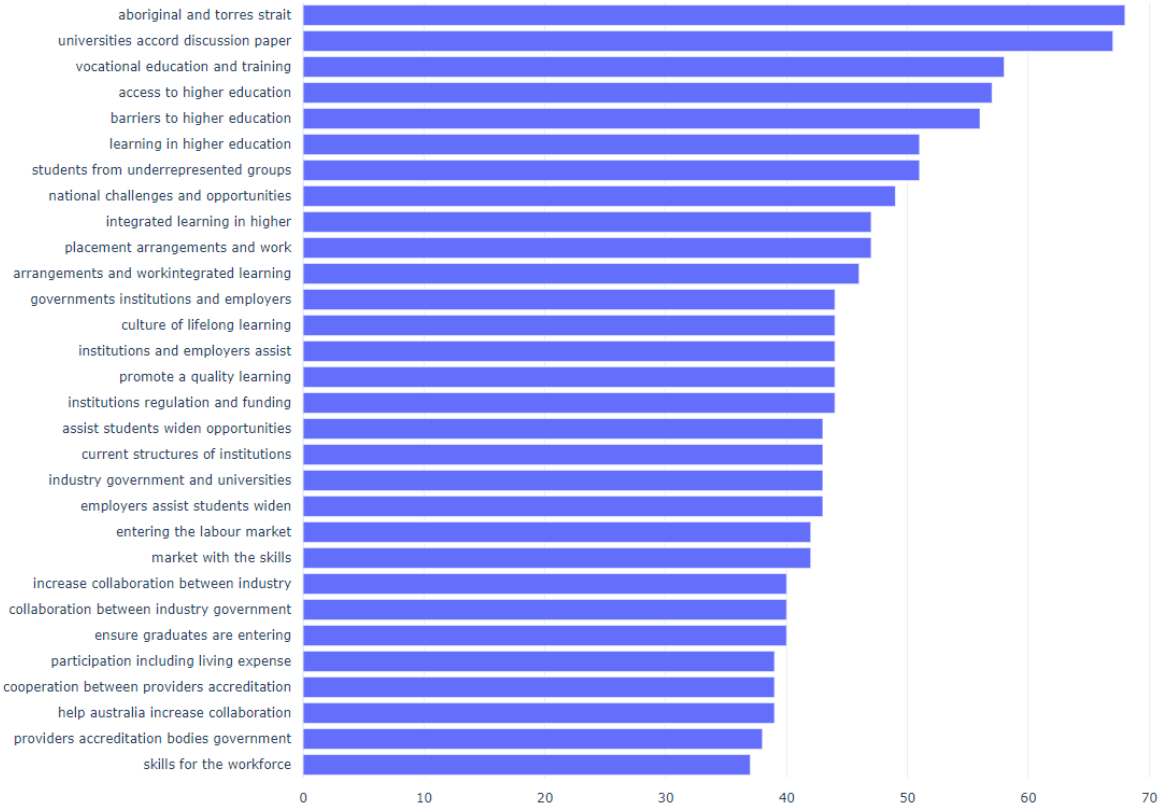
By Frequency Distribution



LEAGUE OF SCHOLARS

Top 30 Quadgrams

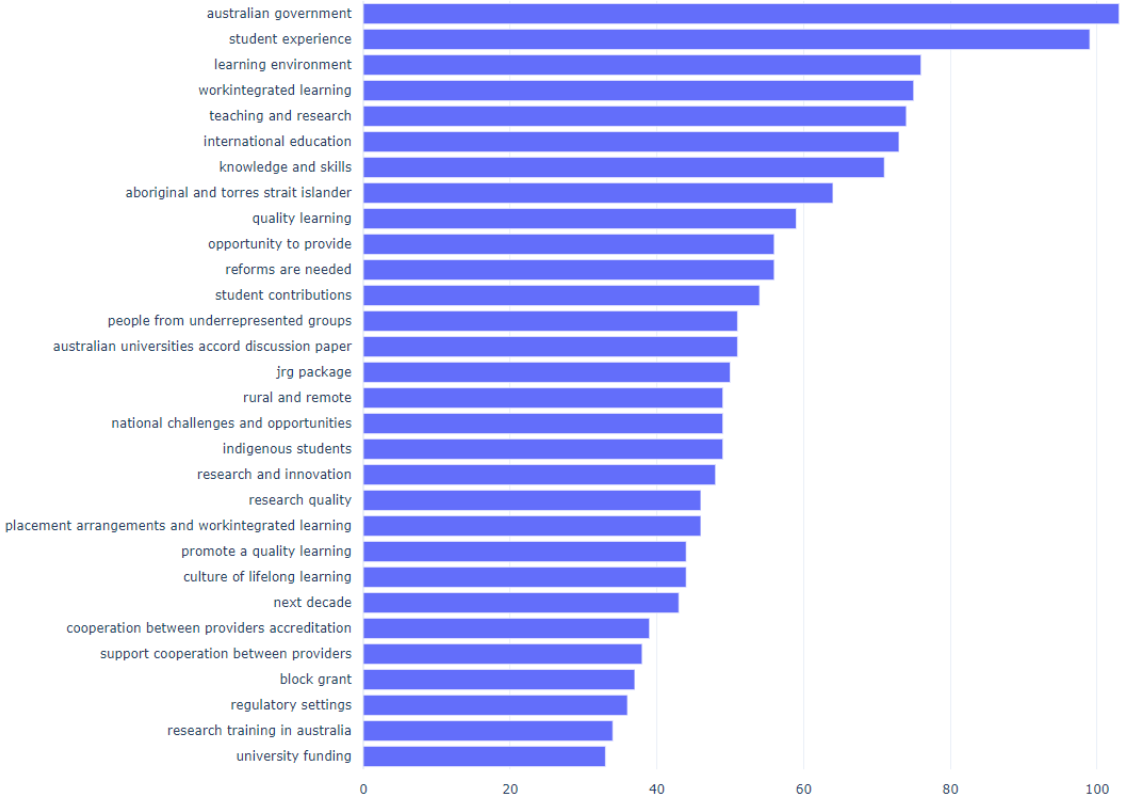
By Frequency Distribution



LEAGUE OF SCHOLARS

Top 30 N-Grams

By Frequency Distribution



LEAGUE OF SCHOLARS

Topic modelling

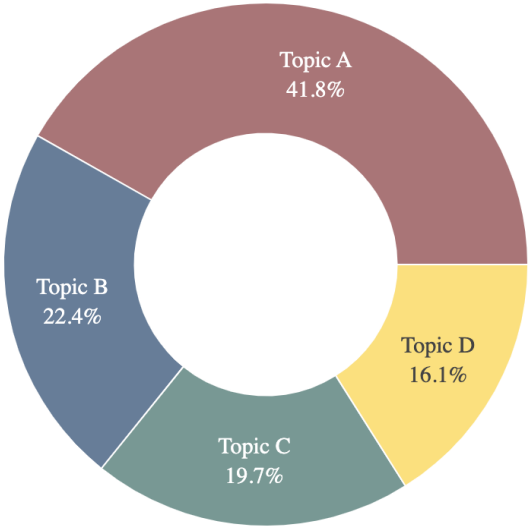
Automatic bottom-up approach

Topic Modelling

Which topics are dominant for each submission?

Using topic modelling*, we:

- Determined optimum number of topics based on a consistency score
- Identified common topics among detailed submissions (n = 299).
- Found Four key topics optimal



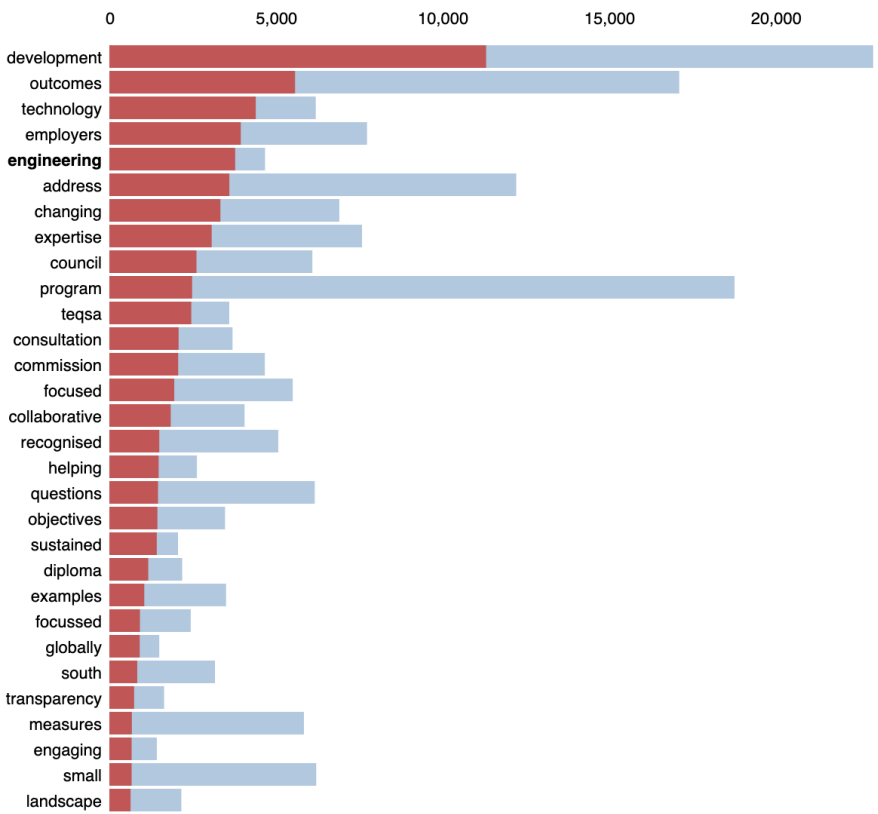
* Topic modeling using Latent Dirichlet Allocation (LDA).

Topic Modelling

Topic	# Submissions	Summary Title	Top Unique Keywords in each Topic*
A	125	Technology and Engineering	academy technological sciences and engineering, choose study maths and engineering, prepares participants for engineering job, experiencing engineering skills supply challenge, delivers outcomes for the nation, development sustainable and enduring outcomes, skills shortages where employers, wage subsidies incentivise employers, employers with training costs met, industry needs will enhanced collaborative, lifelong learning drawing further expertise, enhanced connections expertise can harnessed
B	67	Asia-Pacific	generate annual boost gdp worth, boost gdp worth additional billion, pursuing overseas study compared china, china remains major source region, china and other priority regions, growing the percentage gdp spent, china and india international students, middle class populations such china, partnerships lift capability the asia-pacific, public diplomacy across the asia-pacific, establishing centres for asia-pacific studies, darwin university and james cook
C	59	Disadvantage	review the disability standards, people with disability tertiary education, providers support students with disability, national disability coordination officer program, lived experience people with disability, disability standards for education review, students from low-income families, young people and families break, care services children and families, tuition free and providing debt, recommendation that tuition fee reduction, better living allowances through centrelink, students access centrelink support
D	48	Health	direction express condition the trust, property the university trust apply, turning guthries submissions and financial, longstanding driver dissatisfaction the academic, dissatisfaction and lowered performance students, hospitals offer additional placements, clinical teaching hospitals, brain and mental health disorders, reports increasing mental health issues, mental health and poor remuneration, available placement opportunities across health, align with workforce demands, mental illness and brain conditions

* Distinctive Statistically Improbable Phrases (SIPs) within each Topic ranked by TF IDF.

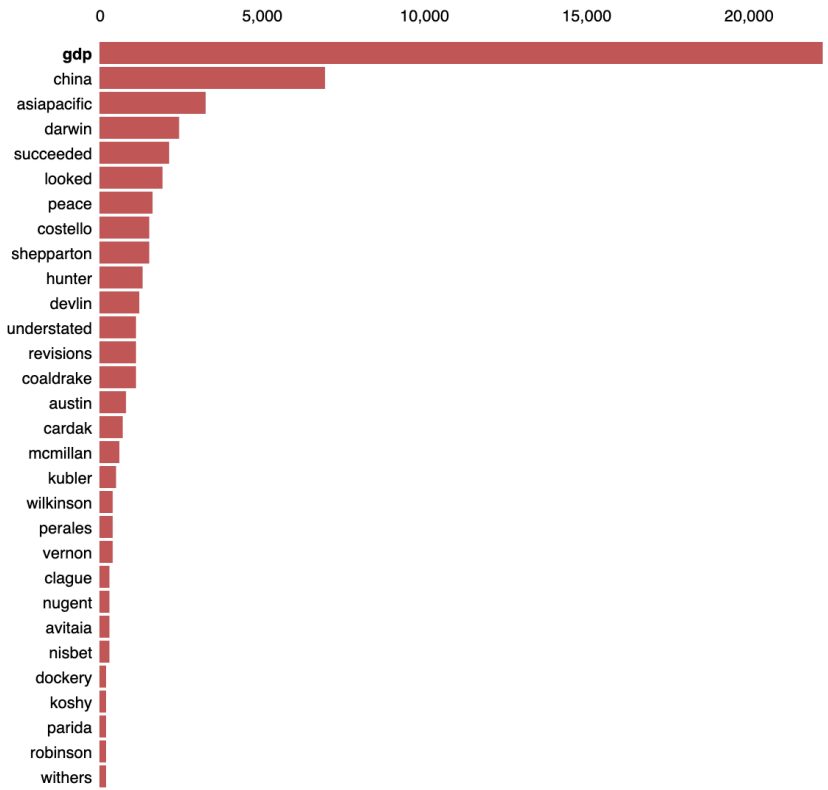
Topic A



Most Significant Topic Terms
Ranked by *term relevance score*

- term frequency within this topic
- term frequency in all submissions

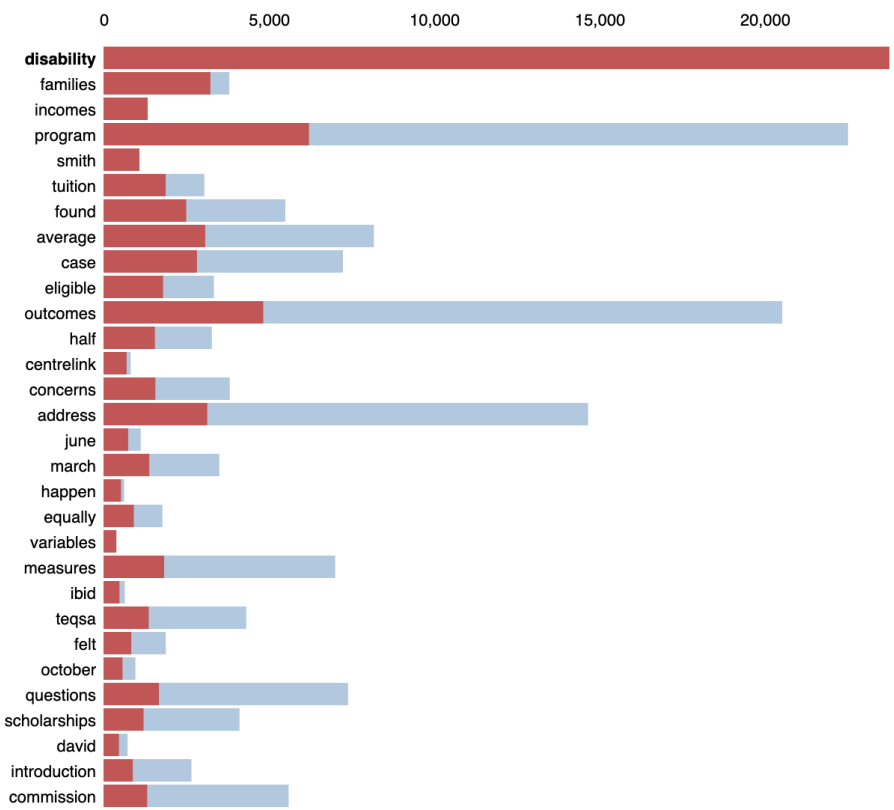
Topic B



Most Significant Topic Terms
Ranked by *term relevance score*

- term frequency within this topic
- term frequency in all submissions

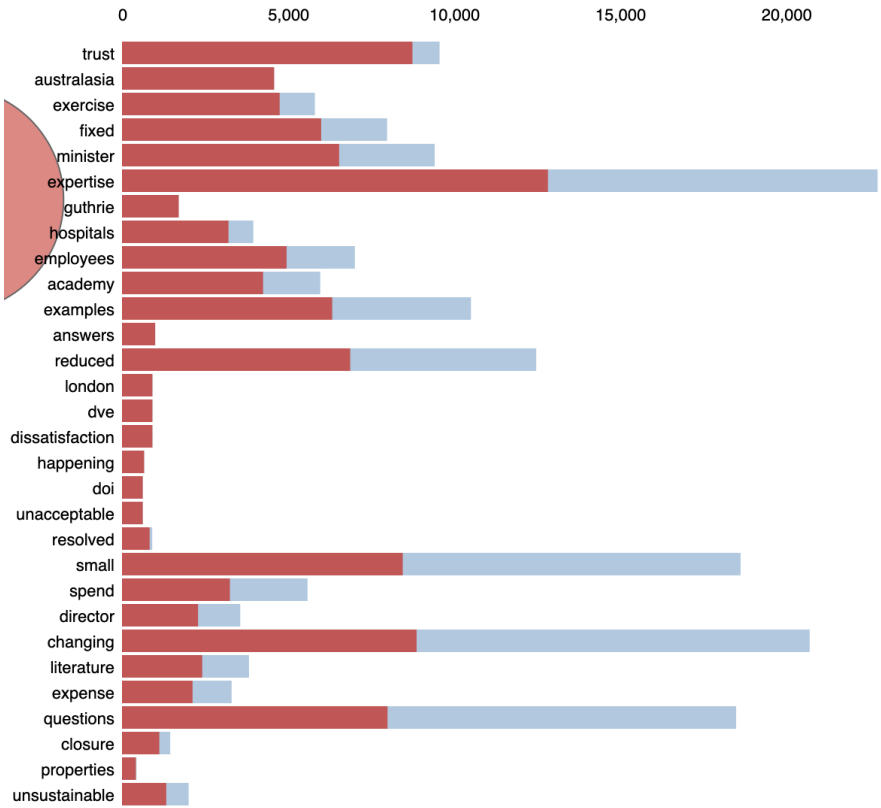
Topic C



Most Significant Topic Terms
Ranked by *term relevance score*

- term frequency within this topic
- term frequency in all submissions

Topic D



Most Significant Topic Terms
Ranked by *term relevance score*

- term frequency within this topic
- term frequency in all submissions

Large-Language Modelling

Document-Level Analysis

What is Large Language Modelling

Overview

A large language model (LLM) is a language model consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabelled text using self-supervised learning.

We are using a neural network training model to generate numerical representations of two sets of documents — themes and submissions.

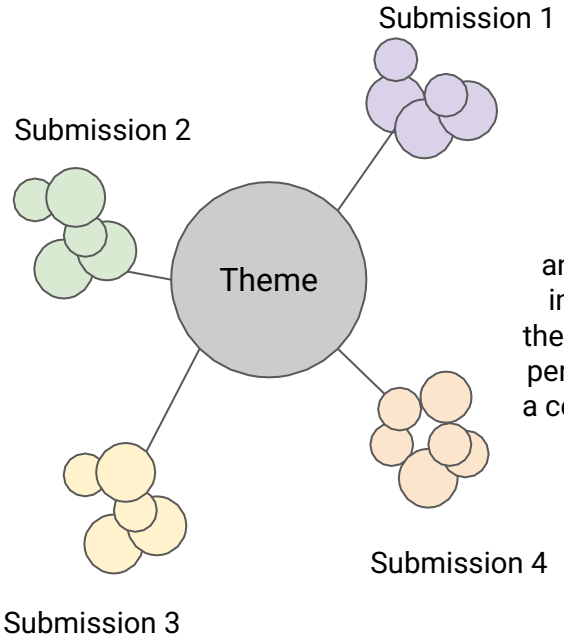
- We parse and clean each document
  a. Submissions (source)
  b. Descriptions of Themes (target)
- We vectorise each document - transform the documents into a numerical representation.
  a. This considers contextual relationships between words not just frequency
- Each document both source and target are represented in multidimensional space as a series of numbers or vectors.
- Similarity between themes and submission can then be easily compared using cosine similarity.

LEAGUE OF SCHOLARS

The Universities Accord is a multi-themed review

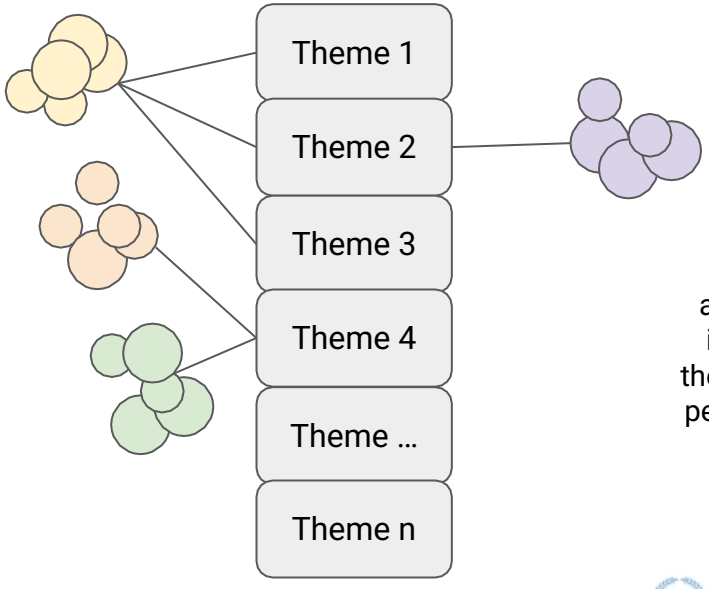
Unlike others there are many themes under review

Single-Themed Consultation



Submissions analysis offers insights about the key differing perspectives on a central theme.

Multi-themed Consultation



Submissions analysis offers insights about the key differing perspectives on a variety of themes.

LEAGUE OF SCHOLARS

LLM Pipeline

Design and test with initial submissions

Developing an approach to parsing and preprocessing of submissions.

1. Creation of reference corpus based on Discussion paper.
2. Dividing into content related to headings and subheadings of topics relating to the nine Themes.
3. Pre-processing of initial submissions parsing and cleaning text (n=49).
4. Splitting by paragraph.
    - Noting submissions are in different formats.

Vectorisation of content and maps

1. Creation of embedding vectors for each paragraph.
2. Creation of embedding vectors for each heading and subheading themes.
3. Map the two together in the same vectorisation space.

Two mappings are then possible:

- Paragraph-level analysis
- Document-level analysis

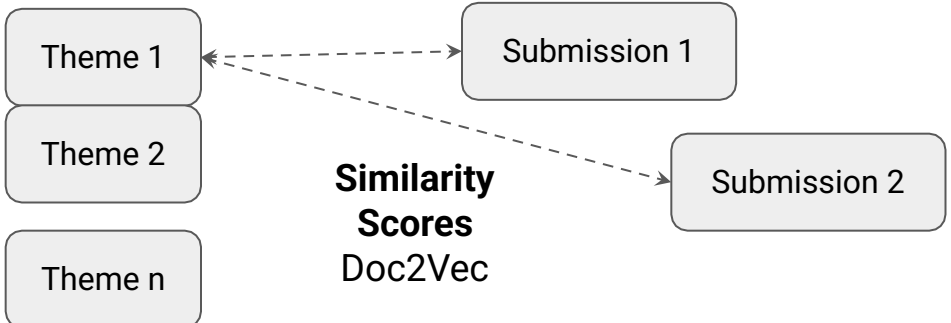
Paragraph-level vectors can be summed to provide a document-level analysis where we can discover the main theme for each submission.

LEAGUE OF SCHOLARS

Mapping submissions to themes

Nine Themes from the discussion paper and corresponding chapters

Themes	Challenges and opportunities for the higher education system	
• Teaching	3.1	Quality teaching delivering quality learning
• Skills	3.2	Meeting Australia's knowledge and skills needs
• VET education	3.3	Connection between vocational education & training and higher
• Innovation	3.4	A system that delivers new knowledge, innovation and capability
• Access	3.5	Creating opportunity for all Australians
• Accountability	3.6	Governance, accountability and community
• Quality	3.7	Quality and sustainability
• Intl. Ed	3.8	The role of international education
• Economics	3.9	Investment and affordability



LEAGUE OF SCHOLARS

Submission to Theme Matrix

Submission similarity to nine themes

Submissions_Index	3.1 Quality teaching delivering quality learning	3.2 Meeting Australia's knowledge and skills needs	3.3 Connection between the vocational education and training and higher education systems	3.4 A system that delivers new knowledge, innovation and capability	3.5 Creating opportunity for all Australians	3.6 Governance, accountability and community	3.7 Quality and sustainability	3.8 The role of international education	3.9 Investment and affordability
1	0.06	0.09	0.05	0.13	0.01	0.09	0.12	0.10	-0.01
2	0.18	0.38	0.20	0.23	0.37	0.25	0.08	0.34	0.24
3	0.11	0.01	0.05	0.07	0.21	0.10	0.22	0.15	0.26
4	0.17	0.04	0.02	0.13	0.04	-0.01	0.31	0.07	0.06
5	0.20	0.32	0.25	0.33	0.44	0.30	0.19	0.29	0.26
6	0.06	-0.07	0.18	0.03	0.16	0.22	0.12	-0.03	0.18
7	0.16	0.37	0.23	0.26	0.36	0.32	0.07	0.28	0.19
8	0.18	0.01	0.02	0.05	0.01	-0.04	0.15	-0.09	0.03
9	0.17	-0.03	0.05	0.00	0.05	0.27	0.18	0.16	0.25
10	0.23	0.07	0.06	0.10	0.17	0.02	0.11	0.07	0.03
11	0.26	0.18	0.17	0.06	0.12	0.25	0.18	0.29	0.12
12	0.15	0.18	0.06	0.04	0.12	0.10	0.21	0.24	0.14
13	0.08	0.27	0.05	0.21	0.25	0.05	0.01	0.15	0.11
14	0.19	-0.02	-0.03	0.08	0.13	0.02	0.01	0.08	0.32
15	0.01	0.24	0.15	0.11	0.24	0.22	0.16	-0.02	0.33
16	-0.01	0.20	0.13	0.15	0.05	0.07	0.26	0.14	0.09
17	0.11	-0.01	0.06	0.16	0.00	0.12	0.10	0.25	-0.08

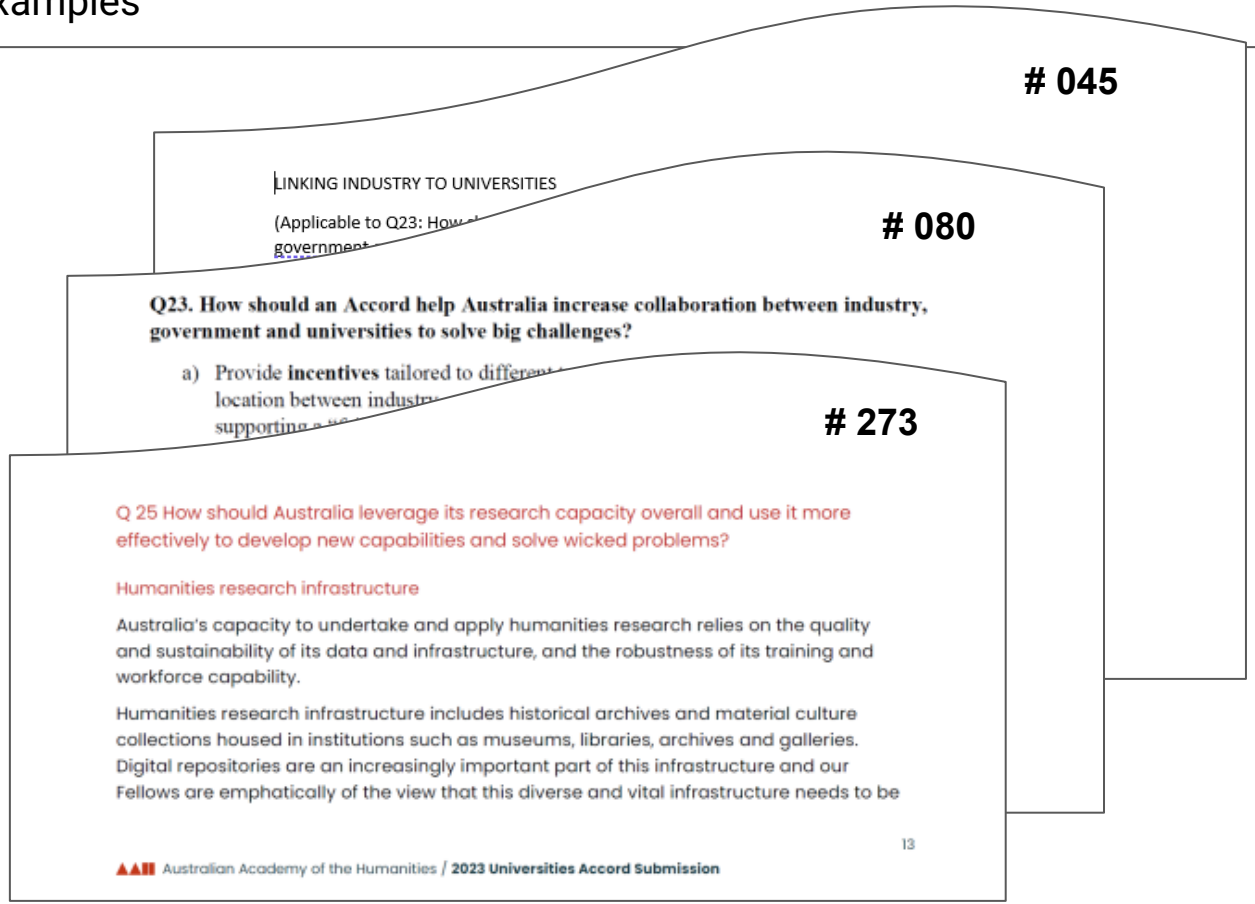
Top Submissions Ranked by Theme Similarity

Here we show top submissions aligned with Capability & Innovation theme

Submissions_Index	3.1 Quality teaching delivering quality learning	3.2 Meeting Australia,Äôs knowledge and skills needs	3.3 Connection between the vocational education and training and higher education systems	3.4 A system that delivers new knowledge, innovation and capability	3.5 Creating opportunity for all Australians	3.6 Governance, accountability and community	3.7 Quality and sustainability	3.8 The role of international education	3.9 Investment and affordability
273	0.21	0.08	0.10	0.37	0.16	0.19	0.13	0.10	0.08
80	0.09	0.01	0.11	0.35	0.10	0.11	0.12	-0.02	0.05
45	0.10	0.20	0.14	0.34	0.12	0.17	0.05	0.01	0.05
86	0.05	0.01	0.21	0.34	0.10	0.12	0.17	0.35	0.15
283	0.04	0.17	0.08	0.33	0.15	0.12	0.12	0.19	0.09
5	0.20	0.32	0.25	0.33	0.44	0.30	0.19	0.29	0.26
221	0.18	0.24	0.03	0.32	0.26	0.07	0.15	-0.06	0.13
23	0.23	0.28	0.23	0.31	0.16	0.15	0.04	0.24	0.14
104	0.17	-0.01	0.01	0.31	-0.01	-0.03	0.02	-0.06	-0.11
71	0.07	0.10	0.26	0.30	0.19	0.08	0.08	0.25	0.01
249	0.09	0.13	0.28	0.29	0.07	0.07	0.12	0.01	0.12
81	0.15	0.13	0.08	0.29	0.00	0.07	0.09	0.04	-0.05
153	0.04	0.08	0.08	0.29	0.05	0.09	0.00	0.02	0.06
162	0.16	0.11	0.22	0.29	0.18	0.49	0.08	0.20	0.18

Capability & innovation themed submissions

Top examples



LEAGUE OF SCHOLARS

Top 10 Submissions

By similarity to nine themes

3.1 Quality teaching delivering quality learning	3.2 Meeting Australia's knowledge and skills needs	3.3 Connection between the vocational education and training and higher education systems	3.4 A system that delivers new knowledge, innovation and capability	3.5 Creating opportunity for all Australians	3.6 Governance, accountability and community	3.7 Quality and sustainability	3.8 The role of international education	3.9 Investment and affordability
29	203	102	273	5	66	165	37	163
73	196	199	80	105	162	114	156	181
25	121	213	45	203	214	44	122	238
96	2	248	86	196	7	66	86	102
95	7	261	283	57	69	253	2	30
108	213	111	5	87	95	4	121	91
165	220	262	221	181	159	18	233	15
19	5	141	23	121	5	24	223	198
116	266	48	104	50	238	127	134	14
179	56	280	71	2	234	42	98	203

Large-Language Modelling

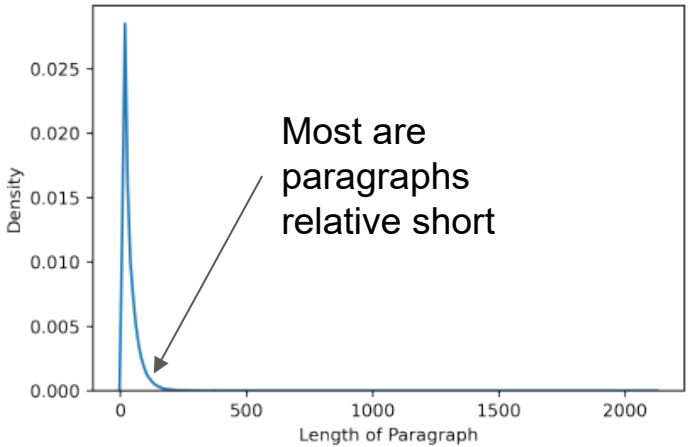
Paragraph-Level Analysis

LEAGUE OF SCHOLARS

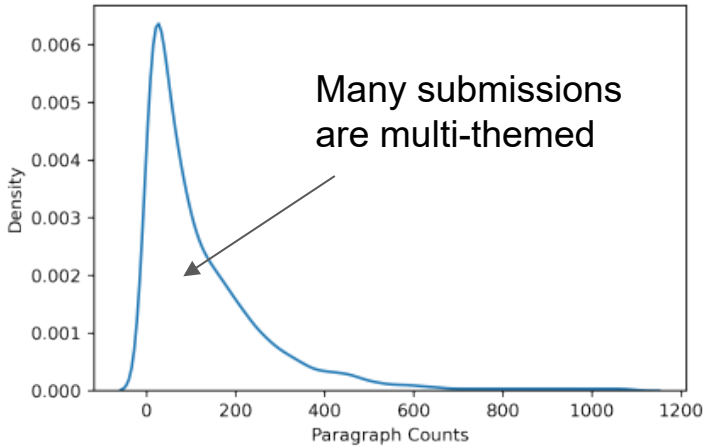
Short paragraphs validate mapping of multiple themes

Parsing submissions to paragraphs

- Length of Paragraphs
  - Between 11 words up to 2000+ words
  - Short paragraphs with 10 words or less omitted.
  - Most of paragraphs (95%) are with less 100 words.



- Number of Paragraphs
  - Documents include between 1 paragraph up to 1100 paragraphs,
  - Most of documents (90%) have less than 300 paragraphs.



Submission to Theme Matrix

By paragraph similarity to nine themes

Submissions_Index	3.1 Quality teaching delivering quality learning	3.2 Meeting Australia's knowledge and skills needs	3.3 Connection between the vocational education and training and higher education systems	3.4 A system that delivers new knowledge, innovation and capability	3.5 Creating opportunity for all Australians	3.6 Governance, accountability and community	3.7 Quality and sustainability	3.8 The role of international education	3.9 Investment and affordability
001 - 1	-0.07	0.13	0.09	0.09	0.09	-0.08	0.21	0.15	0.13
001 - 2	-0.03	0.11	0.09	0.20	0.00	0.13	0.18	0.17	0.03
001 - 3	0.14	0.30	0.29	0.18	0.30	0.19	0.28	0.18	0.36
001 - 4	-0.03	0.18	0.07	0.16	0.12	0.15	0.14	0.01	0.16
001 - 5	0.10	0.23	0.15	0.21	0.26	0.06	0.17	0.02	0.15
001 - 6	0.07	0.29	0.15	0.27	0.13	0.29	0.22	0.16	0.25
001 - 7	0.02	0.17	0.10	0.16	0.12	0.36	0.22	0.07	0.34
001 - 8	0.07	0.25	0.16	0.35	0.00	0.25	0.27	0.25	0.21
001 - 9	-0.08	0.09	0.05	0.03	0.10	0.03	0.04	-0.08	0.00
001 - 10	0.02	0.23	0.10	0.11	0.14	0.05	0.02	0.00	0.17
001 - 11	0.05	0.20	0.15	0.09	0.00	-0.01	0.08	0.06	0.14
001 - 12	0.18	0.09	0.18	0.14	0.29	0.10	0.20	0.01	-0.11
001 - 13	0.14	0.25	0.17	0.21	0.14	0.17	0.25	0.07	0.17
001 - 14	-0.14	0.20	-0.04	0.07	0.17	0.08	-0.09	0.05	0.31
001 - 15	0.04	0.19	0.16	0.06	0.18	0.11	0.18	0.07	0.12

Visualisation

Interactive visual
representations of data

Interactive Visualisation

Overview

Visualisation benefits

- Many people can see and explore patterns in submissions
- This may lead to new insights as some people may see different associations and linkages
- Topics, themes and submissions may have interdependencies, overlaps and complementarity

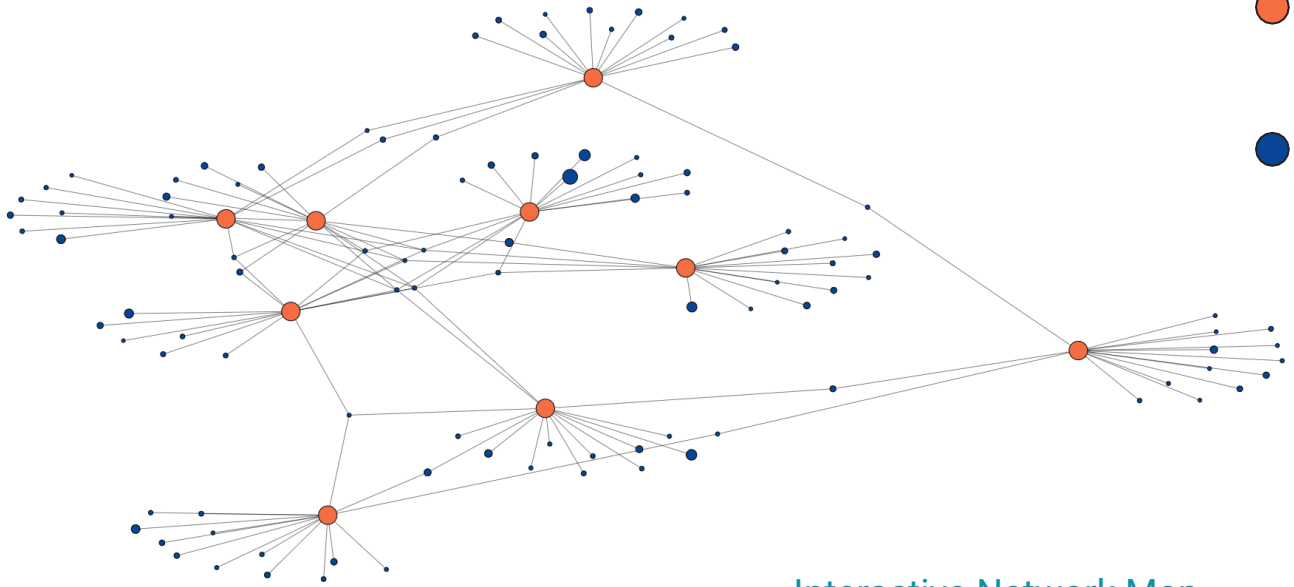
Reveals

- Interconnections between themes
- Theme-bridging submissions
- Key submissions by theme

LEAGUE OF SCHOLARS

Network Map

Here a network of Top 15 submissions closest to each of the Nine Themes.



Themes from Discussion Paper

Submissions most aligned to each theme.

[Interactive Network Map]

Size indicates of nodes indicate submission length and distance from theme its similarity to that theme.

Contact

Paul X. McCarthy
CEO and Co-founder League of Scholars
Industry Fellow, The Data Science Institute, UTS
Adjunct Professor of Computer Science, UNSW Sydney
Phone +61 418 608 224 • Email paul@onlinegravity.com • Twitter @paulxmccarthy

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

UNSW SYDNEY

LEAGUE OF SCHOLARS

Australian Universities Accord

Australian Government

Potential Follow-on Work

Cohort analysis

Subtopic modelling by theme

LEAGUE OF SCHOLARS

Cohort Analysis

Classifying the relative scale of representation of organisation submissions

Many inquiries accept submissions from both individuals and organisations. In some cases, such as the NSW Floods Inquiry most of the focus and submissions were from individuals rather than individuals.

When exploring analysis of submissions from organisations in some situations it may be important to consider the scale and submission of the stakeholder groups that formal organisational submissions represent.

For the Higher Education Accord, we may see submissions from organsations that may include:

- Universities
- Businesses
- Unions such as National Tertiary Education Union
- Associations such as the Academy of Sciences or the Group of Eight

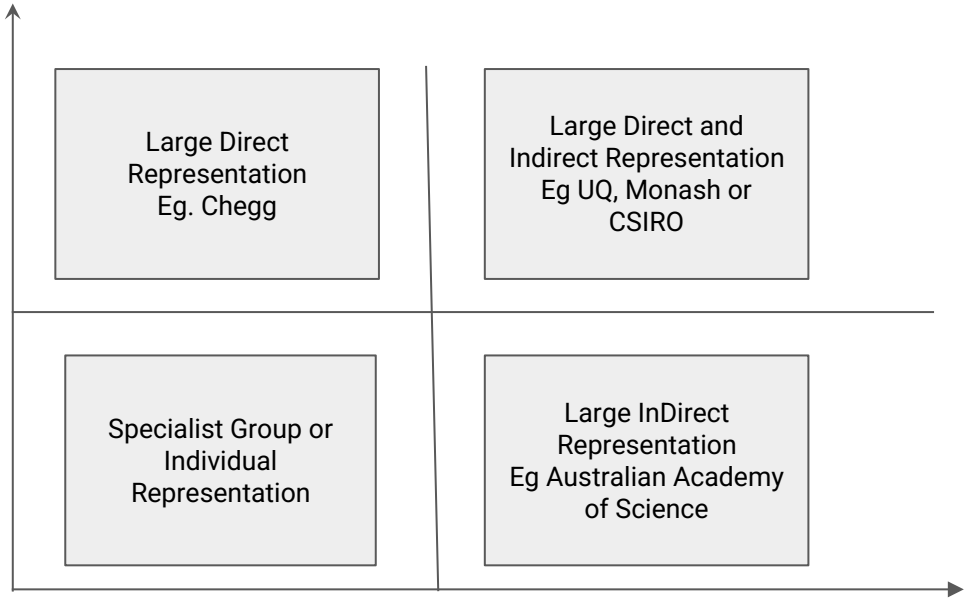
In some cases one formal submission from these organisations may represent the voices of hundreds or in some case thousands of individuals so having a sense of the scale of this representation could be useful in helping the inquiry evaluate the responses.

Organisation Submissions

Classifying the relative scale of representation of submissions

Direct Representation

Metrics as Web Traffic, News Mentions



Influence

Metrics such as Network Centrality

LEAGUE OF SCHOLARS

Subtopic Modelling by Theme

A further analysis of paragraph-level data could reveal what the key topics are within each of the nine themes.

This could reveal more complex and subtle patterns as well as specific policy or reform suggestions related to each theme such as international education.

Here we would propose create nine corpora of data based on the paragraphs most aligned to each theme and run topic modelling and LLM mapping on each of these.

Specific issue and sentiment analysis

Further analysis can focus on one or all of the nine themes or another more specific issue such as Job Ready Graduates (JRG).

It would be possible too to group responses into those supportive of a particular issue and those against it given exemplars of each can be identified we can train a machine learning classifier to group similar for, against and neutral responses.

Paragraph-level visualisation with scatter plots and dendrograms may also provide insights about topics at a paragraph level.