Can AI Preserve Our Science Legacy?

The Challenge

The NASA Technical Report Server (NTRS) includes hundreds of thousands of items containing scientific and technical information (STI) created or funded by NASA. Imagine how difficult it can be to locate desired information in such a large repository! Your challenge is to develop a technique using Artificial Intelligence (AI) to improve the accessibility and discoverability of records in the public NTRS.

Background

The NASA Technical Report Server (NTRS) includes hundreds of thousands of items containing scientific and technical information (STI) that were created or funded by NASA. As of June 1, 2022, a total of 381,547 of these documents include an organizational Cluster Data Management System (CDMS) tag. Many of these documents with a CDMS tag were scanned and Optical Character Recognition (OCR) was applied to produce Portable Data Format (PDF) files. The NTRS records for these PDF documents contain a summary and a subject category. To enable searches of this large NTRS database, potential users such as members of the scientific and historical research communities could use an application that can read a collection of PDF files, summarize those files, produce statistical reports of the language usage, and list topic key words. Future researchers would be able to use this information to find desired historical data quickly and easily.

Objectives

Your challenge is to develop an AI application to improve the accessibility and discoverability of records in the NTRS. For example, you could use Natural Language Processing (NLP) to automatically read NTRS documents, summarize them, generate text analytic data, and produce a list of topic keywords to help researchers find the documents they need. Think about what types of information future researchers will need to locate desired documents. What would be the best data to aid them in their search for relevant information?

Potential Considerations

As you develop your project, you may, but are not required to perform the following steps:

- Build a Corpus: A corpus is a body of work or a collection of documents. To build a test corpus, you can navigate to the NTRS home page and click the Start Search button under Publicly Available Content. If the Filter form does not appear automatically on the left hand side of the page, look for the "hamburger" button icon (it's a stack of three short horizontal lines). Click the hamburger button to open the Filter form. On the Center drop-down list, select Legacy CDMS. Review several of the files to verify that they are searchable PDF files and download them to create a corpus. Participating teams that decide to construct queries may want to use the NTRS OpenAPI (see Resources section for the OpenAPI link and documentation) to develop a corpus.
- Research and select open-source code libraries and examples (e.g., NLP): Develop an AI application that can open and read a collection of PDF files in a folder (i.e., your corpus) and generate a report of relevant information. Desirable report features can include, but are not limited to, a summary of each document and a list of topic key words found in each document, including the frequency with which those key words appear. The NASA Scope and Subject Category Guide (refer to the Resources tab at the top of the page) contains a list of potential topic key words.
- Demonstrate your NLP application: Generate reports and post them where judges can access them (e.g., on your Space Apps Project Page). Consider including a list of the documents or links to the NTRS records your application has analyzed and a link to the repository where your code is stored. If you produce a web application, you can also provide a link to it.

You may, but are not required to, consider the following as you develop your application:

- Notional products from this challenge may also include source code for interpreted languages such as Python, R, or JavaScript.
- Space Apps Judges cannot download and run programs from an external site; if your team develops a web app consider hosting it on a free server. To locate such a service, you could use your favorite search engine to

search on key terms such as code repositories, cloud platform, free hosting, and free web hosting sites.
- If you develop a desktop application, consider providing information and documentation about how to obtain and use your application.
- Consider using your favorite search engine to look for free open-source code libraries that you can employ. Key words to help with your research include: Free open source Natural Language Processing, text-analytics, Python NLP, NLP, NLP software, NLP libraries, NLP tools.

For data and resources related to this challenge, refer to the Resources tab at the top of the page. More resources may be added before the hackathon begins.

NASA does not endorse any non-U.S. Government entity and is not responsible for information contained on non-U.S. Government websites. For non-U.S. Government websites, participants must comply with any data use parameters of that specific website.

https://2022.spaceappschallenge.org/challenges/2022-challenges/science-legacy/details

NASA Resources

- NASA Technical Report Server
- NASA NTRS OpenAPI
- NASA Scientific and Technical Information (STI) Repository OpenAPI Data Dictionary
- NASA STI Repository OpenAPI Documentation
- NASA STI Scope and Subject Category Guide