Australian Government
Jobs and Skills Australia

Training product similarity analysis:

Using machine learning to analyse the similarity of qualifications in training packages

This paper outlines the methodology used in the exploratory analysis of the similarity of VET course across and within training packages in Australia. This analysis employed natural language processing to calculate similarity scores for each national VET qualification to all other qualifications across all training packages. The outputs of the model used is available in the Qualification Similarity Technical Appendix on the website. The model used in this preliminary analysis will continue to be refined. This analysis will ultimately assist to build a richer picture of the training market.

9 July 2021

Contents

Introduction

At the height of the COVID-19 pandemic in 2020, the key focus for policy makers and industries was to assist Australians back into work in jobs that were in-demand and resilient to the immediate economic shocks of the public health crisis.

As Australia emerges from the economic impact of the pandemic, there needs to be a shift towards ensuring the education and skills system delivers the skills and knowledge for the economy now and in the future. There is an inherent uncertainty in estimating the future skills needs of the economy, therefore it is crucial to ensure that our education and skills systems are resilient enough to deliver the skills we need despite this uncertainty.

To do this, it is important to better understand the pathways from education to employment. Before the pandemic, the Australian labour market had been progressively shifting towards higher skilled employment and this trend is likely to continue as we recover from the economic impacts of COVID-19 (NSC, 2020). This emphasises the importance of post-school qualifications in the labour market going forward. There is a need for a more detailed understanding of how to better match skills developed in education and training to the skills required by employers. A granular view allows policy makers and education providers to design better qualification pathways that are timely and cost-efficient for users and employers.

The nexus between education and employment is an important aspect identified by the Strengthening Skills: Expert Review of Australia’s Vocational Education and Training System by the Honourable Steven Joyce (Joyce, 2019).

The review identified a need for better careers information and data on skills needs and future demand. This is an important part of the linking piece between the education and employment spheres to drive better advice in careers information, VET sector outcomes, and future skills demand. Qualification reform and rationalisation is an important part of building a resilient VET system to deliver the skills required by employers and industries.

Machine learning techniques have been used previously to explore the labour market. The World Economic Forum (2018) and the former Australian Government Department of Employment, Skills and Small and Family Business (2019) used machine learning and natural language processing techniques in their work to compare occupations. And Kern et al., (2019) analysed twitter data to assess the common personality traits and values associated with a range of occupations. This research can assist students, job seekers and other members of the public to determine occupations that they would be suited to.

Machine learning can handle large quantities of data in a way which could not be done with other techniques. Machine learning techniques can also be applied to better understand the education and skills system to compare the similarity of qualifications. This paper analyses the similarities between VET courses across, and within, training packages in Australia. This has been done using natural language processing, which is a branch of machine learning, to calculate similarity scores. The analysis relied on language models that represent text as vectors which allows the calculation of similarity scores based on the text contained in the course names, the units within the course, the course description and the frequency of keywords within these texts.

The language embedding model, data sets and the model outputs and validation are discussed in detail in the next section.

Training similarity analysis

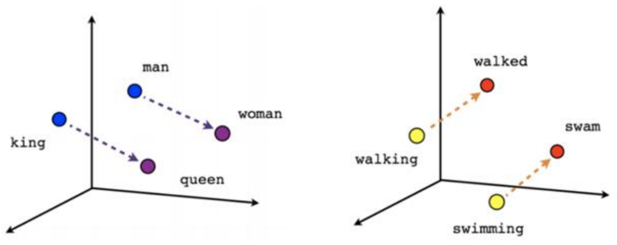
Language Embeddings and BERT

One of the core breakthroughs in natural language processing (NLP) is the invention of language embeddings. Unlike naturally numeric data, text cannot be directly used by machine learning models.

First it must be transformed into numerical representations with the meaning of the text embedded. Language models achieve this by representing text as vectors. These are series of numbers, which reflect different aspects of the original text's semantic meaning. In practice these vectors are created by machine learning models which are trained on comprehension tasks.

The canonical example of language embeddings, word-2-vec, is a shallow neural network model that is trained to perform simple language comprehension tasks (Mikolov et al., 2013). An example is predicting the missing word in a sentence. Since this can be applied to any text, a large set of text (a corpus) is used to train the model, which helps the model understand a broad range of contexts. The vectors representing the text start randomly initialised. The model adjusts these during training to best capture the meaning of the text. The final output is a dictionary of word vectors that can be used for further NLP models and applications. As seen below, these vectors encode semantic meaning in vector-space so that, when plotted, neighbouring words have related meaning and intuitive distance properties.

Figure 1: Word-2-Vec encodes words in vector-space with intuitive distance properties



Male-Female

Verb tense

Source: *Google, 2020*

While ground-breaking at the time of release, the original word-2-vec model had several limitations. This was due to a combination of its algorithmic simplicity, training data size, and its failing to consider the context of words in a sentence as well as words with multiple meanings.

The current state-of-the-art in language embeddings are attention-based transformer models. Compared to word-2-vec, these transformer models are much larger. For example, the recent Open AI's GPT-3, released in June 2020, has 175 billion machine learning parameters, requires computing infrastructure worth around $50 million in order to train, and used a scrape of the entire internet as training data (i.e. 45,000GB of text) (Brown et al., 2020).

GPT-3 is a research model and is not practical for most small-scale NLP applications. However, the previous-generation transformer models such as Google's Bidirectional Encoder Representations from Transformers (BERT) model, which broke several NLP benchmarks upon its release in 2018, is available for public use and is maintained by Google (Devlin et al., 2019). Transformer models provide high performance to ordinary users by taking advantage of the asymmetry in resources required for model training and model inference (use in applications). Model training requires dedicated large-scale infrastructure and enormous datasets. However, once trained, these models can be downloaded pre-trained and model inference performed on a typical high-end laptop to be used in applications.

Data

The model uses VET administrative data to determine the degree of similarity between each qualification against all other qualifications. The analysis included all training package qualifications current as at December 2020 listed on the National Register of VET. Accredited courses and nationally accredited skill sets are not currently included due to the lack of detailed course information currently published on the national register. These could be included in a future update if the data is made available. The current version of the model has 1312 qualifications in-scope.

The information available about VET training package qualifications can be broken down into several components that can be used in this analysis. As displayed on the National Register of VET, a qualification has a title, a description of a few paragraphs, and contains a list of core and elective units which each have their own titles and descriptions.

Model

BERT can encode multi-word text, unlike its predecessor word-2-vec. Therefore, each text components mentioned above can be encoded by BERT directly, resulting in 5 sources of embeddings:

- Qualification Title
- Qualification Description
- Qualification Keywords (extracted from the description)
- Core Unit Titles
- Elective Unit Titles

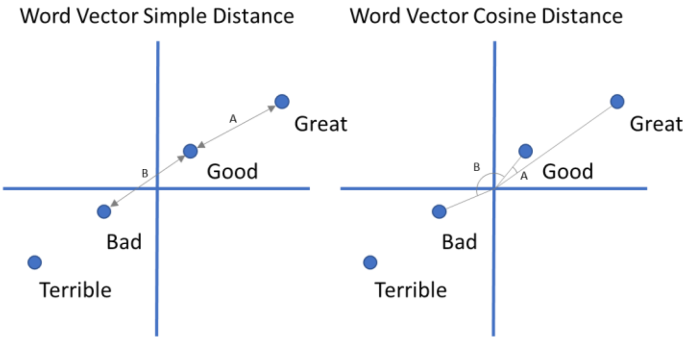
In cases where there are multiple vectors (such as several core units within a qualification), vector averaging was applied.

Each qualification's 5 components are then used to calculate a weighted average, and these weights were fine-tuned to improve performance. This results in a single BERT vector for each qualification which represents a combination of the original text.

The final step is to compare qualifications using vector distance to quantify how similar the vectors are. While this could be achieved with ordinary vector distance, cosine similarity is usually the preferred method in NLP. This is because cosine similarity captures not just the difference between texts, it can also account for nuances within the strength and meaning of language. It achieves this by calculating the difference in angles between the vectors rather than their distance.

In the example below, the vector distance between ‘good’ and ‘bad’ is the same as ‘good’ and ‘great’. This would indicate a similar level of difference in meaning between these pairs of words despite ‘good’ and ‘bad’ being opposites while ‘good’ and ‘great’ both indicate varying levels of positivity. However, the angular differential between ‘good’ and ‘bad’ is much larger than that between ‘good’ and ‘great’.

Figure 2: An illustration of vector distance and cosine distance



Source: *NSC analysis, 2021*

Hence the overall methodology is to encode the components of qualifications using BERT vectors, average them together using a weighted average, and then perform cosine similarity on the final vectors. This produces a final measure of similarity between qualifications.

Model output and validation

The final output consists of an overall similarity score for qualification compared to every other qualification. The output also includes the similarity scores for each text component so that users can identify how the overall similarity score was calculated. These can be ranked, searched through, and filtered by both input and output as demonstrated in the interactive appendix. Further details of the interactive appendix can be found in Appendix A.

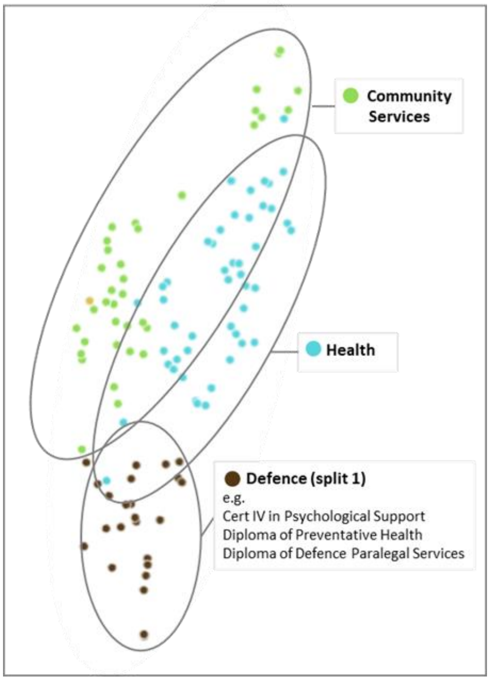
There is currently no single model which compares the similarity of course design within the Australian tertiary sector. Therefore, the results were manually checked by analysts to assess model performance. The top 20 most similar qualifications for the most popular training package courses were manually assessed. The weighting of each text component has been adjusted to place less emphasis on keywords and elective units which will be discussed in further detail in the next section.

Discussion

The output of the model is the similarity between pairs of qualification. To visually compare these, it is useful to see how large groups of qualifications relate to each other. This can be achieved using the dimension reduction algorithm, TSNE, which is able to compress the similarity of all qualification pairs into a 2-dimensional projection (van der Maaten & Hinton, 2008). This is shown below, for the example training packages, in figure 3.

In this projection, some distinct clusters appeared. Not surprisingly, some of these clusters mirror the existing training packages. This is due to the way VET training packages are bundled. Each training package services an industry, or a number of related industries, which require related skills such as construction or community services. Training packages also contain qualifications across multiple qualification levels (Certificate I up to Graduate Diploma) that focus on the same area of study.

Figure 3: The t-SNE algorithm produces a scatterplot that shows the relationship between Community Services, Health, and Defence



Source: *NSC Analysis, 2021*

There were some interesting findings that emerged. Some training packages were much more closely linked than others. For example, courses from the Health training package and the Community Services training package share a significant intersection. Examining this intersection in more detail shows this intersection with links to both Health and Community Services includes qualifications in areas such as disability, population health, allied health assistance and mental health. On the other end of the spectrum, there are courses which focus strictly on Health or Community Services such as pathology collection and youth justice respectively.

There were also some training packages that split into distinct clusters which indicates a diverse range of skills taught within the same training package. For example, the Defence training package splits into two clusters – with split one containing qualifications related to health, legal services, and psychological support, while split two focused on defence technology such as explosive ordinance. Further details can be seen in Appendix B.

Like all NLP models, the output is not perfect. NLP models consistently find words with multiple meanings challenging to process. One example in this analysis were qualifications in strata community management within the Property Services training package. Due to the high frequency of the word ‘community’ used within the course description along with words like ‘facilitate’ and ‘support’, these were considered by the model to be highly similar to the Community Services training package. To address this issue, the weighting on keywords was lowered.

The preliminary output also produced some unusual similarity pairings due to the elective units of competency within the qualifications. Some qualifications have a very high number of elective units and relatively few core units. For example, the Certificate III in Agriculture has 133 current elective units and only 2 current core units. Therefore, it would be impossible for anyone completing this qualification to undertake all the elective units on offer. To overcome this, the weighting on individual elective units was reduced.

Another limitation identified was due to the information available about VET courses. Course descriptions often contain words such as assessment, training, certification and regulatory requirement. This means that qualifications in the Teaching and Assessment training package have a lot of false positive matches across many other VET qualifications. This issue is unique to the Training and Assessment training package qualifications.

Conclusion

The Australian VET system is complex. There are over 15,000 units of competency across the 56 training packages. Analysing this using traditional methods would be time consuming and difficult. Traditional methods of manually analysing and codifying text are time consuming and prone to sampling errors and biases based on the reliance of human judgement.

The machine learning techniques and NLP described in the paper can be used to better understand the qualifications and skills being taught through the Australian VET system. Jobs and Skills Australia will continue to build on this exploratory training similarity analysis. It has the potential to help identify similar training products which can assist in the simplification of the VET system and qualification design. This work will also deepen the current understanding of the links between jobs, skills and training. The sharing of intelligence aims to enrich Australia’s capacity to better understand and adapt to the changing labour market.

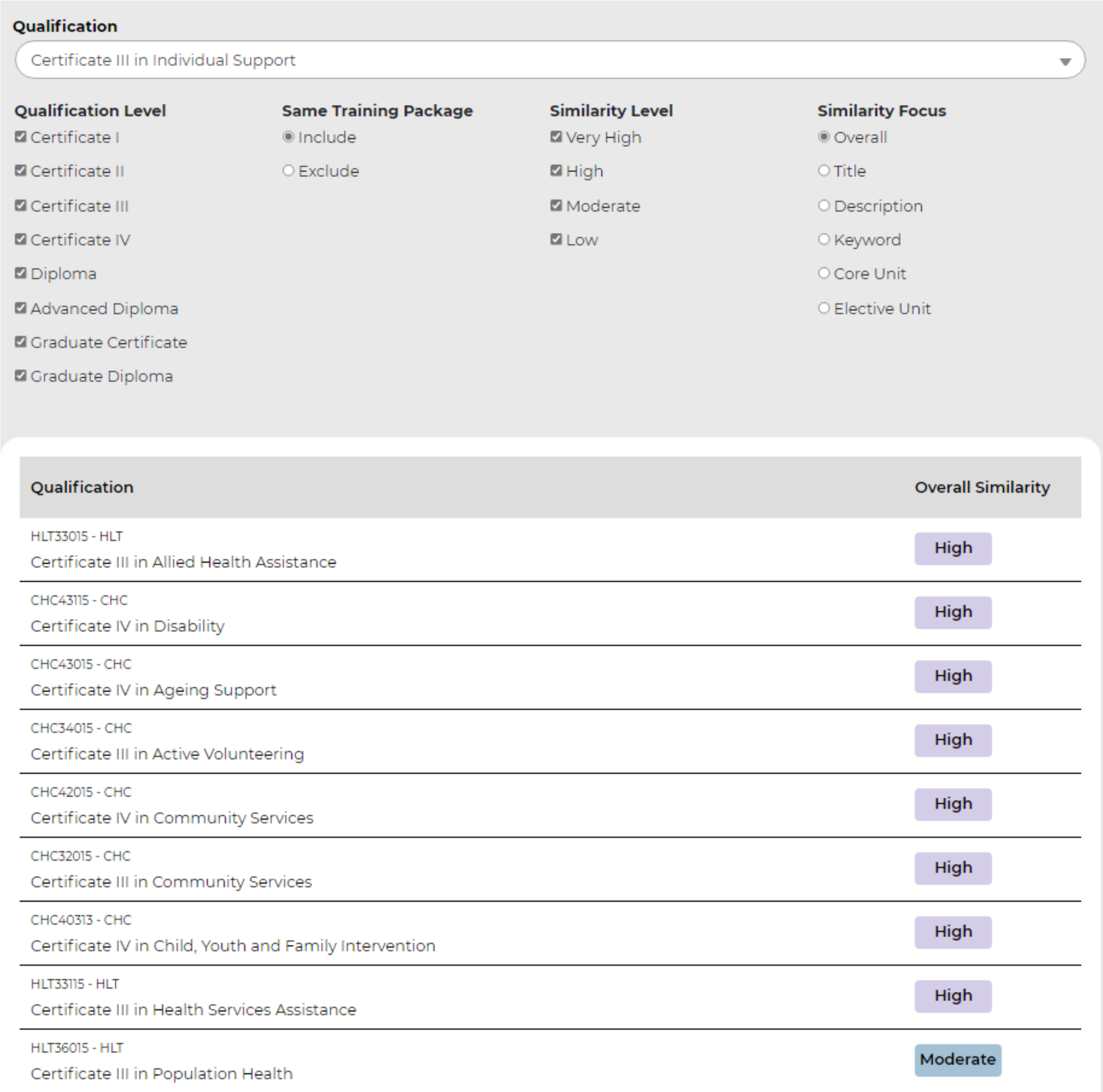
All feedback regarding potential use cases and model improvements are welcome via: skillsintelligence@jobsandskills.gov.au

References

1. National Skills Commission (2020), *The shape of Australia's post COVID-19 workforce,* Commonwealth of Australia, Canberra, Australia.
2. Joyce, S. (2019), *Strengthening Skills: Expert Review of Australia's Vocational Education and Training System,* Commonwealth of Australia, Canberra, Australia.
3. World Economic Forum (2018) *Towards a Reskilling Revolution: A Future of Jobs for All*, World Economic Forum, Geneva, Switzerland.
4. Department of Employment, Skills, Small and Family Business (2019), *Reskilling Australia - a data driven approach,* Commonwealth of Australia, Canberra, Australia.
5. Kern, M.L., McCarthy, P.X, Chakrabarty, D., & Rizoiu, M. (2019) *Social media-predicted personality traits and values can help match people to their ideal jobs,* Proceedings of the National Academy of Sciences, Melbourne, Australia.
6. Google (2020), *Embeddings: Translating to a Lower-Dimensional Space*, Google Developers Site: https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space
7. Mikolov et al., (2013), *Efficient Estimation of Word Representations in Vector Space*, Cornell University, arXiv:1301.3781
8. Brown et al., (2020*), Language Models are Few-Shot Learners*, Cornell University, arXiv:2005.14165
9. Devlin et al., (2019), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Cornell University, arXiv:1810.04805
10. van de Maaten, L. & Hinton, G., (2008), *Visualizing Data using t-SNE*, Journal of Machine Learning Research

Appendix A: Qualification Similarity Interactive Appendix

Figure 4: Screenshot of the Qualification similarity interactive appendix displaying the overall similarity of Certificate III in Individual Support to other qualifications across training packages



Qualification

Certificate III in Individual Support ▾

Qualification Level

- ☑ Certificate I
- ☑ Certificate II
- ☑ Certificate III
- ☑ Certificate IV
- ☑ Diploma
- ☑ Advanced Diploma
- ☑ Graduate Certificate
- ☑ Graduate Diploma

Same Training Package

- ● Include
- ○ Exclude

Similarity Level

- ☑ Very High
- ☑ High
- ☑ Moderate
- ☑ Low

Similarity Focus

- ● Overall
- ○ Title
- ○ Description
- ○ Keyword
- ○ Core Unit
- ○ Elective Unit

Qualification	Overall Similarity
HLT33015 - HLT Certificate III in Allied Health Assistance	High
CHC43115 - CHC Certificate IV in Disability	High
CHC43015 - CHC Certificate IV in Ageing Support	High
CHC34015 - CHC Certificate III in Active Volunteering	High
CHC42015 - CHC Certificate IV in Community Services	High
CHC32015 - CHC Certificate III in Community Services	High
CHC40313 - CHC Certificate IV in Child, Youth and Family Intervention	High
HLT33115 - HLT Certificate III in Health Services Assistance	High
HLT36015 - HLT Certificate III in Population Health	Moderate

Source: *NSC website, 2021*

The qualification similarity interactive appendix is an interactive tool which compares the similarity of one qualification to all other qualifications in training packages in Australia.

The interactive appendix is available on the Jobs and Skills Australia website.

Accredited qualifications (those outside training packages) are not included in this tool.

The similarity of a qualification has been graded as Very High, High, Moderate, and Low.

These gradings are based on cut-offs of the similarity scores to all other qualifications based on the overall similarity focus.

The top 20 matches for each qualification were segmented into percentiles for their similarity scores:

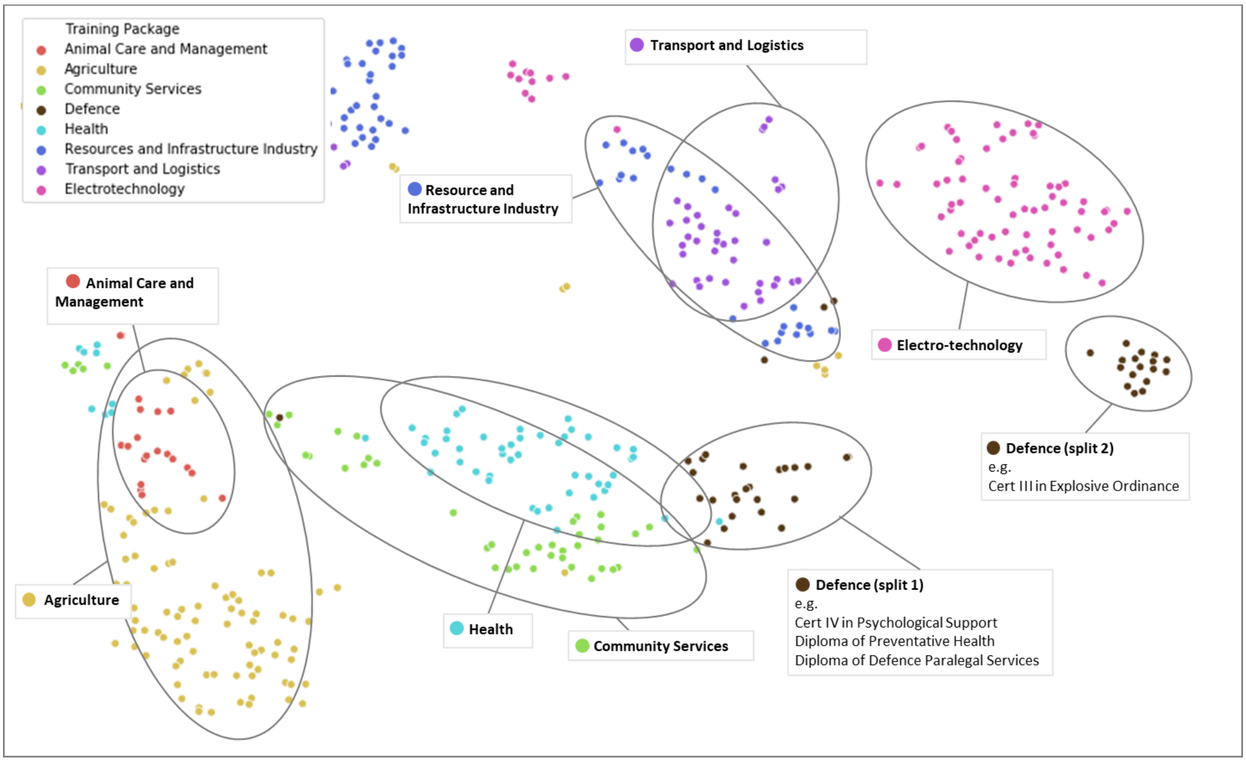
- top 25 per cent of similarity scores graded as ‘very high’,
- second 25 per cent of similarity scores graded as ‘high’,
- third 25 per cent of similarity scores graded as ‘moderate’
- remainder of similarity scores graded as ‘low’.

Filters can be applied to view the rankings of similar qualifications based on:

- Title similarity
- Description similarity
- Keyword similarity
- Core unit Similarity
- Elective unit similarity.

Appendix B: t-SNE algorithm output for the training product similarity analysis

Figure 5: The t-SNE algorithm produces a scatterplot showing a selection of training packages and how their qualifications relate to each-other in terms of similarity.



Source: *NSC analysis, 2021*

The t-SNE scatterplot in the discussion section (figure 3) highlighted links in qualifications for the training packages community services, health, and defence.

Figure 5 above presents a fuller picture of the qualification landscape by showing some of the training packages with the most qualifications.

As in figure 3, this shows how individual qualifications relate to each other in terms of similarity and how clusters of qualifications intersect within and across training package.