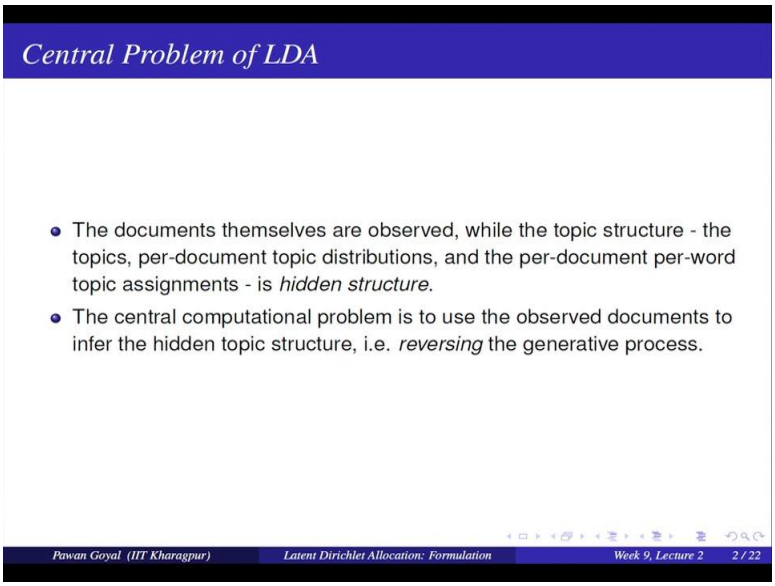
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 42
Latent Dirichlet Allocation: Formulation

Welcome back for the second lecture of this week. So, we had started a formulation of topic models in the last lecture so, but we mainly focused on the basic intuitions. So, today we will do the mathematical formulation of latent Dirichlet allocation. So, that is the main sort of topic models that are used; also known as LDA.

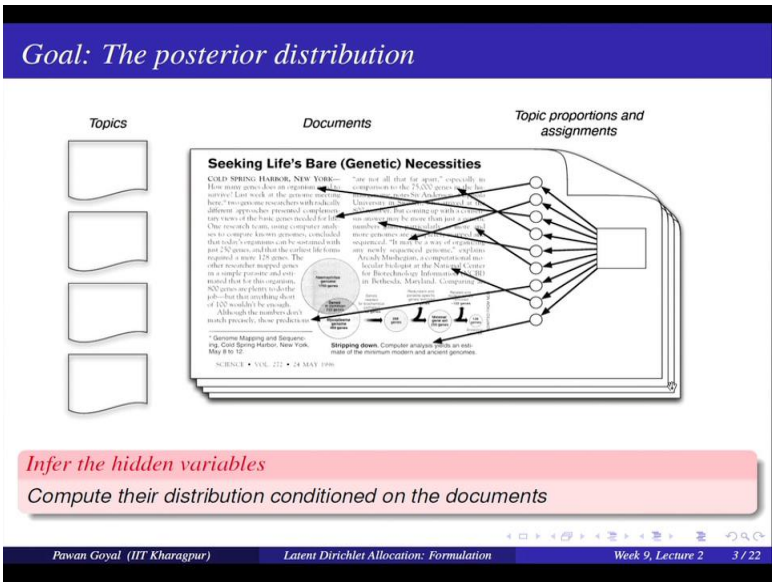
(Refer Slide Time: 00:45)



We were here and we were seeing that what is the central problem of LDA. So, problem we were saying was that we have observing only the documents, but we know nothing about all the parameters of my generative model. So, I do not know, what are my topics? What are the topic proportions of the document and I do not know what is the per document or per word topic assignment, I do not know that.

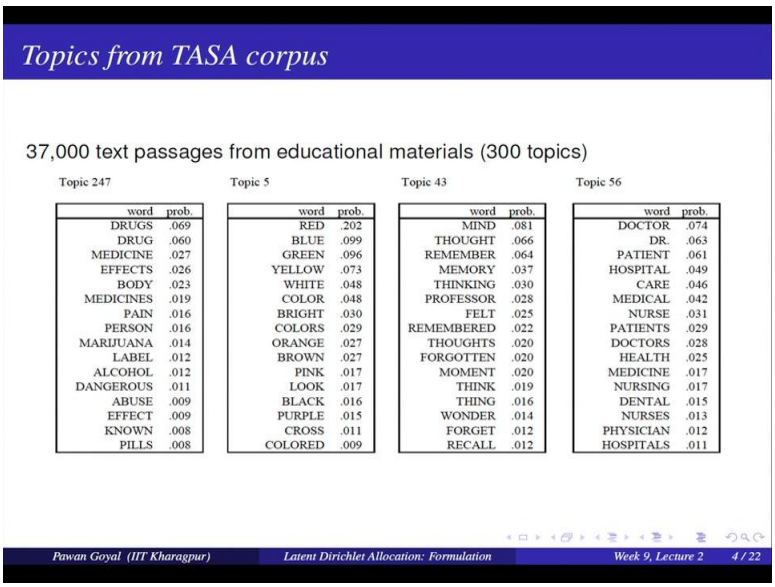
My real problem is from my observation I want to reverse the generative model and come up with all these probabilities.

(Refer Slide Time: 01:24)



So this figure will help you in understanding this goal. So, remember we saw an earlier figure where we had all these topics given to us, we knew what are the proportion; topic proportion for this document and then we were drawing topic for each word in this document, but now so in reality, I will only have the or the documents. So, I know all these documents. So, I know what are the words, but I do not know anything about what are my topics what are these proportions and what are these topics assigner for individual word.

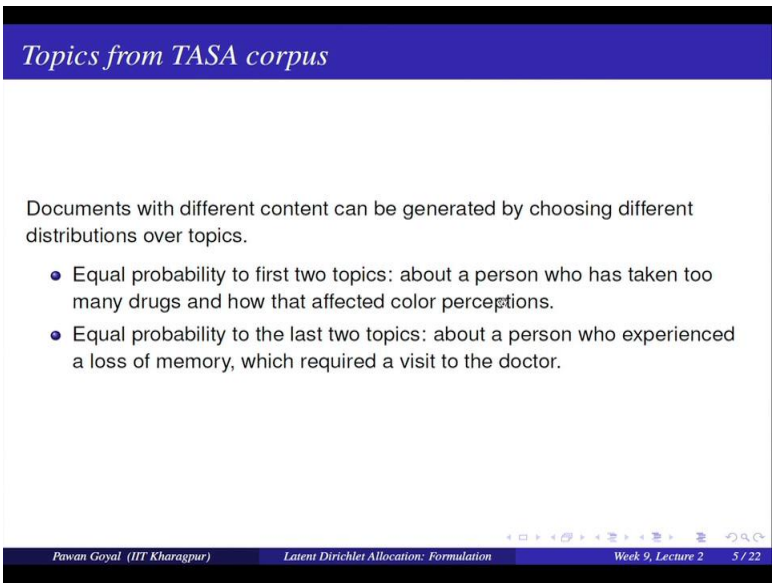
(Refer Slide Time: 02:09)



I have to compute the distribution conditioned on all the documents that I am seeing in my data. So, another simple example to listen the addition behind the generative model, so, here you have some 37000 text passages from some educational material and suppose you run LDA you found roughly 300 topics. So, here in this slide, what you are seeing? You are seeing 4 different topics. So, you are having the topic like 247 which has what like drugs, drug medicine, effects, body, etcetera; another topic having words like red, blue, green, yellow, white, another word about mind, thought, remember, memory, thinking and another topic about doctor, patient, hospital, care, and so on.

Therefore, the topics that are there in the corpus by running LDA over these 37000 text passages, now to get intuition about the generative model. So, what was in the generative model? You have some topics, now how to generate topics? We take some of these topics in some proportions and you start generating words for that. Now, suppose you try that.

(Refer Slide Time: 03:18)



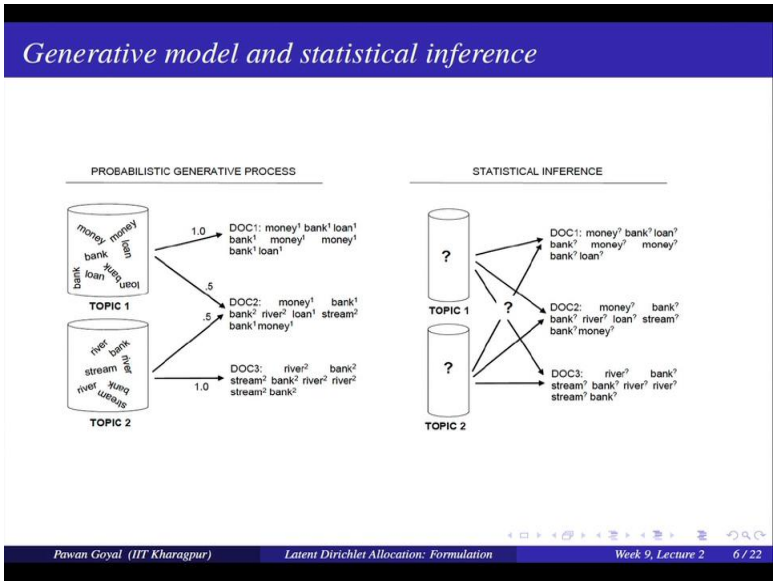
Suppose you take topic 247 and 5, so, suppose I give an equal probability of the first 2 topics, so, what can a document you can generate? This thing, we will see these topics. So, first topic is about drugs, medicine, person, marijuana, second and another topic is red, blue, green, colors. So, suppose I blind these 2 topics together. So, what kind of document can I generate? So, there can be some different sort of documents you can

think of, so one could be that someone who was taking a lot of alcohol marijuana and so on and it affected its color perception.

Suppose you want to document general document like that we will blend these 2 topics together and write that. Similarly suppose you want to blend these 2 topics together. So, this topic is about mind, thought, remember, memory, forgotten and this is about doctor, medical, nurse, patients and so on. So, here if you blend these 2 topics together, you may generate the document that says someone who had suffered some sort of memory loss and it is led to visit of the doctor. So, like that you are having different topics you can blend some of these topics and then generate doc document. So, this is the generative view.

We can give equal probability to the first 2 topics that gives some sort of documents and equal probability to the last 2 topics that give me another sort of document.

(Refer Slide Time: 04:46)



Now this single picture explains both the parts; the generative part and the inference part together. So, what you are seeing in the left figure? So, this is the generative model. So, you are generating some documents and how are you doing that? So, we are having some topics, suppose you are having 2 topics, topic 1, topic 2, each topic is nothing but a distribution over words. So, you are having some words like bank, loan, bank, money, loan. So, these mean this words coming multiple times here. So, these words are having

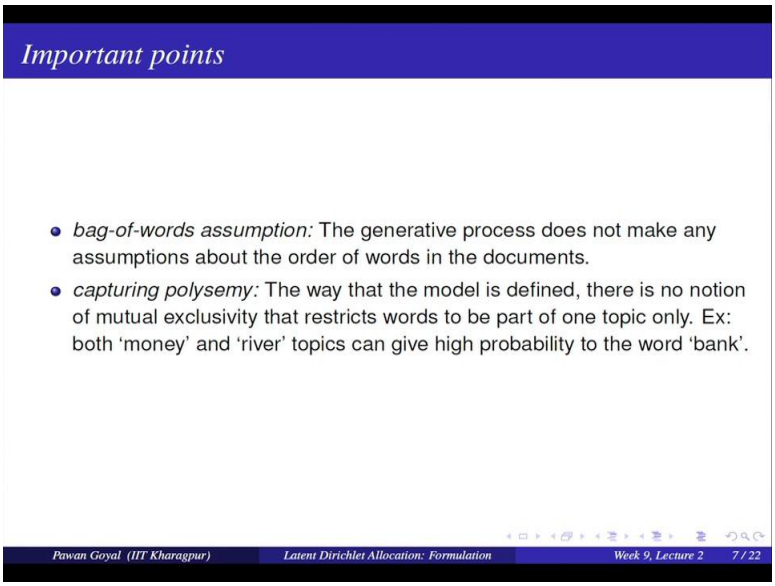
high probability on the other hand second topic is river, stream, bank, river so on. So, these are 2 topics.

Now, using these 2 topics, I am trying to generate my documents. So, how can I generate the documents? I will mix these topics. So, suppose I only take topic 1 and I can have a generate document that contains words like money bank loan and so on, I can just take topic 2 and generate words like river, bank, stream and so on, but I may also take both these topics instead, say equal proportion and generate document like money, bank, river, loan, stream, bank money. So, what you seen here, in this document to the words are labeled with topic 1 as well as 2 because the words could have come from another topic.

On the other hand in document 1, all the words are labeled with topic 1, doc 3 also all the words are labeled with topic 2 because we have the only topics possible in these documents. So, this is my generative model idea. So, here everything here, you have the topics, you have per document proper proportions and you also have per document per word topic assignment. So, you know the topics for each individual word also.

Now, what is the statistical inference part? So, inference part, you only know your 3 documents. So, you know my doc one contains these words, doc 2 contains these words, doc 3 contains these words especially, but you do not know, what are your topics? And you do not know, what are the proportions of various topics that are represented in each document? So, all these numbers you do not know and you also do not know for each word, what is the topic assignment? And all these you have to infer. So, I hope with this figure, this is clear, what do I mean by my generative model and what is my problem that is to infer all these probabilities of my generative models only from my observation.

(Refer Slide Time: 07:23)

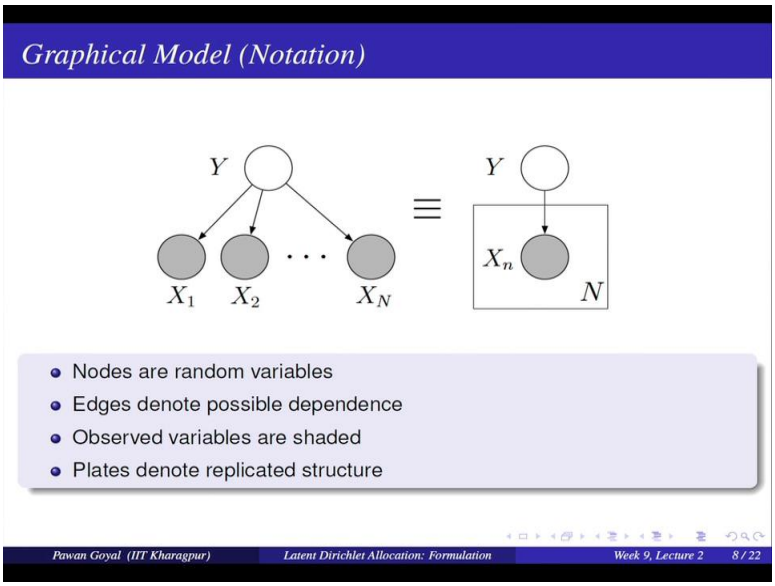


Now important points about LDA, so, first of all it uses a bag of words assumption. So, I hope by now you understand, what is a bag of words assumption? That is I am not looking at the order in which the words occur. So, this is like a bag forming. So, I am taking all the words and putting them in a set it is not like in list where some order is important. So, LDA does not model the word order as it is. So, it takes only what are the words present in any order.

Second and I guess you would have also noticed that in the previous example, LDAs are also good at capturing polysemy. So, remember what is polysemy? Polysemy is a given word might have multiple senses. So, in the case of topic models, what can I translate that; that means, the same word might correspond to multiple topics and this is perfectly allowed because each topic can have its own probability distribution; that means, the same word can come in 2 different topics also and remember in the previous slide, we are having the word like bank that was coming in both the topics and that is perfectly allowed. So, that way, topic models can capture better this word is coming from topic 1 and topic 2 in its different senses.

The way the model is defined there is no notion of the words being mutually exclusive to the topics. So, for example, money and river can give high probability to the word. So, both money and river topics can have high probability for the word bank and that is perfectly fine.

(Refer Slide Time: 09:03)

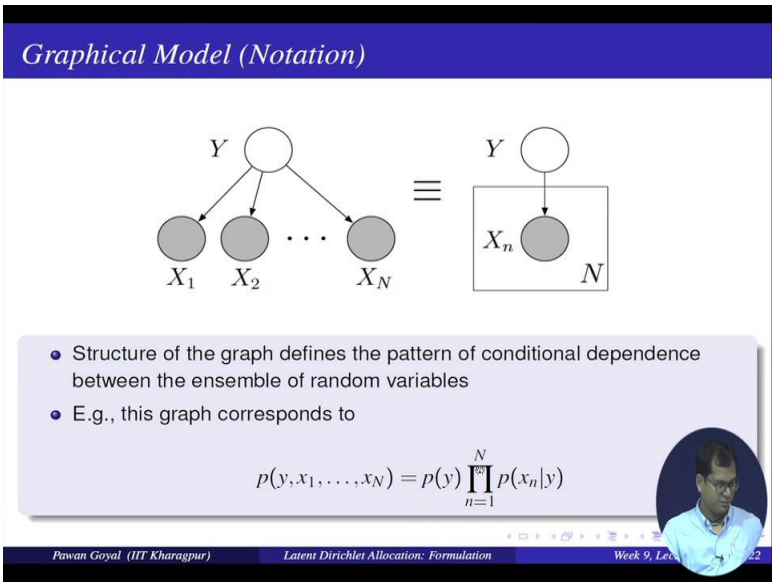


Now, to understand the LDA model, so what is the LDA model? So, you have to first see, what is a graphical notation? So, if you are not gone through some of these graphical notations, so this slide tries to explain, how do we interpret the graphical notations? So, here, what you are seeing? I am having some variables. So, having a variable y and some variables x 1 to x n, so all these nodes when that I see my graphical model, these are random variables. So, I have random variable x 1 to x n.

Now, what are these edges? These edges will denote the possible dependence. So, here I know that x 1 depends on y, x 2 depends on y and x n also sub 2 x and all these depend only on y, but there is not depend on each other. So, that is why there is no edge from x 2 to x 4 and x 1 to x 2. So, these depend only on y. Then there is a difference that some are shaded; some are not. So, what is that? So, observed variables are shaded so; that means, these are the variables that I am observing and this is a hidden variable that is not shaded. And to simplify these notations, what we can use? We can also group some variables together. So, some max instruction that is being repeated, you can put it in the plate notation. So, this is you are seeing in the right hand side. So, you are having these and different variables and you can group them together by x n and you write here capital N; that means, x n is repeated capital N times.

This and these are equivalent and these further simplify these notation. So, once we have seen this. So, how do we interpret graphical notations then we can look at look at the graphical notation for LDA.

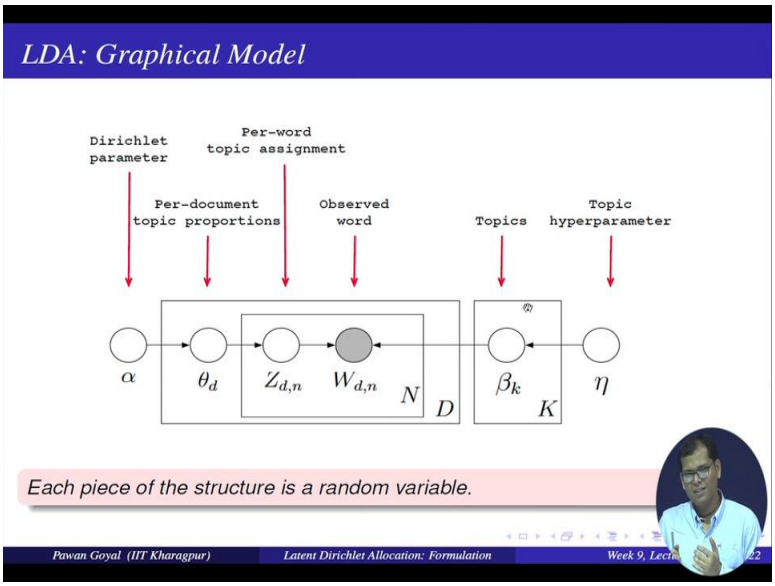
(Refer Slide Time: 10:51)



So just one more thing, that once we have this notation, we can also compute the probability of this the whole graphical structure. So, I have variables y x 1 to x n. So, probability would be I have the probability of y and each of these depend only on y. So, the probability of this whole structure can be probability of y times probability of x 1 given y x 2 given y up to x n given y.

This graphical; the structure of my graph also defines, what are the conditional dependence between various variables and that I can use to write my joint probability distribution. So, this is my joint probability distribution over all these variables.

(Refer Slide Time: 11:39)



Now let us see, what is the graphical model for LDA? So, you know from the previous slide, how do interpret graphical model? So, all these nodes are random variables and observed variables are shaded. So, the only shaded part is W d n, these are my object words, everything else is hidden.

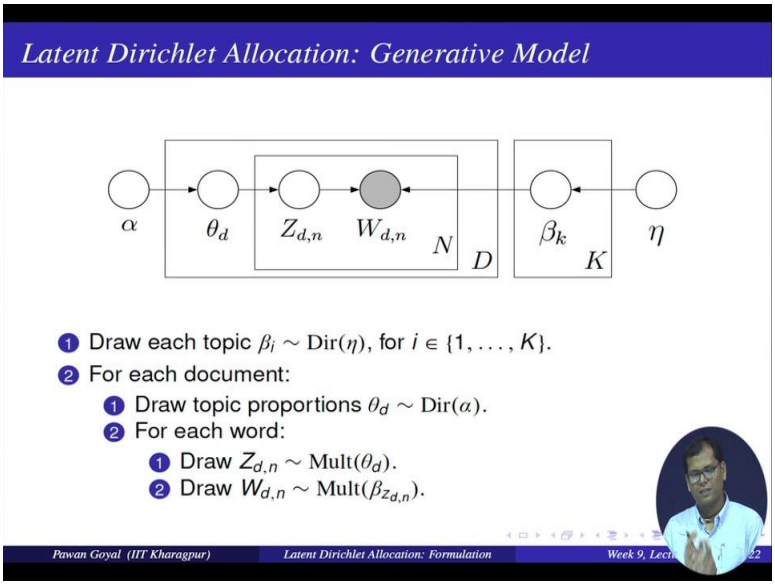
Now, what are everything else? So, there are some plates here. So, let us see this one capital K and what are these topics? So, what I am saying here? There are capital K different topics and each topic is a probability distribution. So, this is the probability distribution over top over different words for a given topic, for a topic it is small k. So, beta k is what is the probability of word one what is the probability of word 2 etcetera for the topic k and this I am doing for all the capital K topics.

Then let us look at this part. So, I have capital D, different documents in my collection. So, for a given document small d, I have figure d variable that denotes per document topic proportions. So, for this document is small d, what are the proportions in which you are blending different topics together? So, this will be different for different topic; different documents. So, you are having capital D, different documents and there is again a Dirichlet parameter here and a topic hyper parameter here. So, we will discuss what are these? But right now, you can understand by simply this beta k is the distribution over words for a given topic and theta d is a distribution over topics for a given document.

Now, let us go inside. So, here, now we are going to 1 particular document and this can have capital N words and what are these? So, you are having per word topic assignment Z n, yes, each word will have only 1 topic and this is the actual word that you are observing and these are all hidden. So, I do not know what my betas? I do not know my theta; I do not my Z n. So, this is only to explain, what are different nodes here? What are the different plates here? So, your plates corresponding to topics documents as well as the words in a single document and you have all the variables that we were using, all the notion that we were using earlier, there are variables for each of these.

Now, let us see how; what is the actual generative model? How we generated the words using this structure?

(Refer Slide Time: 14:26)



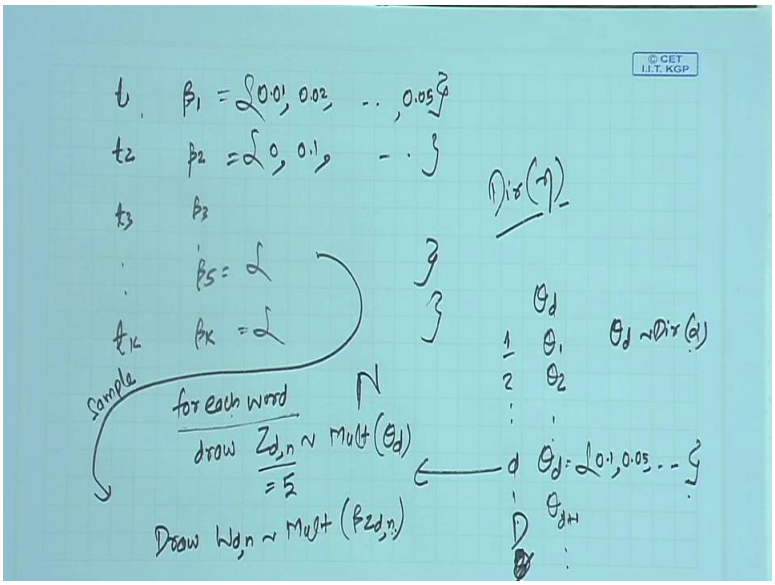
What is the generative model? So firstly, you draw each topic beta i as for the Dirichlet higher hyper parameter eta and you do that for all the K topics. So, first of all, we generate the model, you are drawing your topics. So, fine I have my capital K topics, now you go to your documents. So, for each document that is small d, draw topic proportions. So, you find out, what are the topics that are involved in this document as Dirichlet distribution over alpha. So, we will talk about these Dirichlet distributions, what do I mean by these Dirichlet distributions?

You draw a topic proportion and then for each word in my observation. So far; I am now going to this document for each word, you draw Z n as a multiple from a multinomial

distribution over theta d and draw a word from this. So, draw a topic and then draw a word from this multinomial distribution over beta Z n.

Now, yeah let us try to understand this a bit more. So, let us go one by one, draw this topic beta i is your Dirichlet distribution over eta.

(Refer Slide Time: 16:05)



What I am doing here? I have k different topics. So, I am trying to draw the proportion the probability distributions from my vocabulary. So, what I am saying? So, I have k different topics; topic t 1, t 2, t 3, up to t k, for these topic, I am drawing beta 1, beta 2, beta 3 and beta capital K and what are these? They are nothing but the probability distribution over my words in my vocabulary. So, it can be something like yeah this is 0.01, 0.02, 0.05, so on, this can be 00.1. So, these distributions are drawn by using a Dirichlet over eta.

So, Dirichlet; it is a distribution over distributions. So, it helps you decide, what kind of distribution will be preferred over another, so that we will again discuss. So, this we will discuss later, but the idea is which kind of distribution, you will prefer and this helps you draw some distributions for all the k topics.

Now this you have drawn, now the next line says for each document, draw topic proportion of theta d using Dirichlet over alpha. So, now, these are your topics, but in your collection you have again capital E documents. So, document 1, 2, small d, capital

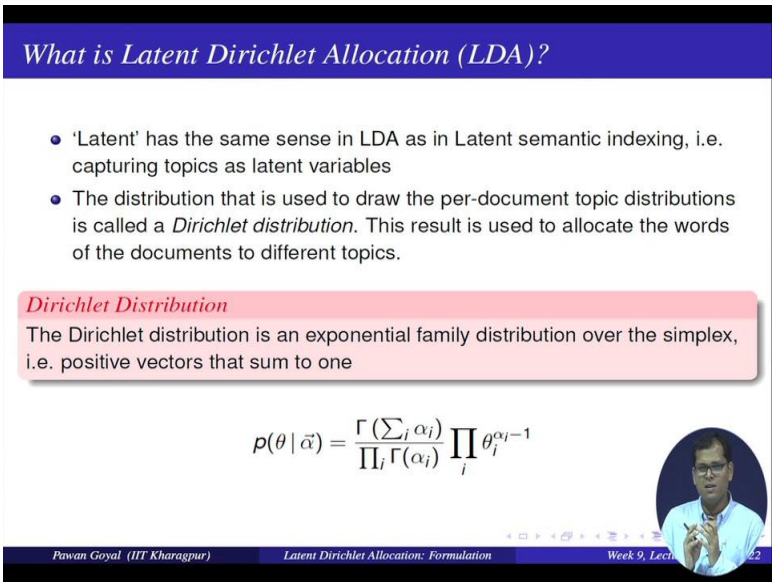
D, so they are capital D documents in your collection, now what is it? What are you drawing for each document? You are drawing theta d, what is theta d? Theta d is a probability distribution over the topics. So, it will be something like what is the probability of topic 1? What is the probability of topic 2? What is the probability up to topic capital D?

For each document, you are drawing these distributions and what are theta d? Theta d are coming from the Dirichlet over alpha. So, again this alpha is telling me, what kind of topic distributions will I prefer over others? So, these are my, so first drawing the topics then drawing the topic proportion for the documents, now what else, now it says now suppose I am going to this document then for each word, how do you about words, draw Z d n as in from a multinomial over theta d. So, theta d is a probability distribution and it is a multinomial distribution from there sample one topic. So, this distribution k topics, suppose the Z n is equal to 5 that is I am taking topic 5. So, for each what I will draw 1. So, it can be whatever that that comes according to this multinomial distribution.

Suppose it is 5, now how do I generate the word? Now I know the topic, now I have to generate a word, so, I go to so it says draw W d n from multinomial of beta Z d n, what is Z d n? Now Z d n knows my 5 that is the topic you that you draw and now you take multinomial over beta 5 so; that means, for beta 5, fifth topic I will find out what is the. So, I will have the distribution and from there I will sample one word. So, I will sample about from this probability distribution and that is what I am generating.

This you will do for each of the capital N words in my document and is the whole generative model first you are generating your topics then for each document topic proportions then you are going to the individual topic going to the individual word and generating that word fine so that is what we were saying in this slide.

(Refer Slide Time: 20:20)

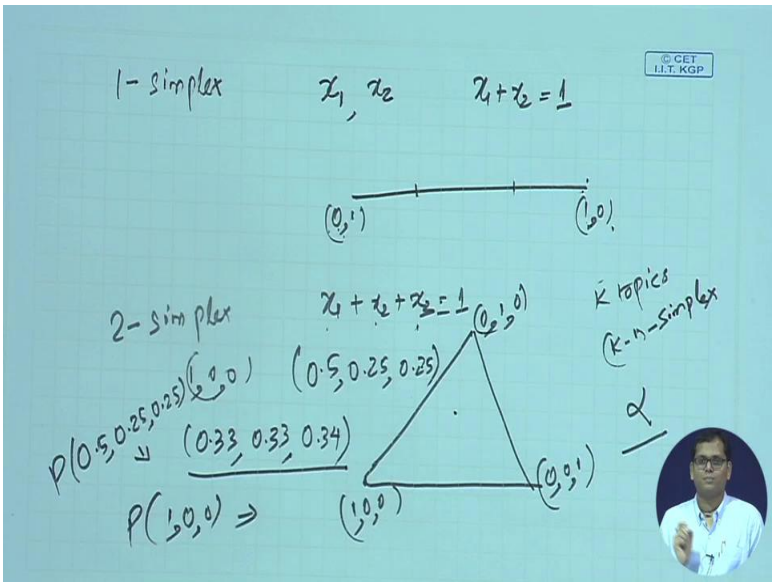


Now so 1 interesting point here is why do we name it at latent Dirichlet allocation? So, we are doing some allocation what why latent and Dirichlet coming. So, latent in this LDA, say has a same sensing energy latent semantic indexing. So, that is all these topics are some sort of latent variables. So, there are some sort of hidden topics, some latent topics, I do not know, what are these topics as such? I cannot give them good maybe good labels also, but there are some distributions some hidden distributions over my words and these are called latent.

Then what is Dirichlet here? So, you saw the Dirichlet distribution at 2 places. So, that is the distribution that is used to draw the per document topic distributions is a Dirichlet distribution. So, that is how am I sampling topics for a given doc document that is coming from a Dirichlet distribution and this result is used to allocate the words of the documents to different topics and that is why the word allocation is also coming. So, you are having latent Dirichlet and allocation.

Now we will look at this Dirichlet part in bit more in bit more details. So, what are the Dirichlet distributions? So, if you think about it, so if you incredibly try to understand that this is a distribution over probability distributions what do I mean by that? So, Dirichlet distribution is an exponential family distribution over the simplex that is the positive vectors that sum to 1.

(Refer Slide Time: 22:11)



When I talk about simplex so you can talk about say 1 simplex would be 2 elements say x 1 x 2 such that x 1 plus x 2 is equal to 1 positive vector that added to 1. This is one simplex. So, you can see that they are infinite solutions here and you can read it as a line, it is a line and this might correspond to say 0 1, this might be 1 0. So, that is topic 1 has 0 topic to each or yeah x 1 is 0; x 2 is 1, here x 1 is 1, x 2 is 0 and any point you can accordingly give some definition what are the values of x 1 x 2.

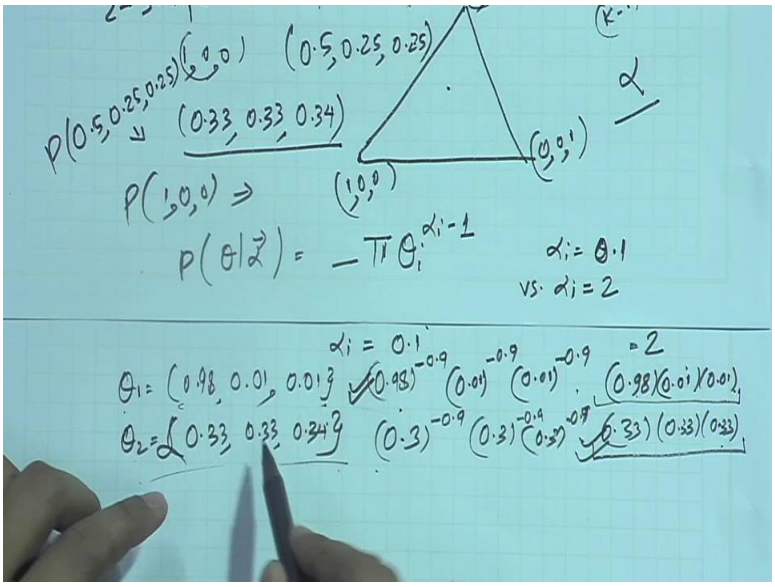
You can see that this will be a line, simple line x 1 plus x 2 is equal to 1, this is my 1 simplex then you can have 2 simplex then you are having 3 things x 1 plus x 2 plus x 3, there are 2 1, in this would be some sort of triangle. So, this point might correspond to say x 1 is 1, x 2 is 0, x 3 is 0. So, this might be x 1 is 0, x 2 is 1, x 3 is 0. This might be 0 0 1 and then each point, we will have some proportion such that all 3 add up to 1, this is my 2 simplex like that you can define of any and simplex. So, if you are having k topics you can think of it as k minus 1 simplex.

Now, what is my Dirichlet distribution? Now, we are saying it is an exponential family distribution over the simplex that is a positive vectors that up add up to 1. So, again let us try to understand that intuitively first. So, what I am saying here? Suppose I am having a 2 simplex then lot of different value is my 3 x 1 x 2 x 3, they can take a lot of different values. So, I can take values like 1 0 0, I can take values like 0.33, 0.33, 0.34 and things like that they can take a different values like 0.5, 0.25, 0.25, etcetera.

So, what my Dirichlet distribution does? It gives me a probability of this probability distribution. So, what is the probability of getting distribution like this? What is the probability of getting a distribution like this? So, that is what kind of distributions I will prefer. So, I can use that to say that I will prefer distributions where one topic one of the topics as a high probability and others have very low probability or I will like to have a distribution where all the topics have equal probability.

This kind of constraints you can put by using your alpha. So, that is where the formulation is. So, what is the probability of theta that is a distribution over these topics given my alpha that is some drawn function? So, even if we forget about this term. So, this is multiplication over theta I to the power alpha I minus 1. So, let us look at this term only. So, what is being said here?

(Refer Slide Time: 26:02)



Probability theta given alpha is multiplication over theta i to the power alpha i minus 1. So, let us try to understand that suppose my alpha i is say 0.1 versus alpha i is equal to say 2, what would happen in the 2 cases and let us say I look at 2 different thetas. So, my theta 1 is 0.98, 0.01, 0.011 topic has a high probability others have roughly 0 and theta 2 is say 0.33, 0.33, 0.34 and now you can see what kind of alpha will prefer 1 theta over another 1.

Let us take the case with 0.1 alpha i 0.1. So, what would you is the probability of this getting this distribution it will be 0.98 to the power minus 0.9, 0.01 to the power minus

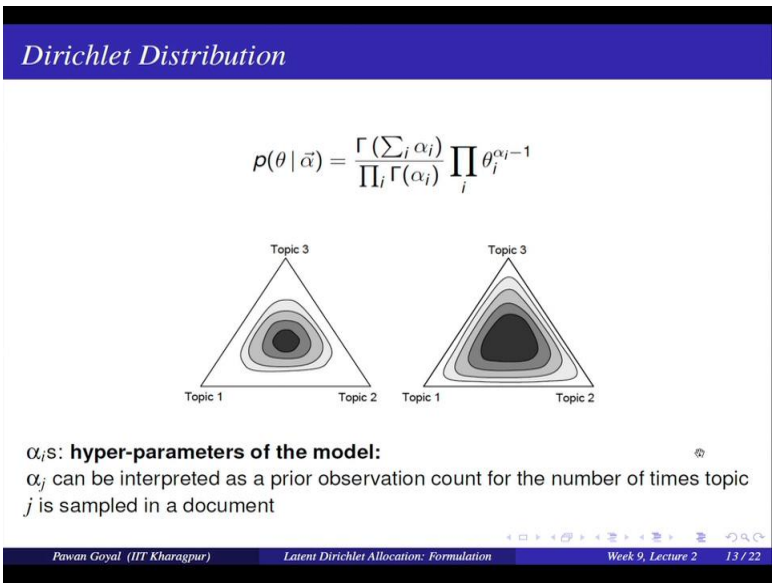
0.9 0.01 to the power minus 0.9 and this probability would be same 0.3 to the power minus 0.9 0.33 to the power minus 0.9 and so on. On the other hand, if I take alpha i is equal to 2 then this would be 0.98 times. So, alpha i minus 1 will become 1 the power is 1.01 times 0.01 and this will become .033 times 0.33 times 0.33.

Now, what is your observation here? So, one thing we see it that if you take alpha is equal to 2, if you take alpha is equal to 2 then the topics where so the distribution where one topic is having very small probability we will get a overall very small probability, your multiple 0.98 by 0.01 times 0.01, this will become very small only when this will be this is like 1 by 3 times 1 by 3 times 1 by 3. So, as you increase alpha, it will prefer to have topics or the distributions where each topic the word probability is roughly equal.

It will prefer this 1, but if you having a smaller alpha then what you are seeing? So, now, 0.01 to the power 0.9, this will be now can written as 100 to the power 0.9, this will become very large and this will be roughly this will not be large. So, what will happen as your alpha becomes small they will prefer the probability distribution where one topic is having high probability and others are having low probability?

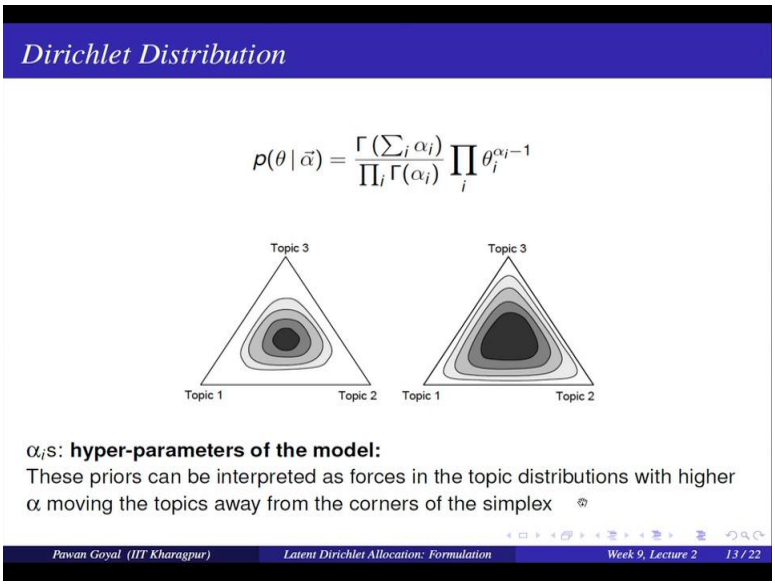
By tuning your alpha, you can prefer 1 topic probability, one sort of distribution over the distribution.

(Refer Slide Time: 29:59)



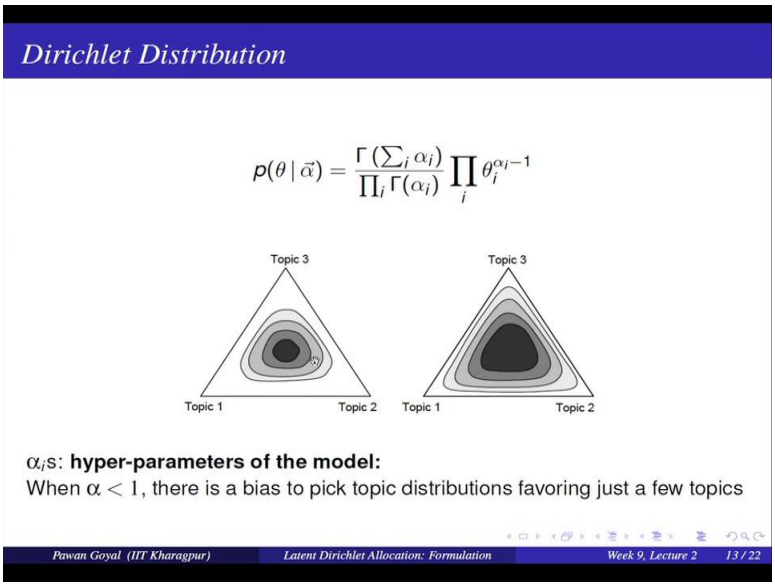
This is my Dirichlet distribution. Now, again to give you some visualization, so here are 2 different sort of simplex. So, alphas as such, you can interpret it as some prior observation count on the number of times a topic j is sampled individual alpha j so that is how many times this topic will be sampled?

(Refer Slide Time: 30:23)



But and you can also think as some forces and higher alpha will move the topics away from the corner of the simplex. So, let me come back to this point again. So, you are saying 2 different simplex, one where the topics are moved away from the corners, these are being moved away and here you are going towards the corners and this is for because of higher values of alpha and we will come back to this again.

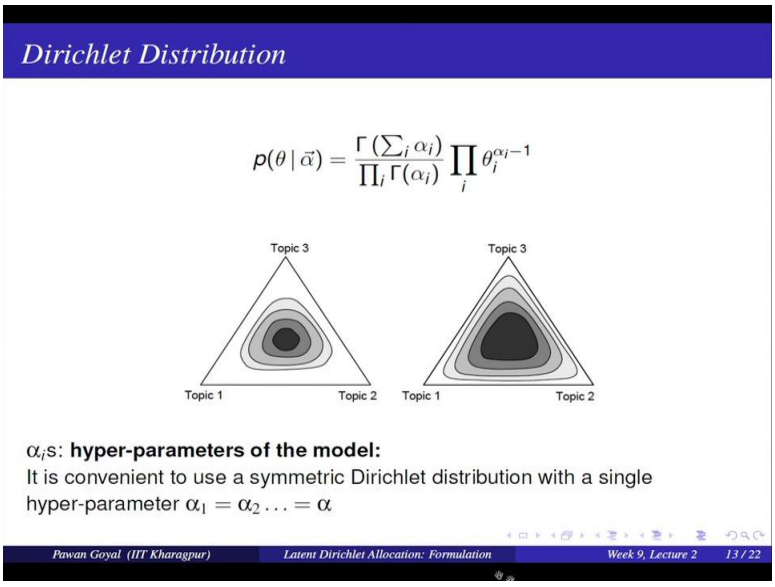
(Refer Slide Time: 30:51)



Now when alpha is less than 1, there is a bias to big topic distribution is favoring just a few topics and this is what we saw just now on paper that when alpha is equal to is less than 1, it tends to prefer the distributions where only a few topics as a high probability and others have lower probability.

Now, while in general, you can take different alphas for your different topics, what is convenient effect? You take all help us to be rough to be the same.

(Refer Slide Time: 31:30)



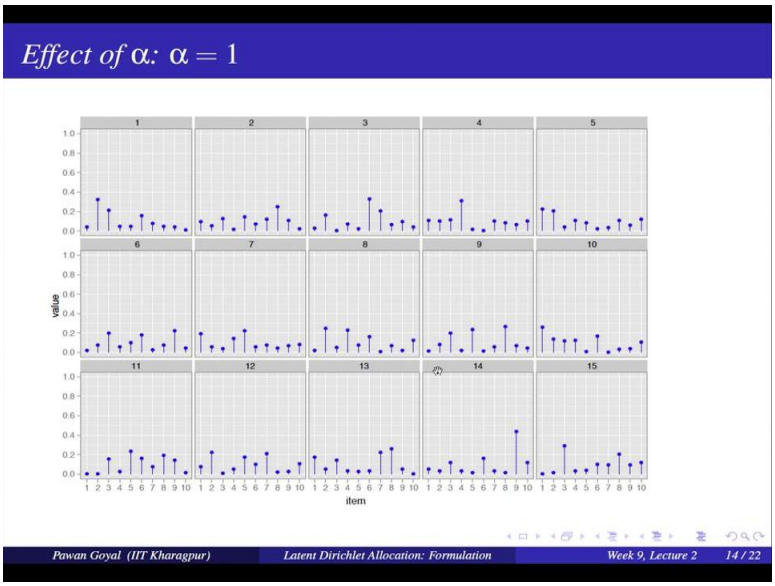
You have a singular hyper parameter alpha. So, I will all alphas are same now, what matters now is what is the value of this alpha? So, the relatively they are the same all alpha 1 to alpha and alpha capital D for all the documents, they are the same sorry alpha 1 to alpha call capital K, they are same, but what is the relative number is it like what is the number is it like 0.125, there will be a matter.

Now if alpha is small, what will happen? You will tend to prefer topics with way which sorry, you will tend to prefer distribution where 1 topic will have a higher probability. So, remember how do we define simplex? So, in this case what this it is boundary means. So, these colors denote, what is the probability distribution? So, black means a high dist probability distribution and so on and as it diminishes so the color becomes so faded then, you are seeing that the probability is decreasing.

In the left hand side figure, the probability is mostly centered for in the center of the simplex by in right hand side, it is moving towards the corners. So, that you can interpret it as if in this simplex whenever all 3 topics have the same weight it is given a higher probability here even if one topic is having in more proportions than others it is getting a high probability. So, this is not moving away from the center of this simplex.

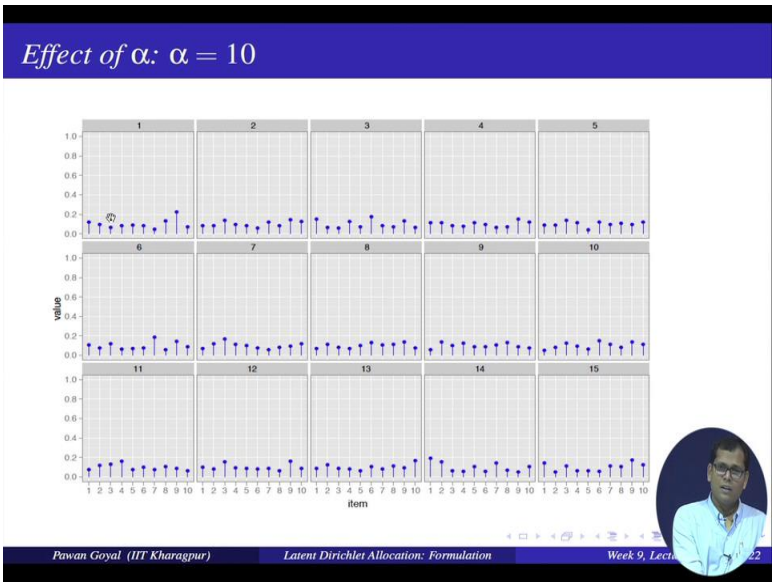
It is going towards the corner. So, if you go to the corner; that means 1 topic has the high probability, other 2 have the small probabilities. So, this is favoring also to have topics where sorry, also to have distributions where one topic has high probability than others while this we favor the distributions where all 3 topics have roughly equal probability. So, now, you can easily tell which one is corresponding to higher alpha lower alpha. So, the one that corresponds to lower alpha will be like that it will favor distributions where 1 topic is having a high probability than others.

(Refer Slide Time: 33:57)



Now, this is some from simulation, what happens if you take different values of alpha? So, if you take alpha is equal to 1, what kinds of distributions are preferred. So, there are 15 documents here that are being shown and there are 10 different topics.

(Refer Slide Time: 34:27)



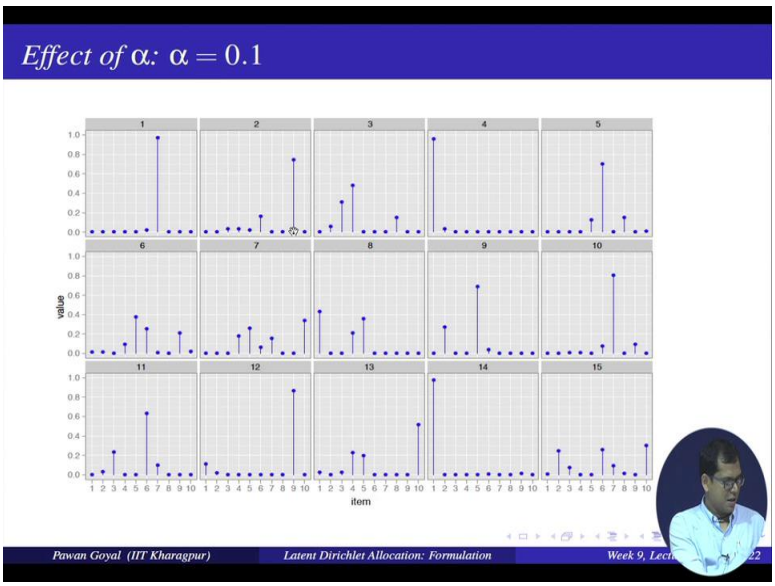
We see the distributions are so you are having some distribution for topic t 2 t 3 and so on for these 15 documents, they are different different distributions, but suppose you try to now increase the value of alpha as you go to 10. So, as you increase the value of alpha, it will start favoring those distributions where all the topics are roughly same,

probability you see now this is getting flattens. So, you are having, you are now seeing all the topics and if you try to increase alpha to 100, you will see all the topics have same probability and it is not something that you will there.

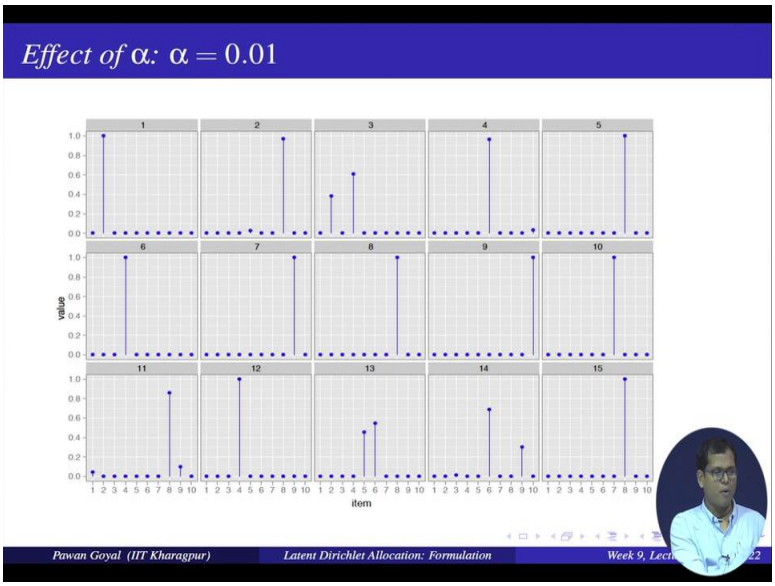
(Refer Slide Time: 34:49)

It is not good that each document has all the topics in same probability then if the topic model is does not put in any sense. So, alpha is equal to 1 was looking ok.

(Refer Slide Time: 35:12)



(Refer Slide Time: 35:24)



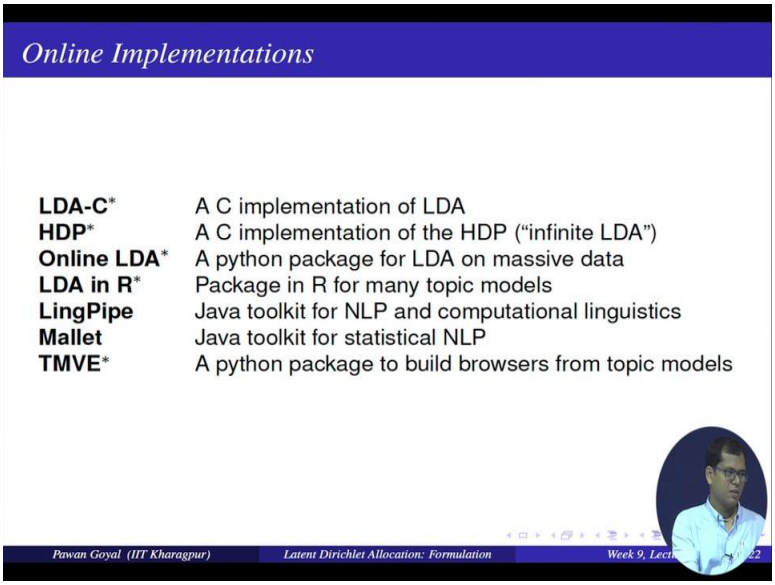
This kind of distribution we would like, but now suppose I want to decrease the value of alpha, if suppose I go from 1 to 0.1, now what you are seeing? It will start favoring the distributions where one topic has a high probability or 1 or maybe 2, but suppose I increase, I decrease it further to 0.01.

(Refer Slide Time: 35:34)

Most of the problems of probability 0, only 1 topic comes as a probability 1 or here 2 topics are coming. If you further reduce this value to 0.01, you get only 1 topic in each document. So, that will give you some idea of how if you modify or change this

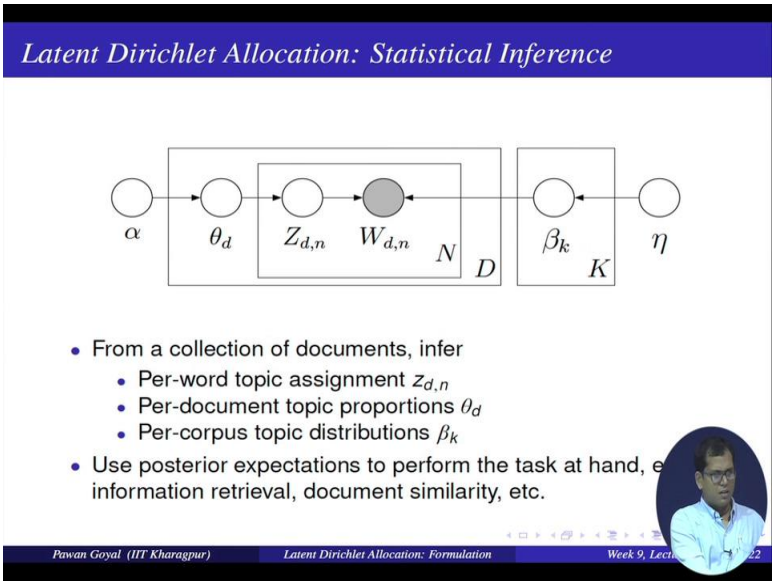
parameter alpha, how does this affect the overall topic distribution in my corpus, certain kind of distribution, we will get reference over others.

(Refer Slide Time: 35:54)



Now, for LDA, there are a lot of interventions that are available so and these are like very very popular, you can also use implementations that are available in gensim.

(Refer Slide Time: 36:06)



Now so we are again, so we have discussed all the in genetic part, but now I am full details that how do we; what is the generative model of LDA? But remember, what is your problem? How do we infer all these probabilities? So, that is I am given a

correction of documents and I want to infer per document topic assignment Z d n per document topic distributions theta d and per corpus topic distributions beta k, I want to infer all this.

Now, once I am able to infer all this, I can use this to find out say to use to for information retrieval, document similarity and many other task, but the question is once I am given these observations of the of the words, how do I infer all these probability values. So, there are different ways of doing that. So, we will discuss about one such method in the next lecture.

Thank you.