# NCBI Document Recommender

Andrew Y

## Abstract

For students, finding relevant research papers for their studies is a problem faced in every assignment outside of their expertise. The goal of this project is to create a useful tool to find papers useful for their desired topic. Scope-wise, the model will work with documents from the National Center for Biotechnology Information, or the NCBI, as they have the an extensive database of documents relating to biology, mainly bio-informatics, bio-medicine, and bio-technology. The desired outcome is for the user to be able to input a keyword, and receive a list of documents that would be relevant to their work. To do this, our project will use the NCBI API to pull in documents from the database, and then utilize an unsupervised learning model to find a list of relevant documents to return.

**ACM Reference Format:**
Andrew Y. 2018. NCBI Document Recommender. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

The project is a recommender system model (tentative) where a user will be able to input a keyword to find a list of documents whose topics are relevant to the keyword. The problem this model attempts to solve is to help minimize the effort required to find useful sources or references to use when one is attempting to do educational work outside of their field of expertise. For example, if a student was writing an essay about different allergies for a class, then they would be able to enter the keyword "hay fever" and be able to receive numerous published educational pieces around hay fever. While not life changing, this would help reduce the time spent browsing numerous documents and allow the user to allocate more time to other tasks of their assignment. To help keep the scope manageable, the project will be working with data from the National Center of Biotechnology Information, or NCBI for short. They are a research institute that is a branch of the National Institutes of Health, and is known for housing a multitude of databases containing tools and services relating to biotechnology and biomedicine. They also house databases dedicated to bio-related literature, which is what this project will be pulling the data from.

## 2 Related Works

As a recommender system, it can utilize obvious references to the renown Google search engine. In addition, NCBI already have their own search function for documents, a tool named Entrez, which is their dedicated text retrieval system. However, their existing search function only has four different ways to sort their result, which are article order, publication date, journal number, or PMC live data. The goal of the project is to create a different way to receive the results, one that utilizes the document contents rather than their metadata which is different from the existing methods. In doing so, we can ensure that the resulting list of documents will be of more use to the user than scrolling through documents listed in chronological order.

## 3 Proposed Work

The way the project will be tackled will be to first utilize Entrez to pull a list of documents from the database. Next we will commit data wrangling onto the document abstracts. We do so on the abstracts as with the average document length and number of documents, it is more feasible to utilize the abstracts as they can be seen as shorter summaries of each document. This could be treated as a dimension reduction technique as well. Afterwards we will feed the abstracts into a TF-IDF vectorizer to put it into a form more suitable for machine learning.

We will then have to deal with the task of determining if a document will be significant enough to warrant a recommendation. One way is to tie significance to a metric that each document has which is reference number, or how many times it has been referenced by other documents; highly referenced documents should be of higher significance if they are often used by other authors. This could also work as a way with dealing with the problem recommender systems commonly face of cold starts. The next major task would be to determine how to find a group of documents to return. The two methods to be explored would be to create a recommender system that creates a list of documents that are similar through cosine similarity, or to utilize clustering methods to cluster the documents into groups and return every document in the cluster.

The main issue would be the lack of a way to evaluate a recommender system as there will not be any feedback or ground truths to utilize for measuring metrics. If an idea for

*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

a proper metric does not appear, most likely will have to swap to utilizing clustering methods.

## 4 Evaluation

As mentioned prior, while a recommender system seems the most useful, there is a need to find a way to evaluate its performance when there is no feedback. On the other hand, if a clustering model is used we can measure metrics such as silhouette score or calinski harabasz score to view our model performance. In terms of clustering methods, K-Means clustering will most likely perform better than hierarchical clustering due to the fact that we can preemptively set our K value as the number of documents we want to return, though both methods will be tested.

## 5 Discussion

The first priority of the project timeline is to get the data ready. This will involve utilizing the Entrez API to pull data from the database, and doing extensive cleaning of the text data to better prepare it for modeling. The second milestone will be to ensure that our metric of determining whether or not a document can be considered significant enough to be considered for returning. The third milestone will be to be able to create a model that will return a list of significant documents. Once theses steps are finished, we can undergo model optimization and performance tuning.

The potential challenges mainly resolve around evaluating the recommender system, in which the clustering method will act as a suitable backup plan as a way of improving performance.

## 6 Conclusion

In the best case scenario, the project will end up as a useful tool for users that are undergoing biology related studies, where they can simply enter a keyword and receive an extensive list ( or short list depending on the user's wishes ) or potential documents for the user to study and utilize for their work. < Add key findings later > < Add future work later >