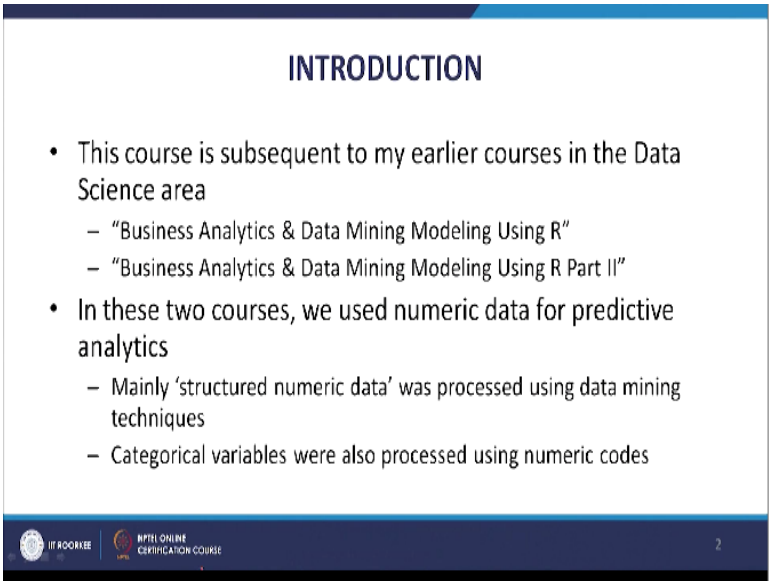
Business Analytics And Text Mining Modeling Using Python
Prof. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology-Roorkee

Lecture-01
Introduction-Part I

Welcome to the course business analytics and text mining modeling using Python. This is the very first lecture and we are going to cover the introductory aspects of this particular course. So lets start as you might be familiar this course is subsequent to my earlier courses in the data science area. So previously I have taken two courses business analytics and data mining modeling using R.

(Refer Slide Time: 00:51)



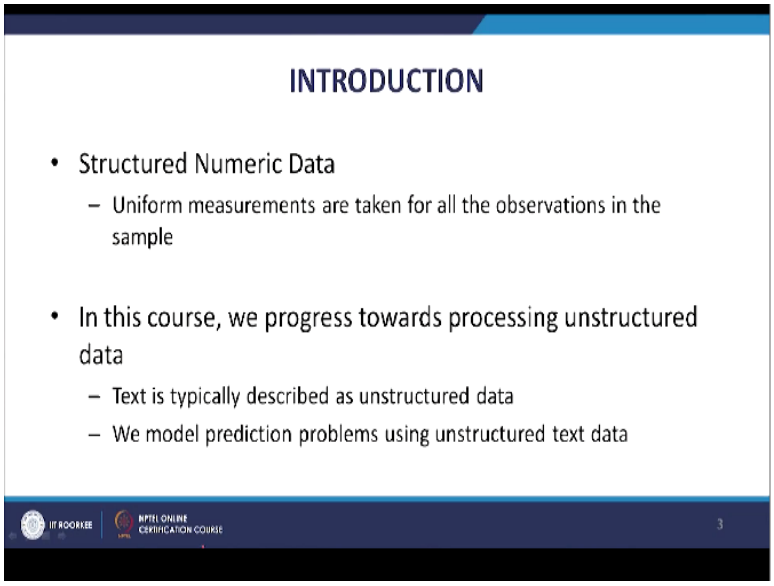
And business analytics and data mining modeling using R part II, so these two courses focused on the data mining aspects where we work with numerical you know data which is highly structured in nature and we apply various statistical data mining and mathematical techniques to solve you know prediction and classification problems. So in these two courses as you can see in the slide also we are talking about the numeric data very structured and processed using data mining techniques.

You might also you know from first few lectures of these courses you would also get an idea about especially from the first course business analytics and data mining modeling using R

where we have talked about the you know variable types like you know continuous variables, categorical variable and there we have also discussed about categorical variables and how they are processed you know in data mining techniques.

So typically we use numeric codes to process to define categorical variables. So you know so that kind of analysis takes place in data mining techniques and that we covered in these two courses. Now when we talk about a structured numeric data that we have used in the previous courses.

(Refer Slide Time: 0two:two5)



It is about uniform measurements taken for all the observations in the sample, so in a tabular format, in a matrix format we have columns and rows, so variables on the column side and observations on the road side. So for a given column we have to take same measurement for all the observations. So that is how we get our you know structure numeric data.

In this course we progress towards you know processing unstructured data, text is typically described rather it is naturally you know unstructured data. So we are going to be dealing with text data in this course and that is why the name is business analytics and text mining modeling. So in this course another thing that is going to be slightly different is the uses of data science platform.

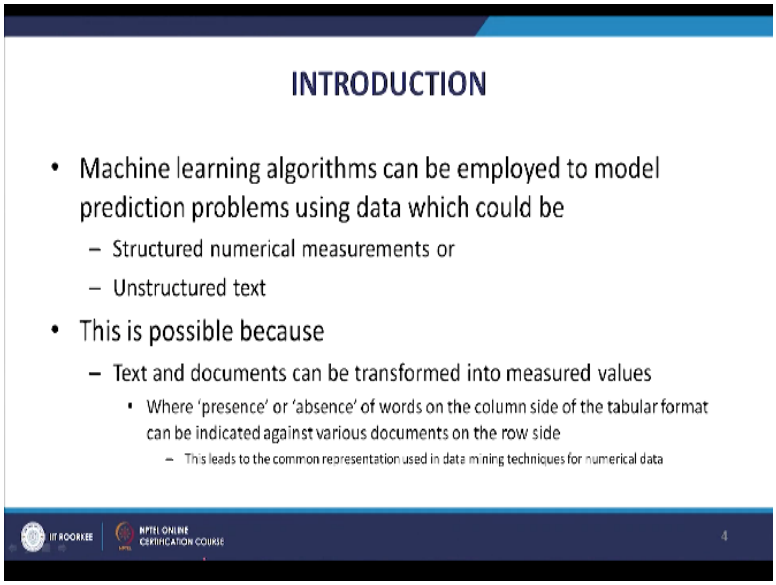
So in the previous two courses we had used R platform for performing various exercises writing

over codes and you know doing our model building process. In this course we would we would be using Python platform, so python is as we have talked about in previous two courses python is slowly catching up and becoming has already become second most viewed placed form after R in data science arena.

And in future it might even overtake R platform, so we are going to introduce Python platform as well in this course and much of the work that we you know plan to do in this course would happen using that platform. So as I said text is typically described as unstructured data. So we are going to model prediction problems using unstructured text data. So we will learn those aspects in this course.

Now in our previous two courses we have talked about machine learning techniques as well, for example decision trees, artificial neural networks. Now some of these techniques were earlier employed to model predictive prediction problems using numerical data.

(Refer Slide Time: 04:4two)



Now many of the and some of the techniques some of the machine learning you know can also be used to model prediction problems in you know using unstructured text. So the same thing is mentioned here that machine learning algorithms can be employed to model prediction problems using data which could be structured numerical measurements as we did in the previous two courses or unstructured text that is something we are going to cover in this course.

So how this is going to happen because when you imagine you know a structural numeric data you are always going to imagine the tabular format, the matrix format, the columns and rows. However, when we talk about the unstructured data we talked about the text data it is you know series of words you know composed together to create a text document.

So how I wanna seen learning algorithm are going to be used to process this kind of data where we know that you know machine learning algorithms they typically work with the highly structured data. So this is now we are going to discuss what is typically done to these to that label where machine learning algorithms can be applied.

So the next point is so how this is possible is because texts and documents can be transformed into major values. So what can be done is presence or absence of words can be shown on the column side of the tabular format. So any text document it would be in essence composition of a set of words. Now if we are able to identify the unique words out of you know all the text documents that we might have.

And if we are able to so that you know if we are able to represent those unique words on the column sides just like we did you know for you know just like we did for in the structure numeric data set in the previous two courses where we had variables they of course most of them were numerical in nature. So on the column side we can have these unique words, so in the tabular format and on the row side for each document we can record whether these terms are present or absent in that particular document.

So in essence we would be capturing the presence or absence of words in all the documents that we might have in our collection. So in this fashion we would be able to transform the text data into a format into a tabular or matrix format just like the one used in you know data mining techniques and we apply machine learning algorithms. So that kind of transformation can be done.

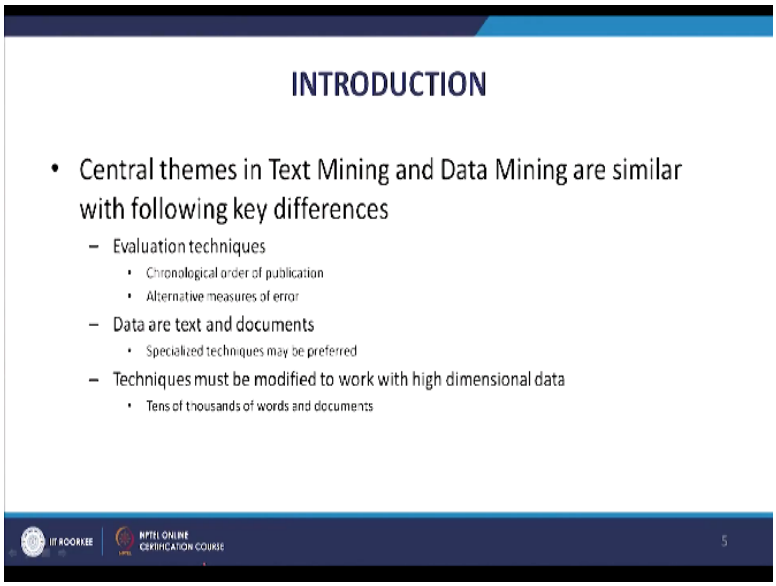
And therefore you know we can say that you know some of the you know machine learning algorithms could be found really useful to solve to deal with to process unstructured textual data

and solve in your predictive text analytics problems. So if we do this kind of arrangement where we are indicating the presence or absence of words for various documents. So this will lead us to a common representation used in data mining techniques for numerical data.

If we are going to apply machine learning algorithms and the finally the presentation format, the tabular format, the matrix format is going to be used in both text mining and data mining. So will understand that central themes are going to remain similar. So eventually data is you know transformed into the particular you know highly structured tabular form where you know machine learning algorithms can be applied.

So central themes remain similar in both the domains, text mining as well as in data mining. However, a few key differences are there.

(Refer Slide Time: 08:58)



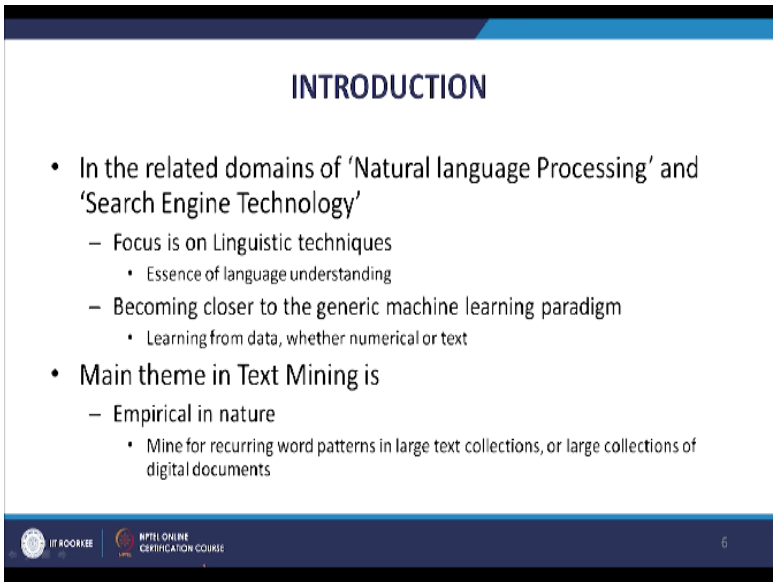
For e.g. evaluation techniques, so these techniques will have to adapt to the chronological order of publication, this is different from the numeric data, we will have to use the alternative measures of error, because the way we are used to in terms of measuring accuracy and error of our models in data mining that might not be really you know an relevant or meaningful to that extent in text mining.

So these are going to be some differences, now data are texts and documents, so some of these specialized techniques which can process this kind of data are might be preferred. So this is

going to be another key difference, then when we talk about text data because you know any even a small document might be having a huge number of unique words which would essentially are going to be treated as you know you know vectors attributes or predictors.

So therefore we are going to be dealing with high dimensional data. So because of that you know if we have tens of thousands of words and documents so we are going to be dealing with you know we are going to have a lot more number of you know unique words and therefore attributes and will have to process the dimensionality of the data set would be on the higher side. So techniques will have to be scaled up, will have to be modified to be able to process that you know large amount of data.

(Refer Slide Time: 10:5two)



So these are some of the key differences that are definitely going to be there. Now there are some other you know related domains like natural language processing and search engine technology which are you know very close to what we do in text mining. However these two domains natural language processing and search engine technology they focus on linguistic techniques. so essence of language understanding that is very important there.

Because you know they are trying to learn from that semantics of the language and then they are trying to use that in their tasks. So however because of the increased uses of machine learning paradigm in general the fields are becoming more closer and you know learning from data whether it is numerical or text, it is you know becoming more easier. So you know these fields

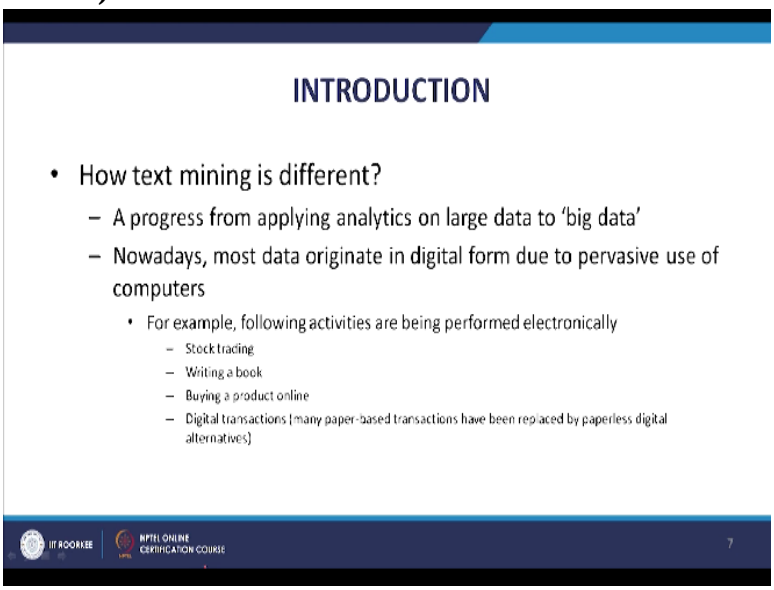
natural language processing.

And search engine technology which were you know earlier dominated by linguistics and with the help of you know computer specialists who will help them in terms of you know processing you know this kind of data. So now those fields and text mining they are becoming closer because of the computation abilities that we have and our ability to apply some of the machine learning techniques for text mining tasks.

Now if you look at the way text mining is structured the main theme is very empirical, so we are not going to apply the language understanding like that is done in natural language processing in text mining, there is very empirical you know theme that we follow in text mining and that is what it makes a closer to data mining rather than natural language processing. So we typically mind for recurring world patterns in large text collections or large collections of digital documents.

So essentially just like data mining where we are mining numbers, here we are going to to be mining words so we are looking for patterns in words, there we were looking for patterns in numbers. So the mainly the main theme is empirical in nature and there is no focus on the language aspects of you know the data which is text.

(Refer Slide Time: 13:10)



So this brings us to the question how text mining is different, so few more points apart from what

we have discussed till now. We are this is in essence a progress from applying analytics on large data to big data, because when we define big data it is not just large datasets of highly structured you know numeric data, but you know it will also include so much of you know texts, textual information unstructured data in the form of text.

That is being generated in the internet where we have so many of ecommerce websites with comment section people talking about products, obsess, commenting, writing blogs and there are technical you know websites focusing on the technical content, there is so much of you know interactive data in the text format that is being generated. So all that is also considered to be part of big data.

Now naturally this kind of data you know that is this kind of data is unstructured in nature. So we are in a sense processing we are you know in terms of applying our analytics from highly structured data to you know to big data. So that is how text mining is different from you know others you know data mining specifically. So it brings us closer towards you know big data analytics. So a few points about you know the big data the way you know the nowadays the amount of data that we have to deal with.

So nowadays most data originated in distort form do you do pervasive use of computers so few examples are given here following activities are being performed electronically, stock trading you know earlier you know traditional way we used to call agents and they used to help us and do the paperwork for us. So that is you know more or less gone and we typically rely on electronic mechanisms to for stock trading, writing a book so earlier you know pen, paper and pencil we used you know we used to write our you know our ideas our thoughts to in a book form.

But nowadays we use word documents, Microsoft word documents to write our books, buying a product online, digital transactions. So if we look at the you know most of the governance space whether in the private sector or public sectors we are moved away from the paper-based transaction to you know paperless digital alternatives. So in that sense you know digital transactions have increased almost you know most of our transactions are now happening in the

digital form.

So once **so** most of our activities are you know nowadays being perform electronically, so that is going to generate a lot of data and then you know it will you know there are going to be services and people are going to be talking about those products and services. So a lot more text data is also going to be generated. So text mining in a sense is different, in a progress towards you know a step towards progress in the sense where we are also trying, where we are also making an attempt to analyze the unstructured data.

(Refer Slide Time: 16:47)

Let us discuss a few more points about data mining versus text mining. So we will discuss these similarities as well as dissimilarities. So both text mining and data mining they are about finding valuable patterns in data, in data mining we you know try to find valuable patterns from the numeric data, in the text mining we are going to you know we are we attempt to find valuable patterns from the text data.

So that is the similarity, now if we look at the data mining domain and analyze how it has progressed. Now at this moment it is in its maturity phase, so we are not going to expect any significant development it might happen but you know we believe that it has least into the maturity phase and incremental development of course that is that will continue, no longer an emerging economy so emerging technology.

So it is being widely used in industries you know and the you know public sector as well, so incremental development will happen but we are not expecting very significant things. Techniques are highly developed as you might have you know experienced in our previous two courses where we were fairly you know sure about what is supposed to be done given a data set what techniques are to be applied.

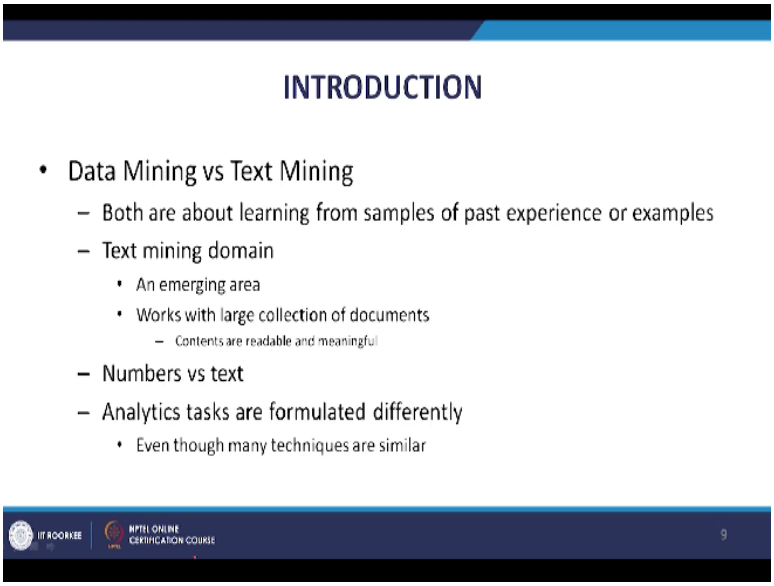
How the results are to be analyzed, how the results are to be interpreted you know typical data mining modeling process we followed, so many things that we discussed in the previous two courses. They tell you that you know this feed data mining domain access has moved into the maturity phase. It requires highly structured numeric data so that is true that we have learned from previous two courses also.

Everything is in tabular format all the variables that we have columns they are very well defined, how they are to be measured that is also you know very well defined. So it of course when this kind of we are dealing with this kind of highly structured data it will require extensive data preparation. So from various sources we collect and gather data and then we will have to process that we have to transform that to finally arrive at a highly structured form.

So that is a you know very labor intensive effort, so much of what we do in data mining you know much time off what we do in our data mining is actually involved in this data preparation of his because of the requirement of highly structured data. Lacks universal replicate applicability, this is mentioned mainly in the sense that because we will we would typically be working with the numeric data which is not the case in the overall you know big data space that we have.

So therefore you know until and unless you know we process some of that data we cannot apply in order to mining techniques. So therefore there is lack of universal applicability in that sense.

(Refer Slide Time: 19:49)



Another point which is common to both data mining and text mining is that, both domains are about learning from samples of past experience or examples. So in both the domains we try to learn from the you know past experiences past record past examples that we have.

We build our model based on that and then we try to project our learnings on the new data that we might have to score and that we might have to project. So in both the domains this thing is common. Now points specific to text mining domain this is an emerging area, so this is though a lot of work has already happened, but this is this particular area is still considered to be an emerging area.

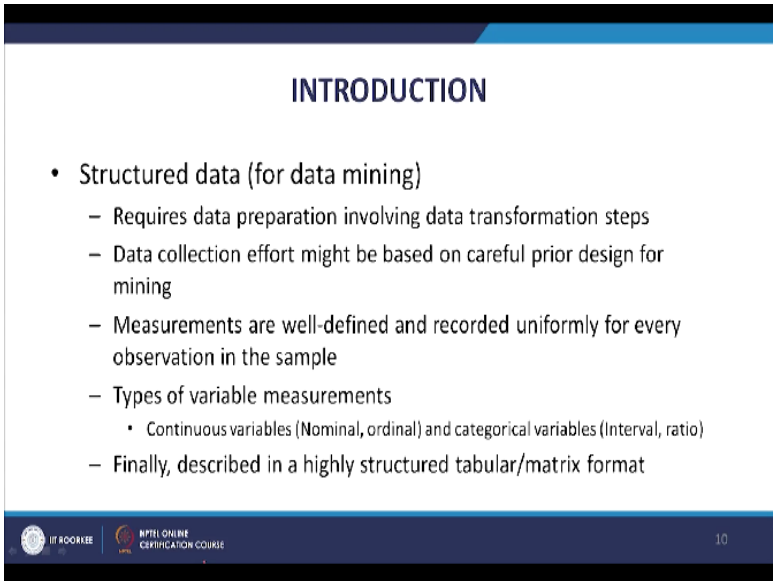
And in this we work with large collection of documents, so documents when we talk about documents they are going to be in our spoken language, so the words the contents that are there they are in the form of several words which are readable, which are meaningful which is not the case in data mining. So we are dealing with a large collection of documents in text mining which have the contents are meaningful, contents are readable.

So it is like numbers versus text whenever we talk about you know data mining versus text mining it is about numbers versus texts in a brief sense. Now if we look at the tasks the kind of tasks analytical tasks that we have to perform. They are going to be formulated differently so many techniques are going to be similar in nature because as we discussed even in text mining

we eventually transform our you know textual data our text documents into a format which is closer to you know tabular or matrix layout being used in data mining.

So some of the techniques are very similar, but the tasks the way the problems are you know formulated that is you know different.

(Refer Slide Time: two1:58)



Now let us talk about a structured data which is typically used in the in a data mining effort and the various aspect of a structural data and then later on we will focus on the unstructured data and mainly text and once we learn this then we will be able to understand the differences between the two you know. So let us start so as we have discussed you know any kind of structured data it requires data preparation involving data transformation steps.

So from various sources we collect data, we process them we prepare and you know present into finally a tabular or matrix format as we have discussed. So much of the data collection effort you know that goes into this might be based on very careful prior design. So we might have a business problem, we might you know reduce our you know our analytics component from it.

And whatever is required, whatever kind of data, whatever kind of variables are required to you know to model that analytics problem we try to collect that so in a sense there is a you know prior design in place in our data collection effort you know when we are trying to you know gather structured data, collect structured data. Measurements are well defined so as we have

learned in our previous two courses the variables that are used in the column side in the tabular format they are very well defined.

We define them properly and how they are going to be major that is also very well defined, so and the recorded uniformly for every observation in the sample. So for every observation for every row for a given column for a given variable in the same fashion we measure that particular variable. So these variables might also be representing to a higher level concept, so that concept is to be measured you know uniformly across the observations in the sample.

Now we look at types of these measurements types of variable measurements these are two types mainly continuous variables you know categorical variables. Again continuous variable you know continuous variable two types you know interval and ratio and the categorical variable two types nominal and ordinal. So these are the types of variable measurements, finally we get a highly structured tabular or matrix format.

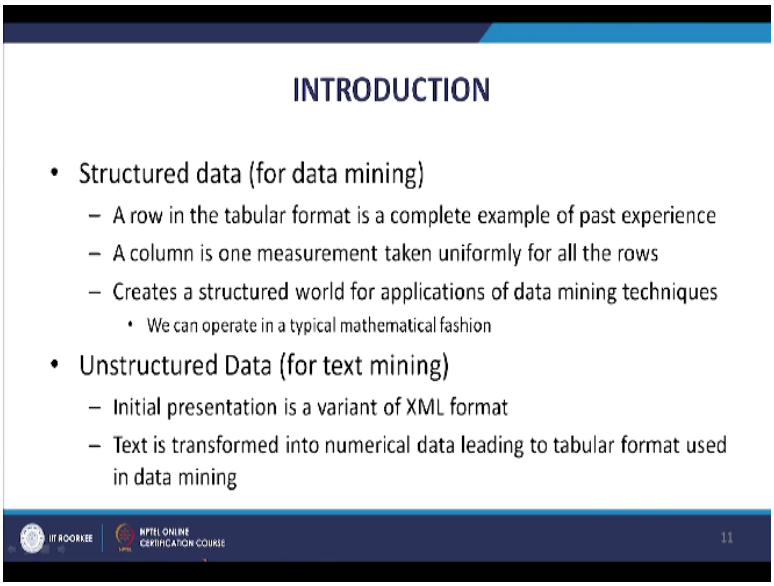
Few more points about a structured data so now if we look at the rows and column side in a structured data a row in the tabular format is a complete example of past experience. So an observation could be about individual could be what form you know could be about you know household so that is you know and whatever you know in from whatever information is being measured along the column side for different variables, that is going to convey a you know a complete example of past experience.

So that is how row can be you know described. A column is one measurement taken uniformly for all the rows. So a column is representing a variable a predictor it might be higher level you know concept, so that is measurements it is well-defined measurements are taken in a uniform fashion across all the rows. So all this creates a structure word for applications of data mining techniques and typically we operate a you know a typical mathematical fashion you know we you know partition the data.

Based on sample that we have taken we partition one you know training partition we take that and you know build our model then we apply that model to score the test partition and you know

check you know how the performance is there training partition verses you know test partition and all that. So a typical you know step by step fashion in mathematical fashion, all these steps are performed using various functions that are available in our data science platform.

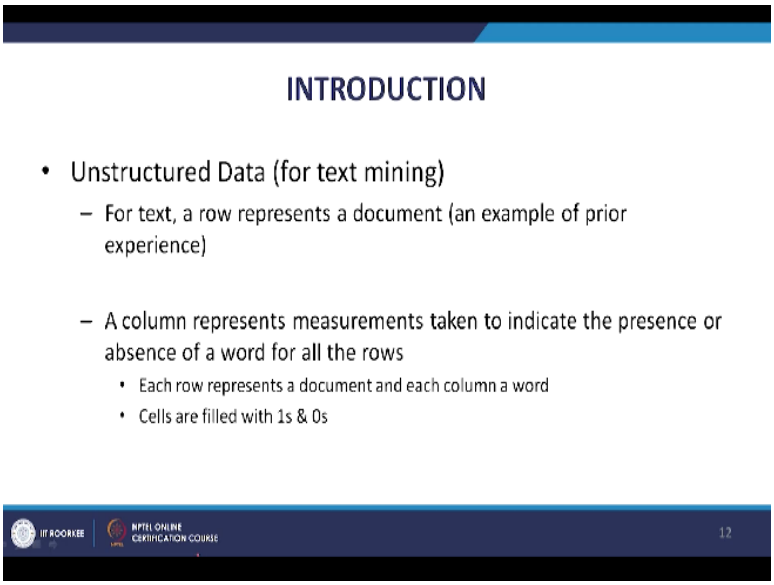
(Refer Slide Time: two6:01)



Now let us talk about unstructured data, so this is mainly in the context of text mining, so initial presentation is a variant of XML format. So of course when we talk about textual document text documents it is this composition of words so sentences paragraphs, all those things are going to form a particular you know text document. So it is very much similar to you know any XML document that we see in the internet.

So now for processing in text mining this text is transformed into numerical data leading to tabular format using data mining. This is what we have mentioned quite a few times.

(Refer Slide Time: two6:43)

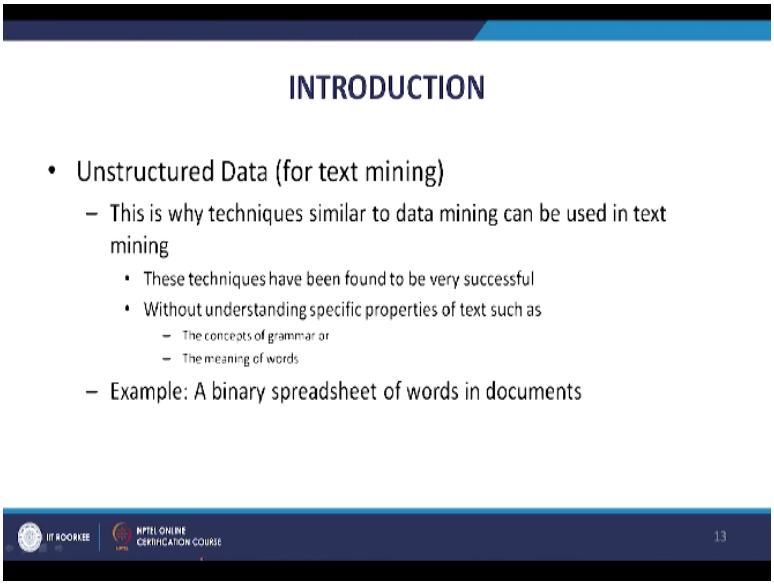


So when this transformation is done when we arrive at a tabular format, so for text a row is going to represent a document. So in data mining typically the row is an observation on individual or form or you know industry or country or household here a row is representing a document. So that document is in essence is an example of a prior experience and column is going to represent measurements taken to indicate presence or absence of words.

So the words are going to be there in the column side and you know the column is going to represent the words measurements. So each row represents a document and each column a word and now the cells in the tabular layout they are going to be filled with 1s and 0s. So if that word is present in that document than 1 if it is upset than 0. So the whole you know the layout all the cells is going to be filled with 1s and 0s which is slightly you know different from the layout that we might have in data mining.

Where actual values because the variables are representing an actual you know a higher level concept and those values are real in nature, here it is either 1 or 0 you know indicating presence or absence. So one when we do this kind of transformation to apply text mining techniques you know it is very easy for us to understand why techniques similar to you know data mining can be used in you know text mining.

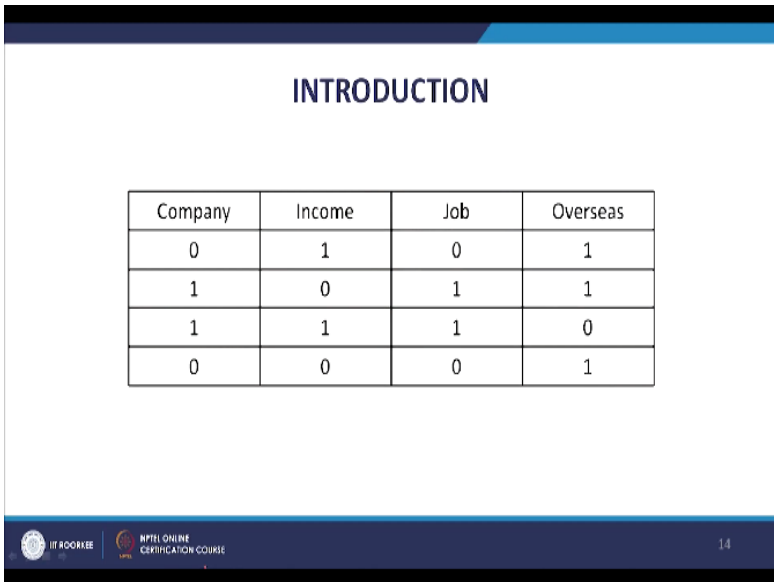
(Refer Slide Time: two8:two7)



Because we can arrive at the you know similar kind of format and these techniques have been found to be very useful and this goes you know this is happening, this success is achieved even without understanding a specific properties of text like concepts of grammar meaning of words, so without even focusing on the grammatical aspects or the you know meaning of those words.

We are applying you know in text mining domain you know our machine learning techniques in a very empirical fashion and still you know these techniques have been found to be very successful. So let us have a look at a binary spreadsheet of words in documents to give you an idea.

(Refer Slide Time: two9:17)



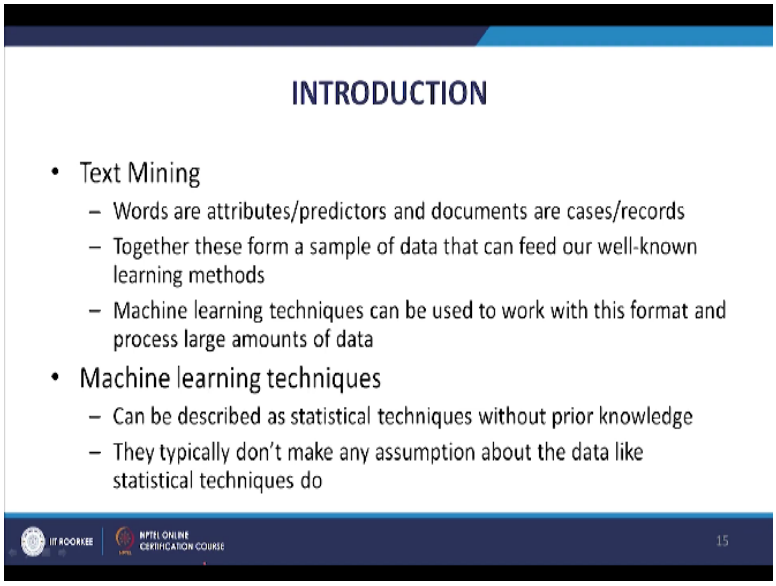
So this is one example you can see here we have you know first column company, then second column income, third column job, fourth column overseas. So these are words, so they might represent variables well in this case you know these are terms are representing words, words in our collection of documents and you know first row that we have here is representing you know is indicating whether these words.

For example company whether it is present in the first row that means in the first document or not. So it says 0 that means it is absent, then in the second column we have income, so whether this word income is present in the first document or not. So the value is 1 so it indicates that this income word is present in the first document. So on so forth similarly for document number two for company it is 1.

For income is 0, for job and overseas it is 1, so in this fashion whether these words you know as indicated in the column side whether these are present or absent in the you know corresponding documents on the row side, that is how this is being represented. Now in a data mining context when we are dealing with numerical you know variables mainly you know these values are going to be very different.

For example if income was considered to be a variable will have actual income values here, like 10,000, two0,000, 5,000, 1,000. So those kind of numbers will have, these numbers are you know quantification of the income they will they signify something here you know these are just words and we are just you know indicating you know the presence or absence. So this is a very rudimentary way of representing our collection of text documents. But even with this when we apply machine learning algorithms we get very successful models.

(Refer Slide Time: 31:twotwo)



So in text mining words are attributes or predictors and documents are cases or records, together these form a sample of data that can be that can feed our well-known learning methods. So learning methods unsupervised learning methods that we have learned in previous two courses. So this kind of you know arrangement you know can be used to use to apply some of the methods that we have used in the previous courses.

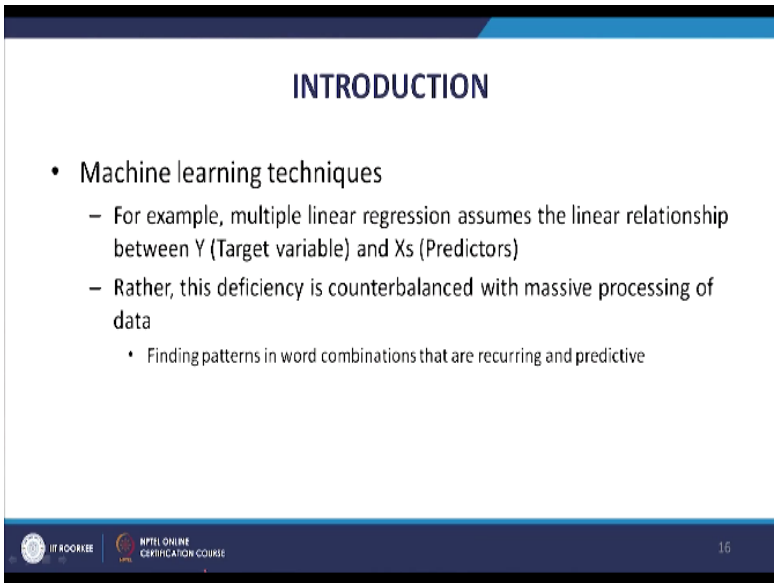
Machine learning techniques can be used to work with this format and process a large amounts of data, now let us discuss few more few points about machine learning techniques in general. So how we can describe machine learning techniques, they can be described as statistical technique techniques without prior knowledge. So in statistical techniques we are always you know going to you know make certain assumptions about data or variables.

And how they are going to be related and based on that based on certain rules you know we do our modeling. So but in machine learning techniques we do not make those kind of you know assumptions about the structure of the data and we try to you know process massive amount of data and try to counterbalance that you know that you know that lack of assumptions. So that is why you know we can refer you know machine learning techniques as a statistical technique without prior knowledge.

So we are not assuming anything about the data, we are trying to learn whatever is there from the

data itself. For example if we take this statistical technique multiple linear regression so it is going to assume that a linear relationship between Y target variable and X's which are predictors.

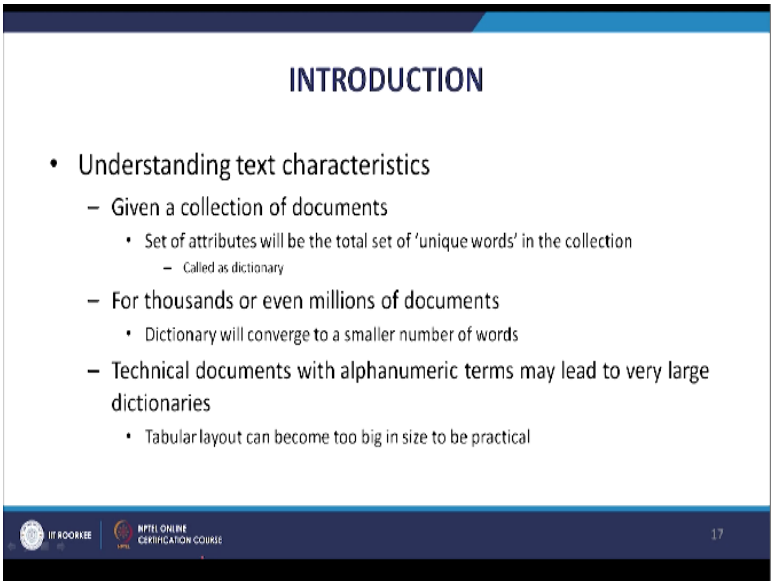
(Refer Slide Time: 33:two5)



So between Y and X's we are going to assume linear relationship and multiple linear regression which is not the case in the machine learning techniques, where you know we are you know whatever kind of a relationship is there it has to be learnt from the data. So do not make apriori assumptions about those structures. So but you know like of these assumption might create your most certain deficiency in machine learning techniques.

But that is very well counter balanced with massive processing of data. So essentially we are finding in using machine learning techniques we try to find patterns in world combination that are recurring and predictive. So any patterns that can help in our prediction tasks. So that we are always looking to identify using machine learning techniques. Now to further understand the difference between the unstructured data and text data numerical versus text.

(Refer Slide Time: 34:30)



Let us understand few text characteristics, so if we given a collection of documents the set of attributes you know or predictors set of attributes that we are going to you know show in the column side, they are going to be you know total set of unique words in the collection. So in the collection we might have you know tens of thousands of documents and those documents are going to be composed of you know a set of unique words.

So these set of unique words are going to be you know set our set of attributes that is going to be used you know for our you know analytics in text mining. So for thousands or even millions of documents essentially. So this set of attributes is also referred as dictionary and when we are dealing with thousands or even millions of documents this dictionary will convert into a smaller number of words.

Because there can be you know if we look for meaningful words there can be very you know a limited number of words as we can see in any you know dictionary belonging to any you know particular language. So eventually irrespective of the number of documents that we are dealing with we will converge to a smaller number of words. However, there might be situations where you know this thing might not remain practical in terms of number of words that we will have to deal with number of attributes that we might have to deal with in text mining.

When we are analyzing technical documents because technical documents are also going to

include the alphanumeric terms. So therefore you know the dictionaries the size of dictionaries can be very large and therefore the tabular layout is going to be very large in size and therefore it might not remain practical to process it. So this also can happen in the text mining context. So you know there are few more characteristic that we would like to discuss which will make it easier for us to process data. So we will stop here and we will discuss some of these you know characteristics in the next lecture, thank you.

Keywords: Text mining, Data mining, Prediction, Classification, Attributes, Regression.