Serverless Vocabulary Extraction and Curation for Cross-Divisional Search in NASA's Science Mission Directorate

Ashish Acharya[1], Carson Davis[1], Stephen McNeal[1], Andrew Weis[1], Emily Foshee[1], Kaylin Bugbee[2], Rahul Ramachandran[2]

University of Alabama in Huntsville[1], NASA Marshall Space Flight Center[2]

Why Serverless?

- Removes the need to manage servers
- Uses resources on an as needed basis, eliminating costs to keep servers running all the time
- Scales easily
- Allows for small, decoupled, replaceable, and individually testable components
- Makes deployments faster

Why Extract Vocabulary?

Let's say we're building a search application that aims to find the user's search term across N sources. In order to give the user intelligent hierarchies to sort their search results, the application would benefit from some kind of categorization of concepts across the N sources. For example, if you're purchasing a television, you'd want to be able to filter results by manufacturer, video resolution, size, and so on.

However, in order to intelligently create these categories, the application managers must first obtain a list of terms used across each source. We call this the vocabulary of a source.

Source-level vocabulary extracted in this way can be used to help the search user in various ways. For instance, we may be able to use Natural Language Processing (NLP) techniques to determine synonyms and acronyms which can help expand search results.

Supporting Cross-Divisional Search

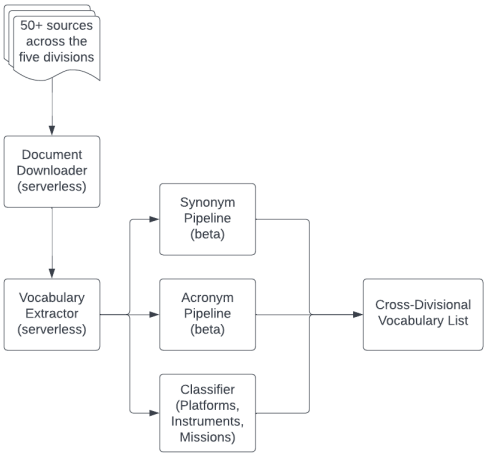
Science Discovery Engine (SDE) aims to be a cross-divisional search portal across the five divisions – Earth Science, Planetary Science, Heliophysics, Astrophysics, and Biological and Physical Sciences – in NASA's Science Mission Directorate (SMD). Each division has its own vocabulary and idiosyncratic metadata.

In order to conduct a cross-divisional search, we needed to build filters that span across the five divisions. The work presented here allowed us to filter concepts such as Platforms, Instruments, and Missions, which have become important parts of the search experience in the Science Discovery Engine (SDE).

Methodology

We started off by making a list of 50+ important sources of documents and metadata spanning across the five divisions. Most of these were web-based while some were JSON APIs. Using Python, we wrote web crawlers to fetch content from these sources, filtering pages that were of value for search purposes. Libraries for natural language processing, such as nltk, helped us extract important nouns from these documents, which we then stemmed and made consistent.

Having compiled a list of over 9 million vocabulary terms, we then wrote pipelines for synonym detection, acronym expansion, and classification into one of our three categories – Platforms, Missions, and Instruments. We then wrote lambda functions on AWS cloud to bring in these components together and run a full pipeline periodically, allowing us to update our vocabularies.



Future Work

A major caveat of this work is that the results always need to be manually verified by Subject Matter Experts (SMEs) before we can confidently update them on the search portal. Therefore, an idea is to create a web-based portal for SMEs to log in and validate classified vocabulary terms, synonyms, and acronyms, as they become available.

In addition to web crawlers and JSON, we also plan to support other data source types in the future, such as XML and direct database connections.

Moreover, the synonym pipeline and the acronym pipeline are in a beta stage at the moment. Our plan is to improve the algorithms used in them and make them fully production-ready.