Technical Report

Building Subject term Taxonomy - APO

SWINBURNE UNIVERSITY OF TECHNOLOGY

Created by: Jihoon Woo
jihoonwoo@swin.edu.au

Table of Contents

Introduction

APO (Analysis & Policy Observatory), the organisation providing essential policy literature resources for people engaged in policy work, uses a combination of some portion of subject terms drawn from Faceted Application of Subject Terminology (FAST) and the subject terms that the domain experts have manually identified.

Currently, APO has collected around 5,700 subject terms over a number of years to index the grey literature of policy, research and information management. However, there are some issues in facilitating these terms to make effective searching in APO. Firstly, APO has reported that there are irrelevant and non-reusable subject terms in the controlled vocabulary. One factor behind the high frequency of non-reusable subject terms comes from the fact that many subject terms were added as part of a bulk metadata import project. Another factor is that curators sometimes had difficulties to review and choose relevant subject terms against the existing subject terms. Secondly, due to the above situation, it was easy to inadvertently create new subject terms that overlapped semantically with existing subject terms. Thirdly, the subject terms could not be navigated due to a lack of subject term reference structure.

Building a taxonomy of subject terms enables us to represent highly related subject terms together, and their paths indicate how these terms are semantically associated with one another.

In this report, we introduce (1) identification of missing subject terms to refine subject terms, (2) building word2vec model with APO documents, (3) merging infrequent subject terms to primary subject term, and (4) inducing a subject term taxonomy from the primary subject terms.

1. Identification of Missing Subject Terms

This section will describe the process of identifying missing subject terms. The process begins with exploring input datasets and introduce a string matching method to identify missing subject terms and finally, showing statistic results after identifying missing subject terms.

1.1 Input data exploration

APO provides two datasets 'apoDescriptions' and 'apoSubjects'. The 'apoDescriptions' contains information of collected literature such as document id, title, description, summary and subject terms which are manually assigned to a given document by domain experts While 'apoSubject' contains around 5,700 of subject terms that are used index the grey literature of policy, research and information management. Below table shows the number of rows and columns for two input data files.

	apoDescriptions	apoSubjects
Columns	**Nid, Title, Description, summary, Subject,** new subject	Term Id, **Term**
No of rows	40,553	5,725

Table1.1.1 Statistic information for two input data files

Columns names	Number of missing values
Nid	0
Title	0
Description	48
Summary	31633
Subject(s)	2370
New subjects	40552

Table1.1.2. Identify missing values in apoDescriptions data

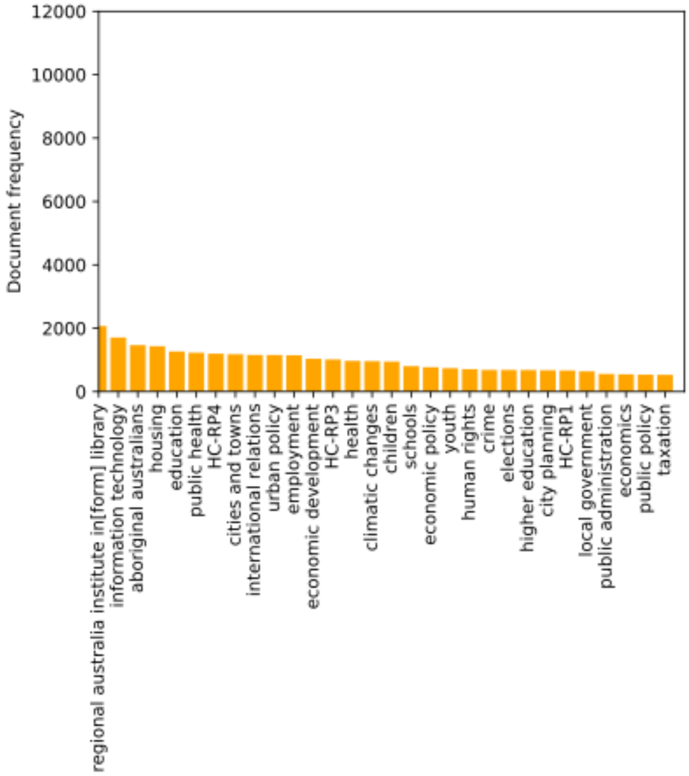
As 1.1.2 table shows that "New subjects" column is meaningless as all values are missed. Therefore, "New subjects" column is removed for our tasks. Meanwhile, all other columns are considered to be used for analysis even though 78% of values in the summary column are missed. On the other hand, there are no missing values in apoSubjects data file.

1.2 Identify unique subject terms and their usages

In apoDescriptions data file, “Subject(s) column” contains the subject term(s) which are assigned by domain experts for a given document. There are **5,654** unique subject terms exist over all documents. We identify unique subject terms and their usages by counting their document frequency. 1.2.1 table and 1.2.2 figure show top-n of the most popularly used subject terms over all the documents.

Subject Terms	Document Frequency
Regional Australia Institute In[Form] Library	2070
Information technology	1707
Aboriginal Australians	1466
Housing	1433
Education	1267
Public health	1228
HC-RP4	1197
Cities and Towns	1180
International relations	1157
Urban Policy	1155

Table1.2.1 top 10 subject terms with the highest usages



Bar chart 1.2.2 top 30 subject terms with the highest usages

1.3 Procedure to identify missed subject terms

Although we found 5,654 of unique subject terms exist in apoDescriptions data file, we need to identify subject terms that are potentially relevant but previously missed as indexed subject terms by a string-matching technique. Our assumption is that if a subject term appears in a document, it can be a candidate of the document's subject terms. Thus, we check if the document contains each term in existing subject terms in its text content.

The process of identifying missing subject terms are described as follows:

1. In apoDescriptions data, combine values in three columns (Title, Description and Summary) and saved in new column call 'text'. We search for terms in this 'text' column.

2. In apoSubjects data, divide subject terms into two groups. Uppercase subject terms (i.e. VIC, VCE and 5G) and lowercase subject terms (i.e. class, government, and public).

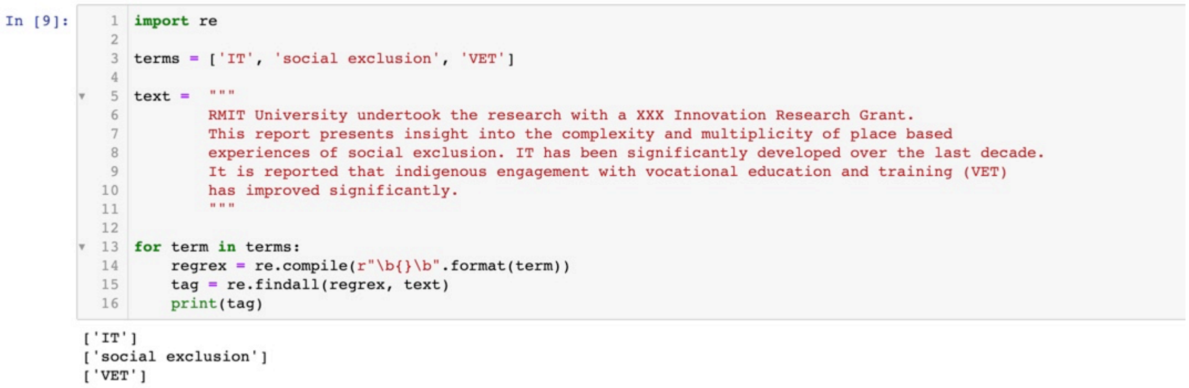
3. Remove all special characters in text (i.e. non-alphanumeric characters, or numeric characters)

4. Use Regular Expression to find subject terms in the text contents.

Suppose that the subject terms 'social exclusion', 'VET', and 'IT' exist in the text which are denoted in boldface in the text.

"RMIT University undertook the research with a XXX Innovation Research Grant. This report presents insight into the complexity and multiplicity of place-based experiences of ***social exclusion****.* ***IT*** *has been significantly developed over the last decade. It is reported that indigenous engagement with vocational education and training (****VET****) has improved significantly."*

Below screenshot shows how the Regular Expression works to find subject terms in text.



```
In [9]:
import re

terms = ['IT', 'social exclusion', 'VET']

text = """
        RMIT University undertook the research with a XXX Innovation Research Grant.
        This report presents insight into the complexity and multiplicity of place based
        experiences of social exclusion. IT has been significantly developed over the last decade.
        It is reported that indigenous engagement with vocational education and training (VET)
        has improved significantly.
        """

for term in terms:
    regex = re.compile(r"\b{}\b".format(term))
    tag = re.findall(regex, text)
    print(tag)
```

```
['IT']
['social exclusion']
['VET']
```

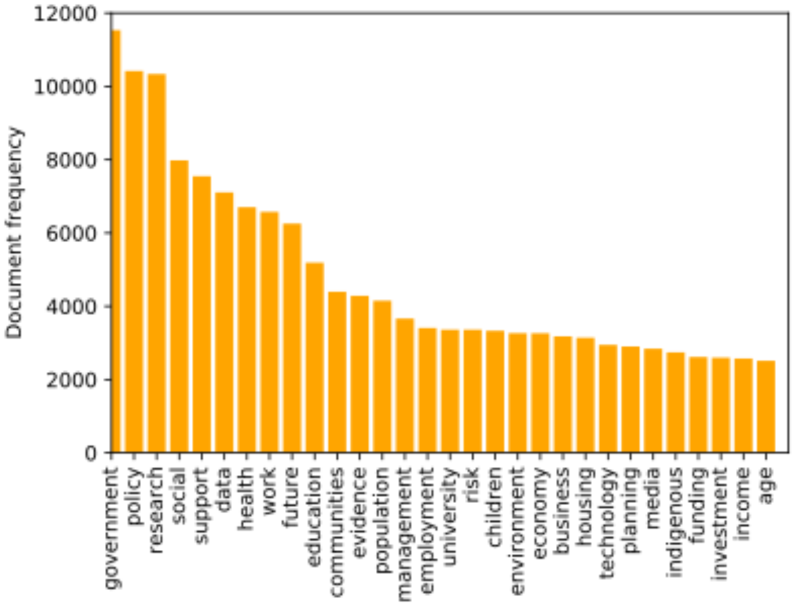
Screenshot 1.3.1 Example of exact matching by using Regular Expression in python

1.4 Missed subject terms

Total 48 of additional unique subject terms are found over all documents. Below 1.4.1 table and 1.4.2 graph shows the top-n of the most popularly used subject terms after identifying missed subject terms. As the table shows, the usages of subject terms (Document Frequency) are dramatically increased compared to 1.2.1 table, and figure 1.4.2 also shows new top 30 subject terms with the greatest usages.

Subject Terms	Document Frequency
'government'	11534
'policy'	10420
'research'	10337
'social'	7982
'support'	7549
'data'	7107
'health'	6703
'work'	6576
'future'	6256
'education'	5190

Table1.4.1 top 10 subject terms with the greatest document frequency (After identifying missing subject terms)



Bar chart 1.4.2 top 30 subject terms with the greatest document frequency (After identifying missing subject terms)

1.5 Visualizing comparison before and after identifying missing subject terms

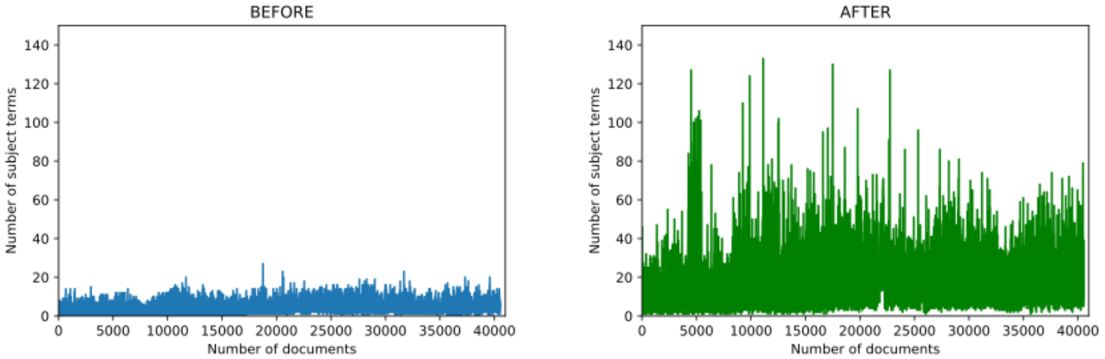


Figure 1.5.1 comparision between two distributions of subject terms used to index the APO documents

Figure 1.5.1 shows a comparison between two distributions of subject terms used to index the APO documents (Before identifying missing subject terms (left) and After identifying missing subject terms (right)). Y-axis represents the number of assigned subject terms over each document and x-axis denotes the document index. We calculate the average number of subject terms used to index each document as well as MAX and MIN numbers (see table 1.5.2). Both figures showing that using the proposed method can identify additional subject terms that can be used to index documents by

	Before identifying missing subject terms	After identifying missing subject terms
Number of unique subject terms	5,654	5,702
Avg number of subject terms used to index each document	4	15
Max number of subject terms used to index each document	27	134
Min number of subject terms used to index each document	0	1

Table1.5.2 statistic information of comparison before and after identifying missing subject terms

2. Generate Word2vec Model

Word2vec takes a corpus of text and as the output, it produces a vector space where similar words are positioned close to one another. In order to calculate the similarity between infrequent subject term and frequent subject term for merging (It will be described more detailly in section 3), we need to have word2vec model that is trained by APO corpus.

Initially, we remove stop words and apply to lemmatize to APO corpus, then train the word2vec model with words appearing in the corpus. As the output, each word in the embedded model is represented by a numerical vector so that we measure the similarity between words using the cosine similarity between their corresponding vectors.

We use genism library to train word2vec model where required parameters are set as follows:

- Size = 100 (length of converted vector)
- Window = 5 (Maximum distance between the current and predicted word within a sentence)
- Min_count = 3 (Ignores all words with a total frequency lower than this)
- Workers = 32 (Use these many worker threads to train the model (=faster training with multicore machines)
- Sg = 1 (skip-gram algorithm)
- Negative = 5 (how many noise words should be drawn)

3. Merging Subject terms

3.1 Determine merging subject term group

Although we have identified missed subject terms, all of the subject terms may not be usefully for indexing the APO corpus. Figure 3.1.1 shows the frequency distribution of the subject terms used for indexing the APO corpus. The x-axis shows the number of the subject terms sorted by their document frequencies, and the y-axis shows their document frequencies. The document frequency indicates the number of times a given subject term is used to index the documents in the given corpus. As we can see, the imbalance ratio of the document frequencies of the subject terms is very high, where some terms are dominantly (very frequently) used (see the subject terms < 500 (x-axis value)) but the majority of the subject terms are rarely (very infrequently) used (e.g. see subject terms > 1000 (x-axis value)). Also, as observed, there is a very long tail of the subject terms that are rarely used.

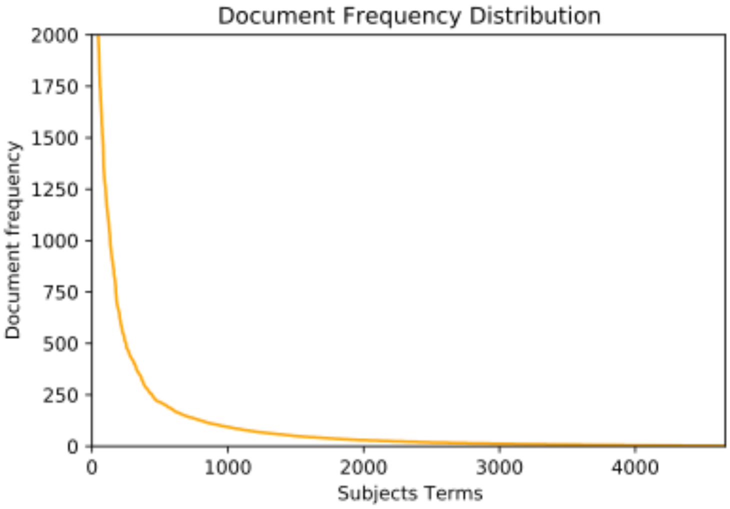


Figure 3.1.1 frequency distribution of the subject terms

In fact, table 3.1.2 shows percentile of document frequency where 80 percent of the document has less than 78 document frequency. We define **min-df** to 78 that it can be used to distinguish frequent subject terms (document frequency equal or greater than 78) and infrequent subject terms (document frequency less than 78) based on their document frequency.

Percentile	Document Frequency
50	14
60	24
70	41
80	78

Table3.1.2 percentile of document frequency

3.2 Judge Secondary subject term's usefulness

Frequent subject terms (Primary subject terms) will be used for inducing subject term taxonomy as they are very frequently used for indexing APO corpus while infrequent subject terms (secondary subject terms) need to be determined whether it can be merged to one of the primary subject terms or discard to reduce the size of the vocabulary. Our approach is using wordnet and word2vec techniques to calculate the similarity between the secondary subject term and primary subject term.

3.3 Merging infrequent subject terms by using wordnet

Wordnet (NLTK) is a freely available software package that makes it possible to measure the semantic similarity and relatedness between a pair of concepts (or synsets). However, we only measure similarities between the single length of subject terms (e.g. government) as the wordnet won't be able to search similarities for multi-length words (e.g. Victoria government).

Wordnet process:

1. Import wordnet from NLTK
2. Load an information content file from the wordnet_ic corpus (ic_brown.dat)
3. Lemmatize all subject terms (e.g. books -> book)
4. Calculate similarity with "Jiang-Conrath Method" between each of primary subject term and secondary subject term. (Jiang-Conrath Similarity Return a score denoting how similar two-word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node) and that of the two input Synsets)
5. Choose a primary subject term for each of secondary subject term with the highest similarity score.
6. If the highest similarity score is greater than **0.7** then we merge the secondary subject term to the matched primary subject term.

3.4 Merging infrequent subject terms by using word2vec

Using generated word2vec model in section 2, we can convert all subject terms to vector values then measure the similarity between words using the cosine similarity between their corresponding vectors. As same as wordnet processing, we choose a primary subject term for each of secondary subject term with the highest similarity score and discard secondary subject terms if it has no matched primary subject term with similarity over 0.7

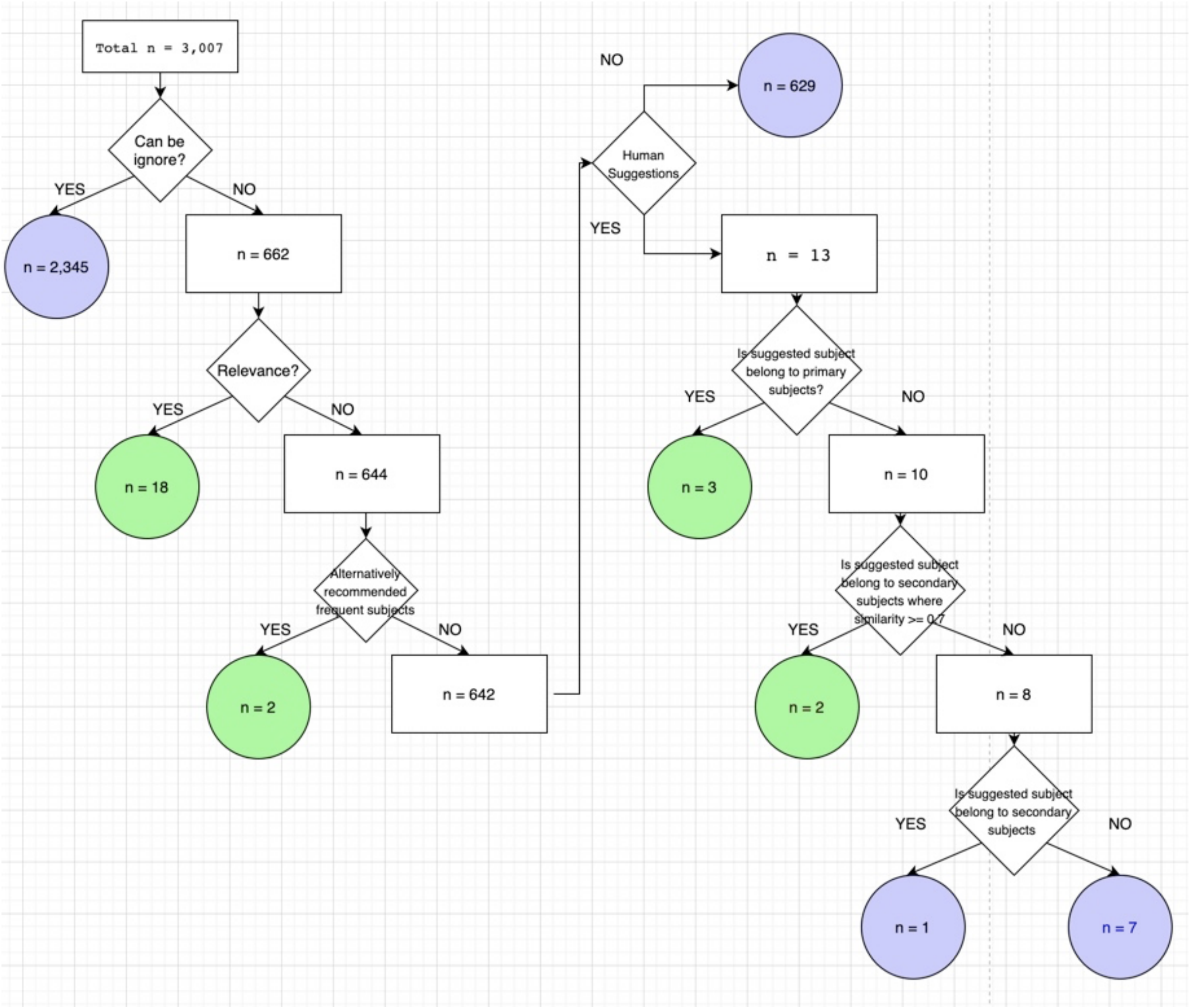
Total number of subject terms	5702
Number of secondary subject terms	4556
Number of secondary subject terms can be merged into one of the primary subject terms	3007

Table 3.4.1 Merging subject terms result

Table 3.4.1 shows total 3,007 of secondary subject terms can be merged into one of the primary subject terms after we apply the wordnet and word2vec model to identify the similarity between primary/secondary subject terms.

3.5 Evaluation of merging subject terms

For evaluation of merging 3,007 pair of subject terms, we request APO to give judgement their relatedness.


Total n = 3,007
Can be ignore?
YES
NO
n = 2,345
n = 662
Relevance?
YES
NO
n = 18
n = 644
Alternatively recommended frequent subjects
YES
NO
n = 2
n = 642
Human Suggestions
NO
YES
n = 629
n = 13
Is suggested subject belong to primary subjects?
YES
NO
n = 3
n = 10
Is suggested subject belong to secondary subjects where similarity >= 0.7
YES
NO
n = 2
n = 8
Is suggested subject belong to secondary subjects
YES
NO
n = 1
n = 7

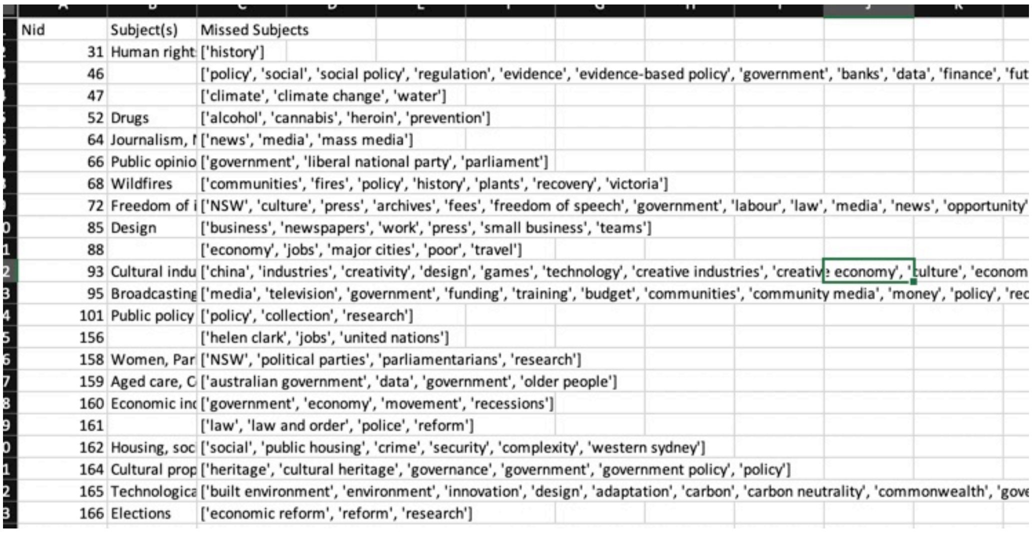
Above diagram shows the result of analysing merging evaluation. There are 3,007 secondary subject terms in the evaluation dataset, and 2,345 of secondary subject terms (78.0%) are considered to be ignored while 662 of secondary subject terms (22.0%) are not to be ignored. Thus, we directly ignore those 2,345 of secondary subject terms and only focus on 662 secondary subject terms. 18 out of 662 secondary subject terms are judged to be related with our suggested primary subject term while 644 does not. Further, we found 5 (2+3) cases where domain experts select one of the primary subject terms is related to the secondary subject terms and lastly, 2 cases found where domain experts select one of the secondary subject terms is related to the secondary subject terms. Therefore, we merge 25 of secondary subject terms (sum of numbers in green circles in the diagram) to one of the primary subject terms and add two secondary subject terms to primary subject term group.

4. Build subject term taxonomy

Building a taxonomy of subject terms enables us to represent highly related subject terms together, and their paths indicate how these terms are semantically associated with one another. This section describes the process of building subject term taxonomy, the process consisted with three parts (1) generate taxonomy input data, (2) build subject term taxonomy, and (3) find the optimal value of threshold to build a taxonomy by analysing depth and quality of taxonomies.

4.1 Generate taxonomy input data

In order to generate taxonomy input data, we need two data files. First one is a result of chapter1. Identification of missing subject terms that it consisted of subject terms defined by APO and identified missing subject terms (see figure 4.1.1 screenshot). Second data file result of analysing merging evaluation that contains primary subject terms and secondary subject terms.

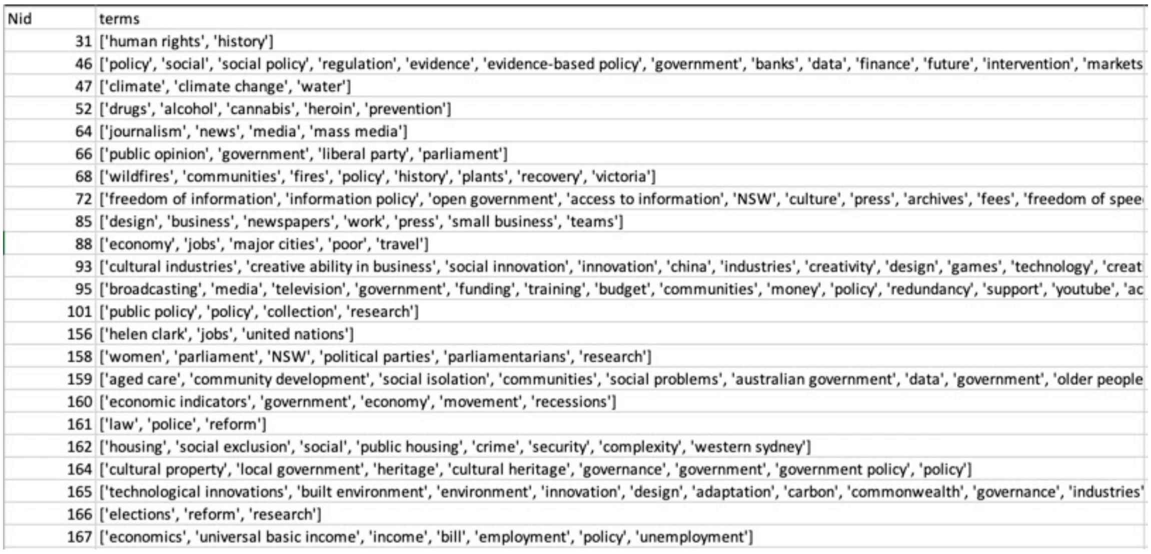


Nid	Subject(s)	Missed Subjects
31	Human right	['history']
46		['policy', 'social', 'social policy', 'regulation', 'evidence', 'evidence-based policy', 'government', 'banks', 'data', 'finance', 'fut
47		['climate', 'climate change', 'water']
52	Drugs	['alcohol', 'cannabis', 'heroin', 'prevention']
64	Journalism, I	['news', 'media', 'mass media']
66	Public opinio	['government', 'liberal national party', 'parliament']
68	Wildfires	['communities', 'fires', 'policy', 'history', 'plants', 'recovery', 'victoria']
72	Freedom of i	['NSW', 'culture', 'press', 'archives', 'fees', 'freedom of speech', 'government', 'labour', 'law', 'media', 'news', 'opportunity'
85	Design	['business', 'newspapers', 'work', 'press', 'small business', 'teams']
88		['economy', 'jobs', 'major cities', 'poor', 'travel']
93	Cultural indu	['china', 'industries', 'creativity', 'design', 'games', 'technology', 'creative industries', 'creative economy', 'culture', 'econom
95	Broadcasting	['media', 'television', 'government', 'funding', 'training', 'budget', 'communities', 'community media', 'money', 'policy', 'rec
101	Public policy	['policy', 'collection', 'research']
156		['helen clark', 'jobs', 'united nations']
158	Women, Par	['NSW', 'political parties', 'parliamentarians', 'research']
159	Aged care, C	['australian government', 'data', 'government', 'older people']
160	Economic inc	['government', 'economy', 'movement', 'recessions']
161		['law', 'law and order', 'police', 'reform']
162	Housing, soc	['social', 'public housing', 'crime', 'security', 'complexity', 'western sydney']
164	Cultural prop	['heritage', 'cultural heritage', 'governance', 'government', 'government policy', 'policy']
165	Technologica	['built environment', 'environment', 'innovation', 'design', 'adaptation', 'carbon', 'carbon neutrality', 'commonwealth', 'gove
166	Elections	['economic reform', 'reform', 'research']

Figure 4.1.1 Screenshot of 'missing_subject_terms.csv'

Firstly, we read '*missing_subject_terms.csv'* file and combine two columns Subject(s) and Missed Subjects and rename the column to 'terms'.

First, we identify the list of unique subject terms per each document then we replace secondary subject terms to the primary subject. Once all secondary subject terms replaced to matched primary subject term and all other infrequent subject terms are removed, the final output will look like Figure 4.1.2 taxonomy input data where a list of unique primary subject terms assign to each document. The result saved as 'taxonomy_input_data.csv'.



Nid	terms
31	['human rights', 'history']
46	['policy', 'social', 'social policy', 'regulation', 'evidence', 'evidence-based policy', 'government', 'banks', 'data', 'finance', 'future', 'intervention', 'markets
47	['climate', 'climate change', 'water']
52	['drugs', 'alcohol', 'cannabis', 'heroin', 'prevention']
64	['journalism', 'news', 'media', 'mass media']
66	['public opinion', 'government', 'liberal party', 'parliament']
68	['wildfires', 'communities', 'fires', 'policy', 'history', 'plants', 'recovery', 'victoria']
72	['freedom of information', 'information policy', 'open government', 'access to information', 'NSW', 'culture', 'press', 'archives', 'fees', 'freedom of spee
85	['design', 'business', 'newspapers', 'work', 'press', 'small business', 'teams']
88	['economy', 'jobs', 'major cities', 'poor', 'travel']
93	['cultural industries', 'creative ability in business', 'social innovation', 'innovation', 'china', 'industries', 'creativity', 'design', 'games', 'technology', 'creat
95	['broadcasting', 'media', 'television', 'government', 'funding', 'training', 'budget', 'communities', 'money', 'policy', 'redundancy', 'support', 'youtube', 'ac
101	['public policy', 'policy', 'collection', 'research']
156	['helen clark', 'jobs', 'united nations']
158	['women', 'parliament', 'NSW', 'political parties', 'parliamentarians', 'research']
159	['aged care', 'community development', 'social isolation', 'communities', 'social problems', 'australian government', 'data', 'government', 'older people
160	['economic indicators', 'government', 'economy', 'movement', 'recessions']
161	['law', 'police', 'reform']
162	['housing', 'social exclusion', 'social', 'public housing', 'crime', 'security', 'complexity', 'western sydney']
164	['cultural property', 'local government', 'heritage', 'cultural heritage', 'governance', 'government', 'government policy', 'policy']
165	['technological innovations', 'built environment', 'environment', 'innovation', 'design', 'adaptation', 'carbon', 'commonwealth', 'governance', 'industries'
166	['elections', 'reform', 'research']
167	['economics', 'universal basic income', 'income', 'bill', 'employment', 'policy', 'unemployment']

Figure 4.1.2 screenshot of taxonomy_input_data.csv'

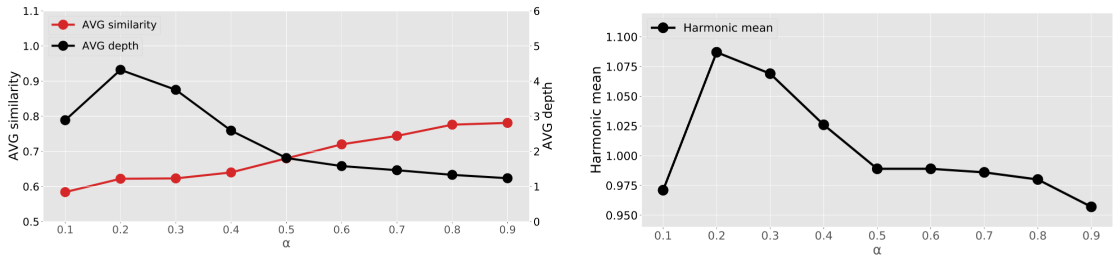
4.2 Build subject term taxonomy

To automatically build a taxonomy from primary subject terms, our approach is to use the subsumption method that is an unsupervised approach for building taxonomy. The fundamental of this method is to use the co-occurrences of subject terms for indexing each document. From the co-occurrence knowledge, we can induce that a subject term A subsumes another subject term B (i.e. A is the hypernym of B) if the documents indexed with B are a subset of the documents indexed with A. However, this method can produce multiple subsumers for a subject term which violates the structure of a taxonomy. Therefore, we calculate the subsumption score of a subsumers for a given subject term and select a subsume with the highest subsumption score.

4.3 Find the optimal value of threshold to build a taxonomy by analysing depth and quality of taxonomies

The key parameter in the subsumption method is the value of the threshold. The higher value of the threshold is, the lower the average depth and the higher the quality of the induced taxonomy. A trade-off thus needs to be considered between a higher average depth and a higher quality of the taxonomic relations.

We apply harmonic mean of the average similarity between all the parent-child nodes in taxonomies and the average depth of taxonomies that taxonomies generated with value of threshold from 0.1 to 0.9. Figure 4.3.1 (a) shows the average similarity between all the parent-child nodes as well as average depth in taxonomies which are generated with the value of threshold from 0.1 to 0.9. Red line represents average similarity, and the black line is average depth. The average similarity getting increased when the value of the threshold is increased (positive relation) while the average depth is decreased when the value of the threshold is increased (negative relation). Figure 4.3.1(b) is a distribution of harmonic mean values where the value of the threshold 0.2 has the highest harmonic value. As the results, we can assume the optimal value of the threshold to build a taxonomy is 0.2



(a) *(b)*

Figure 4.3.1 Distribution of the AVG similarity/depth of taxonomies(a) and distribution of harmonic mean values(b)

4.4 Taxonomy Visualization

Below screenshot shows visualized taxonomy, we use “collapsibleTree” which is one of the R packages to visualize built taxonomy. The R script is saved in /build_taxonomy directory.

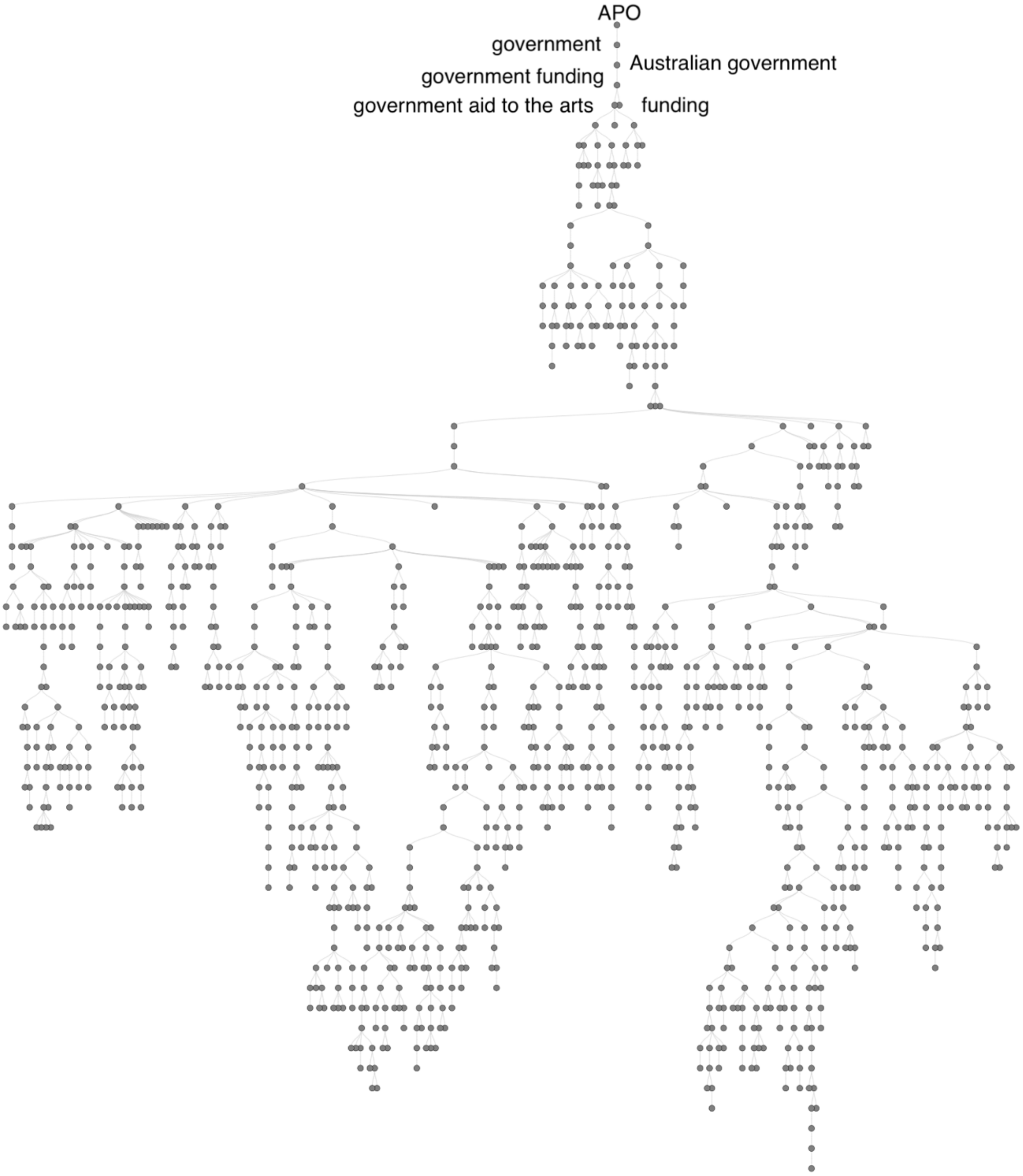


Figure 4.4.1 screenshot of visualized taxonomy

Conclusion

In this report, we presented a methodology for refining an existing set of subject terms that have been already used to index APO documents as well as methodology to build a taxonomy from the refined subject terms. More specially, identification of missing subject terms process leads potentially useful subject terms that have been missed for indexing. And merging infrequent subject terms reduce the imbalance ratio of the document frequencies of the subject. Lastly, building subject terms allows us to identify semantic relation between subject terms which potentially contribute document classification.