PLOS: Encouraging Open Science

MSDS 2018 Capstone Proposal
University of Virginia Data Science Institute

Jack Prominski
Pragati Shah
Dr. Phil Bourne, Advisor
Public Library of Open Science, Sponsor
The Bourne Laboratory, Sponsor

Summary

The Public Library of Science, or PLOS, is on the forefront of the Open Science movement. They believe that this free-distribution model promotes accelerated scientific discovery, public enrichment, and improved education.[1] In an effort to provide the best reader experience possible in support of this goal, we will develop a recommendation engine to better connect PLOS' readers with relevant content. In addition to distributing for free seven peer-reviewed journals, PLOS makes available the underlying datasets associated with each scholarly article. While this has additional benefits of improving reproducibility of research and further analysis, there are certain risks associated with making these datasets open to the public. Before publishing, PLOS must review each dataset to confirm that its publishing will not cause harm. In an effort to improve the efficiency of PLOS' operations and to mitigate the risk of potential harm, we will also investigate and develop algorithmic methods to perform automated data checks on these submissions.

Background

The Open Science Revolution and PLOS' Open Data Policy

PLOS was born in 2000, when Harold Varmus, Patrick Brown, and Michael Eisen circulated an open letter urging the academic community to embrace open science and make their research freely available. Though the letter gained 34,000 signatures, in which the signees committed to only "publish in, edit and review for, and personally subscribe to only those scholarly and scientific journals that have agreed to grant unrestricted free distribution rights to any and all original research reports," few actually adhered to these guidelines. The next year, Varmus, Brown, and Eisen, recognizing that little progress had been made, established PLOS as a publisher and advocacy organization for the Open Science movement. Since then, PLOS has published more than 165,000 articles from authors in 190 countries and has grown to seven academic journals.

PLOS requires authors to make all data underlying the findings described in their paper fully available without restriction, with rare exception. Authors are required to provide a *Data Availability Statement* affirming compliance with PLOS's policy. If authors did not collect data themselves but used another source, this source must be credited as appropriate. PLOS not consider manuscripts for publication if authors do not share data either out of personal interests, such as patents or potential future publications, or in cases where the conclusions depend solely on the analysis of proprietary data. PLOS believes that data availability allows and facilitates:

1. Validation, replication, reanalysis, new analysis, reinterpretation or inclusion into meta-analyses
2. Reproducibility of research
3. Efforts to ensure data are archived, increasing the value of the investment made in funding scientific research
4. Reduction of the burden on authors in unearthing old data, retaining old hard drives and answering email requests
5. Easier citation of data as well as research articles, enhancing visibility and ensuring recognition for authors[2]

However, open science does not come without its drawbacks. Critics of the philosophy cite that:

1. Too much unsorted information may overwhelm scientists, and influence them by exposing them to the notions of consensus.
2. Risk of misuse of scientific data as sensitive information in the wrong hands can lead to creation of biological weapons, etc.
3. Unintentional misinterpretation of data by the general public, who may not have the contextual knowledge to interpret it correctly
4. Post publication peer review have been criticized to inadvertently contribute to production of low quality papers. Currently, there is no framework to ensure any basic quality of standards.[3]

PLOS are particularly concerned with the potential for harm arising from their publishing of articles and datasets openly.

Problem & Objectives

First, PLOS would like to improve the reader's experience by improving their article recommendation engine. Currently, PLOS recommends related articles to readers based on similar MeSH terms, which are a set of keywords that describe and characterize the contents of the article. They would like to extend their recommendation engine beyond these terms to deliver more value to researchers and the public.

Second, PLOS would like to improve the efficiency of their internal operations. They have grown to the point where they receive 25,000 paper submissions a year, a volume that has strained their internal resources and review process. In accordance with their data policy, PLOS requires the underlying data set associated with the research to be included with any article submission. PLOS has an ethical obligation to do everything in

their power to minimize potential harm arising from the publishing of this data. Currently, PLOS performs manual checks on these data sets to ensure adherence to their Data Policy, a process that is slow and tedious. Automating this process would allow PLOS to speed up their publishing cycle and more efficiently operate their organization. PLOS currently does not do data science and has little automation in their processes. What we can leverage, though, is PLOS' expertise in the publishing industry. As discussed later in this proposal, they will provide valuable insights into their readers' motivations and needs, which will drive our recommendation engine, as well as frameworks to aid in our developing of automation algorithms.

Thus, our objectives are as follows:

1. Develop a recommendation engine to improve related article suggestions for the PLOS reader.

2. Help improve PLOS' manuscript evaluation process by investigating and developing algorithmic data checking techniques.

Technical Approach

The PLOS Capstone team will begin by gaining an understanding of the corpus. PLOS articles are accessible in XML format through their own API and on PubMed Central, a digital repository operated by the the National Institutes of Health's National Library of Medicine. Accessing the articles through PubMed Central is preferred, as they have formatting requirements and procedures that ensure JATS-compliance. JATS is a tag suite for XML that standardizes elements and attributes to characterize journal articles.[4]

Next, we will overlay an index on the corpus. This index will cover article metadata, such as the authors' names and affiliations, citations, journal information, MeSH terms, as well as the full text of the abstract and article. Indexing will allow for fast searching and information retrieval and will facilitate further analysis. We will explore different indexing engines, including Apache Lucene/Solr, Sphinx, and MySQL.

To begin our text analysis of the corpus, we will begin by exploring a document modeling approach called a knowledge graph to enrich our semantic understanding of the text. The key to this approach is the combination of a fine-grained relation vocabulary with information theoretic measures of concept associativity to produce a graph-based interpretation of texts leveraging large amounts of structured knowledge. Edges and Nodes are the building blocks of a graph. Edges in the semantic graphs are

thus weighted so as to capture the degree of associativity between concepts, as well as their different levels of specificity. We will develop a new measure, based on graph edit distance techniques, in order to compute document similarity using our semantic graphs.[5]

We will then use this indexed corpus and knowledge graph to generate visualizations that map the network connections of PLOS' corpus. Several different variables could be visualized, including citation data, author/co-author data, textually and semantically similar papers, etc. This will be a useful step in understanding the corpus and is an analysis that PLOS has not done but has expressed interest in. This will serve as an early deliverable to PLOS and an opportunity to gather feedback and insights from our sponsor.

Recommendation Engine

In the past several years, recommendation engines have been the focus of much research and development. We will explore the two main approaches to these systems, content-based filtering and collaborative filtering.[6] A third hybrid approach, combining aspects of content-based and collaborative filtering, will also be investigated.

Content-based filtering relies on using characteristics of the content of an item to develop an understanding of that item and recommend similar items. This is how PLOS' current "Related Content" feature works. It uses the MeSH terms associated with the article to suggest other articles with similar MeSH terms. This current system is limited in scope and relies on accurate and thorough MeSH tagging. A content-based approach that we will explore is to employ a semantic analysis of the text to extend the recommendation capabilities beyond the simple MeSH terms.

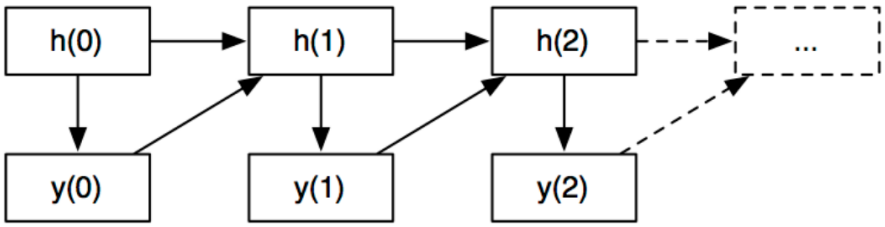
Our content-based filtering approach will potentially use a standard approach for term parsing to select single words from documents. The vector space model and latent semantic indexing are two methods that use these terms to represent documents as vectors in a multi dimensional space. Relevance feedback, genetic algorithms, neural networks, and the Bayesian classifier are among the learning techniques for learning a user profile. While some of the learning methods represent the user profile as one or more vectors in the same multi dimensional space which makes it easy to compare documents and profiles, other learning methods such as the Bayesian classifier and neural networks do not use this space but represent the user profile in their own way. Each learning technique has its pros and cons, and a suitable technique can be chosen on further exploration. Note that, the over-specialization problem has to be combated by making the following between two types of information delivery: exploitation, where the system chooses documents similar to those for which the user has already expressed

a preference; and exploration, where the system chooses documents where the user profile does not provide evidence to predict the user's reaction. [7]

Collaborative filtering recommendation systems, in contrast, rely on past user data to make recommendations. For example, a movie review website may look at ratings that a user has made and suggest new movies that other users who have similar tastes have also rated similarly. For PLOS, we will explore several techniques and factors to quantify a basis for recommendation. One possible metric to explore is to access users' traffic history. We can identify users with similar tastes based on the articles they view, and recommend new articles that similar users have also viewed. PLOS also allows users to share and save articles, which may be strong indicators that a reader is interested in a topic and could be used to generate very relevant recommendations. Any sort of collaborative filtering approach will require additional user data from PLOS.

Another plausible approach is using a recurrent neural network for collaborative filtering. In essence RNNs are models that predict a sequence of something. Here we can predict the next article given the user reading history. More formally, let's assume we have time steps 0 to t−1. The model has a "hidden" internal state corresponding to these time steps. These are generally vectors of some dimension k. Every time step, we have two things going on

- Predict the output given the hidden state. We need to model a $P(yi|hi)P(yi|hi)$ for this.
- Observe the output yt and feed it back into the next hidden state. In the most general form, $hi+1=f(a(hi)+b(yi))$. In practice, f is generally some nonlinear function like sigmoid or tanh, whereas a, and b are usually a simple linear transform.



[8]

Automated Data Submission Checks

Developing automated checks is a very large and challenging problem to tackle. Some forms of harm may be easy to identify and others will require very complex semantic understanding of the issues and may be impossible. This portion of the project will require close collaboration with PLOS. We will need to gain a thorough understanding

of PLOS' current manuscript and data evaluation process. In essence, what does PLOS do now to identify potentially harmful datasets? We will explore solutions to the following methods of harm that PLOS would like to identify, and ultimately come to an agreement with PLOS after discussing the requirements and feasibility of each.

- **Human subject data:** Can we detect the presence of personally identifiable information? Does the data set include phone numbers, social security numbers, social media accounts, etc. Additionally, can we identify risk of the ability to re-identify de-identified data?
- **Pseudoscience:** Can we detect pseudo-scientific terms or methods? These include intelligent design and creationism, homeopathy, cosmology, time travel, "alternatives" to well-established theories and physical world laws.
- **Image Manipulation:** Can we run forensics on an image to detect if it has been manipulated? Potential manipulation techniques include splicing, background repetition, contrast issues, and excessive cropping.
- **Endangered species/environments**: Is there risk for these to be revealed with sufficient accuracy to potentially lead to further harm?
- **Dual use research of concern**: For example, could the data be used to engineer gain of function in viruses, reconstruct extinct pathogens, or incite bacterial resistance?

The feasibility of developing success for these topics is largely dependent on our ability to train our models to understand the underlying characteristics and markers. As we are without extensive scientific backgrounds, in some circumstances this will pose a challenge. To develop these algorithms, we will likely need PLOS to supply training data. Much of the semantic understanding techniques that are utilized for the recommendation engine portion of the project may be repurposed here. Additional dictionaries obtained online, for example, to recognize pseudoscience, may be employed. We anticipate using Regular Expressions to identify personally identifiable data.

Deliverables

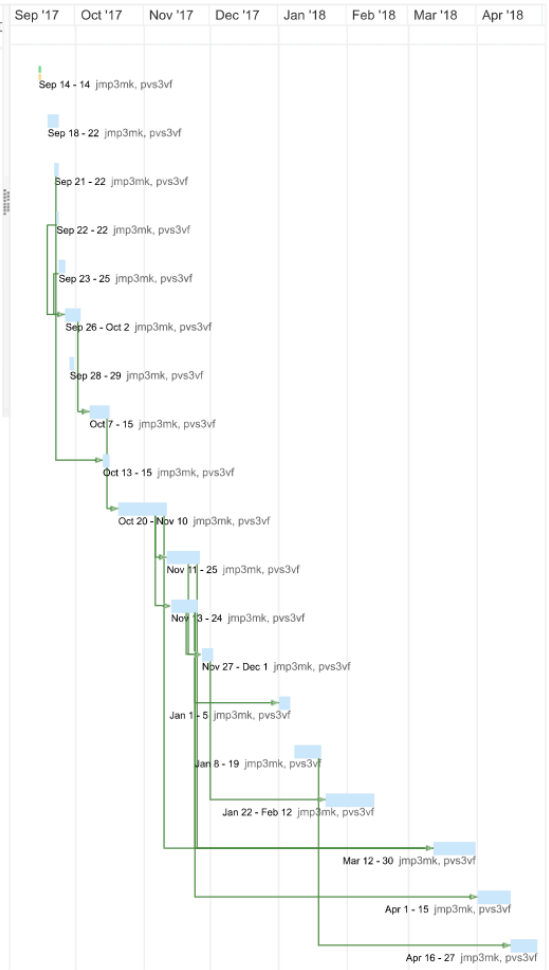
- Network analysis of PLOS' corpus
- Recommendation engine algorithm for suggesting related PLOS content to readers
- Algorithms and implementation recommendations for automated data checks of PLOS submissions
- SIEDS conference paper and presentation
- PLOS sponsor presentation

Resources

We will begin by analyzing subsets of the data on local machines, but once we start working with the entire corpus, we will likely require more powerful resources. We will primarily use Amazon Web Services cloud based server architecture. There is also an opportunity to utilize Rivanna, UVa's high-performance computing cluster.

Schedule

TITLE	%	DURATION	START D
~~Download Corpus~~	100	1 days	09-14-2017
Data Exploration on Corpus	0	5 days	09-18-2017
Ethics In-class presentation	0	2 days	09-21-2017
Set up computing infrastructure on AWS	0	1 days	09-22-2017
Observation Log post Data Exploration	0	3 days	09-23-2017
Perform Indexing on Corpus	0	7 days	09-26-2017
Create Project Proposal	0	2 days	09-28-2017
Analysis using Index Computed	0	9 days	10-07-2017
Ethics Paper for Capstone Project	0	3 days	10-13-2017
Extend MESH terms using semantic analysis	0	22 days	10-20-2017
Build initial knowledge graph	0	15 days	11-11-2017
Visualize networks in knowledge graph	0	12 days	11-13-2017
Develop Recommendation Engine	0	5 days	11-27-2017
Develop Personally Identifiable Data check	0	5 days	01-01-2018
Develop image manipulation check	0	12 days	01-08-2018
Explore and implement additional publisher tools	0	22 days	01-22-2018
Write Paper	0	19 days	03-12-2018
Develop User Interface	0	15 days	04-01-2018
Prepare for Final Presentation	0	12 days	04-16-2018



Budget

The proposed budget for this project is $500 for fees related to the SIEDS Conference. No travel expenses are expected.

Qualifications

Jack Prominski

Jack graduated with a B.S. in Commerce from the University of Virginia in 2013. Following graduation, he worked in New York for the e-commerce fashion website Gilt Groupe in retail planning and analytics. While at Gilt Groupe, Jack collaborated on data science projects to recommend price changes and optimize inventory purchasing. After a brief sabbatical in Nepal, Jack joined Myers-Holum, a data architecture and systems consulting company. Jack has experience in R, Python, and SQL.

Pragati Shah

Pragati graduated with a B.E. in Electronics and Telecommunication Engineering from the University of Pune in India. Her professional career includes 2 years as a Data Analyst with a Bank of New York Mellon company, 6 months as a Technology R&D Manager at Bajaj Finance Limited. She has also completed a Big Data Analytics and Optimization certification program at INSOFE, India. She has experience with R, Python, SQL, Excel, Tableau and MATLAB.

Sources

[1] "Why Open Access? | PLOS." https://www.plos.org/open-access/. Accessed 29 Sep. 2017.

[2] "Data Availability - PLOS ONE: accelerating the publication of peer ...." http://journals.plos.org/plosone/s/data-availability. Accessed 29 Sep. 2017.

[3] "Open Science" https://en.wikipedia.org/wiki/Open_science. Accessed 28 Sep. 2017.

[4] "JATS: Journal Publishing Tag Set - Journal Article Tag Suite - NIH." 8 Jan. 2016, https://jats.nlm.nih.gov/publishing/rationale.html. Accessed 29 Sep. 2017.

[5] "Knowledge-based Graph Document Modeling - MADOC - Uni ...." 28 Feb. 2014, https://ub-madoc.bib.uni-mannheim.de/35464/1/schuhmacher14a.pdf. Accessed 29 Sep. 2017.

[6] "Recommender system - Wikipedia." https://en.wikipedia.org/wiki/Recommender_system. Accessed 28 Sep. 2017.

[7] "Recommender system- Content-based Filtering" http://recommender-systems.org/content-based-filtering/. Accessed 28 Sep. 2017.

[8] "Recurrent Neural Networks for Collaborative Filtering · Erik ...." https://erikbern.com/2014/06/28/recurrent-neural-networks-for-collaborative-filtering.html. Accessed 29 Sep. 2017.