Tokyo Workshop on Genomic Medi...eutics, and Health 20240413-03

Sat, Apr 13, 2024 11:17AM • 1:14:07

**SUMMARY KEYWORDS**

topic, patient, icd, call, document, data, learn, modality, work, words, top, model, code, multimodal, characterize, based, problem, predict, normalize, sum

00:03

Maybe we can call 30 divided by the author Alright, so I appreciate you guys so again, the second part was switch gears up

00:38

so the learning objectives

01:03

So, the basic concept of poly modeling well how many of you know about latent Dirichlet allocation

01:18

your facility familiarized with basic concepts on the monitor and then we want to learn how to do multimodal profit using the electronic health record. And now we want to dive into the specific value for value which can solve the identifiability problem very often you have difficulty identifying what pocket

01:49

so, and then the fourth one, we're going to talk about a another reason to be using so called embedded within leveraging deep learning and then the knowledge work right after lunch and then we talked about the applications and then include the ICD code and we see the healthcare data as well as in itself reports. So, these are available from the type of data so we make use of the code. So, what is the lifetime health record? So, there's really a kind of very broad umbrella that includes pretty much all phenotype related metrics, starting from clinical notes, these are surprisingly were very often difficult for patient nurses. These are by far the patient states call me putting a non prevailing post. Usually what's available is largely an appealing qualified diagnostic tool so that's what type of individual has many, very often different new for a building purpose, at least in North America. When you visit a hospital, you feel for your house insurance, very often the hospital purpose for that word ICD code, report and what's the purpose of this visit? Because all cases are or what's the purpose of your stay in the hospital inpatient? So this is a very systematic metric. That is a little bit detail on that. So the really cool also include this a procedural code right? It's called the CPT Current Procedural Terminology. So this, let's say your patients are going through surgery, right. So there are specific holder that have asked for a funeral. And the laptop also has a standardized protocol, like abbreviations that have various so you can use that in many different hospitals. So that means it's a visual and it's not. So you can train a model on one dataset and then apply to others. And then the pharmaceutical company, sorry, pharmaceutical data includes sort of medication. So the RX norm is that one of the popular format that people use and then there are other ones like geology and ride that's already grabbed for these are also under for the guitar because he gets really bad words type of thing and multiple so this is a all data survey increase off the box in 2008. Mostly United States, United to have pretty much all the states has adopted, at least a very basic clinical note that few other modalities, and then the comprehensive 30 by one. So that kind of motivates the incentive, divided method that can really leverage it's a massive amount of human nature across hobbies. So ICD codes, probably many of you already know. So, currently, ICD 10 is the most universal that people use. We have used earlier version of ICD nine. So now ICD nine and 10 they have a similar structure. So starting with the first three digits, so these are the categories and then the festival is a more specific copper location or subtitle. So that's, I think, I think you can see that I think the nine has the three digits to the pattern I started with automatic and then with a few minutes and then followed by for your parents who are similar that can be the maintenance has a unit rate depends your aeration. So our understanding of the disease has to continue before gonna it's gonna be at the level to be as a default. We have a more fine grained understanding of these to have a better characterization. Hopefully we can contribute to that. Later versus more AI approach.

06:22

So, but solely rely on the ICD code is not adequate, very often phenotyping people up as a new base that is, for example, diagnosed later, up call to worry, that's a category for three days. When I will do it, are you going to maintain low health record and see whether they have a full report and now you're also going to pull other ones that are mutually good practice? So that's the rule so whether you can see that this is.

07:20

So, our understanding of the disease is not complete. So manually classifying patients based on defined rule based approach is error prone. It

07:41

one simple approach is to have multiple offers is when you have 1000s of people. Your model doesn't scale, right? You don't have that much for example. And then often for a disease like our early onset, Alzheimer's is leading weekly label, right even the label itself is not confident because people will know her hurting outset. So that's the challenges of using Google is what we want to do. We want to actually learn from mistakes, and then really find the phenotypes and also to personalize predictions based on this individual patient's record. What's the risk metric field? So the the process that we're going to use that we're going to automatically characterize the topics. So what we do so the assumption is that these phenotypes are not standalone, they're not in isolation, they are actually corded. So we call this a comorbidities allow you here this is a field of medical home into one of these modules. Now, Bob is going to correspond to some type of psychiatric. Well, there are many different types that fall into this. We want the model to learn from the data to accurately characterize differences. So we call this sort of main topic and that's probably the one or two too many different topics. This topic characterized for curves that offer a set of thesis. So now once you have these main topic, you can then use them as a template. Now for occasion, have a maybe have record one, two or three and then you mash with it. And now you want to express your probabilistic sort of belief or for whether this individual have these holdings. So that's sort of we call this a mixture of memberships that you can have a membership for a gym membership for a membership. So that's a sort of mixture of numbers. You don't have to subscribe exactly one or you can have multiple. That's like a disease. You should individual patient might have a multiple

10:07

excuse me See All

10:12

right. So this, the address here is that if there belong to the same topic, the ads tend to be have a sticker, not with actually I have to admit that this was taken maybe in the long run while my memory on this is just a high level of what we're trying to achieve, right. So.

10:43

So basically, you have this kind of phenotype if you don't know, yes or not. So now you have this a four five and you can see that these are the molecules just like the gene, you have. Right. So you move into this, you have this layer, you have this sort of the edges are computed by the correlation between the highly correlated evidence they tend to co occur in the same side of the field. So the Yeah, so we're gonna do the lender cases. I'm gonna briefly introduce what is LDA. So, once again, this lecture is very, I'll say it is a little bit dense, but this is really expose you to these techniques. When you really want to learn detail and you want to do more offline research. I'm happy to discuss so here is a high level concept of commonality. So now you saw you get in a coffee shop and then you have descriptions of these how they work because a lot of a document as you know, it can contain multiple topics. For example, you hear a article of all MIDI sequences. So this article, so the highlights are colors that correspond to specific topics. For example, here, you have a two parter colors. So this come from this topic. That is talking about in DNA united. Now I have this thing COVID is coming from another positive while Eliza Your mood is coming from computer I believe it is about magmatic right from from computer. So you started to kind of have a mixture of a lot of course, whether you would just look at an article we don't know which you know which topic is open conference. You find we don't even know the topics. So it was really hard to figure out what this topic distribution is. That each of the following is a categorical multinomial distribution, the distribution mean that you have a probability for each of the words under this topic. So this the number of words is the vocabulary size for the number of adults, so the probability will come one under this and now you can see that the top works for different topics, there'll be different for each topic that they characterize a distinct theme. So that's the topic or you're trying to infer the topic introductions. So this is a global, this is independent from specific document but at the same time, we also want to differ for each topic with popular culture. The two of you facts are interdependent, right. So once we know which topic the opponent can run with and then make use of that to generate sufficient statistics to compute the topic distributions and obviously we know that poverty is huge. And now we can do that to defer the probability for these COVID were which topic, so the two hours so that's the difference. All right. So, back to the ICD code. What we do is, let's say we, we have sort of, for each patient record, we have the ICD count. So let's say you know, it's the count backwards. So here are for individual index, index resulting in a document per patient, or EHR document. And then we have the year we treat this as a bag of words. That's what the back of revival orphan is IVs work, we don't care about or we hear it equalizer decide they don't have much more or less than last year, not going to have a cup of water, or let's say people process. Now you can be mostly you have only like a first visit most of the patients, but that's the sort of process so now sort of first for example

15:03

one reference, same word, I think a lot of

15:12

times. Now what we do is that we're gonna load it into a lot of accents. If you've heard three times. We've repeated it so far. I think before we do that, you might wonder, Well, the reason we do that is we kind of do want to simulate like for example, I can go back to the previous so that they were accounted for multiple times right in the document. But you have that sort of thing, work. You can occur in different places. You might come from different topics, right depending on the context. When you hear we kind of go this is a long bytes. And then we saw that therefore we can infer which topic these topics. This is the whole one and then with the new topic as well. So yeah, just to kind of summarize we have patients we treat it as a document. So this is an analogy for you. So I think you call it this is probably in the document I have to be called from. So the total count of ICD nine. Let's say you have the word W for you guys who are being face to face in the summer. That's a great thing. So I'm up in New York where you can go from there. So now we have our natural breeze Colton. We have to like sample for a minute Hey, baby. Let's take this auto truck only. We've actually assigned to one where the level is exactly one. But then at the patient level is the Mitch remember. Last so given the name lasercut. Okay, so probably pharmacists that work for general view of what they're introducing in 33 that Mexico but here's why it's a multivariate pencil. So the nice thing probably about the multivariate distribution is that it will sum to one without the accurate attributes between zero and one. So now there's room for for one variable to have one dimension will have high probability. Other ones will have a low profit. So you hear let's say in this case, the hyper parameter you find is the density, location allocation, right So like here, alpha is this hyper parameters. So, these are three dimensions, you have three values for alpha, where alpha is we call this a symmetrical alpha that is established when you have the same value of alpha 1.5. So the density are mostly kind of uniform. On the other times you've gotten Is it all right or is it another 555. So now the density is concentrated around here. So that means the model tends to compensate for the density over a specific region of bases. So now on the other side of that density is the alpha hyper parameter is asymmetrical. So that is the first time answering 1.0 And then the other two you have two. So that means you have a density over the dimension for the divergence where the lower density for the first time and now this is another example where you have high downloads and sort of third, sorry, high value first, third and third variable, right. So that is the sort of disconnect. So we're gonna make use of this to capture the mixture membership of the agent, you guys you come here that hot area you can do within the first year. So that means your membership itself. There's a certain probability for each individual belong to different phenotypes. So there I think I would have selected Okay, depending on the alpha, you have a different priority over all the major memberships

19:37

to work with dramatic, this is the general process of allocation. So now we we have these global problems. So this is independence across Pacific party. So Derek, by the way, I use the open eyes and index surgery. So for let's say for Kaufmann, do you have a sense of the number of folks and then Cherokee is next? Index, okay. So now for five or five different IDs reduced over the vocabulary. So in the previous example, I saw that the four different topic genes in a computer, right so that's the word for the fiber optic cables. We sample from this. There's different types of dimension, sizes same as the vocabulary size. Large size coverage

20:42

for now, for patients that information, so that's what's in this plate. So this is tvm. With a plate for characterizes a repeated pattern, where do you have a fee paid directly to the defendant to medication so usually we write it at the plate, and then we use the index to say, go from one to the last data. Maturity isn't the latent variable, right? Because I'd say they are the variable Sadie and those are the author of the topic. We don't know that the making here should sort of meet your membership for coffee to be unshaded but we assume that it is follow another and then didn't wait for it because they haven't seen that face see that as the topic for you saw the folder. Are you going to sample from this categorical distribution. We're kind of our political category with a with a rate categorical race set to be the same as say that. So now it's one of the topics in high school now the probability that you sample from that topic behind. So that is the ID for the assignment. But that's why we have arrow without didn t see is the topic assignment for each topic. So now you have the focus, right. So now you have the right answer the topic sentence. And now given D now you can then use a specific topic, let's say a T ID for that open equal to Now I gotta do that. That's fine. You have K. We have a big number of topics. We'll have a big five sets of five will not even be you know which topic counsel not even to sample the actual diagnostic code. The actual word from this type of this car called the new building is characterized by the OSI index. So that's the tip is it is something that we alter right because we know the diagnostic code for that patient. We know we have the words for that patient. So here we see you have a three sets of eight variables. So the new for his problem is trying to maximize the likelihood of the observed words with his bathroom

23:34

Okay, well that's what we're trying to do. So this is a big data transfer process, right but that's also our assumption or our scientific hypothesis of the ICD for the vendor. So, so now, what we're trying to do is pretty low, we don't have to worry about the first four or five is the kind of guys we can do with your eyes. So therefore, we kind of just focusing on your 30 topics

24:10

we didn't talk about so that is what we're trying to get further here.

24:17

Right so the idea is this week, I talked about one of the topics and then for each activity for each other. So you already know the topic assignment for all the tokens over all the dolphins What do you think?

24:36

Or what do you think the FBI just sort of facing you by call you for this call? I do. Well, here's Cohen from my publisher conference. Call three months you're probably wondering, well, what is the answer? For? We're not.

25:12

Saying it's the theta d. So it's a pain free number to sample properly. So he told you for that probably one. So what's the probability? Phase and D we'll have them Mitch remember to have that membership. We have a tool you have. Obviously. That also means it's a proportion. So you've got all your topics to say that will be just before that I think that's the focus number our focus. Let's say I want to compute poverty poverty, one, we're talking on your account, your how many focus has already been assigned to one and then divided by the total number of focus. That is basically your favorite color image, but that's Word document. For fighting for five for a copy of one, five global, you have a theta v already. You have a portrait for this document. For probably one and then you're just sum up your proportion that will give you that overall you should use it. Well, sorry. So you have the topic of time for for probably one right. So for confused to introduce it for five which is over to the entire vocabulary, right for topic one. You just kind of how many opponents have already been assigned to copy one. And now you see okay, what's the word for that? Okay. So that so five paid over to metabolism, right. So that's it for work. You know, Gene. So now maybe you have five filters that are left over across all the documents, but then three out of five assigned to one. So that's your five for the for the for the top one rep for June, will be three divided by the total number of folders in your corporate over all the documents. In other words, you can all be you can learn to write. So therefore, if you have a way to defer, just treat the data in the fire has a new set of practice that we want to read, you wanna just like during the training, and we focus on your brain after that become recovered. So that's what we're trying to do. So in here, we've kind of write down the base, right? So this is a conditional X. But the problem again, when you do babies always in the denominator, because you want to try this as a priority to use those as a priority. And you need is a marginal denominator. Where does it on there's always a problem because this is often new tractable. Right. So theory you want to marginalize over all of it the topic of assignments, because you don't know the answer to the variable at the same time you have never here or you have never over data, you have never over thought. So this part is a problem. So when we do this kind of difference. There are two ways to approximate Bayesian inference. One is the classic sampling class. That means that you can actually do that without beta. And the reason you're staying on fire because it's cheap that we talked about 1000s and 1000s. It turns out reasons they're not an issue they because mathematical convenience, congregate to market and that's fine. But there's as well, so actually this module so far as you can actually integrate our data. And then the margin is also if there's a monkey flag, and then for data module, it's also yours. Right so I don't want to get into the details. But the point there is that once you have an update on the file, you cannot cope with the actual inferences that are asked of the barrier. So, here are the most you can do have it all safe and fun. And all you have is to write all these neat variables. And now you deal with this kind of tensor objects. So this tensor object you have this documents as the one that mentioned the worst of the other elements and the main topics as the server. So so what we're trying to do is doing this conditional deferring what's the topic? What's the probability for work Holton, is involved in data science and politics, when you condition the rest of the world and then also of course, the, the actual word identity. It turns out you just meet so this is a proportion for this. One, what are the first term here? So you can see that this is based on based on two lines of evidence? The first evidence is that because this is all talking, right now we're talking about topics. We want to count how many has already been assigned. That so if there are a lot of tokens already been assigned to the topic, but now we have a higher this also needs to be assigned to. The second is that I want to look at what's the word we're holding. So this is a real alternative. So this is XIV. So that's the state of North here. That's what we noticed if it were been offered. I work I want to know for other tokens that have the same work, how many has already been applying. So this is a global, right? So this is far more all the documents that has the polka dot network. So you have that happening. And then the normal life over that summation over all that. So that's gonna face so you have these two lines, right? So this is proportional, but this denominator is easy to read, because here you're looking at pawikan. You want to normalize it while you normalize by the sum. Over all deposits. So that the topic was Psalm one, the probability for tokenized document. Okay. So this normalization so now, you might think, Well, okay, how do I got it? Well, the thing is, you know, here again, we're conditioned, right? So we conditioned being that we pretend that we know that information. So let's look at this document. We basically sum over all the tokens. So here I have this. So we sum over all the tokens in the in the document it sent me over when the Pokemon Cara uses the wild card. There are used as a wildcard encoded meaning that you know for Civ do have these vectors. One time code means that one answer is one, the rest are zeros. So now we've been taking the sum over that already kind of describe this as it sounds. And here it is. It's a kind of a W. W. Without a loss of generality. Because the W and our saw me over documents also we saw me older folks so now we say because we're dealing with public health. Yeah, the UK type one. And then because we're dealing with vocabulary, we're doubling XIV. So this is a full indicative to our tool. So this will be one that will contribute to the counts as one count. One of them is false. It will not right, because we're counting how many tokens has been already assigned to product when the token is equal to work. So in other words, you have these a tensor tensor if your class each time he named it a sum over each time that you've asked for a number. If you saw me over the latest topics. If you had this information that you saw, more or less what you got out of that it's talking that that works. That's this document that works what you can do, you can use your interface right. So you can do that or say, you know, if the if some diagnostic code is the absence in the record, based on the topic application, I can compute the expectation of whether that diagnostic call should be present, in that case, visual right. Now, if you saw me over the word, now you have 1005 main topics. Will you ever characterize a patient's right characterize for each topic? What's the risk of this case it will be allowed for that field? Right? So that's a useful information. Now if you saw me over documents, I've been doing damage. Now you got the excellent topic distribution that is for the shallower so

35:06

whether it is worth too long for that topic. So now for the columns are popping across the word. If there's a multiple tubers belong to that topic, that means they're they have a comorbidity kind of connection between each other under the same topic. None of that help you understand, what is the core of it and the condition for profit. Both are the sort of information in other words, the kind of things you have to eat for token assignments. You can derive these kinds of information that compute these useful information. So that inference is actually very simple. So in France, if you cycle through the Pope, just like we talk about foreign in a sense, this is actually you can do this you can get samplings or you can sample from it, or the variation on infants also extremely similar. So that is the difference guys this computer virus will be used when you condition on the rest of the person whose immune system is bad that photos are independent under a virtual law you cycle through each of us

36:16

that's abuse. And stuff, you done some sufficient statistics, WD when the PE, class, and then that's answered, and then you go out and use that you recycle them again. But that's the basic overview.

36:39

I was just gonna ask a question. If you go from ICD nine to ICD 10, just changing the number of words when you're increasing the number quite a bit in the slides.

36:56

The front lines of tender but copper is mostly just like for example, you have document translated from the Japanese. So there's so so then you can just operate so this is a one modality.

37:21

It doesn't have a huge impact when you're actually increasing the size.

37:27

Right. So the whatever it doesn't have the impact on the phone. Because this is based on this model works well on Spark. That's the major thing that you want to learn is all of it or not, as far as I'm concerned, you take one article the fraction of the work, sure, but probably on average. We work with only the observable. We don't actually tire is not observed in that document. We don't care because you can see the multinomial distribution we work with. We really care about author. That's the like. The counterpart is the sort of the process of the frost or the balloon. The balloon is that like, if this were for modification, but then you have to also account for whether it works absence babbacombe come knocking for this document, you have to enumerate all of the cabinet. But then, you know, that's a very special, but all here we only care about afterwards. There's only a small fraction of work. So that's that's what makes this very efficient. And that's why we chose also, here are several factors that contribute to the success of this model. So after you do this alternate e&m and not another union, you can compute the expectation of As always, this is basically the fracture of the bones that has already been decided to say that. Hey, that you have a big data will be feeding a Bayesian because outside is coming from the hyper parameter from there. So the correct equation is that you have the hyper parameter, sum over the later I'm going to show you in the guide if possible, you can make this alpha personal. That's based on specific information. Alpha for right now we treat alpha as just a flat. The same for all public health and same for the law, the denominator, it's because it's often the same for all hobbies k times plus the sum over all the four Yeah, for all. And now the other part was the same at this network. So once you have this web page, that's why you can't do it here to recover five uses the normalized data. It's also come from that there's four possible sides and then normalize over data is a flat vocabulary size is a big mean times beta plus the sum over all the worst over topics. So that's the following. Once you do this alternate iteration, and after that, you're you're now paying as I mentioned, you can clap the late topic you can obtain the expectation of the document all the work goes into writing something into quotation, that's a fallacy that you can copy. So here is another way to show you kind of think of this as a matrix factorization. If you're familiar with some of the techniques. For example, you here you have the ICB five patients features three examples with the key components into ICD features by topics or topics, your latent factors and then the topics by patients. So that's a tool to make up for all of this. The ICD is tight and answer. Usually you have a much fewer number of topics than the number of the size of the vocabulary. So it's sort of you're doing dimensionality reduction. Now that five is this topics right? So each column is topics. So now if you're learning the remodel learning some distinct topics, we also saw is that we take the top 10 words per topic and then we show this a Hema. So now, if your topics are very much distinct, for the same hours for copy one, because you know this has a very high intensity, high probability up off the top, but it has very low probability for the rest of the topics that doesn't get done is you have this ladder pattern. At the top is our graduate students. That means the model is alerting somebody but also your document doesn't have a very interesting information. And then when you actually visualize it doesn't look like this. It's gonna look like the same words also have some habits or other problems. Right, so there's gonna be a lot more noisier than this. So last on the topic of distributions, and on the patient side, you can also take a one slice here. Now you can see there's our let's say, I take a Firefox, but for now, the columns are below if I know how to one maybe this is a copy with angles all starts with small so this is all based on the top works top at the vehicle. Now I probably won't be launching all the relays when I when I look at the column different patients, but I can take the top three or maybe top 20 patients without the poverty one and I take the top two, top three etc. Now I see this ladder pattern. So that means the top patient would belong to the top one is top 20 patients, they have a low probability for the rest of the talk. So that implies these are high risk patients for the office. So here I did a first subsampling from one of the new char dataset so this I can also look for ICD a call 331. So this is ICD nine code first well for Alzheimer's disease. So you kind of see that the actual patient, the black bar indicates these patients has the ICD code two to one. So these are high risk patients. Now if you look at some of the individuals that don't have that there isn't one, although they don't have to be one, they might have somebody else. So that may indicate these are high risk patients that have missed the fact that something that you could follow up. Also, you know this topic one and topic five, they all correspond to all because probably five because either these talks about effects and also have quite the rich of vehicles. That is a you have to assign one topic one disease. In fact there are identifiability problems. Usually earlier, you have five problems. At the beginning, you will know what topic for as long as you're gonna have several different levels and one of the five to one and then topic one topic for fourth 12434434 Multiple Sclerosis from New York and then the other 1332 is there. So this is based on the real data. So once you train the model you can visualize the topic distribution and that topic specific news and now you have operational applications. So that's all good. So, so that's nice, but there are several limitations where the basic LDA so here we try to address with our research. So the first one is doesn't distinguish different data packs as a multi modality University for lab pass medications. How do you account for a multimodal auto confirmation? Sometimes the missing data now obviously, we don't have time to talk about that. But, for example, place an order a lab test for that individual that is based on the suspicion that test results are going to turn off from abnormal otherwise they wouldn't make money off the test where you wouldn't want to run that test. Done. The timing wasn't right. When you actually did work. The results going to likely it will be normal. So that's now missing a random problem that we tried to account for in one of our paper allow the following models uncivilized, as well as new that in the end, you have been the target, whatever you want to predict. What are you going to predict maybe come down the party, predict all that so that that itself doesn't give you the prediction accuracy? Right? So you could do a dimensionality reduction and I deal with your reduced dimension. I know you're dealing with a classifier very often to to Article coffee. Right. So when you do unsupervised learning the model focusing only on the general distribution on the data, right? It doesn't know that in your home you want to predict this current label. So in our confusion, the second one is we propose more or less where you go dimensionality reduction and you learning the different views at the same time. You want to predict the current available. So that's also offer forms. The basic pipeline I just realized I don't have time to talk about it. And then the third one is a topics are not identifiable. As I mentioned before you trade 20 topics you don't know what they correspond to. So this turns out to be a really prevalent problem. That's not really your trade principal component analysis, PCA. PCA is an inactive until you actually analyze it. So we learned that in PCA. PCA corresponds to acid. But now you have to know what if you have fallen poverty, right. So how do you know? So that's the idea. So we got to make use of genotype concepts we call the free code is actually there's about 1500 People First round to us. We use that as a concept to learn 1500 This topic has exactly corresponds so that's the we have what is called a guided copy model. And then, recently, we have another paper that was on survival. So you have the Cox regression going on within the topic. You want to predict survival rates, individually individual stuff that is more applied to the patient in the ICU. When you're first admitted to the ICU based on first today's observations, you want to predict what's the risk of mortality rates. Getting back to you in the next few days. So that's really survivable for the model. And then we we have a recent ongoing work that we want to defer to Sofia. During the fall, baby, assume there's a fire they will probably want to sub specify the field. How can you refer to a nested sub field hypothesis is in our masks. And then the lastly, we use a neural network based approach to also leveraging the biomedical knowledge graph that's a lot without cannot leverage Knowledge Graph and also cannot leverage

50:00

in terms of batch effects, here, integrating data that's coming from two different hospitals, right? systematic differences and how they're

50:15

using yes ed code. But that's, that's the point. So let's say you do have a queue we have read multiple hospitals, right? So you can't have the same time you can't have a global coffee. So what do you want is your one the topic that are specifically popular is the classroom the topic that our hospitals and all the rest of the topics that are globally shared across hospitals, and you can use that to actually capture that nutrients. And that was actually another product that we had this Saturday the learning topic because we're familiar with Taiwan. Try not bipolarity but that's one positive.

51:10

So yeah, so from now on, so these are very short facing that kind of made me stop at some river. And now we're going to talk more about high level this kind of expose you to different applications right through the way and trust that I will share with you the flies and then you can read the papers and so first we want to talk about how to new for multimodal projects. So the here's a bit of motivation. So from here you can see this, that the blue bar is the comprehensive EHR. Hrm is the basic EHR, okay, meeting permission. I'm aware of that. problem This comprehensive has a track record overall, including laboratory test. Results. There are many different modalities. So you can see that through all the years, the hospital has adopted comprehensive EHR also by downloading, machine learning, that kind of integrity. So here I'm not throwing your genomic information but that's another important different modalities, right. So besides that ghrp multimodal, you also have a genomic kind of genotype. You have gene expression, many different genomic so we first we started work with this thr based data sets is a multimodal it has a different modalities. It's a publicly available, so this is created by Fisher on that under so this data primarily generated from the hospital ICU. So in total, we what we use is we first we knew this patient with a foot or so for the same patient wait a minute once in that dataset 20 to 30 minutes, short amounts $39 patients, and then the test patient and model. So this is a patient of you predicting mortality. So basically based on the first admission of the test patient, can you predict when the patient died? Maybe the patient died in the second region. Right readmission is probably the first example a patient have cancer first, and maybe it's not that bad. But then after a few weeks, they rehabilitate. So that's a big cost of medical care. So now is the time to predict the risk of readmission. You might as well add the patient's state. So that's kind of what we're trying to create for the modality size, so that we have a sense our 7500 patients for the modality size here and you have that we model the clinical notes. So this the number of unique features is that the cognitive science theory removes almost stop words. So our syntax, the grammar books, they don't have any semantics. So after that, we have 33 unique and now we have ICD codes involves $6 unique procedure code as well because 100 lab tests I wonder it for a couple of quick books and medication called consolidate and then the DRG for DRG is similar ICD is being used also the fees related to cost about 1500. So, together the unique feature that mentioned as well 46,000 certainty that the number of features is actually bigger than 20. So this is a very high damage. And then the total records that means the total number of tokens. It's about 4 million, 4 million. So although this document has very sparse representation, we only observe very small fraction of the words occur in that specific patient. Very small fraction of STD calls. So that is multi modality kind of off together. And then the for the reference. So so this is a universe that is based on real data on the medic data. So what we do is we treat each of these modalities modality as a method. So we call this a vertical integration. Vertical integration means that you have this same patient. I have laptops that are not defaulting. So you take a vertical slice over these columns. You can have a multiple, multiple matches. And then the rows are the features. So in a nutshell, what we do is we decompose this five matrices into five small sub bases. There's like a PCA basis and then here you have this is Fisher and Amanda Collins are now fleeing, multimodal so each of these modalities to find you on the topic in their own lives. We do know about target we normalize within modality itself to one and now that's how you account for different modalities. So that's sort of simple training for multimode. So they will normalize we didn't know that they're supposed to normalize or cross modality because each modality might have different haversacks normalize the process. You kind of said you know what I do have more vocabulary tend to be overwhelming to others. And now in theory have the patient so here as we are added to talk about how do we interpret it? So that is your take on this last year. Now we're talking about one talk the one target user I have a signal that I always wanted to add. So here we have a signal that we have this DRG for disease related to the diagnostic procedure called Test notes. So these are characterizing the same topic. Here we're showing the bar plot. So these are called holes under the same time. So now we have a new theory for the DRG. You can see there's fliesen poisoning psychosis. So these has some sense on top of that the number of small because you have a margin of a cabinet size. So these two are significantly higher than uniform. Uniform is a one lower. So we have C called monomers. They fall off tomorrow. It's like they're late. So you hear how these appointed from now and you're here you see the ICD 10 diagnostic code. The top ones are bipolar disorder, schizophrenia, and bipolar disorder, different subtypes. Another outlet schizophrenia. We can see that various literally this just based on this modality this happens about psychiatric, it's based on you know, psychiatric evolve bipolar, either bipolar or schizophrenia. The two are very much they have a very significant genetic origin. And now the procedural code doesn't mean much because I think psychiatric patients don't have a lot of procedural knowledge, each modality kind of sound without importance, Heart Foundation, zero quality hormones, etc. And now this is not for laboratory tests versus more for guys and easy these are for psychiatric and then both also see the abbreviation so these are written by physicians nursing. Prescriptions also guys. So this laptop is a test the level of your classes before they prescribe you those things. So now from here, now you're kind of in front of here. Why is there a poisoning? type filter? I remember your thoughts on yesterday, a patient overdose, right so psychiatric patients. Remember that causality right? So whether the patient the overdose, call psychiatric venue, or by the schizophrenia calls the overdose. Now from here you can see that the patient overdose so they have a drunk point. That's why they admitted to the ICU. And that is the reason they're here. Because they have bipolar and schizophrenia. Now of can see that just from this modality, you will learn a lot from what I have soldered on these cars, right. So now that's for this topic. So now now we know this topic is What psychiatric, the amount of days I need to allocate these topics. Now if you take a look at the patient level. So now I can do the same as I mentioned, for each of these topics, I can choose the top pages. But there are two topic occasions for this topic. And part 50 Patients are 3531 Now, here's your topic 50. You can see that these are the diagnostic codes for leukemia. So if I'm here you can sort of obtain the high risk health implications. Verified, Mr. membership for, let's say time the fact that if I had a patient that this attack, they will have

1:01:07

here's a popular vandalism I'm hearing Cicero's see that these topics are very much the top patient often where money works for this. where that is and then, yeah, so here you can take the patient by topic. Or you can let's say you have a target label, right you kind of swaying on linear classifier, you learn to request implementations of predict. So this has vanished off but dealing with a low dimensional so because if you're otherwise you will offer it on this kind of management. For dissolving forensics on expanding overfitting not going to work on the size of the training, data size training data size, a smaller number of issues, nationality production you operate on here, right and now you feel the need to request the number of regression coefficients is the same number of these topics. There are usually 55 or 75. Now TGM with the multi modality you have more stuff going on.

1:02:19

And that's a question is also is if the goal was just to do this prediction could have applied the same things I did in the first lecture. Right? Yeah. So in terms of if that was the only objective but here they're getting a lot more information from this but in terms of the overall performance, the goal is just this prediction that you can gain something from nuclear physics and

1:02:48

I think that's a very good point, right. So, so now, I mean, this can also be applied to the first now where you learn fertilizer and you learn to apply this LD here to the genome. And what are you gonna learn you're gonna learn all the blocks, that snip occur under the same products, that people like the fact that the LDA was isolated so we have historical LDA it was personally I went all the way from the outside inside the LDA corpus. Now there was first class that you might get for discovery, assess pretty blocks, right. So yeah, so they're on the other side and vice versa, right. So now you can also apply the lifers, right? Although I don't know anyone have done it, the either manager and now you've trained it to predict. The downside is that you put them all over again, right, you treated the input feature as it is. And now, now you to directly acquire as well firstly, later you can look at the quotations and see which color is important.

1:04:11

So here we were done about a mouse is the area trying to look at what topics are related to age. But oh boy, do we have some topics, and then we just do for regression, but here we just correlate. Basically, we got correlation coefficients, we have some data that says according to the now we take the top three topics that are positively correlated eight, and the top three topics that are our bottom three topics that are negative. The topic here, we want to characterize as a patient as you know, what type of phenotype risk the first one is associated with a symptom associated with heart failure. So now, this poll here, Kotter by the sycamore house, Academy brought the book, the first two is related to the procedure of heart failure, your procedure, pacemaker, etc, our cardiac test and then the second one is associated with cardiovascular. It's also it's an age, that's the rest of the developing this typically says the third line is with the dementia. So by the way, here we manually because these targets are not identifiable on their own. So we take a look at the sort of hot coals and car business card bathroom. And then this is obviously with a mental disorder. And then here's a bit of kind of comorbidity or with the topics of the inflammatory kind of awesome under this topic, of the majority is that we believe that the more dementia, the third one is often mentioned, behavior. So the last three, so that's the main difference. And common negative correlate with the age is a kind of trivial finding, right? Because they already have that this is basically related to the nuisance. Of course it's negative. So here is a neonatal. So these are all it turns out in this mimic data ICU. There's a six often babies that seems as the Minister for sake saw the baby among them there are there within that given birth. You also for those artists, will be Patrick, somebody, some people they want to analyze a reward or six year as it comes, I will say that I am all clatters on the desk. And then we did a similar analysis in predicting mortality. As I mentioned, we can make use of the patient with a multiple they should apply the topic sentence by topic over the past patient and now we can have a trading tactics that we identify the topics that are for positively correlated with mortality, building your quotation, the bottom three topics that are negative. So now what are those topics the first one is based on the top calls, here we concatenate them together but these are all coming from different categories without without. The first one was with a with a modem. Second one with a tactic right? Well narrows the house so that's 31 with the most I've ever after, it's often difficult three lethal pockets. And then the bottom is radar. Super important. Newborn normal baby they're not gonna die and that's why they are negative. So I'm not gonna support a miss hrs or we normalize the topic within the school. Now, we might ask, What if I just apply all the door token coming from different modalities? You know that you look at the top quotes under each of the topics. These are different topics. So we identify again, that only by the top three polyfluorinated popping out and now we see that across the board so there are one single notes. The reason that is the case because the nose has a 33 It's All finishers. It basically overwhelmed. The rest of the without that when you look at the top that has the worst being on the top consume normalizing of fossil fat. Right then then other code from other one, because the size of the coverage is different. Well, nonetheless, you can see that it's actually comforting some of the topics that we find here. So here the first topic is obviously with metastasis, tumor carcinoma, cancer chemotherapy. The second one we have these abbreviation of corresponds to the third one is with melanoma and then the the other one for example, you hear the bottom the negative party well with a newborn, and the end is rubbish. So that's what we have here is effort and care. We also have a poll that come from different adults so that's the kind of multimodal topic just giving you a very brief overview of the analysis.

1:10:05

So now the limitation as a manager, so the topics are not viable. So therefore you need a manual annotation, so only limited by progress. For Humanity, I work with one mentor doctoral students who has medical expertise to help me so the number of topics and also determining arbitrary s&p 500 So we did across a number of topics. So it's hard to train over ones on topics we could train this model, such as the outcome of that is that so you'd maybe you have sent me a topic that has some meaning the rest of the 950 or so because like all the topics, have the same problems or whatever, what happens. The problem would come to the agenda, right? You don't use circuit prior to God. It doesn't take too long to load information or take the patient days to find the data there are different disease onset. So if you can take that into account also can help interpret what are the bases that we did during the topic as possible. I think I'm gonna stop here. We kind of wrap up I think it's too late. I've gotten tired or exhausted standing here for three hours.

1:11:50

Maybe also we want to get some comments about the tools they need to install.

1:11:55

Oh yeah. Yeah, so started Usami that

1:12:01

Yes. So yeah, for for the first second, I think for the constructions you'll need to install. Yeah, for that stuff at intersections, you'll need to have Python, like at least version of 3.8 and you also need Jupiter lab. So all the packages that you didn't need to install is for my section, which is at 1pm. That install command is at the beginning of the notebook. So you already will share the link to the notebook on GitHub. So please try and like clone the repo or download it on your machine and try to do an installation. I'll be around during lunchtime. So if you have any issues with installation or be happy to help. The Notebook should be like, I put in a lot of efforts and document everything and explain everything. So even if like this solution doesn't work, you should be able to at least follow the sort of practical guidelines. But yeah, if it wouldn't be better if you could do so like and then like, also like, you know, play with some of the parameters and run things to get better intuitions about the material that were covered in the first half. Yeah, I think yeah, we'll put the link on the on the slides. We're going to share on those slides

1:13:21

and then we start to get at one or maybe one

1:13:27

or the other. I'm going to wrap up slides or we're Sally is one of the first enough to that second, so maybe you can start with your participants. I think the rest are hopefully this already we have this concept. We want to have more friends.

1:13:54

Okay, let's start with that one. Okay, let's start one