# Summarization of PubMed Articles

### Datasets

- We train the model on the medical research papers in the [Open Access Subset of the PubMed Central Corpus](https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/).

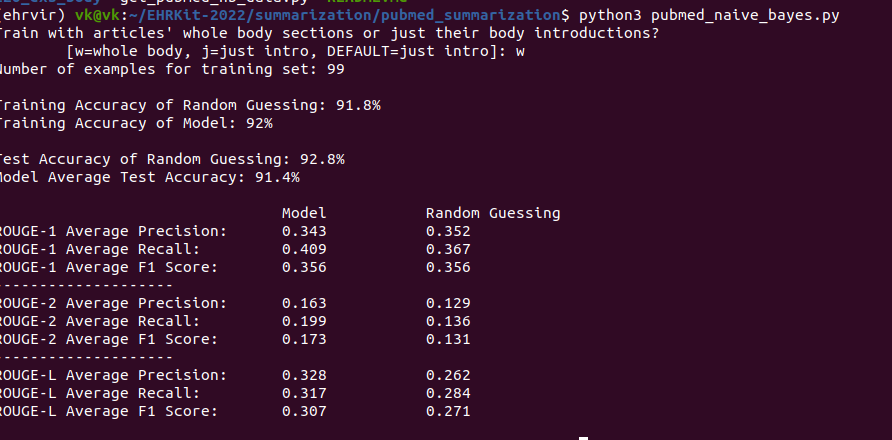
- We can apply the trained model to the EHRs in the Medical Information Mart for Intensive Care ([MIMIC](https://mimic.physionet.org/gettingstarted/access/)) database.

### Extractive Summarization with Naive Bayes

Running pubmed\_naive\_bayes.py performs Naive Bayes summarization on the parsed Pubmed articles. To fit this model, feature vectors are constructed the first time it is run for a given number of articles and body type.

On running the pubmed\_naive\_bayes.py and giving the number of training set and specifying the type of content we get the model trained with training and testing accuracy

This creates a new directory, /PATH/TO/EHRKIT/summarization/pubmed\_summarization/N\_exs\_BODY\_TYPE/. This contains files with the articles selected for training and test sets. Inside this, the directory nb/ contains the feature vectors (feature\_vecs.json and test\_json/), as well as the model's predicted summaries (test\_summaries/) and their ROUGE scores (ROUGE.txt) on the test set.



Each feature vector represents one sentence, taking into account its number of nouns, its length, the average frequency of its words in the document, and the inverse sentence frequency. (Inverse sentence frequency measures the average importance of the words to that particular sentence relative to the document in general.) The sentence's label is 1 when it is in the abstract, and 0 otherwise.