Pinakin Nimavat

pnimavat@hawk.iit.edu

Topic Modeling Report
(Machine Learning)

1 Problem Statement

Using given data uncover the common features for a given medical condition.

What I inferred reading the above statement.

- By skimming the problem statement and looking at the given data I figured that my task is to find the underlying common factors for a medical condition using topic modeling techniques. Where each medical condition will be my topic and the keywords that represents that topic will be the underlying common factors. Also, this is an unsupervised machine learning task.
- I have been given patients' electronic health records and its annotated files in a .ann format which has used BRAT schema to annotate the key classes.
- .txt files which are patient's EHR(Electronic Health Record) contains vital information about a patient's history from the time of admission to the discharge period.

2 Solution and Analysis

As the key points that we are interested in an EHR is Chief Complaint which gives us the medical condition, History of Illness and Discharge Diagnosis which contains situation of patient when they were under treatment with a doctor, observation from the doctor's perspective, symptoms and other vital signs.

Here I propose two approaches and will share in detail why this is the best way to fulfill our task.

1. Converting the unstructured data to structured format (a pandas dataframe) and Using pre-trained weights to extract the desired information (such as texts which talks about medical condition, symptoms, medicines etc.) and perform any of the topic modeling techniques (either rule based or ML).
2. By creating a custom NER model to extract and tag medical condition names and supportive classes, train another model on publicly available datasets (such as MIMIC dataset) to extract particular portion of EHR and then use them as input to train our topic modeling algorithm.

I chose to go with the first approach due to time constraints.

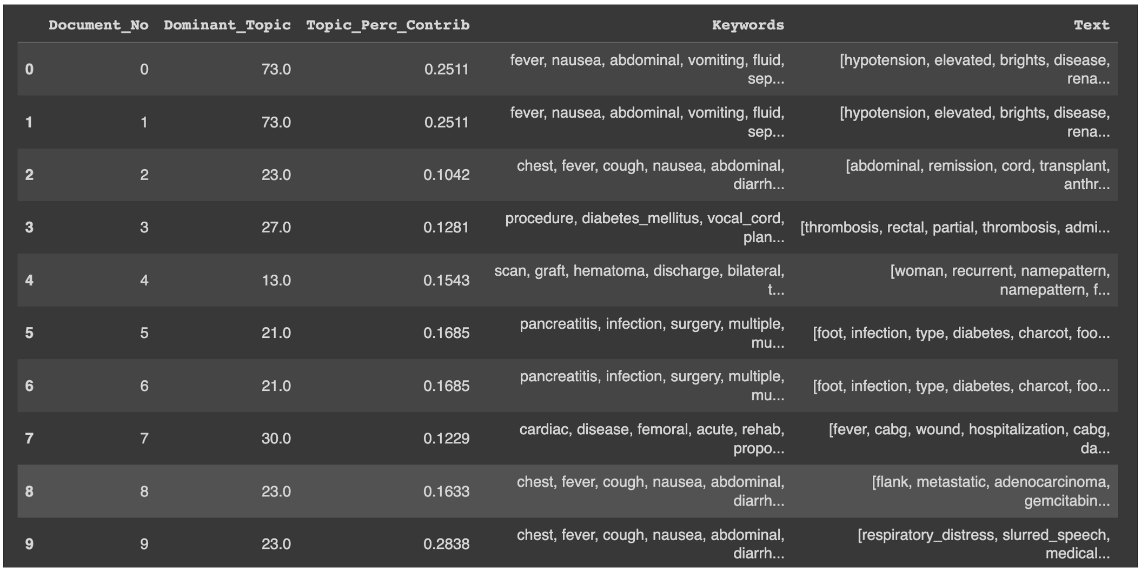
What is my solution?

- By analysing the problem statement and looking through a few texts files, I developed an extraction function which takes these raw text files as an argument and returns a dataset that contains three features; 'Chief Complaints', 'History of Present Illness' and 'Discharge Diagnosis'.
- I chose scispaCy's **en_core_sci_md** as a spaCy model to process and extract the biomedical related textual entities from the clinical texts. Reason behind choosing **en_core_sci_md** is that it has larger vocab and 50k word vectors. I had to consider which ones are small, concise and takes less time to load and then **en_ner_bc5cdr_md** which is trained on BC5CDR corpus and contains entities which are more related to cellular component, gene product and cancer related items.
- At this point I have preprocessed texts which are curated by removing unwarranted patterns and ready to feed to a model.
- Choosing a perfect model to seek out the best performance factors on a lot of things. There are lots of available algorithms through which we can find the relation between a topic and underlying words.
- I have used Latent Dirichlet Allocation, Hierarchical Dirichlet Process and BERT topic modeling.

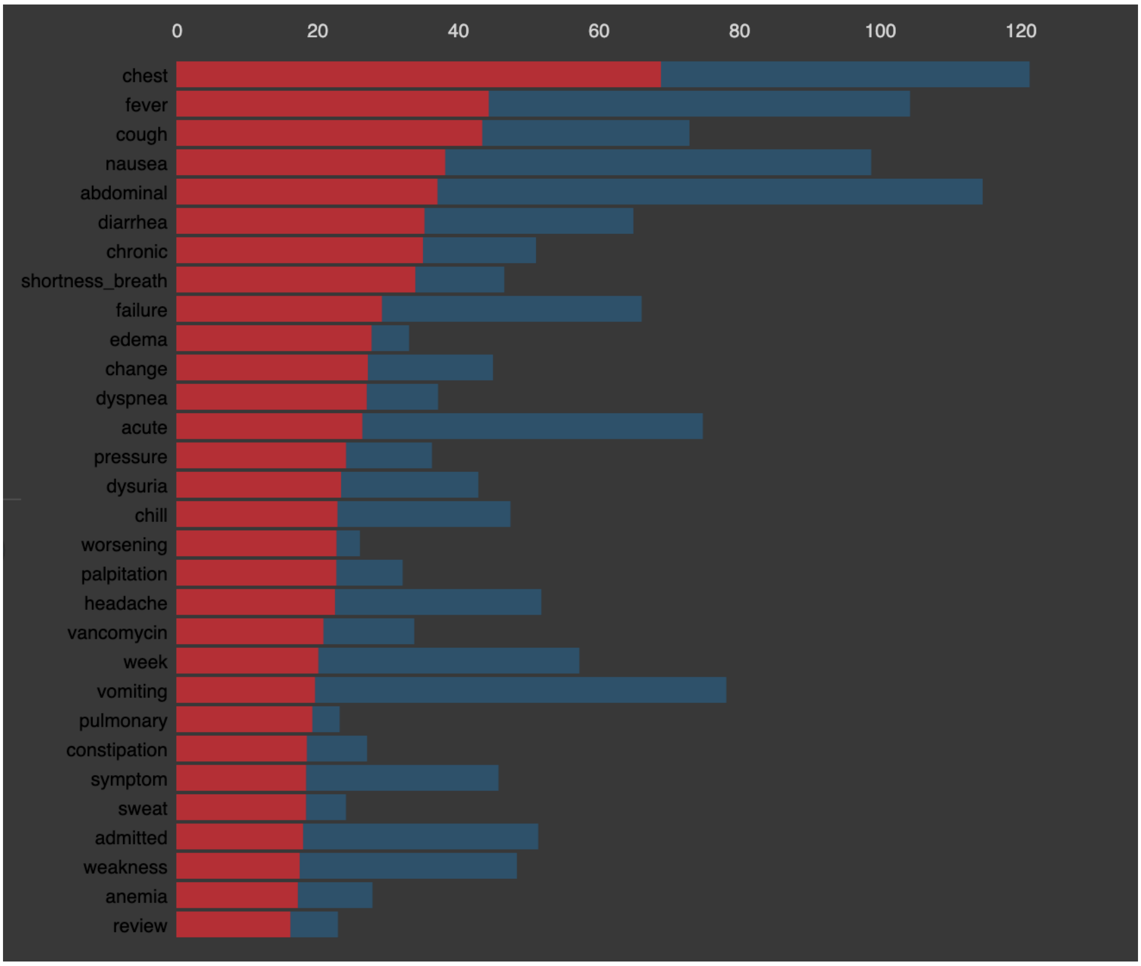
Why LDA, HDP, BERTopic and not LSA

- First, I will argue as to why I have not used Latent Semantic Analysis. LSA computes only one representation for each term. In a way, LSA “selects” what meaning of the term is more prevalent.
- Also primary drawbacks of LSA are lack of interpretable embeddings (as we can’t know what the topics are and the components can be arbitrarily positive or negative).
- LSA has less efficient representation and requires large set of documents and vocab to get accurate results.
- From this point, I will argue as to why what I have implemented works well. LDA is generative probabilistic model of a corpus where documents are represented as a random mixture over latent topics where each topic is characterized by a distribution over words.
- In laymen terms, LDA is better in generalizing topics because LDA sets α a prior on the per document topic distribution and β a prior on the per topic word distribution.
- In my models I have set alpha = 1 which means that density of topics in a document is higher or in laymen terms I am assuming that the documents share many topics.
- ETA in my LDA model is 0.10 which means that individual documents talk about few topics, which are basically the prior β which sets the per topic word distribution.
- In order to find the topic number I have developed a function which will find the best topic based on the coherence value of LDA model in the range of [20,300]. In my case I got the best value of coherence for 90 topics.
- The thought of using HDP came from an idea that rather than treating each document and its feature independently; what if the features are drawn from a shared distribution and that is what the HDP does?
- We do not have to specify the number of topics to look for, HDP process infers the topic numbers. In this case the it found 142 topics.
- I have also tried BERTopic modeling to see the power of transformers. I have used a pre-trained model weights $'sarahmiller137/BiomedNLP - PubMedBERT - base - uncased - abstract - ft - ncbi - disease'$, which did not perform upto the expectations as the name suggests and that it was trained to find the abstract text from NCBI_dieses dataset. It requires more accurate embeddings and fine tuning.

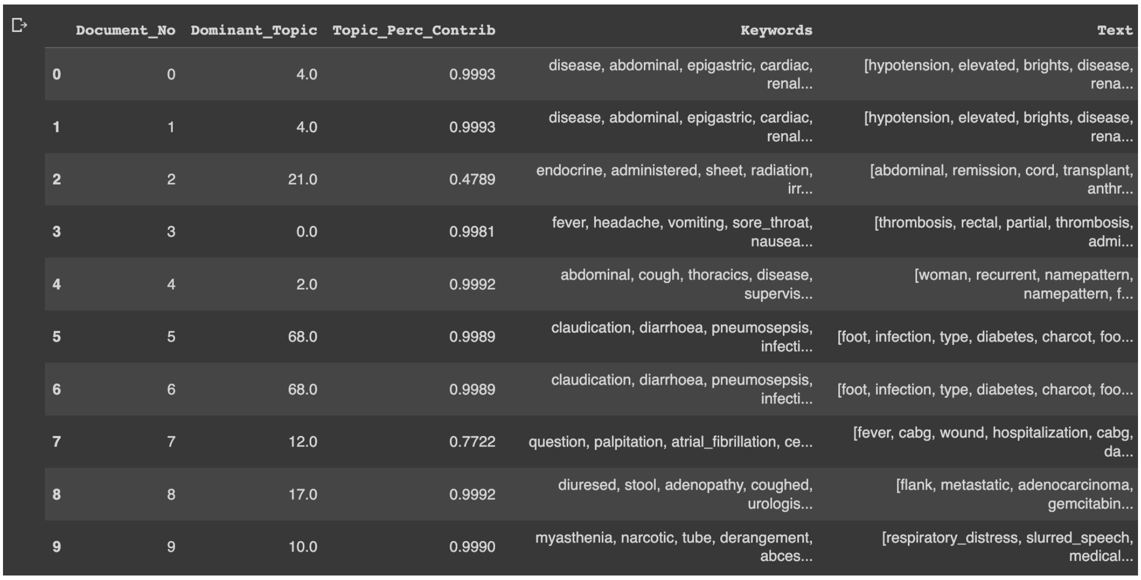
As we can see below for our LDA model, it shows the information about dominant topic and keywords for a document.



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	73.0	0.2511	fever, nausea, abdominal, vomiting, fluid, sep...	[hypotension, elevated, brights, disease, rena...
1	1	73.0	0.2511	fever, nausea, abdominal, vomiting, fluid, sep...	[hypotension, elevated, brights, disease, rena...
2	2	23.0	0.1042	chest, fever, cough, nausea, abdominal, diarrh...	[abdominal, remission, cord, transplant, anthr...
3	3	27.0	0.1281	procedure, diabetes_mellitus, vocal_cord, plan...	[thrombosis, rectal, partial, thrombosis, admi...
4	4	13.0	0.1543	scan, graft, hematoma, discharge, bilateral, t...	[woman, recurrent, namepattern, namepattern, f...
5	5	21.0	0.1685	pancreatitis, infection, surgery, multiple, mu...	[foot, infection, type, diabetes, charcot, foo...
6	6	21.0	0.1685	pancreatitis, infection, surgery, multiple, mu...	[foot, infection, type, diabetes, charcot, foo...
7	7	30.0	0.1229	cardiac, disease, femoral, acute, rehab, propo...	[fever, cabg, wound, hospitalization, cabg, da...
8	8	23.0	0.1633	chest, fever, cough, nausea, abdominal, diarrh...	[flank, metastatic, adenocarcinoma, gemcitabin...
9	9	23.0	0.2838	chest, fever, cough, nausea, abdominal, diarrh...	[respiratory_distress, slurred_speech, medical...



We can see that LDA model was able to achieve generalization of keywords with a cv coherence value of 0.5684. We can say that the model performs good enough. Following is attached topics and keywords from HDP model with a coherence score of 0.74.



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	4.0	0.9993	disease, abdominal, epigastric, cardiac, renal...	[hypotension, elevated, brights, disease, rena...
1	1	4.0	0.9993	disease, abdominal, epigastric, cardiac, renal...	[hypotension, elevated, brights, disease, rena...
2	2	21.0	0.4789	endocrine, administered, sheet, radiation, irr...	[abdominal, remission, cord, transplant, anthr...
3	3	0.0	0.9981	fever, headache, vomiting, sore_throat, nausea...	[thrombosis, rectal, partial, thrombosis, admi...
4	4	2.0	0.9992	abdominal, cough, thoracics, disease, supervis...	[woman, recurrent, namepattern, namepattern, f...
5	5	68.0	0.9989	claudication, diarrhoea, pneumosepsis, infecti...	[foot, infection, type, diabetes, charcot, foo...
6	6	68.0	0.9989	claudication, diarrhoea, pneumosepsis, infecti...	[foot, infection, type, diabetes, charcot, foo...
7	7	12.0	0.7722	question, palpitation, atrial_fibrillation, ce...	[fever, cabg, wound, hospitalization, cabg, da...
8	8	17.0	0.9992	diuresed, stool, adenopathy, coughed, urologis...	[flank, metastatic, adenocarcinoma, gemcitabin...
9	9	10.0	0.9990	myasthenia, narcotic, tube, derangement, abces...	[respiratory_distress, slurred_speech, medical...

3 Future Work

- In future we can make vectors out of these topics and suggested keywords and further use it in the pipeline such as QA model, classification, summerization tasks, chat-bots etc.
- The second approach which I am suggesting in the beginning of the document have three models.

- First one is to create our own embeddings which will contain information about medical condition and symptoms names. We can use publicly available MIMIC dataset to train a Named entity recognition pipeline to extract relation and part of speech tagging.
- Once trained, apply the same weights on our unstructured data to extract the keywords and the data it contains.
- Once we have created a dataset, we can apply neural model to extract topics and keywords. We can fine tune the transformers as per our needs and use it (such as BIOBert) for creation of topics.
- Third model will be our classification task or question answer model which will depend on our specific use case.

4 References

- https://radimrehurek.com/gensim/models/ldamodel.html?highlight=ldamodule-gensim.models.ldamodel
- https://radimrehurek.com/gensim/models/hdpmodel.html?highlight=hdpmodule-gensim.models.hdpmodel
- https://maartengr.github.io/BERTopic/api/bertopic.html
- https://allenai.github.io/scispacy/
- https://huggingface.co/docs/transformers/index
- https://mayoclinic.pure.elsevier.com/en/publications/clinical-information-extraction-applications-a-literature-review
- https://arxiv.org/pdf/1711.04305.pdf
- https://arxiv.org/pdf/1904.03323v3.pdf
- https://people.eecs.berkeley.edu/ jordan/papers/hierarchical-dp.pdf
- https://arxiv.org/pdf/2110.15763v1.pdf