Artificial Intelligence to Develop APO Subject Terms

Data Analytics on APO documents and subject terms

Yong-Bin Kang (ykang@swin.edu.au)
Data Science Research Institute
Swinburne University of Technology

Document Indexing Challenges

- Subjective
- Inconsistent
- Inaccurate
- Incomplete (Important subject terms can be missing)
- Duplicated (Don’t know what existing terms are already defined)
- Time-consuming

Document Indexing Challenges

Example Text

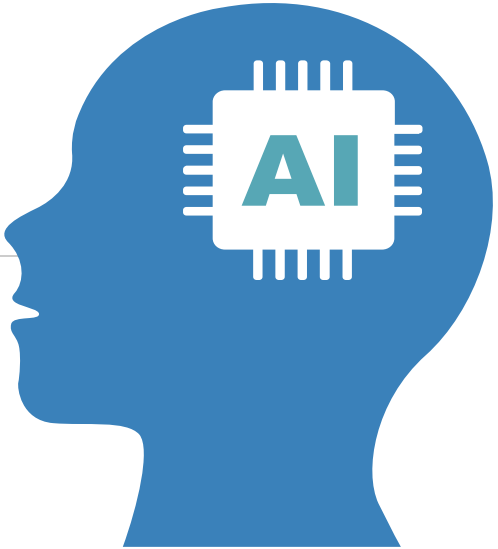
In advance of the United Nations Climate Change Conference in Paris this December, many publics around the world name global climate change as a top threat, according to a new Pew Research Center survey measuring perceptions of international challenges. This is particularly true in Latin America and Africa, where majorities in most countries say they are very concerned about this issue. But as the Islamic militant group ISIS maintains its hold in Iraq and Syria and intensifies its grisly public executions, Europeans and Middle Easterners most frequently cite ISIS as their main concern among international issues

United Nations, survey, publics

United Nations, climate, ISIS

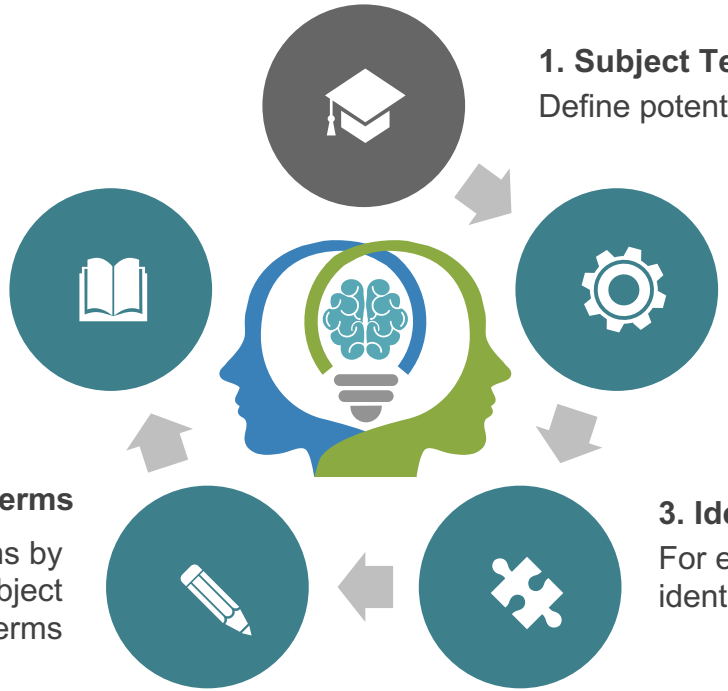
United Nations organisation, Climate Change

Subjective, Inconsistent, Inaccurate, Incomplete, duplicated, time-consuming



How can we solve the problems?

AI Driven Solution



1. Subject Term Dictionary Definition

Define potential subject terms

2. Find Missing Subject Terms

Perform text analysis to identify missing subject terms

3. Identify Synonyms

For each subject term, we identify its synonyms using Wiki

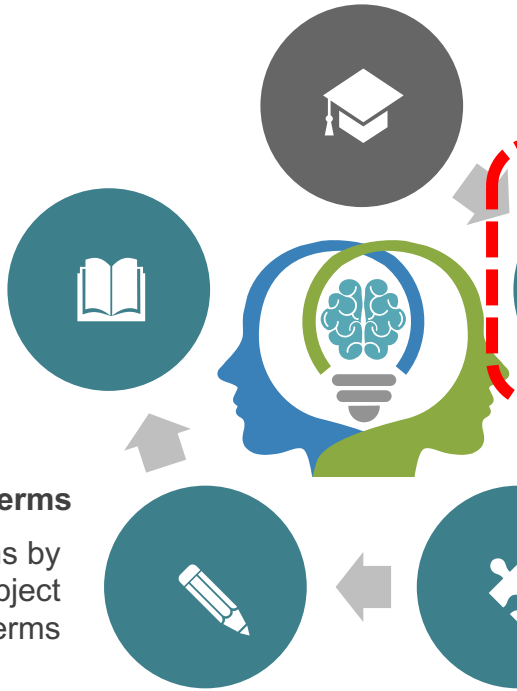
4. Remove Insignificant Subject Terms

Remove insignificant subject terms by merging them with significant subject terms

5. Build Subject Taxonomy

Build a taxonomy of subject terms by understanding their semantic relatedness

AI Driven Solution



1. Subject Term Dictionary Definition

Define potential subject terms

2. Find Missing Subject Terms

Perform text analysis to identify missing subject terms

3. Identify Synonyms

For each subject term, we identify its synonyms using Wiki

4. Remove Insignificant Subject Terms

Remove insignificant subject terms by merging them with significant subject terms

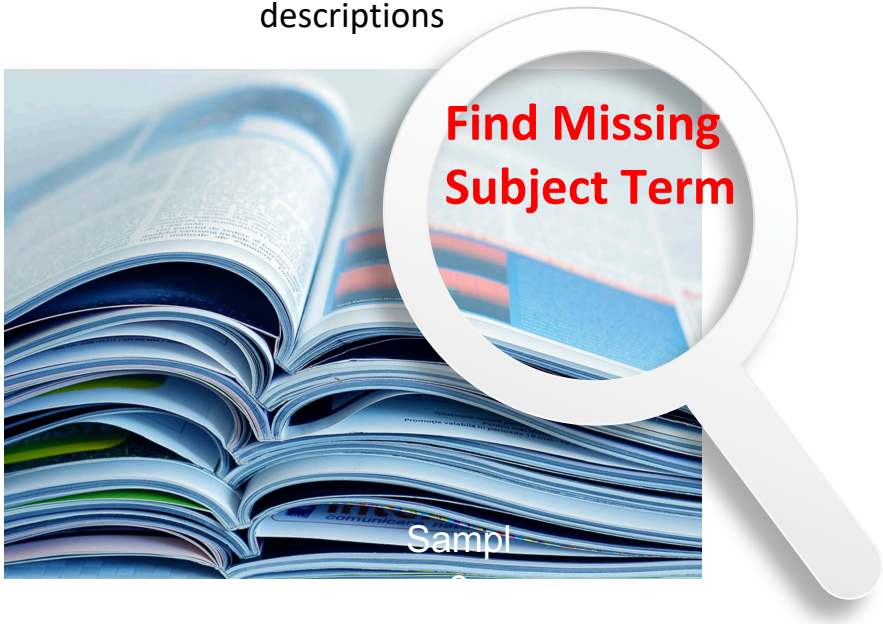
5. Build Subject Taxonomy

Build a taxonomy of subject terms by understanding their semantic relatedness

Finding Missing Subject Terms

Pre-defined Subject Term Dictionary (5,725 terms)

40,554 APO document descriptions

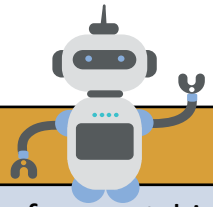


Finding Missing Subject Terms

Example: RMIT University undertook the research with a VicHealth 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.

Subject terms

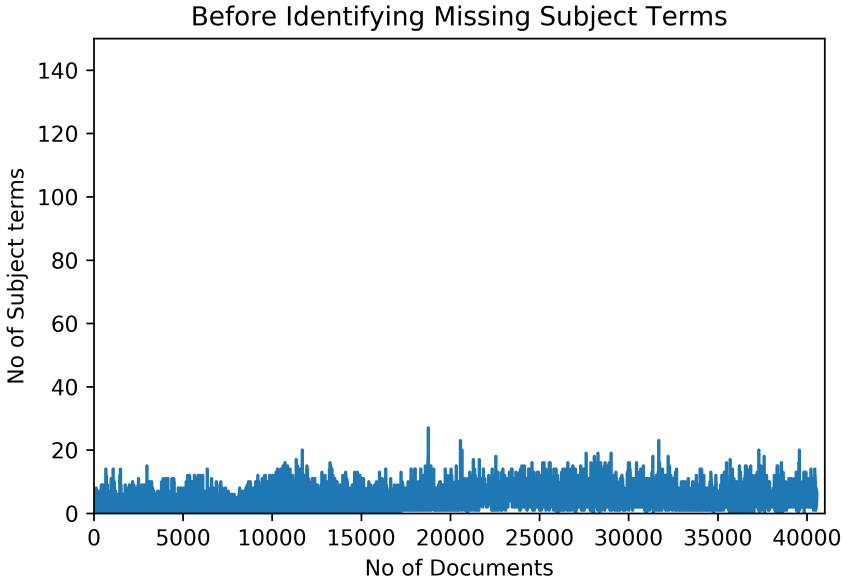
Noise factors

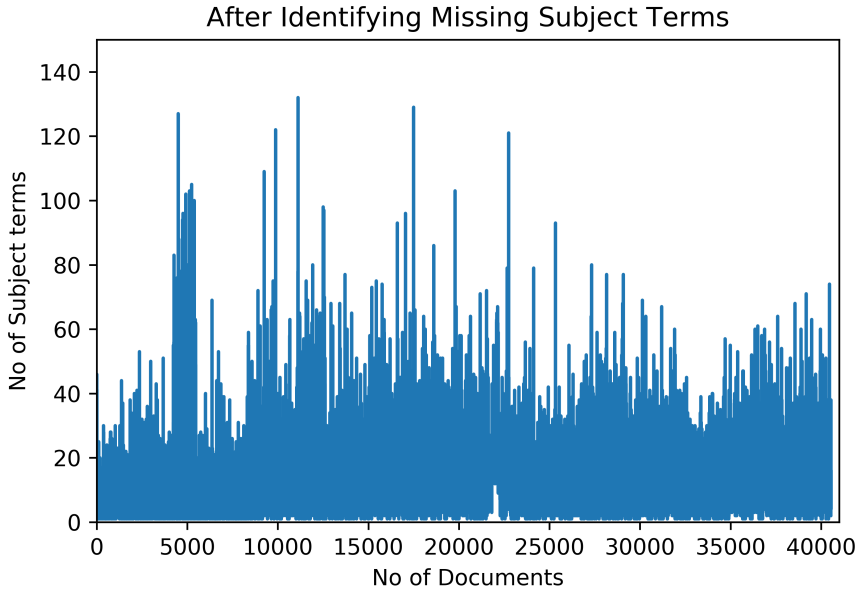


Main Challenges	Solutions
Subject terms can be concatenated with special characters (e.g. "social exclusion" + ".", "(VET)")	Delete all special characters before matching
Some terms consisting of the same letters with capital subject terms (e.g. "IT" vs "It")	Distinguish all uppercases, mixture of uppercases & lowercases and all lowercases letters
Some terms consisting of capital letters can contain subject terms (e.g. "IT" in "RMIT")	Extract Subject terms considering spaces before and after them

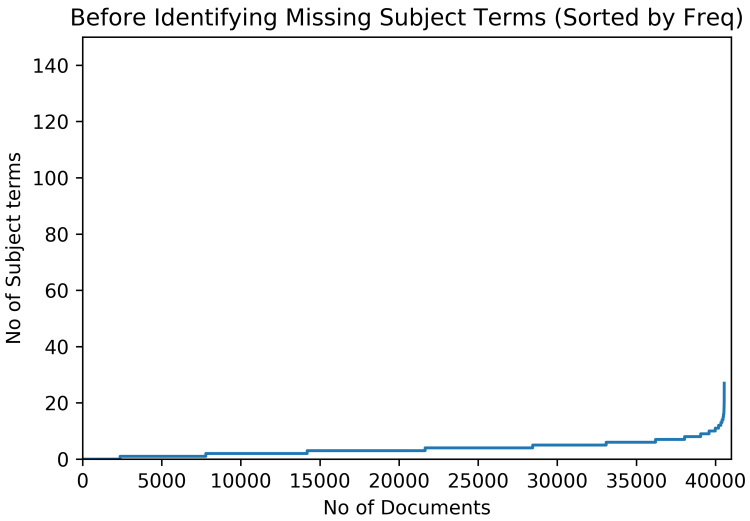
Simple string matching cannot capture all subject terms in the dictionary !

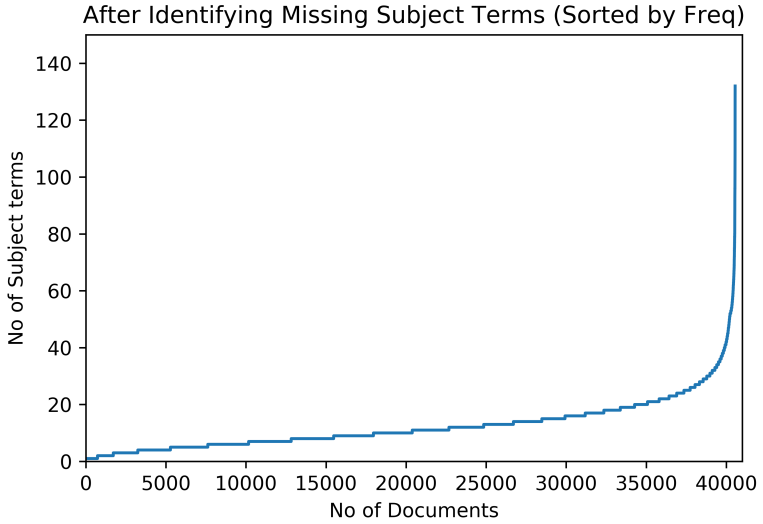
Finding Missing Subject Terms





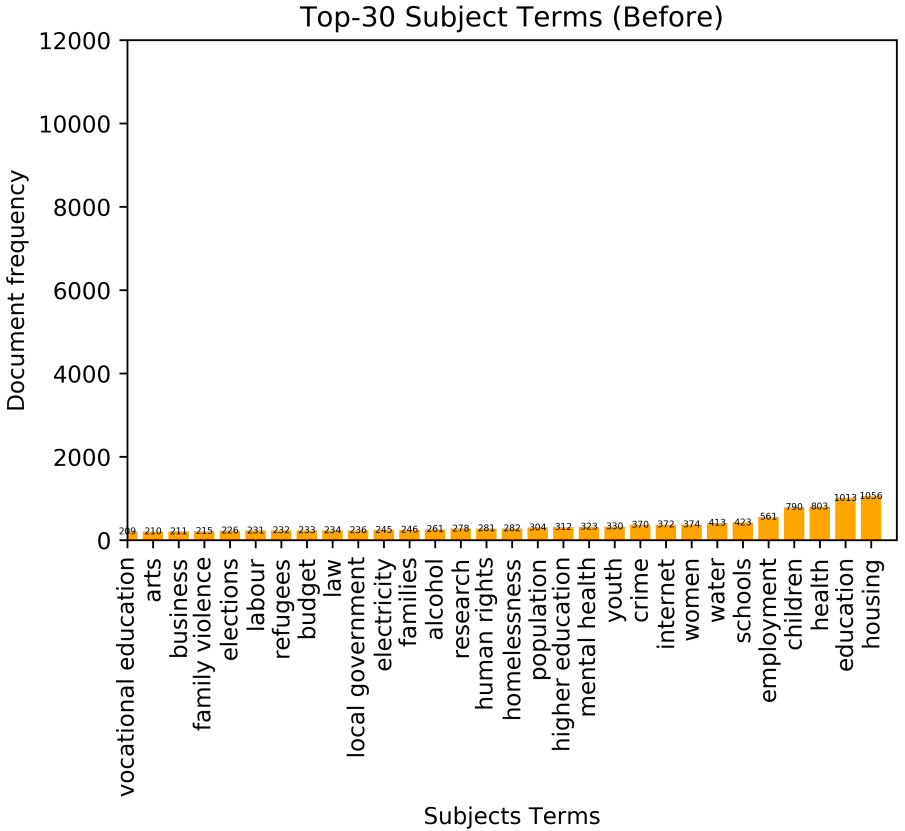
Finding Missing Subject Terms

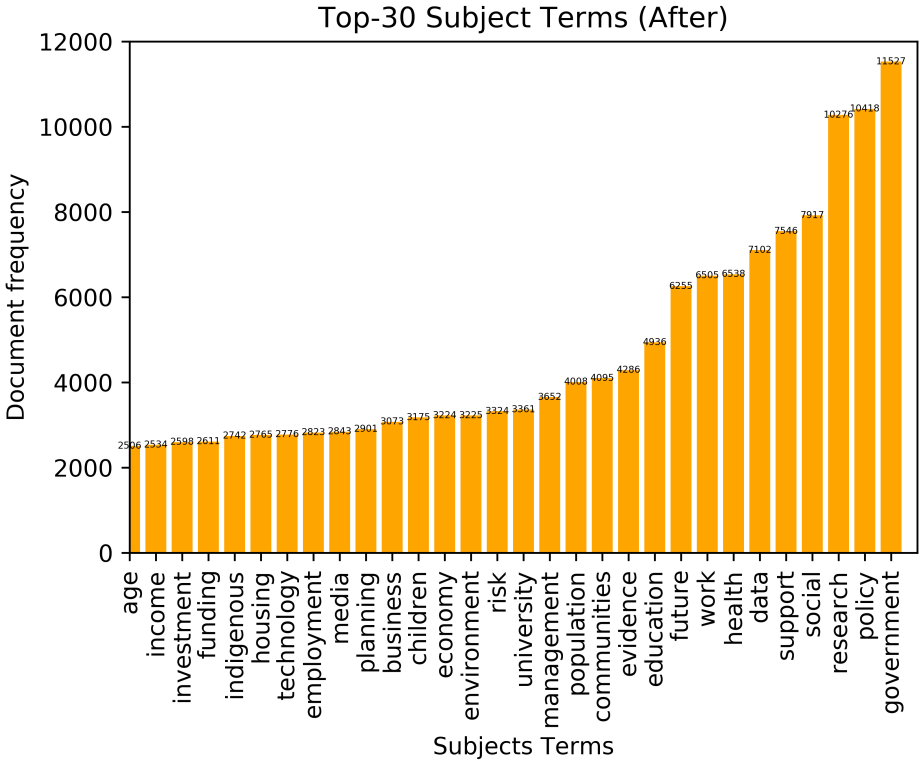




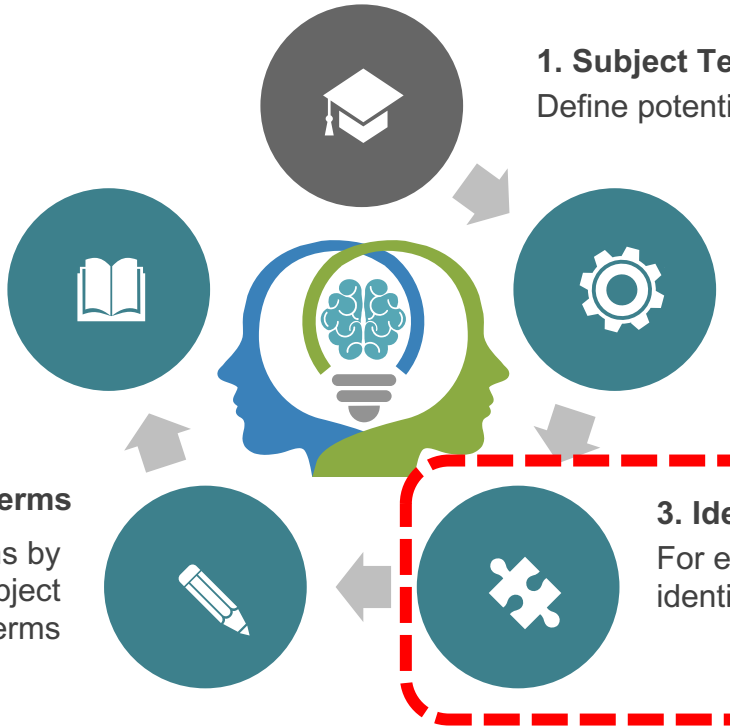
	Before	After
Average # of Subject Terms per Doc	4	12
Max # of Subject Terms per Doc	27	132
Min # of Subject Terms per Doc	0	0

Finding Missing Subject Terms





AI Driven Solution



1. Subject Term Dictionary Definition

Define potential subject terms

2. Find Missing Subject Terms

Perform text analysis to identify missing subject terms

3. Identify Synonyms

For each subject term, we identify its synonyms using Wiki

4. Remove Insignificant Subject Terms

Remove insignificant subject terms by merging them with significant subject terms

5. Build Subject Taxonomy

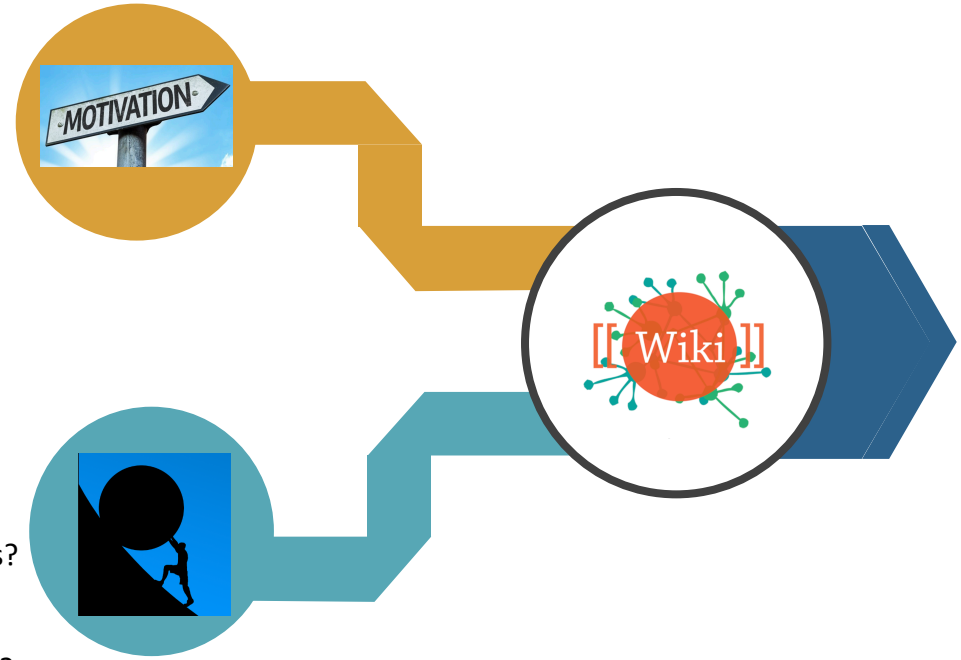
Build a taxonomy of subject terms by understanding their semantic relatedness

Identifying Synonyms

Motivation:

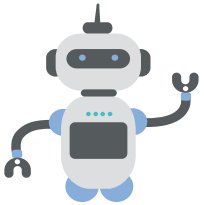
Can we capture similar concepts of subject terms?

Ex. ”social problem” is very similar to “social issues”



Challenges:

1) Is there a comprehensive general knowledge source whose coverage is high enough to find similar terms of our subject terms?
2) How to choose candidates of similar terms?
3) How to measure similarity?



Our approach

1) Wikipedia
2) Wikipedia Titles
3) Using word embeddings of terms in Wikipedia

Identifying Synonyms

Example:

Extended Subject Term Dictionary

"Community Health"

Wiki Titles

"Public health"

"Social Services"

"mental health"

"health education"

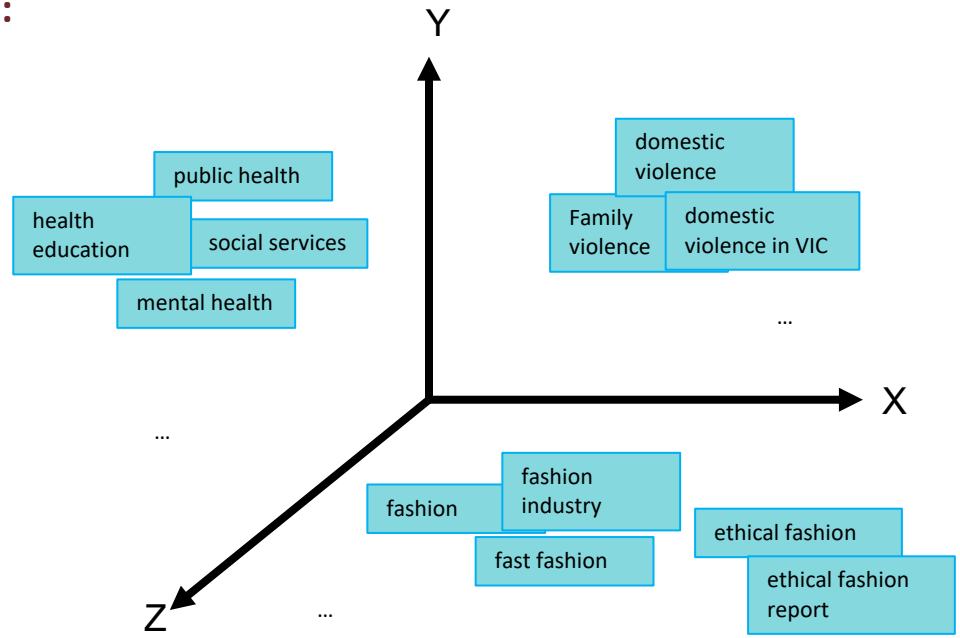
Distance: similarity

Identifying Synonyms

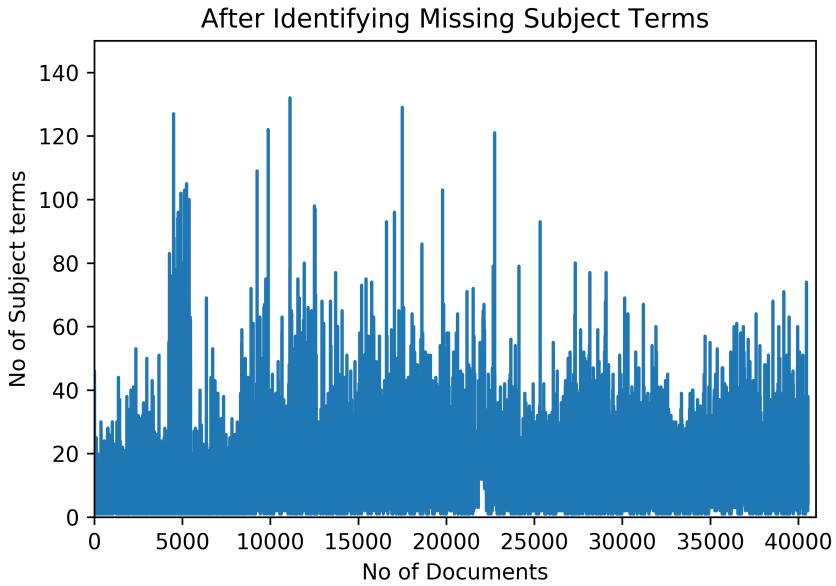
- **How to Measure Similarity between a Subject Term and Wiki Titles?**
  - Map all terms including “Wiki Titles” in “Wiki articles” into a Vector Space.
  - Words with similar meaning have a similar representation.
  - Given two terms, we measure their similarity: their cosine similarity using their corresponding embedded vectors.

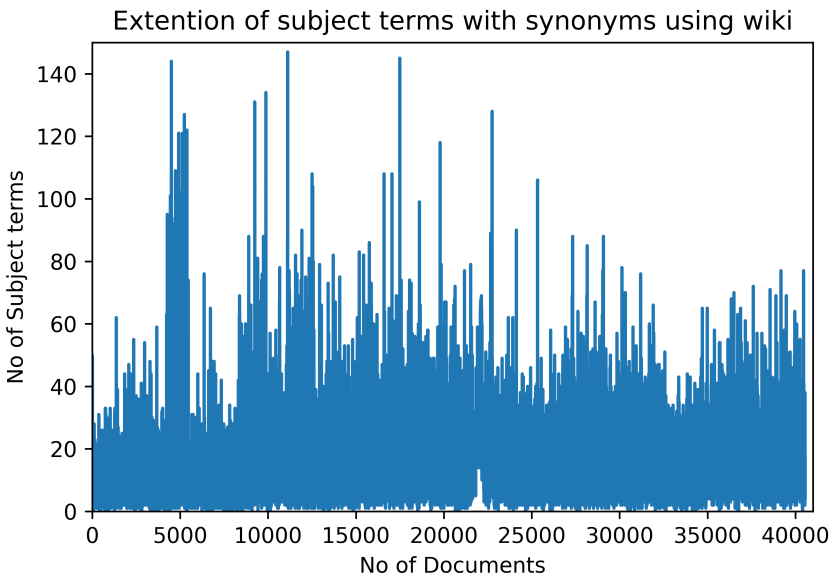
Subject Term Representation including Wiki Titles

Example:

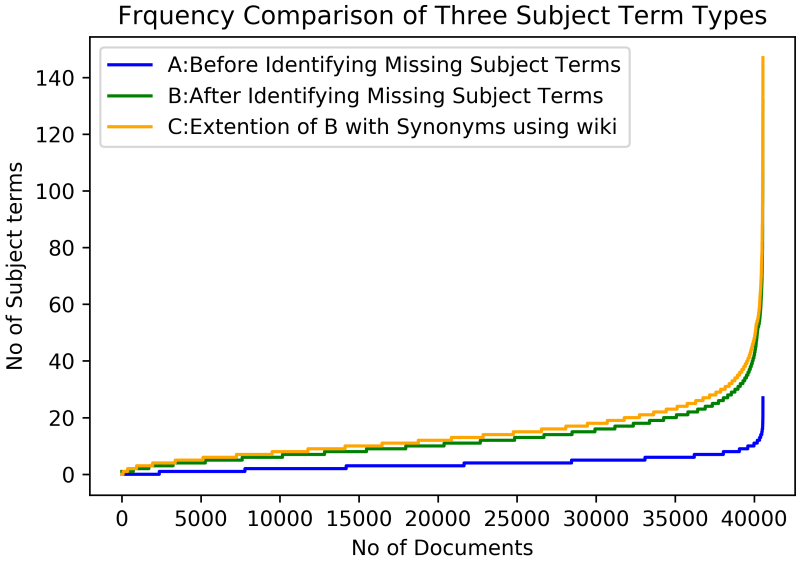


Identifying Synonyms



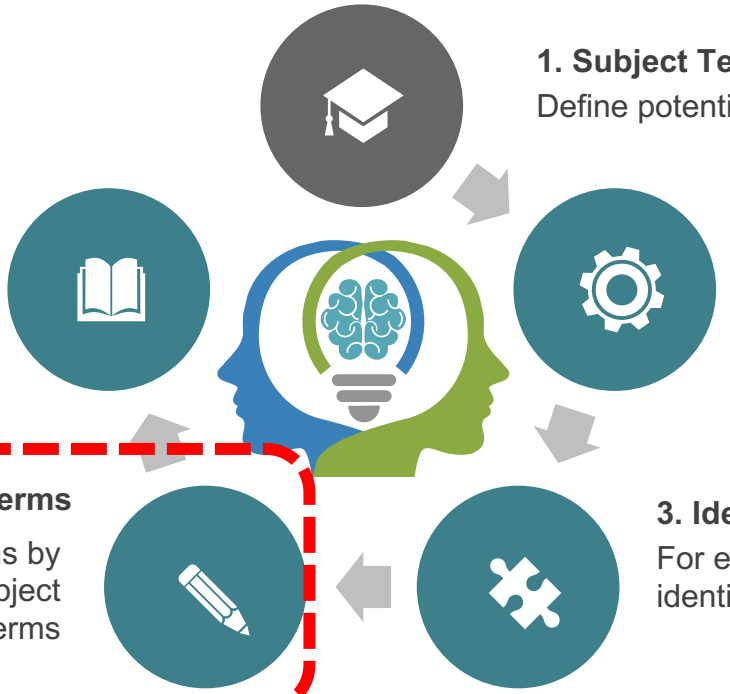


Identifying Synonyms



	Before including Synonyms	After including Synonyms
Average # of Subject Terms per Doc	12	14
Max # of Subject Terms per Doc	132	147
Min # of Subject Terms per Doc	0	0

AI Driven Solution



1. Subject Term Dictionary Definition

Define potential subject terms

2. Find Missing Subject Terms

Perform text analysis to identify missing subject terms

3. Identify Synonyms

For each subject term, we identify its synonyms using Wiki

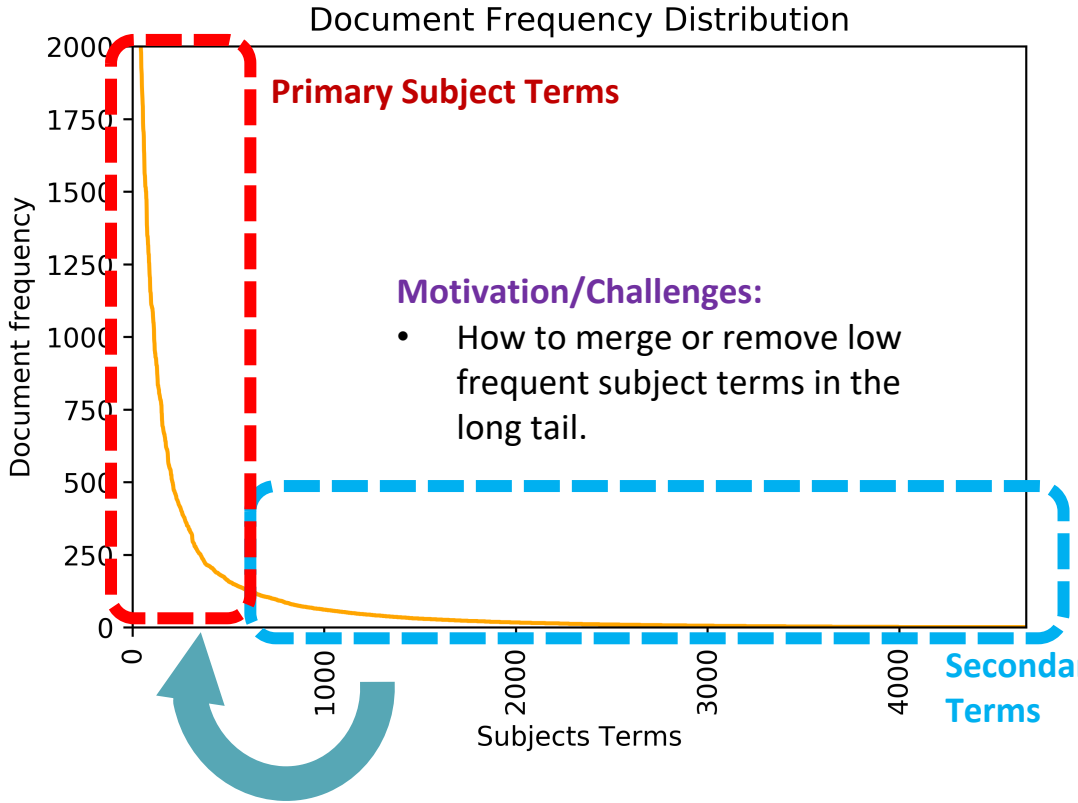
4. Remove Insignificant Subject Terms

Remove insignificant subject terms by merging them with significant subject terms

5. Build Subject Taxonomy

Build a taxonomy of subject terms by understanding their semantic relatedness

Removing Insignificant Subject Terms



Our approach

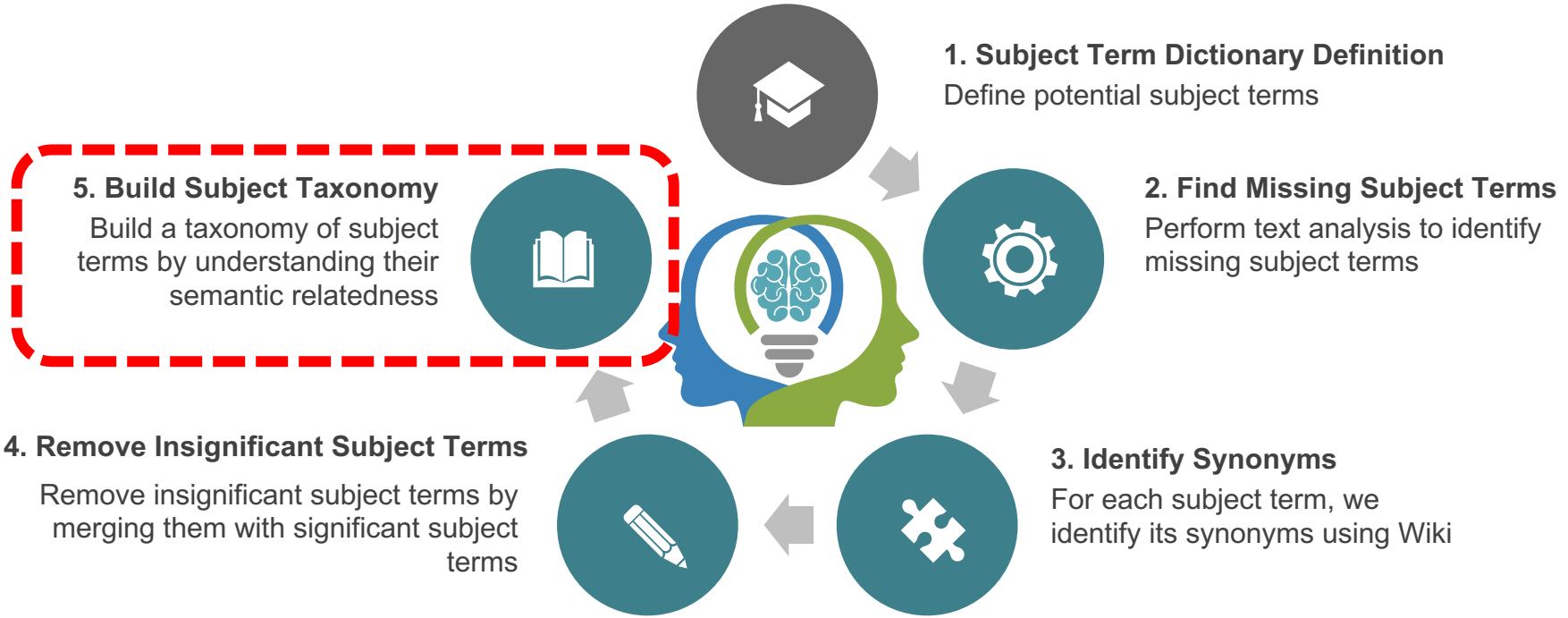
1) Merge or Remove very infrequent subject terms in the long tail using their similarities with the more frequent ones
2) Incorporate user search history (future work)

Removing Insignificant Subject Terms

Summary After Merge: Different Candidate Subject Terms according to Their Percentiles

Target Percentile	# of Target Subject Terms	Document Frequency	# of Merged Subject Terms	# of Ignored Subject Terms	# Final Subject Terms
50	2,327	12	1,148	1,179	2,346
60	2,756	19	1,266	1,490	1,917
70	3,270	35	1,284	1,986	1,430
80	3,731	68	1,154	2,577	942

AI Driven Solution



1. Subject Term Dictionary Definition

Define potential subject terms

2. Find Missing Subject Terms

Perform text analysis to identify missing subject terms

3. Identify Synonyms

For each subject term, we identify its synonyms using Wiki

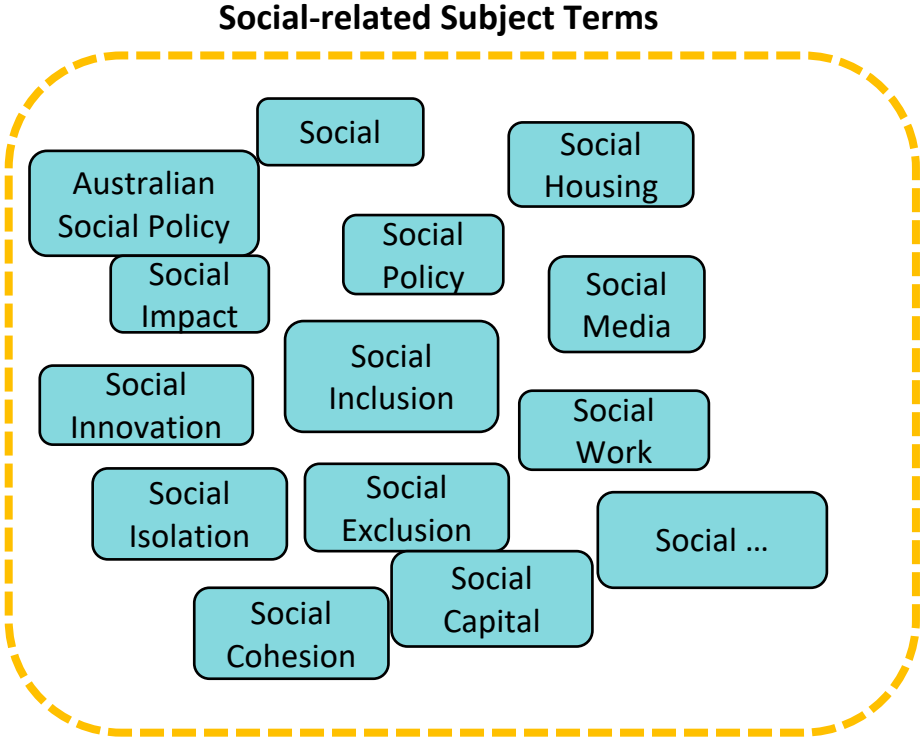
4. Remove Insignificant Subject Terms

Remove insignificant subject terms by merging them with significant subject terms

5. Build Subject Taxonomy

Build a taxonomy of subject terms by understanding their semantic relatedness

Building Subject Term Taxonomy



Motivation:

- Can we group semantically subject terms considering their semantics and make their relationships?

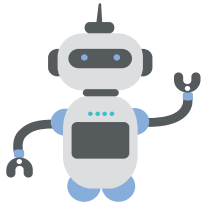
Challenges:

- How to estimate meaningful relationships between subject terms?
- What kind of structure is preferred?

Building Subject Term Taxonomy (Coming Soon!)

Social-related Subject Terms

Social
Australian Social Policy
Social Housing
Social Policy
Social Impact
Social Media
Social Inclusion
Social Innovation
Social Work
Social Isolation
Social Exclusion
Social ...
Social Capital
Social Cohesion

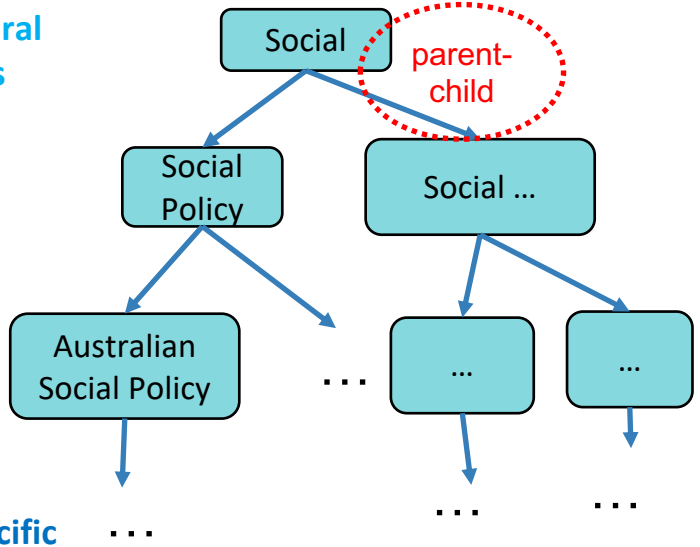


Our Approach:

- Automatic Taxonomy Building

More General Concepts

More Specific Concepts



Note: each node has its similar subject terms (Wiki Titles + Infrequent Subject Terms Merged)

Benefits

Indexing

Automatic Indexing

We can assign appropriate subject terms automatically

01

Automatically identify potential subject terms via text analysis

02

Map the identified subject terms to the primary subject terms

Searching

Improve Searching Capability

We can provide more AI-oriented search capability

01

Exploit Primary-Secondary Subject Terms

02

Exploit Parent-Child Relationships in Subject Taxonomy

Thank You

Find me at
ykang@swin.edu.au