Using NLP to extract information from medical transcripts
Project 5: Natural language processing
By: Jing Lee

I. Question/Need:

What is the question behind your analysis or model and what practical impact will your work have?

Healthcare records are unstructured textual data where the medical professional takes notes at the time of the patient's visit to record information such as biostatistics, overall health, and any new concerns. This information can be helpful in the diagnosis and treatment of future patients. However, since the data is unstructured, we can't do analytics directly on them in order to extract meaningful information for future research.

I'm proposing to use NLP to extract valuable information from medical records of patients. This can be used to build a recommendation engine that will find previously diagnosed patients who have conditions similar to current patients who need medical attention and advice. Medical professionals can use this information in addition to their diagnosis to further improve patient care.

Who is your client and how will that client benefits from exploring this question or building this model/system?

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 Description:

What dataset(s) do you plan to use, and how will you obtain the data? Please include a link!

- Source: Kaggle Medical transcriptions

https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

- Data set: 4,998 documents across 38 different medical specialties

What is an individual sample/unit of analysis in this project? In other words, what does one row or observation of the data represent?

Description of medical record (short summary), medical specialty, diagnosis, transcript and keywords

What characteristics/features do you expect to work with? In other words, what are your columns of interest?

Transcription

If modeling, what will you predict as your target?

N/A

III. Tools:

How do you intend to meet the tools requirement of the project?

- Text processing: Python, text processing libraries (NLTK, spaCy, gensim, scikit-learn)
- Visualizations: matplotlib

Are you planning in advance to need or use additional tools beyond those required?

- Tableau: visualization
- Python: recommendation engine

IV. MVP Goal:

Preprocess the text data resulting in a document-term matrix and further tune model based on early findings.