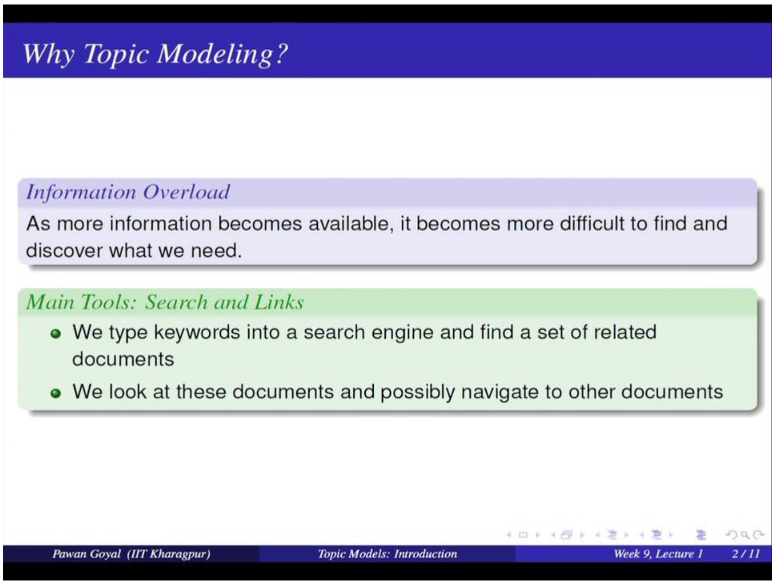
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 41
Topic Models: Introduction

So everyone welcome back to week 9 of this course. So we have been talking about semantics in the last 2 weeks. And we discussed 2 different approaches of semantics: one was using lexical semantics and otherwise using distribution semantics. And we saw how used the corpus or from a lexicon. You can extract semantics you can capture the meanings of words in some sense. That is whether 2 words are similar to each other and many other aspects.

So in this week we will discuss another very interesting way of capturing semantics by using topic models. And they have been very popular tools in NLP for whenever you have to talk about what are the concepts that are there in a particular document in a particular corpus and many other questions that that is around that are around the semantics. So let us see what are these topic models.

(Refer Slide Time: 01:09)



Why Topic Modeling?

Information Overload

As more information becomes available, it becomes more difficult to find and discover what we need.

Main Tools: Search and Links

- We type keywords into a search engine and find a set of related documents
- We look at these documents and possibly navigate to other documents

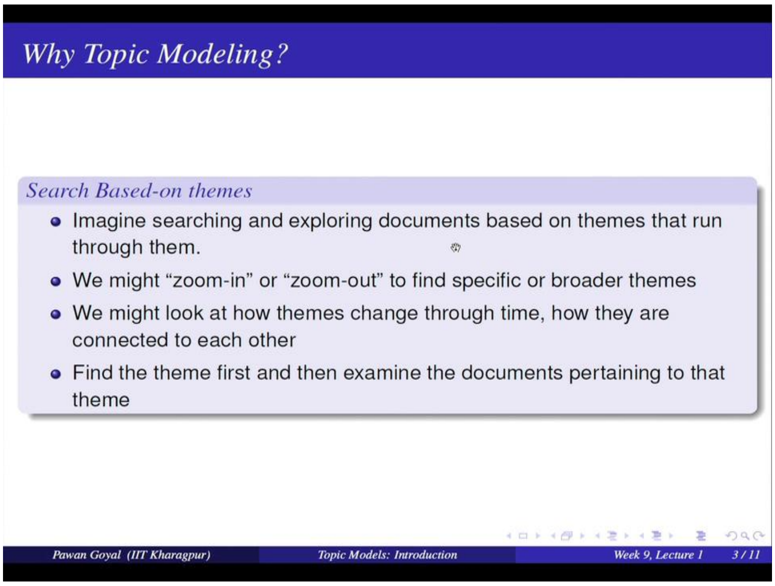
Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 2 / 11

Firstly, let us start the brief motivation that why do we actually need topic models as such. So when we talk about the data, so the text data in general on under that. So there is a huge amount of data available right in form of whatever form. You can think of there are lot of data in terms of news articles scientific articles and blog social media and everywhere.

So you can think of it as some sort of information overload. And you might be interested to capture only certain aspect of the data. So we want capture certain semantics. So what is an easy way in which you can some sort of organize this data and then this can help you to search or browse through the data more much more effectively. So right now what about main tools that we use for doing this search of different information. Either we type some query in our search engine right some human. So free text query we type and we get some reals, and we try to browses the reals and sometimes reals are not relevant or even if they are relevant we go to the pages to which the link to and keep on browsing. So that is some sort of generic behavior of how we try to explore this sort of information.

But there is no easy interface where you can you can say I want to understand these topics only I want to go too deep into this topic I want to go related topics and all that. That is not available with this search and link kind of behavior. So the topic modeling gives you an alternative method of going through this huge amount of information by some sort of searching based on themes criteria.

(Refer Slide Time: 03:00)



Why Topic Modeling?

Search Based-on themes

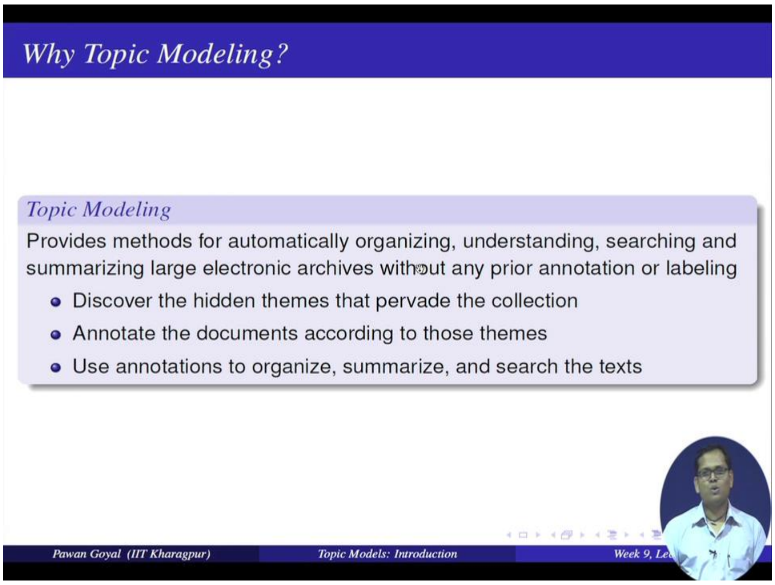
- Imagine searching and exploring documents based on themes that run through them.
- We might "zoom-in" or "zoom-out" to find specific or broader themes
- We might look at how themes change through time, how they are connected to each other
- Find the theme first and then examine the documents pertaining to that theme

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 3 / 11

So you can think of it as if you are having a set of documents and you know what are the themes that are running through this corpus. So you know this documentary about certain themes related to politics this document is about certain theme related to dramatic and so on different concepts.

Now, you are going to a popular thing and then you can either zoom in or zoom out. So that is you want to go inside the theme to very sub themes or you want to go out to a broader theme. So can this be fascinated by it by some sort of modeling? So we might also want to look at how the themes in a corpus are changing over time. How they are evolving over time this happens a lot in for example, scientific article a particular research in in physics might be starting with certain concepts and over that period of time it starting with it is now having new concepts, so by various discoveries. So can we have discovered that what are the themes setup already over time. And you can also talk about this behavior where you have you would select the theme first and then you try to examine the documents that talk about that theme.

(Refer Slide Time: 04:14)



Why Topic Modeling?

Topic Modeling

Provides methods for automatically organizing, understanding, searching and summarizing large electronic archives without any prior annotation or labeling

- Discover the hidden themes that pervade the collection
- Annotate the documents according to those themes
- Use annotations to organize, summarize, and search the texts

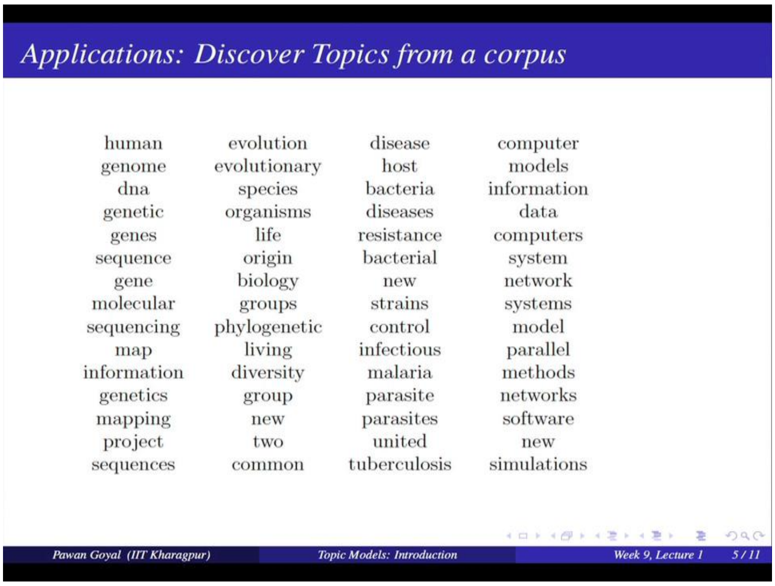
Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Le

So topic modeling, it is a method that gives you this facility by which you can organize all these collections by the themes that are occurring in in those documents and you can then understand search and do summarization and many other applications, without and this is important that for doing all that you do not have to give any prior manual efforts in in labeling the data. So you do not have to tell that this document concerns this topic or this document concerns that topic you do not have to do any labeling. So the interesting aspect is that you give it a corpus an in an uncivilized manner it learns what is the overall structure of different topics and these document concerns which topics and so on.

So this is very, this is some sort of very interesting aspect that made topic modeling very popular that, you no need to have any prior annotation as such. Although there are some variations we will also talk about, where you can give an annotation to get some different sort of topic modeling, but overall in the generic picture you do not have to give any sort of annotations. So it learns on it is own from the huge amount of data.

So what you do here. So you discover what are the hidden themes that are pervading to the collection and the topic model itself annotates the document as for these themes. So it will tell you these are the overall themes this document contains these themes and So on. And once you have these annotations that are given with the topic model you can use that to for validation task like organizing the collection summarizing collection searching when the user query comes by using each annotation. So many different applications, we will cover some of those in this week and, yes we can there is a huge amount of literature around topic models. So you can also feel free to read about it.

(Refer Slide Time: 06:08)



Applications: Discover Topics from a corpus

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 5 / 11

So let us talk about one of the applications before we go into the modeling in in detail. So we talk we are we are consistently talking about discovering topics from a corpus. That is one of the most important application of topic models. So something like that. So you have a large corpus and can you discover that there are many topics and topic look like this. By topic I mean a set of words that occur a lot in that topic. So let us see one topic. Human genomic dna genetic genes sequence gene molecular sequencing map information.

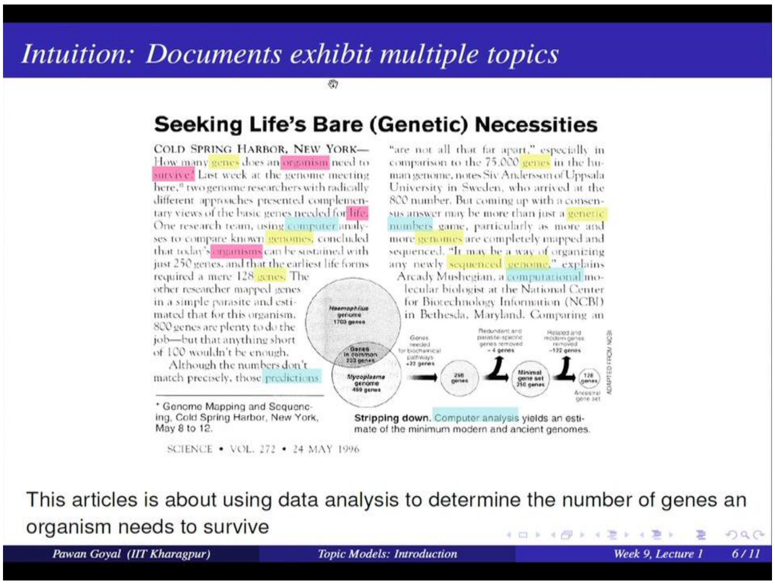
So topic model we will try to get you, this is one sort of topics. It will not give you a label. So although you can see this looks like a genetics topic of genetics it will not give you a label, it will tell you that this is there is a topic in this corpus there is a theme, and in this theme each words are more having higher probability. So these are the more important words in this theme. And then it will tell there is general theme that is going through this corpus, something like this evolution, evolutionary species organization life origin biology. And then you see a set of terms. These are having a high probability in that theme there is no labels that topic model is giving, but you can see it might be evolutionary biology. Then similarly another theme, disease, host bacteria, disease so on and 4 theme computer models information data computers and so on.

So the topic models will tell you in this collection you have these 4 themes in addition to some other themes. So that is your predefined number that you will need to give to the models. So you can say I want 50 different themes in this corpus, or 20 different themes in this corpus. Yes, although there are again variations where you might not have to specify these number a prior we will briefly talk about those also.

But for now let us say we tell the model we need that money themes. So try to discover these are the important themes that I am seeing with a pervading through this corpus. And these are some examples that we are seeing here that are coming from a real corpus by doing the LDS, applying the LDA model.

Now, once you have these themes then the topic model will also tell you this document is about having only these 2 themes out of these 50. This document is having these 5 themes out of these 50, in in what proportions and so on. So like that you can now think about organizing huge your whole collection and also some sort of a summarization or searching through this.

(Refer Slide Time: 08:38)



Intuition: Documents exhibit multiple topics

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

This articles is about using data analysis to determine the number of genes an organism needs to survive

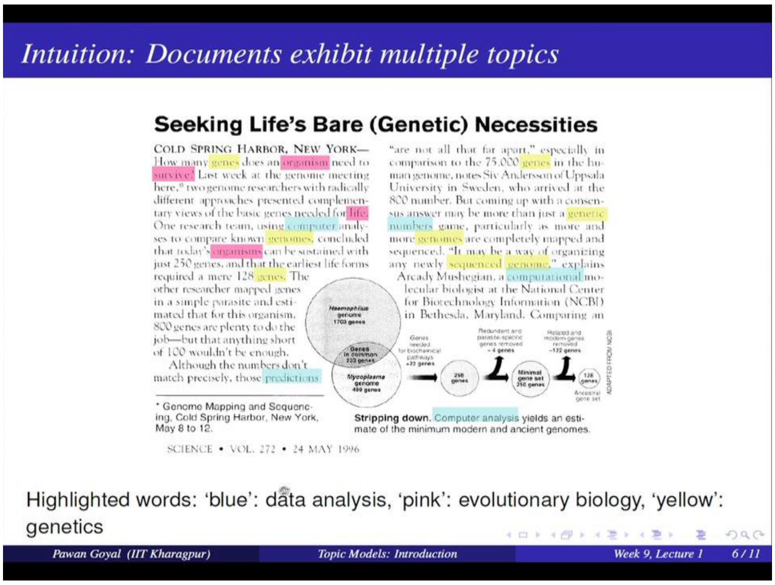
Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 6 / 11

So why is this intuition interesting? So when you look at a particular document in general you will see that yes there are actually some particular themes that are that are pervading this whole document. So that is you have somehow thought of making a document that concerns these themes only. So here is an article from science magazine 1996 Seeking Lifes Bare Genetic Necessities and so if you read this article, this is about using some sort of computational data analysis to determine the number of genes and organisms that needs to survive. So how many genes and organism needs to survive.

So this topic this article will blend some of the topics that are important to convey this message. And these words are denoted in various colors here. So we are seeing some words with pink some words with yellow somewhere to blue and so on. Let us read the words in in pink. So they are organism, survive, life, organisms, yellow, genes genomes, genes genetic, sequenced genome and blue will be computer productions numbers, computational computer analysis and so on. So you can clearly see 3 different themes one is about genetics another is about the evolution another is about this computational data analysis.

So this article is blending these 3 themes together. It is can be capture that in an uncivilized manner without someone has to manually annotate this this is the themes in this document.

(Refer Slide Time: 10:24)



Intuition: Documents exhibit multiple topics

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed -4 genes
Minimal gene set 250 genes
Related and modern genes removed -122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

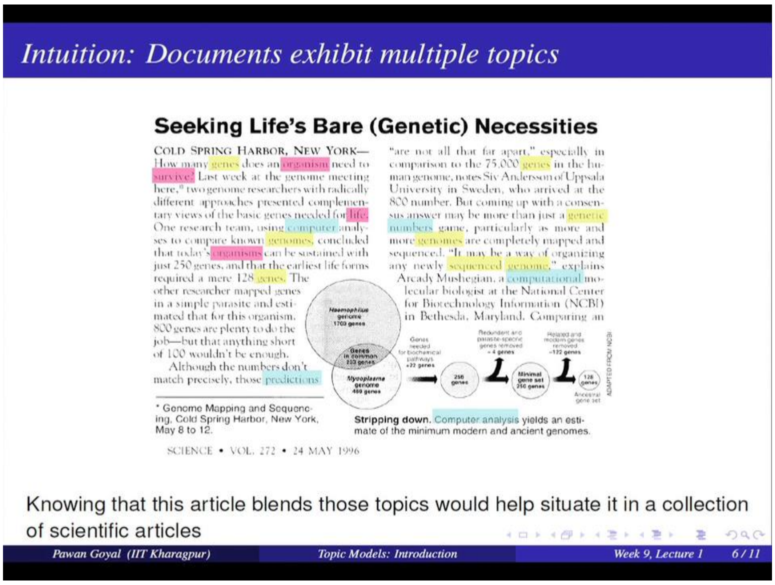
SCIENCE • VOL. 272 • 24 MAY 1996

Highlighted words: 'blue': data analysis, 'pink': evolutionary biology, 'yellow': genetics

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 6 / 11

So if I yes you see blue is data analysis pink is evolutionary biology and yellow is genetics. So this article is blending these 3 topics in different proportions. And that is the motivations that is what is the hypothesis that we are having about your corpus. In the corpus there will be various documents, there will be a there will be some big number of themes that are going through this corpus, but when you look at a particular document it will have only a subset of these themes, in some proportions can we automatically capture those by using some modeling.

(Refer Slide Time: 11:02)



Intuition: Documents exhibit multiple topics

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

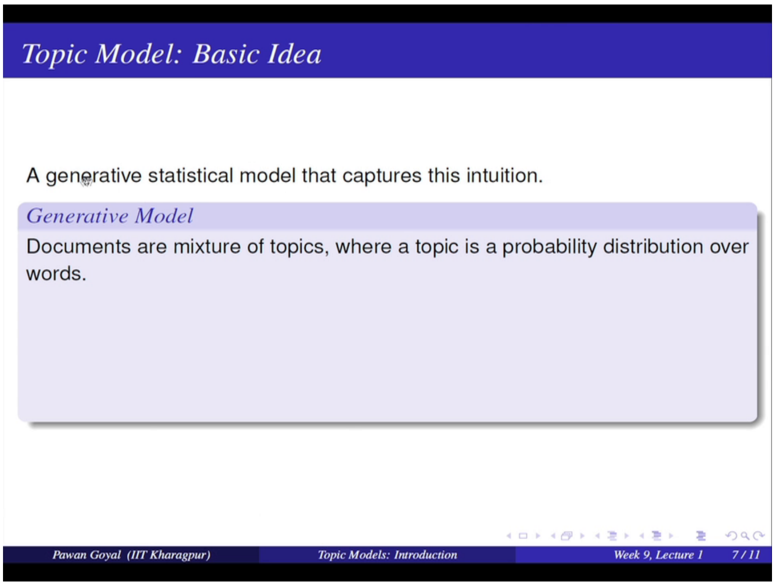
SCIENCE • VOL. 272 • 24 MAY 1996

Knowing that this article blends those topics would help situate it in a collection of scientific articles

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 6 / 11

So once you know that this article blends this topic together you can situate that in an in a collection of scientific articles. You can say where this where this situates what are the articles it is similar to and so on.

(Refer Slide Time: 11:17)



Topic Model: Basic Idea

A generative statistical model that captures this intuition.

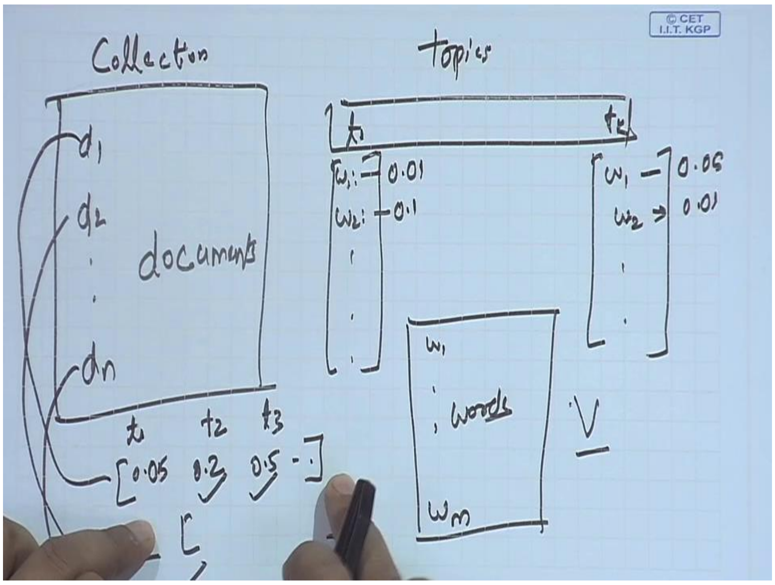
Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 7 / 11

So what is the basic idea? You are not going to the mathematical details that we will cover in the next week, but let us see the intuition. So this is the important idea. So topic model is some sort of generative model. And this just as a model that captures this idea about the collection having topics and topic document having different proportions. So what it says is that. So you are having some documents they are nothing but mixtures of topic and a topic it is nothing, but the probability distribution over words. So what are 3 themes we are talking about?

(Refer Slide Time: 11:55)

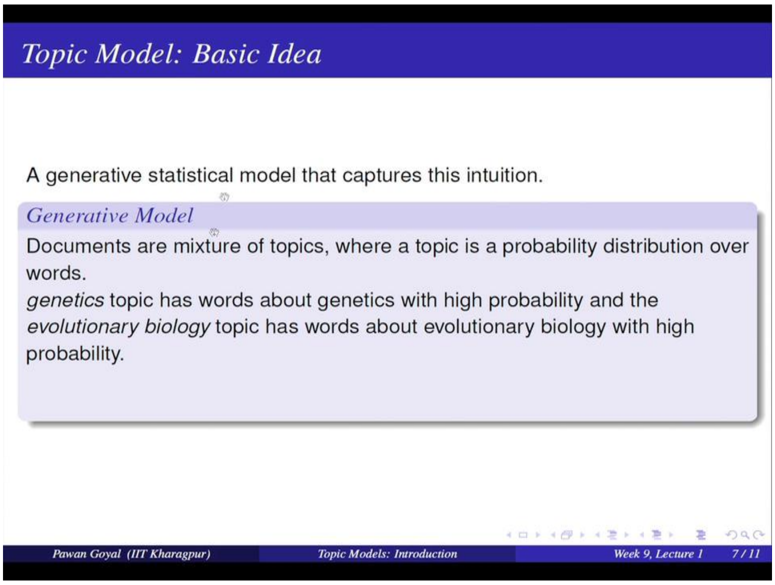


We are talking about a collection, that is having documents d1 d2 up to dn. Here a documents in the collection. Then you are having some topics. So suppose there are some t1 to tk topics. And use your documents in document some words will occur and you say these are my vocabulary. These are my words w1 to some wm.

So here 3 themes, in the collection there are documents, documents some words occur. Suppose you have unique words define your vocabulary and there are some topics. Now what topic models say? Yes topics are nothing but probability distribution over words. So t1 would be something like a distribution. So here word 1 comes the probability of 0.01 word 2 comes with the probability of 0.1 and so on. Similarly, tk will be again a probability distribution word 1 comes with a probability distribution of 0.05, word 2 with a 0.01 and so on. An idea is that probably there will be some part of the themes.

So this is my topics. Topics are defined by probability distribution over words. Now what are my documents? Documents are again some sort of probability distribution over topics. So I say d1, what is d1, d1 is having a distribution over these k topics. So I will say topic one occurs with the probability of 0.05, topic 2 with probability 0.2 topic 3 with 0.5 and so on. So I will say something like topic 3 and topic 2 are more turbulent in this document. And not the other topics and same way I can situate all the documents in some sort of mixture of topics. And this is very important to understand the topic model. So what are my topics - probability distribution over the words; and what are my documents? Again mixture components or topics or you can offset probability distribution over the topics.

(Refer Slide Time: 14:21)



Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

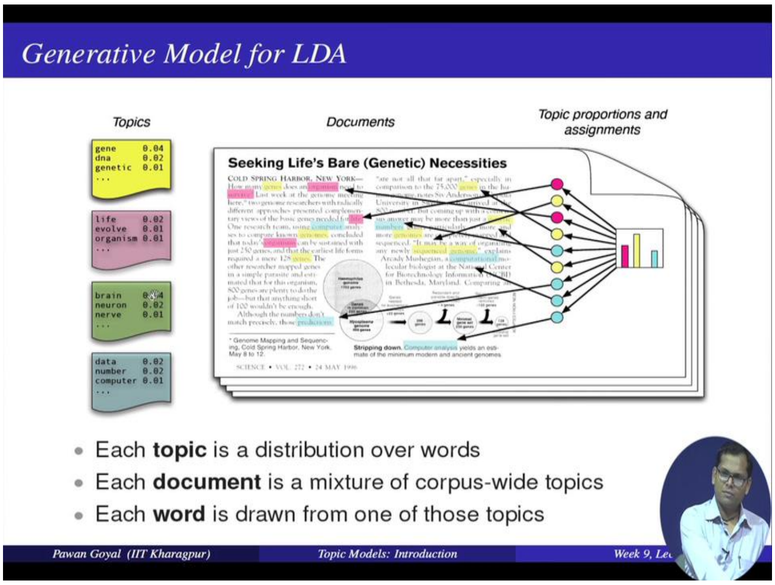
Documents are mixture of topics, where a topic is a probability distribution over words.
genetics topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability.

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 7 / 11

So this is all add up to 1, this will all add up to 1 for all the topics, for all the documents. So what is that mean? So you have a topic like genetics, remember no labeling, but we can see when you see that words. So this is probability word genetics. So if we have topic about genetics this will have words about genetics with high probability, and if we have a topical of evolutionary biology, it will have a topic about evolutionary biological of high probability. So this will be making 2 different themes in this connection.

And the model is a generative model and as we understand generative model. So it will work like from we will first generate the topics, yes we first define the topics. Then when we have the topics, you will now start building the documents, you say document will have some proportion of these topics and then I will write the words in the in the documents. So topic search in that the first and then the documents as per the generative model.

(Refer Slide Time: 15:13)



Generative Model for LDA

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Le

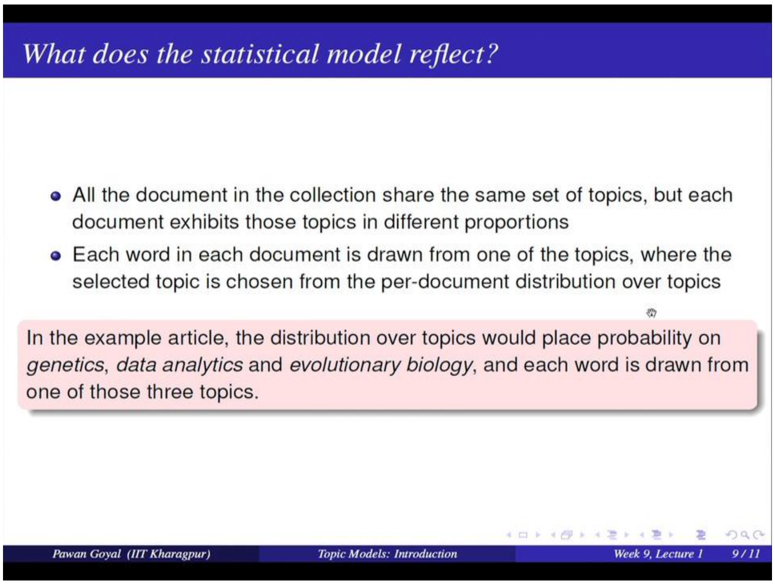
So this picture will make themes clearer. So we are having seen some nice colors here. So, on the left you are seeing some topics right. That is all you are talking about. We are having a set of topics any topic is a probability distribution over the words in the collection. So like here this topic has the word gene with the probability of 0.04, dna with 0.02, genetic with 0.01, life with 0.02. This is different topic brain neuron nerve right data computer number different probabilities. You see seen their different themes.

Now, once we have this corpus wide themes. So each topic is a distribution over words. Do you understand that now what is a document? So you are seeing a collection here and there are a lot of documents, right. And we are being shown one document. So each document is a mixture of corpus wide topics. So these are my corpus wide topics now I will take a particular mixture of these topics. So suppose one of my mixture is here. I take this topic red yellow and blue I will take these 3 topics only maybe others away with a very small fraction. And this defines the topic distribution of my document.

So I have now seen said that my document contains these topics with some proportions. Now how do I generate the words in the documents right. That is important I need to generate the words. So how are these words generated that is again interesting? So you have this you think of it as a multinomial distribution. And from this distribution you sample a topic. Suppose the first sample is this pink that pink topic. That is about an organism life evolve and so on these sample is topic.

Now from the same topic you have to generate a word. How do you generate a word again this topic is what a probability distribution over words? Think of it as a multinomial distribution against a sample a word from here. And that is what you will generate in the document. And this we will keep on going for generating all the words in the document in the in the document. So you will say next word I will again sample a topic I get this yellow topic about gene dna genetics. I use this topic to generate a word by sampling from a multi multinomial distribution. And that is what are you keep on repeating. This is for this document again for next document I will select a mixture of topics. And then once I have the mixture I will again keep on selecting the words that will go in the document.

(Refer Slide Time: 17:51)



What does the statistical model reflect?

- All the document in the collection share the same set of topics, but each document exhibits those topics in different proportions
- Each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics

In the example article, the distribution over topics would place probability on *genetics*, *data analytics* and *evolutionary biology*, and each word is drawn from one of those three topics.

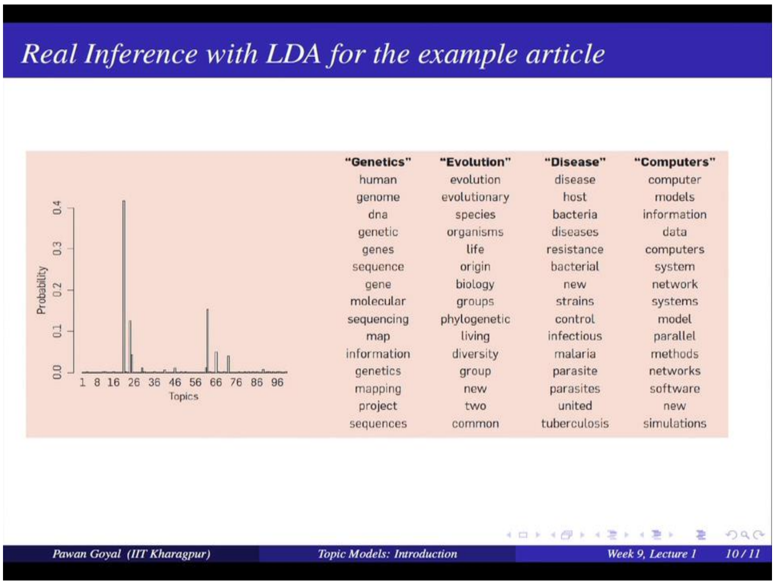
Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 9 / 11

So this is a statistical model also we called it a generative model. So what is this reflect? So it reflects to 2 important things. What are those? All the document in the collection share the same set of topics right. We have an underlying set of topics all documents are sharing the same set, although the proportions are different. So each document exhibits those topics in different proportions. That is one important fact. And then what is the other fact? Each word in each document is drawn from one of the topics where the selected topic is chosen from the per document distribution over topics.

So is that clear? So for each document I have a collection of topics, I have a distribution about topics. So I say this topic is probably 0.2, 0.5, 0.1 and so on. Now each word in the document is sampled from this distribution. How? You sample a topic first and then as per the topics probability distribution you sample a word. This we will cover in the generative model in more details, but that is the intuition that we showed also from the picture. See there are 2 important facts about LDA.

So if we think about the example article from science magazine that we were seeing. So the distribution over topics would place a probability on genetics data analytics and evolutionary biology. And each word will be drawn from one of these topics. So what will happen? When you are doing the influencing you will say, these this document contains these 3 topics and how do we generate a word you will sample a topic from here and as for the topics probability distribution sample a word and keep on generating the words.

(Refer Slide Time: 19:40)



Real Inference with LDA for the example article

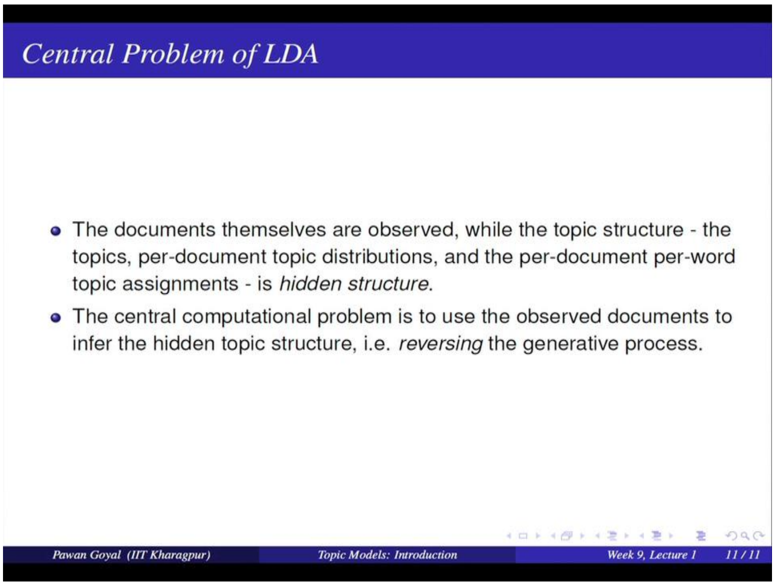
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 10 / 11

So this is the genetic model, but is that what you also see when we have a real corpus we apply LDA and do an influencing. So that is what is being shown. So for this collection of science papers. So in proper model was trained with hundred topics and this is what you see. So these are the hundred topics and for this particular document some topics have a high probability. So this topic has probability roughly 0.4 0.15, 0.12 and the next one having 0.5. So if you see the top 4 topics in this particular document and try to see the most probable words in these topics. So that is what you see. So this topic has human genome dna genetics and so on.

Since our topics that we were showing earlier. And you see this was obtained without doing any manual labeling. So you find out this this document contains these 4 different themes important themes. And this is you can also do some labeling later on manually. This is not done by LDA, but get this is also not very important. So what is important is that LDA can help you to obtain this sort of distribution. This in a very uncivilized manner. So you can find out what are the overall themes and for a document what are the most important themes.

(Refer Slide Time: 21:03)



Central Problem of LDA

- The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments - is *hidden structure.*
- The central computational problem is to use the observed documents to infer the hidden topic structure, i.e. *reversing* the generative process.

Pawan Goyal (IIT Kharagpur) Topic Models: Introduction Week 9, Lecture 1 11 / 11

So we talked about what is the generative model of LDA. But what is the main problem of LDA? So we are saying we will generate the topic first and then we will generate the documents right by sampling a distribution of topics each for each word I will take a topic it is distribution I will take a word and so on. But that is not how we write the documents right. So how will that we used. So it has to be used in a manner that I am observing some data I have the generative model and now I am trying to estimate the parameters of this genetic model by using my observations that is what parameters will maximize the likelihood of seemed observation. So this is like reversing this whole process of LDA that we are talking about.

So we know the documents object in a collection and if the documents, but I do not do the topic structure. So I do not know what are my topics. So whatever distribution of words within each topic. I do not know for each document what will be the distribution of topics. And I also do not know for each word in a document what will be the topic assignment. I do not know any of this a prior. So this is all my hidden structure.

So center problem LDA is to reverses in to process and use the observed documents to infer the hidden structure. So what is the hidden structure can you infer that by seeing only the object documents. And this may also called as some sort of reversing the genetic process. And that is a center problem of LDA. So this is this initial introduction lecture was about to give to give you the intuition, but now that you have some intuition of what LDA is what topic model is, we will next going to details about what is the mathematical model of LDA. And how do you solve this problem of reversing the generative process. So that will be recovering in more details in the next lecture.

Thank you.