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.

Problem Statement Cont.

SOLUTION: Create an unsupervised learning 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 date, journal number, or PMC live date.
    - Goal is to create a model that would recommend documents based on their contents.

Proposed Work

- 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.
    - Project Checkpoint: As we are working with text data, the data has undergone all the typical text-based data cleaning, such as removing stop words or word lemmatizing.
    - Project Checkpoint: The text data has also been run through a TF-IDF Vectorizer for ease of user for later modeling.

Proposed Work Cont.

- Main task is deciding how to find significant documents / deal with cold starts.
  - Project Checkpoint: Decided to utilize the ‘reference number’ value for each document as a metric for determining significance. However, as a good portion of documents are missing reference number, a linear regression model will be created to predict it for documents w/o it.
- Second important task is to create a list of relevant documents.
  - Project Checkpoint: Will create a clustering model to group documents into a number of clusters, and return the cluster with the highest sum total of reference numbers.

Evaluation

- Need to find a way to evaluate a recommender system that returns a list of documents when there is no feedback.
  - Project Checkpoint: Swapped the main model used from a recommender system to a clustering model.
  - Project Checkpoint: 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.
  - Project Checkpoint: 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

Evaluation Cont.

- Project Checkpoint: Ran into a new issue where a regression model will have to be created to help predict missing reference numbers based on their text data, and potentially author.
    - Poisson Loss Regression metric from statsmodels library will be the best choice for this evaluation due to its specialization in count data, which is what the reference numbers are.

Timeline

First milestone: PC: Successfully pulling data, and finish data wrangling

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

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

Fourth milestone: Implement the models

Fifth milestone: Evaluate and Improve Performance

References

Entrez’s Bio Python Package

- https://biopython.org/docs/1.76/api/Bio.Entrez.html