Sustainability research identification method

The University of Queensland is a very large, research-intensive institution with 3680 paid research-active staff. It is unfeasible to use a survey based approach for all our researchers; we would expect the response rate to be low, which risks under-reporting our sustainability-focussed research.

Instead we use Natural Language Programming approaches to our institutional data to identify which researchers are involved with sustainability research.

Inclusion criteria

Projects:

- Research grants (externally or internally funded)
- Received funding during the reporting time period
- At least one UQ paid research, teaching & research or senior executive staff member list as an investigator.

Identify sustainability keywords and key phrases

Our approach is to filter UQ outputs, keeping those containing words or phrases that are relevant to sustainability and doing so in a way that captures as many relevant concepts as possible while minimizing the inclusion of those focussed on non-sustainability concepts.

The first step involved generating as many relevant terms as possible to fully capture the breadth of our sustainability research. For example, including the key term 'renewables' would only fetch UQ outputs that exactly matched that phrase. However, there are many ways to refer to this concept, including 'low-carbon', 'zero-carbon', and 'clean-energy' etc. We used two statistical techniques to generate as many relevant terms as possible: word embeddings and collocations.

1. **Word embeddings to identify single keywords**

   To identify relevant single keywords, we started with 55 seed terms relevant to sustainability, sourced from the UN Sustinable Development Goals (SDG) indicators and the SDSN list of SDG keywords (http://ap-unsdsn.org/regional-initiatives/universities-sdgs/) (Appendix A). Those with obviously ambiguous usage and meanings were not included (e.g. 'carbon', which would pick up many irrelevant chemistry projects, and 'conflict', which could refer to 'violence' or to 'conflicting' ideas/findings.

   We generated new SDG-relevant terms using word embeddings, which are generated using large volumes of text posted to the web and can be used to identify terms that are frequently used in conjunction with an inputted term. Using pre-trained word embeddings (using Stanford's GloVe, Common Crawl (840B tokens) file: https://nlp.stanford.edu/projects/glove/), we retrieved the most relevant terms to the seed terms. We screened out terms that were not relevant to the SDGs or that had obviously ambiguous meaning, leaving 403 relevant terms (including the original 55).

2. **Collocations to identify key phrases**

   Some single terms that are potentially relevant to sustainability, such as 'solar' would pick up irrelevant outputs, such as those focussing on solar astrophysics. However, if they were excluded then outputs mentioning 'solar power', and 'solar photovoltaic cells' could be missed. To generate

relevant key phrases (rather than single words), we generated collocations: the stitching together of words that significantly co-occur next to each other within a corpus of documents. This can be more than two words, e.g. “e waste”, “waste disposal” and “e waste disposal”. To do this we:

a. Download 4,500 research articles from Scopus that used the phrase ‘Sustainable Development Goals” in the title or Abstract. We used this as a seed set to find collocations in a SDG-relevant corpus of documents.

b. Processed the text using text-cleaning packages (removing low-information words, harmonizing spelling variations, de-pluralizing words etc.)

c. Identified collocations that occurred at least 15 times in the corpus and were sufficiently associated with each-other. These were screened to exclude those with low relevance to the sustainability, leaving 1,278 relevant key phrase combinations.

The key terms and phrases are listed in the accompanying file: sustainability_terms_phrases.csv.

We then filtered UQ publications and research projects that contained at least one exact whole term/phrase match. Matches were case sensitive, to avoid case such as ‘AIDS’ matching ‘aids’.

Topic modelling

It was likely that irrelevant outputs were included in the aforementioned filtered outputs. To identify these, we used topic modelling.

Topic modelling takes a corpus of documents, groups highly associated terms into topics, and assigns a proportion of each topic to each document. We can print out a list of the topics the algorithm identifies and isolate any that look irrelevant to sustainability. We inspected a sample of documents for each suspicious topic to check whether these were actually irrelevant, and excluded all documents that were only associated with irrelevant topics (those that were also associated with relevant topics were kept).

We created two topic models: one for publications and one for projects. We used correlated topic modelling for the publications, using their titles, abstracts and keywords. However, correlated topic modelling requires sufficiently long text to give informative results. As our projects system currently only store titles (and not summaries), the text associated with projects is too short to perform correlated topic modelling, so used biterm topic modelling instead.

Identifying staff

Finally, we joined staff details to the remaining sustainability-relevant outputs; any staff member with at least one relevant output is included in the submitted list. The topic modelling steps also enables us to include which areas of sustainability their research encompasses.

Appendix A: Seed terms for word embeddings

habitat	slum	sexism	torture
malaria	biodiversity	water-conservation	indigenous
hepatitis	geothermal	biosphere	clean-energy
zero-waste	rainforest	tuberculosis	wetland
malnutrition	wastewater	vaccine	abuse

drought	ecosystem	empowerment	detainee
arid	renewable	wellbeing	HIV
racism	polio	slavery	sequestration
deforestation	smallholder	cholera	refugee
endangered	famine	urbanisation	violence
pollution	unemployment	farmers	sanitation
recycle	poverty	hygiene	poaching
bribe	trafficking	coral	inequality
low-carbon	green-tech	eutrophication	