SciReader : A Recommender system for Biomedical literature

Desai, P. [1,2] , Lehmann, B.[2], Telis, N.[2], Pritchard J.P[2]
[1] Stanford Center for Genomics and Personalized Medicine (SCGPM),
[2] Department of Genetics, Stanford University

Motivation

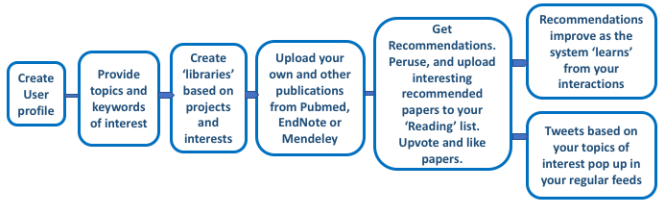
With the recent explosion in biomedical research, it has has become increasingly important and yet challenging to keep up with the relevant literature. SciReader is a personalized recommender system that specifically aims to help researchers and practitioners in the biomedical community parse through the large volume of literature and filter publications that may be relevant and of interest to them.

SciReader was initially developed at the Pritchard lab (Genetics department, Stanford School of Medicine) and is now maintained and operated by the SCGPM. It is currently being migrated to Google Cloud and should be available soon. (http://scireader.org)

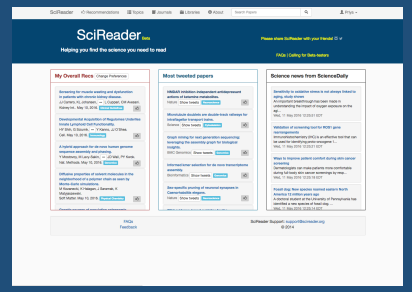
Introduction

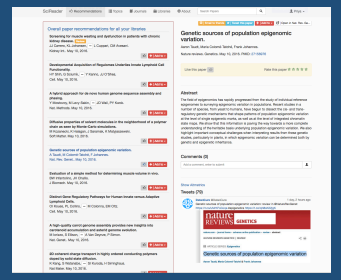
- SciReader is a cloud based service that uses novel algorithms to classify and cluster published biomedical corpora using topic modeling (Latent Dirichlet Allocation).
- Users provide basic info: i.e. topics/keywords of interest and journal papers.
- Best results when user creates a 'library' and upload papers of interest to it. Can create Personalized recommendations based on relevancy, recency, impact factor and sentiment analysis –updated daily.
- Weekly email digests of important publications in your field of research.
- Relevant trending twitter feeds provided in real time.

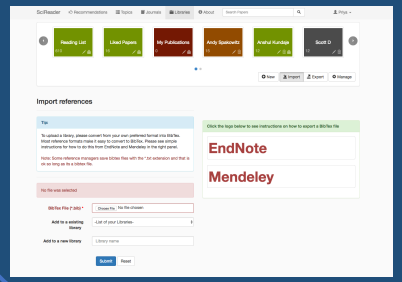
User Work Flow

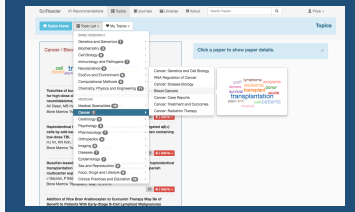


SciReader Screenshots

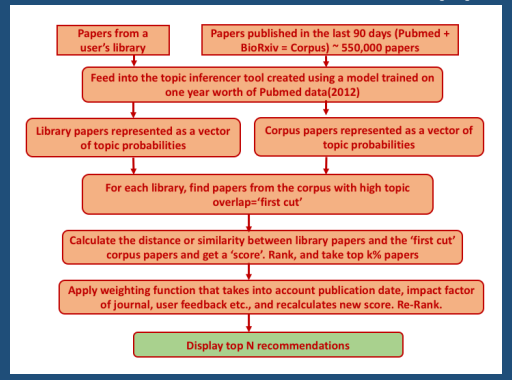








Basic overview of the Recommender pipeline



Topic Modeling of Pubmed/BioRxiv using LDA

- The cornerstone of Scireader is its topic model of Pubmed.
- Topic models represent a class of computer programs that seems to 'automagically' extract underlying themes or topics from large unstructured texts.
- LDA: Latent Dirichlet Allocation, a topic model algorithm
- Mathematically:

$$P(Z|W,D) = \frac{\#\ of\ word\ W\ in\ topic\ Z + \beta_w}{total\ tokens\ in\ Z + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$

- We used Titles and Abstracts from all the articles published in pubmed in 2012 (~1.2 million) to train a LDA model which was then used to create a 'topic inferencer'.
- Our topic model has 150 topics which were grouped into 20 'supertopics"
- All articles from pubmed and bioRxiv have been 'topic modeled' using this inferencer

Example topics 'discovered' by LDA

genomics

Obstetrics

Blood cancers

Summary

- SciReader is a great way for researchers and medical practitioners to stay abreast of advances in their field.
- SciReader's topic model and database can be used as a research tool to perform longitudinal studies on history of disease based on Publication data and other bibiliometric studies.