Project Proposal

Recommendation of similar articles from journal abstract analysis

Misty M. Giles
https://github.com/OhThatMisty

Overview

The corpra of academic research in many fields encompass thousands of documents per year. In 2018, the astrophysics section (astro-ph) of the scientific paper archive arXiv received over 14,000 submissions. Although astro-ph is subdivided into six categories, there's some overlap in the different fields, and scientists don't have the time to read through them all. Scientists at Marshall Space Flight Center have bemoaned the lack of tools to help them find specific related concepts in a field of interest to help advance their own research.

I plan to utilize natural language processing (NLP) techniques on a selection of astro-ph submissions. I'll utilize document similarity in order to recommend similar articles when a user inputs an abstract. This should help researchers find future collaborators and new tracks to pursue in a more focused manner.

Data

ArXiv offers an XML API, and a notebook with code to download data for testing purposes will be provided. Data will be initially filtered to exclude submissions before January 2009 (exact date TBD) and to exclude any article that contains "has been withdrawn" in comments or summary, since that's pretty common when problems are later discovered with methods or data. Submissions will also be filtered for duplicates. Further filtering of conference proceedings, etc. might be necessary to clean the data.

ArXiv removes keywords from the search process. Although the original papers include keywords, I will not be able to use them to sort or group the abstracts.

Analysis

I'll use document similarity techniques to help researchers narrow down a topic to a list of articles to read. Document similarity will be tested with both cosine and TF-IDF methods, but I blindly expect TF-IDF to provide better recommendations. These recommendations will save researchers time when they want to see what work has been previously done in a particular area.

Deliverables

Commented, public code and a short summary paper will be provided at the end of the project.