NCBI Document Recommender

Andrew Y

Problem Statement

- As students, we are often tasked to write essays on subjects that we are not professional experts in.
- To combat this, we instead research work done by people that actually are practiced in the subject in the form of reading their articles / textbooks / blog pieces.
- However, finding relevant pieces of writings can still prove to be difficult with the abundance of research in the field.

- **SOLUTION:** Create a recommender system model that would recommend research articles based on a given keyword.

Related Work

- Will be able to reference the world famous Google search function.
- However it will be much smaller in scope, and be focused on recommending documents from the National Center for Biotechnology Information, or NCBI
- The NCBI has a database dedicated to papers relating to
    - Bioinformatics
    - Biomedicine
    - Biotechnology
- The NCBI database does have an existing search function, but their sorting is based on article order, publication data, journal, or PMC live date.
    - Goal is to create a model that would recommend documents based on their contents.

Proposed Work (Tentative)

- The proposed method is to pull a dataset of documents from NCBI and then returning a list of recommended documents for the user to consider utilizing.
- The process of extracting data from NCBI can be done with Entrez, their dedicated text retrieval system.
- Big task is deciding on how to find significant documents / deal with cold starts:
  - One way is to tie significance to a metric that each document has which is reference number, or how many times it was referenced by other documents; highly referenced documents should be of higher significance if they are often used by other authors potentially
- Second important task is to create a list of relevant documents:
  - Will build the recommender system here, can take a significant document from the previous task and find similar ones?
  - Most likely will use document abstract rather than the actual contents as it is a viable summary, can treat it as a dimension reduction technique.

Evaluation (Tentative)

- Need to find a way to evaluate a recommender system that returns a list of documents when there is no feedback.

- If this method does not work out, might swap to a clustering model where it would return documents in the same cluster.
- If so, silhouette score would be a good metric to evaluate the model, where higher scores would result in a tighter cluster, and thus more similar recommendations.
- K means clustering should prove to be more useful than hierarchical as we can set the number of documents we want returned as the K beforehand, though both will be tested

Timeline (Tentative)

First milestone: Successfully pulling data, and finish data wrangling

Second milestone: Create a metric for finding if a document is significant

Third milestone: Create a method of returning a list of significant documents

Fourth milestone: Evaluate and Improve Performance