**Task Description**

Parse Environmental Impact Statements (EIS) and Records of Decisions (ROD) using Natural Language Processing (NLP) to identify terms from environmental and resilience ontologies.

**Rationale**

EIS and ROD documents that exist today lack provenance. Without reading the entire document, it can be difficult to determine what the environmental impacts where and what decisions were taken. Machine learning, NLP, and topic modeling are proving effective means of parsing these documents and quickly classifying them. By additionally tagging these documents with ontology terms we may be able to enhance the classification and utility of these documents.

For example, the Environmental Ontology (ENVO) tells us that ‘wildfires’ produce ‘hydrocarbon flames’ and occur in the ‘terrestrial biome’. We can expand our classification of any EIS or ROD documents mentioning ‘wildfires’ to now know where they occur and what their chemical outputs are. When searching for similar decisions/impacts, semantic expansion can provide additional context and relevant details.

**Procedure**

This prototype study used the Environmental Ontology (ENVO) and Sustainable Development Goals Interface Ontology (SDG).

ENVO - <https://bioportal.bioontology.org/ontologies/ENVO>

SDG - <https://raw.githubusercontent.com/SDG-InterfaceOntology/sdgio/master/sdgio.owl>

The terms in these two ontologies were compared against EIS and ROD documents from the Housing and Urban Development Exchange web site [1]. Specifically, I looked at all documents pertaining to the Rim Forest Recovery and Reforestation, State of California (2017) and all documents pertaining to Coastal and Social Resiliency Initiatives for Tottenville Shoreline, Staten Island, NY (2017). PDF versions of the documents used are in the docs/ folder.

Python code was written to parse the PDF files and use NLP to break the text into sentences. Each sentence was then inspected for mention of an ontology term. For the matching, I looped through all the classes in each ontology and looked for mention of the rdfs:label, which provided a human-readable label for the cryptic class names, i.e. the rdfs:label is ‘wildfire’ while the actual class name is <http://purl.obolibrary.org/obo/ENVO_0100078>

**Dependencies**

The Python code requires installation of the PDFminer, owlready2, and nltk libraries. In addition, PDFminer is only available for Python2 while owlready2 and nltk are Python3 libraries. As a result, the code uses both Python3 and Python2 installations – the Python2 PDFminer is called as a subprocess using the Python3 subprocess library.

**Results and Discussion**

The semantic tagging was successful and found anywhere from a few hundred to over a thousand terms in each chapter of the ROD and EIS documents. Results can be found as text files in the docs/ directory. However, there were a few practical challenges encountered.

* Ontology construction – The ENVO ontology is continually being updated and expanded. When a term (class) is removed from the ontology the ENVO team leaves that term in the next version, but marks it using owl:deprecated. From an ontology engineering perspective, this is the correct approach. Keeping the term preserves existing tools that reference it, while marking it as deprecated also signifies it should be phased out. However, some libraries, like the owlready2 libraries used here, don’t pay attention to owl:deprecated. Parsing the ENVO ontology from the web resulted in errors as the owlready2 library kept trying to reason with classes that were marked as deprecated. In the end, I had to download the ENVO ontology and manually delete all the deprecated classes. Working with this local ENVO copy worked; however, since I am no longer pulling in ENVO from the web, future running of the code may not have the latest version of the ontology.
  + Solution – the fix is relatively easy, either find a software library that understands owl:deprecated or manually update the local copy of ENVO each time the code is run. Nevertheless, this is an issue to be aware of.
* Ontology labeling – the SDG ontology has rdfs:label values that are too generic for what we are doing here. Labels like ‘participates in’ and ‘is about’ don’t add any value to the results. Similarly, I didn’t see any labels in the SWEET ontology that would be of use to us here. As far as I know, there aren’t many environmental/resilience ontologies available. ENVO may be the only option right now.

**Potential Next Steps**

Currently, the code just outputs the term that was found and the line number where it occurred. We may need to modify the output depending on what the next steps of search and machine learning look like.

It’s currently tedious and labor intensive to find EIS and ROD documents. The HUD web site links to state websites, which then link to additional pages containing multiple PDF files for each EIS (EIS documents are typically broken down into chapters with one PDF file for each chapter). Obtaining data for a study like this is non-trivial. Provenance may be useful here as well. If there was a central provenance repository (e.g. Provisium) then state and local governments could submit a small PROV file simply stating that an EIS (or ROD) had been created and listing of all the PDF URLs associated with it. This would be a huge help for tagging and machine learning projects.

It would be best if the PROV were proactively submitted to a repository. Yet, lacking a PROV repository, embedding PROV as JSON-LD on state and local government web pages would also be useful.

**References**

[1] <https://www.hudexchange.info/programs/environmental-review/environmental-impact-statements/>