Project Proposal
Classification of topics from journal abstracts

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 was subdivided into six categories in 2009[1], each category still contains massive amounts of text. Scientists at Marshall Space Flight Center have bemoaned the lack of tools to help them find articles that have contributed to certain ideas and to find new areas in which to grow their own research.

I plan to utilize natural language processing (NLP) techniques on a selection of astro-ph submissions since the classification system changed to six categories[2]. I'll attempt to first discover if the six categories remain the best choices a decade later. I'll then narrow down the categories that I discover into subcategories and look for overlap.

Data

ArXiv offers an XML API, and a notebook with code to download the 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filtered to exclude submissions before January 2009 (exact date TBD) and to exclude any article that contains "has been withdrawn" in comments or summary, since that's pretty common when problems are later discovered with methods or data.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 from 2009-2019 total about 137,000 before

[1] "The problem with science is that there's just too damn much of it," says the guy who wrote this blog post, more or less taking credit for the astro-ph subcategories. According to his post, the subcategories went into effect on 20 January 2009. This classification, along with many other aspects of arXiv, has been controversial.
[2] I am not a physicist, but I appear as a coauthor on about 20 of these papers as the Assistant Operations Manager for the Gamma-ray Burst Monitor (GBM) onboard NASA's Fermi Gamma-ray Space Telescope.

filtering. ArXiv also contains more than journal articles, so further filtering of conference proceedings, etc. might be necessary to clean the data.

Analysis

I'll use NLP and topic segmentation techniques to classify abstracts into a few categories, and then I'd also like to see those further subdivided into areas of interest where a researcher could potentially narrow down articles to read 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.