Using NLP to interpret medical records

Jing Lee

Abstract

Healthcare records are unstructured textual data holding valuable information about patients and diseases. Since the data is unstructured, we can't do analytics directly on them in order to extract meaningful information for future research.

Natural language processing (NLP) can be used to extract valuable information from medical records of patients. The documents were preprocessed by removing words that are not meaningful and then tokenized into words. A document-term matrix was built using a vectorizer that will assign a value to the frequency of words. Features were reduced to 10 topics using topic modeling. Finally, a recommendation system was built that will find similar medical records based on a current document.

I. Design

The client will be hospitals who could use the NLP model to extract valuable information from their databases of medical records to improve patient care.

II. Data

The dataset was obtained from Kaggle. This dataset contains 4,998 medical records across 38 different medical specialties. Duplicate entries were removed and a subset of medical categories were selected for the analysis. The final dataset contains 1,740 documents.

III. Algorithms

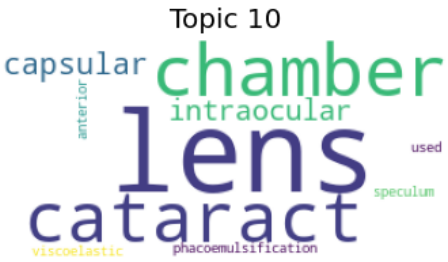
An unsupervised model was built by tokenization of the documents using NLTK. The cleaned document was vectorized using TF-IDF (term frequency inverse document frequency) and then topic modeling (NMF: non-negative matrix factorization) was applied for dimension reduction of 10 topics. The terms associated with the topics were not readily identifiable. Some topics contain terms that were too general or are medical jargons. The results were integrated into a content-based recommendation system. The user will input a document and 3 similar documents will be recommended for viewing.

IV. Tools

- Text processing: python, text processing libraries (NLTK, scikit-learn)
- Visualizations: pandas, matplotlib, WordCloud

V. Communication

Below is an image of 10 terms associated with 1 of the 10 topics. Topic seems to be about the eye.



In addition to the slides and visuals presented, all information can be found on the following GitHub repo: https://github.com/lee-jin81/metis_project_5_nlp