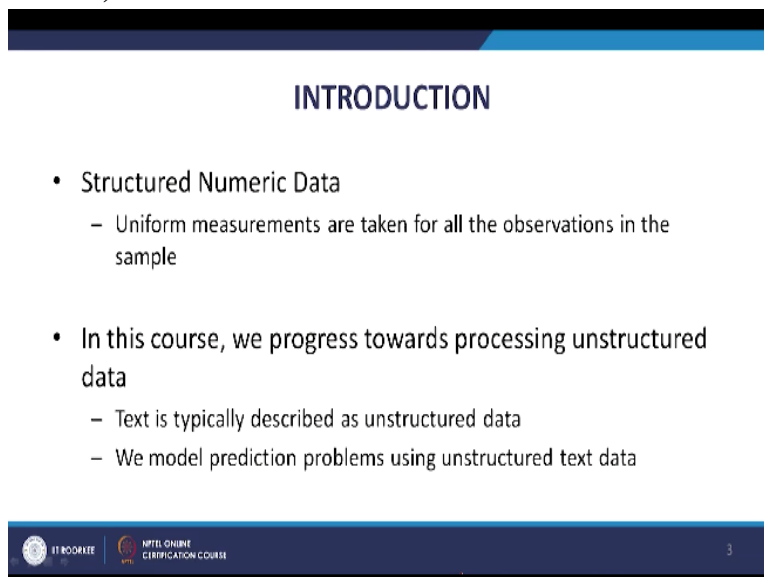


Business Analytics And Text Mining Modeling Using Python
Prof. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology-Roorkee

Lecture-02
Introduction-Part II

Welcome to the course business analytics and text mining modeling using python. So in the first lecture we started our discussion on this topic, we talked about the previous 2 courses that I have done and some of the lectures that might be relevant for this course as well. So we refer to that aspect then we compare how you know this particular course, text mining course is in a sense extension of what we have done in our data mining courses.

(Refer Slide Time: 00:59)



INTRODUCTION

- Structured Numeric Data
 - Uniform measurements are taken for all the observations in the sample
- In this course, we progress towards processing unstructured data
 - Text is typically described as unstructured data
 - We model prediction problems using unstructured text data

ET ROORKEE NPTL ONLINE CERTIFICATION COURSE 3

We talked about structured numeric data and other aspects of processing unstructured data.

(Refer Slide Time: 01:01)

INTRODUCTION

- Machine learning algorithms can be employed to model prediction problems using data which could be
 - Structured numerical measurements or
 - Unstructured text
- This is possible because
 - Text and documents can be transformed into measured values
 - Where ‘presence’ or ‘absence’ of words on the column side of the tabular format can be indicated against various documents on the row side
 - This leads to the common representation used in data mining techniques for numerical data

ST BOORKEE MTEL ONLINE CERTIFICATION COURSE 4

And we talked about machine learning algorithms and their relevance in crossing unstructured data as well. So all these things we discussed in the previous lectures.

(Refer Slide Time: 01:14)

INTRODUCTION

- Central themes in Text Mining and Data Mining are similar with following key differences
 - Evaluation techniques
 - Chronological order of publication
 - Alternative measures of error
 - Data are text and documents
 - Specialized techniques may be preferred
 - Techniques must be modified to work with high dimensional data
 - Tens of thousands of words and documents



ST BOORKEE MTEL ONLINE CERTIFICATION COURSE 5

We compared text mining and data mining scenarios as well we looked at points which are common, the common themes and we looked at certain points which are different in text mining in comparison to a data mining.

(Refer Slide Time: 01:30)

INTRODUCTION

- In the related domains of 'Natural language Processing' and 'Search Engine Technology'
 - Focus is on Linguistic techniques
 - Essence of language understanding
 - Becoming closer to the generic machine learning paradigm
 - Learning from data, whether numerical or text
- Main theme in Text Mining is
 - Empirical in nature
 - Mine for recurring word patterns in large text collections, or large collections of digital documents



 IIT BOMBAY  NPTEL ONLINE CERTIFICATION COURSE 6

We also discussed a bit about other domains like natural language processing, search engine technology and how closely they are related with text mining. So those aspects were also discussed in the previous lecture.

(Refer Slide Time: 01:48)

INTRODUCTION

- How text mining is different?
 - A progress from applying analytics on large data to 'big data'
 - Nowadays, most data originate in digital form due to pervasive use of computers
 - For example, following activities are being performed electronically
 - Stock trading
 - Writing a book
 - Buying a product online
 - Digital transactions (many paper-based transactions have been replaced by paperless digital alternatives)

 IIT BOMBAY  NPTEL ONLINE CERTIFICATION COURSE 7

Data mining versus text mining.

(Refer Slide Time: 01:52)

INTRODUCTION

- Data Mining vs Text Mining
 - Both are about finding valuable patterns in data
 - Data mining domain
 - In its maturity phase
 - No significant development is expected
 - Incremental development will continue
 - No longer an emerging technology
 - Techniques are highly developed
 - Requires highly structured numeric data
 - Involves extensive data preparation
 - Lacks universal applicability



(Refer Slide Time: 01:53)

INTRODUCTION



- Data Mining vs Text Mining
 - Both are about learning from samples of past experience or examples
 - Text mining domain
 - An emerging area
 - Works with large collection of documents
 - Contents are readable and meaningful
 - Numbers vs text
 - Analytics tasks are formulated differently
 - Even though many techniques are similar



(Refer Slide Time: 01:54)

INTRODUCTION

- Structured data (for data mining)
 - Requires data preparation involving data transformation steps
 - Data collection effort might be based on careful prior design for mining
 - Measurements are well-defined and recorded uniformly for every observation in the sample
 - Types of variable measurements
 - Continuous variables (Interval, ratio) and categorical variables (Nominal, ordinal)
 - Finally, described in a highly structured tabular/matrix format


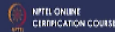
  10

So all these structured data some points here also we had discussed.

(Refer Slide Time: 02:01)

INTRODUCTION

- Structured data (for data mining)
 - A row in the tabular format is a complete example of past experience
 - A column is one measurement taken uniformly for all the rows
 - Creates a structured world for applications of data mining techniques
 - We can operate in a typical mathematical fashion
- Unstructured Data (for text mining)
 - Initial presentation is a variant of XML format
 - Text is transformed into numerical data leading to tabular format used in data mining

  11

And you know then we moved to unstructured data for text mining.

(Refer Slide Time: 02:06)

INTRODUCTION

- Unstructured Data (for text mining)
 - For text, a row represents a document (an example of prior experience)
 - A column represents measurements taken to indicate the presence or absence of a word for all the rows
 - Each row represents a document and each column a word
 - Cells are filled with 1s & 0s

So this part was also discussed in the previous lecture, we talked about how unstructured data is going to be represented you know in the text mining course.

(Refer Slide Time: 02:19)

INTRODUCTION

- Unstructured Data (for text mining)
 - This is why techniques similar to data mining can be used in text mining
 - These techniques have been found to be very successful
 - Without understanding specific properties of text such as
 - The concepts of grammar or
 - The meaning of words
 - Example: A binary spreadsheet of words in documents

And how we are going to take an empirical view of you know textual data how we transform the text data into a numerical you know presentation just like we used in data mining techniques. So these aspects we had discussed.

(Refer Slide Time: 02:37)

INTRODUCTION

Company	Income	Job	Overseas
0	1	0	1
1	0	1	1
1	1	1	0
0	0	0	1



This is one example that we have you know discussed where we are indicating presence or absence of these words indicated in columns, one indicating presence, zero indicating absence.

(Refer Slide Time: 02:48)

INTRODUCTION

- Text Mining
 - Words are attributes/predictors and documents are cases/records
 - Together these form a sample of data that can feed our well-known learning methods
 - Machine learning techniques can be used to work with this format and process large amounts of data
- Machine learning techniques
 - Can be described as statistical techniques without prior knowledge
 - They typically don't make any assumption about the data like statistical techniques do



So this presentation also we had discussed in the previous lecture.

(Refer Slide Time: 02:51)

INTRODUCTION

- Machine learning techniques
 - For example, multiple linear regression assumes the linear relationship between Y (Target variable) and Xs (Predictors)
 - Rather, this deficiency is counterbalanced with massive processing of data
 - Finding patterns in word combinations that are recurring and predictive

ST BUCKLEE NPTEL ONLINE CERTIFICATION COURSE 16

Then we talked about many other aspects and then we were discussing the text characteristics that are going to be really important here.

(Refer Slide Time: 03:03)

INTRODUCTION

- Understanding text characteristics
 - Given a collection of documents
 - Set of attributes will be the total set of 'unique words' in the collection
 - Called as dictionary
 - For thousands or even millions of documents
 - Dictionary will converge to a smaller number of words
 - Technical documents with alphanumeric terms may lead to very large dictionaries
 - Tabular layout can become too big in size to be practical

ST BUCKLEE NPTEL ONLINE CERTIFICATION COURSE 17

So we talked about you know collection of documents and the set of attributes which are you know going to be you know forming the total set of unique words in the collection. So all these aspects we discussed. We talked about you know dictionary that would we you know created in this process. So let us move further, so to continue our discussion on understanding text characteristics.

(Refer Slide Time: 03:37)

INTRODUCTION

- Understanding text characteristics
 - For a large enough collection of documents, the tabular/matrix layout would be too sparse
 - Any individual document will use only a tiny subset of the potential set of words in a dictionary
 - Techniques used to process text expect the sparse data
 - Store only positive cell values in their actual implementations
 - Tabular/matrix/spreadsheet layout is used mainly for conceptual clarity
 - All the values in a text mining spreadsheet are positive
 - Text mining programs use this characteristic to simplify processing

UET BOCKLEE NPTEL ONLINE CERTIFICATION COURSE 18

Now a few points are mentioned here again that you know for a large enough collection of documents the tabular layout the matrix layout that we have talked about you know that is going to be too sparse because on the column side we are going to have a number of you know words and you know those words are not going to be present in each of the documents represented by rows.

So in any column if we take any column there are very few cells which are going to have you know once and many cells are going to have 0s. So that can be now 0s can be you know when the whole sheet we look at as a whole the whole sheet would be filled with you know majority of you know 0s. So that is something that we refer as sparse. So we will have to process this kind of data.

So in the overall sheet and the layout that will have tabular layout that will have there would be few ones which would be indicating the presence of you know those words along the column side in the respective rows, but there would be a big number of large number of 0s indicating the absence. So we would be processing this data and if there are more number of attributes you know more number of words that are to be displayed along the column side.

Then this you know we will be dealing with in a large data set which is going to be too sparse. So the same thing is mentioned here any individual document will use only a tiny subset of the

differential set of words in a dictionary. So techniques so how do we process you know a text that is expected to be sparse. So one mechanism that many programs, many texts you know mining based program that they employ is that they store only positive cell values in their actual implementation.

So any particular you know cell if it is you know indicating the presence of the word, so that would be recorded in the actual implementation and the other you know cells they would be taken as you know their default value will be taken as 0 that means those words were absent absent for that you know particular document. So tabular and tabular or matrix spreadsheet layout is mainly being used for conceptual clarity while our discussion in this course.

And the various examples that we are going to discuss we are always talking about this tabular layout, but in the actual implementation of text mining programs they are storing the whole you know data in a different manner because that would be more optimized it would reduce computational time in terms of processing the data.

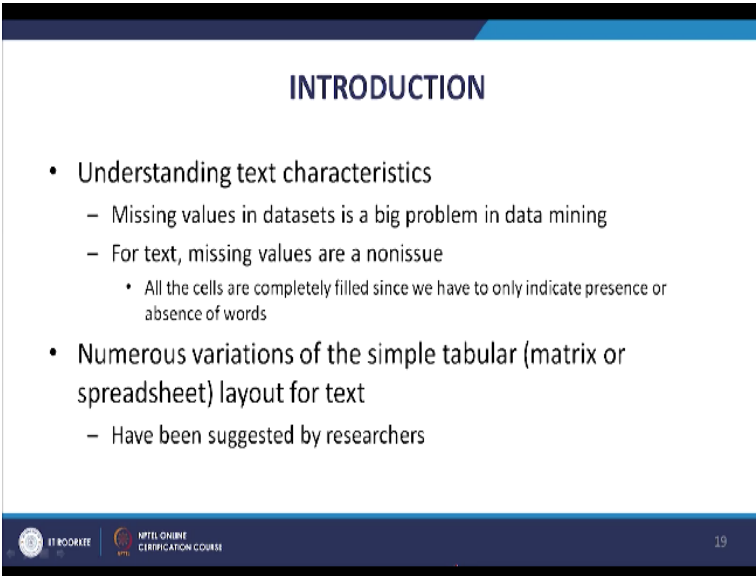
Another important point about you know text is that all the values in a text mining spreadsheets are positive. So if we refer back to you know what we have done in our data mining course is where we are going to deal with the real life variables. So variable like sales you know number of employees and you know many other variables they will have real values. So they will range many of these variables might have negative values, positive values and there would be a big range of values they would be taking.

However, in the case of these layouts that we have discussed for you know text mining there it would be filled with ones and zeros and you see you know mainly you know positive values are there. So these characteristics can be taken advantage of and we can simplify processing of text further. Now next point is about missing values, so missing values in a data set is a big problem in data mining.

So this aspect we have discussed in the previous courses where you know we talked about you know how missing values can be identified and how they can be tackled as well, we talked about

that in a particular column you know if one cell is having missing value we can replace it with the average value for that column. So all those aspects we have talked about, however missing value is a non-issue in a text mining scenario.

(Refer Slide Time: 08:04)



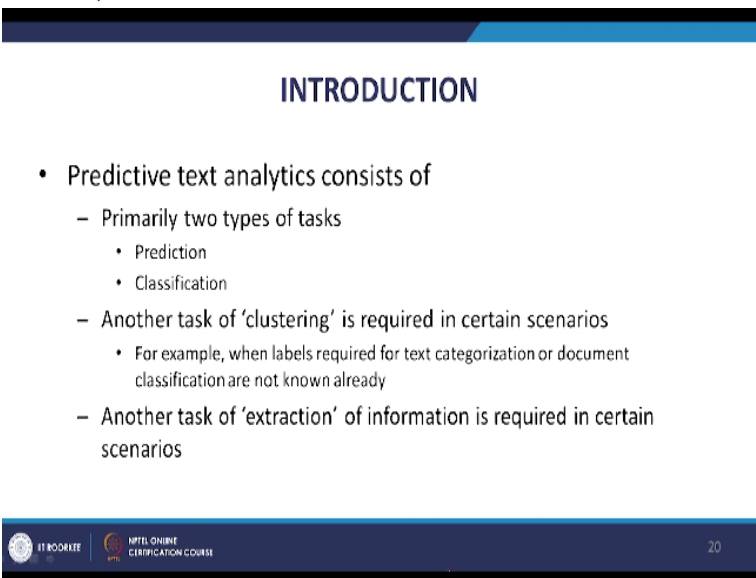
INTRODUCTION

- Understanding text characteristics
 - Missing values in datasets is a big problem in data mining
 - For text, missing values are a nonissue
 - All the cells are completely filled since we have to only indicate presence or absence of words
- Numerous variations of the simple tabular (matrix or spreadsheet) layout for text
 - Have been suggested by researchers

IT BOONKEE MPTEL ONLINE CERTIFICATION COURSE 19

Because what we are indicating is just the presence or absence of words, so typically we would not be dealing with missing values at all, so this is a non-issue, so this is another advantage in text mining. Now here in the previous lecture and in this one also we have talked about tabular you know presentation. This simple tabular presentation that we have talked about there are various you know there are various of variations of this format that are available.

(Refer Slide Time: 08:39)



INTRODUCTION

- Predictive text analytics consists of
 - Primarily two types of tasks
 - Prediction
 - Classification
 - Another task of 'clustering' is required in certain scenarios
 - For example, when labels required for text categorization or document classification are not known already
 - Another task of 'extraction' of information is required in certain scenarios

IT BOONKEE MPTEL ONLINE CERTIFICATION COURSE 20

So that will be discussing you know throughout this course depending on the technique where it is going to be used, so it would not just be you know later on as we will see in coming lectures that it would just not be the presence or absence of words in our tabular layout, it could even be the frequency of words. So, our any other or any other variants that have been suggested by these researchers.

So those aspects could also be there that we will be discussing, then let us come to the kind of types of tasks that we would be dealing with in this course. So, predictive text analytic analytics consists of primarily 2 types of task prediction and classification. So in prediction we would be predicting the value of something and in classification we would be predicting the class of something.

So these are the 2 types of tasks that we will be dealing with, but another task of clustering might be required in certain scenarios one example is given here also so clustering as we have discussed in data mining contexts that it comes under the unsupervised learning methods. So the same thing is here as well, so when we are talking about text categorization we got to know the categories where we are supposed to classify different you know documents.

So if those categories themselves are not known then how do we prevent classification models, so for that we might be using the this clustering task which will help us in terms of identifying those clusters, identifying those labels. So the same thing is mentioned here, another tasks that could be you know that we would be dealing with this extraction of information, so in certain scenarios you know we would like to extract certain information from a given document.

And that information could be connected with a real life, real valued variables and we might be populating a database using this you know particular process extraction process, will be discussing you know these aspects and more detail in this lecture when we talk about different types of problems that we typically solve in you know text mining. So let us start with that. So text mining problems.

(Refer Slide Time: 11:02)

INTRODUCTION

- Text mining problems
 - Document Classification or Text categorization
 - Business Problem: Folder/File Management (online or device)
 - Documents are organized into folders, one folder for each topic
 - Analytics Component:
 - » When a new document is presented,
 - » Task is to place this document in the appropriate folders
 - A binary classification problem
 - » A document can belong to multiple categories
 - Can be considered as a form of indexing
 - Examples: automatically forwarding e-mail to the appropriate team/department, detecting spam mail, future stock movements based on pre- or post-event news articles and financial data

Now going to discuss different types of text mining problems that are typically solved in this area. So our first one is document classification which is also sometimes referred as a text categorization. So in this as the name is suggesting the main idea is to you know classify you know our documents into one of the you know given categories. So the business problem in a generic sense is like folder and file management.

But this could be either online or it could be for your own computing device laptop or you know even mobile. So this is the main problem folder/file management, so as we know that documents are typically organized into folders and you know they might also be organized one folder for each topic. So that kind of organization that is typically we are looking for whether we are talked about the online or offline scenarios.

So what would be the analytics component out of this business problem, so whenever a new document is presented our task is our analytics task is going to place this document in the appropriate folders. So just like in the data mining modeling scenario we have talked about we train a model on you know a training partition and you know any new observation we have to predict either the value or class of that.

Similarly here we would be building models based on the previous information, previous documents and their you know known categorization and once the model is able to learn a few

patterns from there we would be required to classify a new document into the appropriate folder. So this is the main problem, now this problem is further broken down into a binary classification problem set up binary classification problems.

Because a given document can belong to more than one folder, more than one categories. So because we are dealing with text data so when the content is text and we are trying to learn from the words that are there, now different combination of words will indicate you know the categorization of that document to a you know different folder. If there are 2 categories like finance and office then if the more terms are related to you know finance folder, finance theme.

Then probably we will put you know that document in the finance folder, if more terms more words are related to the you know office theme then probably will place you know that document in that you know folder, but it might so happen that you know there are words which are indicating that particular document can be placed in both the folders. So therefore we have to treat this problem as a binary classification problem.

That we will see through an example later on as well, so this problem is also considered as a form of indexing. So as you can see whatever experience that you might have about indexing that you know you know even in your computer systems when you are searching for something so a number of results are displayed when you are searching for a file and folder because files with similar names they might be present in different folders in your computer system.

So all those results are displayed, so that is in a sense based on the indexing process that is implemented in operating systems. The similar kind of thing is what we are doing in this particular text mining you know problem where we are assigning new document to their appropriate folders. Few examples for this kind of you know text mining problem are automatically forwarding email to the appropriate you know team, department typically in organization different teams deal with different kind of work.

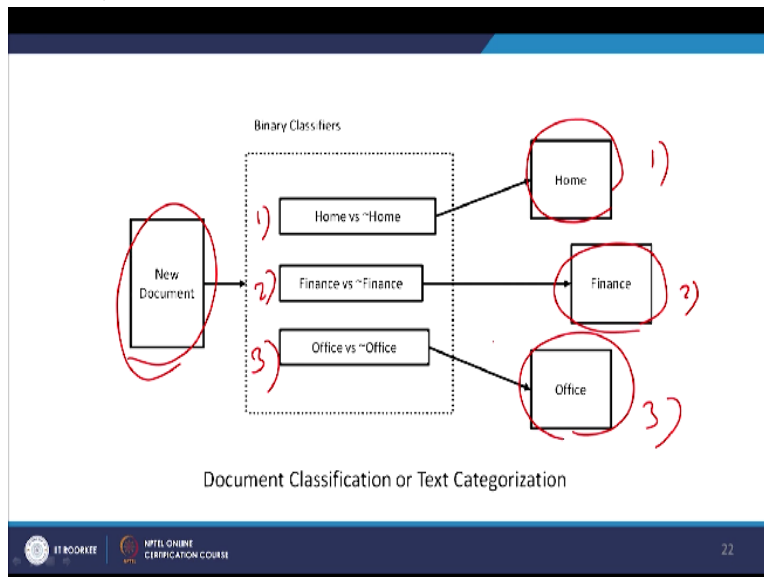
So when a new you know let us say in a complaint from the client side comes so which team is supposed to deal with that complaint. So whether based on you know certain, model whether we

can train that model based on previous instances and whether we can you know predict you know which team should be looking at that particular complaint. So that kind of you know classification can be really useful to improve efficiency in an organization.

Similarly detecting a spam mail future stock movements based on pre or post event news articles and financial data. So typically you know before any major event many people write about what is working well for the you know certain stocks what is not working well. So based on the information, based on the analysis presented and written by various experts in those news articles can we you know predict future stock movement.

And we can also club you know financial data documents as well, so this can also be done you know using this document classification problem.

(Refer Slide Time: 16:27)



Now let us look at the example so we talked about binary classifiers, now in this example we are trying to display exactly how this is how exactly how this is done. So this is example for document classification or text categorization problems you can see a new document is here now this document is to be you know classified into one of these you know categories home, finance and office.


And you can see we have created three binary classifiers here, one is home versus, one is belonging to home, not belonging into home, that we are representing using tilde finance versus

not belonging to finance office versus not office. So you can see the instead of having one you know general classifier for this document to be you know you know classified into one of these categories, one, two, three here again.

We are building three binary classifier where we are trying to find out whether given this document and the model you know the our trained model, based on that will score this document and find out whether it will belong to home or not find out whether it will belong to finance or not find out we will find out whether it will belong to office or not and you know appropriately that document is going to be classified.

So this captures that idea that you know one document can belong to multiple categories, with this let us move to the next data mining problem that we are going to that we typically solve in this area.

(Refer Slide Time: 18:11)



The slide is titled "INTRODUCTION" and contains a bulleted list of text mining problems. The list includes "Text mining problems" with a sub-item "Information Retrieval". Under "Information Retrieval", there are three main bullet points: "Business Problem: Document matcher (online or device)", "Similar to a search engine function", and "Based on measuring similarity as in nearest-neighbor methods". The "Business Problem" section includes sub-points: "Given a large collection of documents, finding relevant documents", "Analytics Component", and two nested points: "Task is to retrieve the relevant documents based on the best matches of input document with the collection of documents" and "New document is compared to all the other rows (documents), and the most similar rows and their associated documents are the answers". The "Similar to a search engine function" section includes sub-points: "A few words are presented, and these words are matched to others" and "Best matches are presented as the responses". The slide footer includes the MIT logo, "MIT BOOKS", "MIT ONLINE CERTIFICATION COURSE", and the number "23".

- Text mining problems
 - Information Retrieval
 - Business Problem: Document matcher (online or device)
 - Given a large collection of documents, finding relevant documents
 - Analytics Component
 - » Task is to retrieve the relevant documents based on the best matches of input document with the collection of documents
 - » New document is compared to all the other rows (documents), and the most similar rows and their associated documents are the answers
 - Similar to a search engine function
 - A few words are presented, and these words are matched to others
 - Best matches are presented as the responses
 - Based on measuring similarity as in nearest-neighbor methods

That is information retrieval, so what is the as the name is suggesting it is about you know matching a particular document with a set of other documents. So business problem is a document can be referred as document matcher, it could be online or you know in your device, in your computing devices. So we might have a large collection of documents and the idea is to find relevant documents.

This is something that we do using you know search boxes that we have in our own personal

systems. So if we break down the analytics component from this problem then our analytics task is going to be to retrieve the relevant documents based on the best matches of input document with the collection of documents that we have. So that this matching document matching is performed and the relevant results are to be presented.

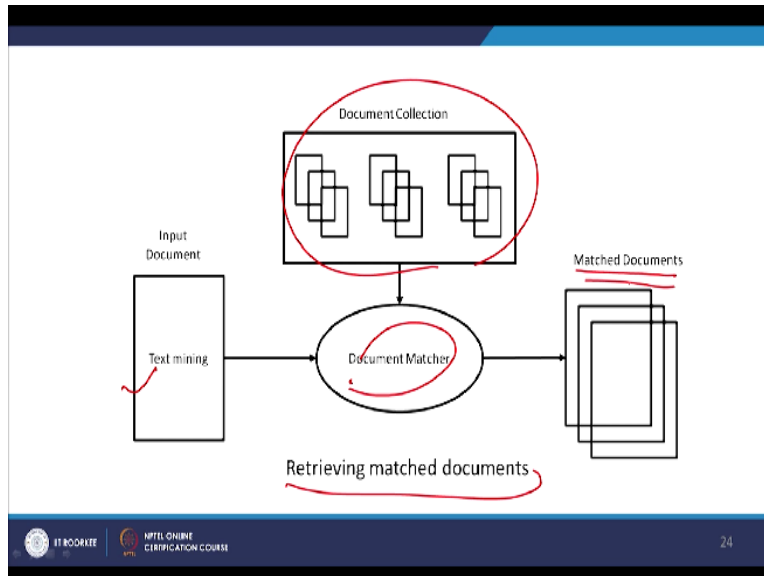
So it will another way to express the same idea is that new document is going to be compared to all other rows where all the rows are representing you know different documents and then most you know similar rows and their associated documents are going to be the you know answers for this. So business problem is going to be document matcher and based on the you know given input document we will be doing document you know similarity kind of thing.

So if we look at this kind of problem information retrieval problem it is similar to a search engine function for any popular search engine whether it is Google or you know or being or any other where typically we are you know typing our keywords whenever we are looking for certain information and based on the few words that are presented there, those words are going to be matched with the information that is there on web.

And best matches are typically presented as the responses, that is the same kind of thing you know we do in this kind of information retrieval tasks where based on the input document we have, we match it with the collection of documents and then relevant results are presented. So again how this is done as I said we typically measure similarity just like this is you know done in nearest neighborhood methods.

You can refer to our discussion on you know K-nn technique in our data mining course , there we are using similarity based measures to find out you know to solve our analytics problem data mining related analytics problem. Similar kind of thing is here also we are going to use you know we are going to measure similarity and some of the nearest neighborhood methods are going to be really helpful in this kind of problem.

(Refer Slide Time: 21:11)



So let us move on so to explain that this information retrieval type of problems you can see we can use this example, so it is about receiving matched documents so you can see we have input document, let us say the name of this document is text mining and we have this document matcher. So it will take the input document here and it will be matched with the document collection that we have here.

And then match documents the relevant document best matches are going to be presented, so this is the whole idea about these kind of problems.

(Refer Slide Time: 21:51)

INTRODUCTION

- Text mining problems
 - Clustering and Organizing Documents
 - Business Problem: Unknown document structure (online or device)
 - Given a collection of documents with no known structure, find a set of folders such that each folder holds similar documents
 - Analytics Component
 - » Task is to cluster the similar documents in the collection and assign labels to each cluster
 - Examples: learn about the categories and types of help-desk complaints
 - Might lead to identification of complaints which have no existing solution

Now let us move to the next problem, this is a clustering and organizing documents, so as we

have said that sometimes we might not be familiar with the labels or categories you know for a given collection of documents. So the document structure might be unknown for us so that becomes a business problem for us unknown document structure it could be online or device. So the whole scenario is we have a collection of documents with unknown structure.

And the idea is to find a set of folders such that each folder holds similar documents. So in a sense the idea is to create appropriate categories to create appropriate labels. So this is something that we can do in you know you know clustering which is an unsupervised learning method just like we did in the you know you know data mining course that we have done. So here also clustering can be useful.

So analytics component out of for this particular business problem can be expressed like this, so our analytics task is to cluster the similar documents in the collection and assign labels to each clusters. So just like you know we learn in our data mining courses that we are looking for you know distinct clusters, distinct you know group of points and then we try to understand you know so then we try to first thing is try to identify those clusters based on you know certain criteria it could be similarities could be some other variable.

And then characterize those clusters, so that characterization is essentially labeling of that you know those clusters, so that is something that is important. Now here in this particular you know task because that once this labeling is done then once labeled you know collection of documents can you know in future be used for in a document classification problem. So one example is also given here.

For example learning about the categories and types of helpless complaints, so any organization might have you know whether they are into you know whether they are selling product or whether they are into services. So they will have their own help desks they will be receiving complaints on a day-to-day basis on a daily basis and how you know different complaints are to be handled by different departments in the organization.

So it would be really helpful for them if they have a mechanism, if they have a model which can

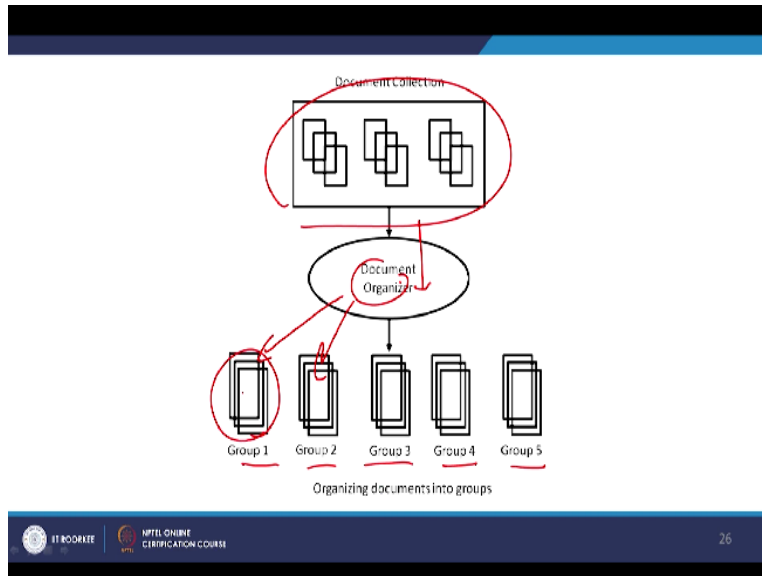
help them in terms of predicting, whether this you know complaint is belongs to the work of team 1 or to the work of team two or team three. So whether it is suppose this is related to IT department whether it is a networking problem, whether it is you know disk problem, whether this is a memory problem or any other issue.

Whether it is a hardware failure or software failure or you know incompatible software versions. So different you know groups might be different teams might be dealing with different types of problems. So whether we can learn about these categories, so that later on it could be really useful if we could you know build the prediction models for the same another advantage you know that might happen especially in the context of this example that we have given.

It might lead to identification of complaints which have no existing solution, so it might happen that the organization might have developed readymade solutions for different you know set of complaints. So this kind of categorization in a sense will also help us, because a new complaint which cannot be easily you know labeled as any of the existing you know categories that we might have.

And therefore will get a problem for which we might not have a solution. So identification of complaints which are having no existing solution through that scenario is also you know well covered by solving these problems. So let us move to you know before moving to the category let us explain the same thing through this graph.

(Refer Slide Time: 26:10)



So we have this document collection, now this is our document organizer here and the idea is to organize you know the whole collection into these groups, group one, group two, group three, group four and group five and each off for these groups are holding similar kind of documents. So that is the idea, now once this kind of labeling is done then later on the document classification problem that we have already discussed you know that can be solved.

(Refer Slide Time: 26:51)

INTRODUCTION

- Text mining problems
 - Information Extraction
 - Business Problem: Populating database from unstructured data
 - Given a collection of documents, automatically filling the relevant values associated with certain defined variables in a database
 - Analytics Component
 - » Task is to extract data from an unstructured format based on words which can be higher-level concepts or real-valued variables
 - » The variable that is being measured will not have a fixed position in the text and may not be described in the same way in different documents
 - Examples: extracting the sales volumes and industry codes from company documents

So let us move on, so the next text mining problem is information extraction, so we talked about little bit about this problem you know in this lecture also. So the business problem is populating database from unstructured data. So there are many business scenarios where the company might be having you know financial statements or other documents where they might have given

details about their sales volume you know industry codes, you know many other things that can identify that form or you know certain values to certain variables.

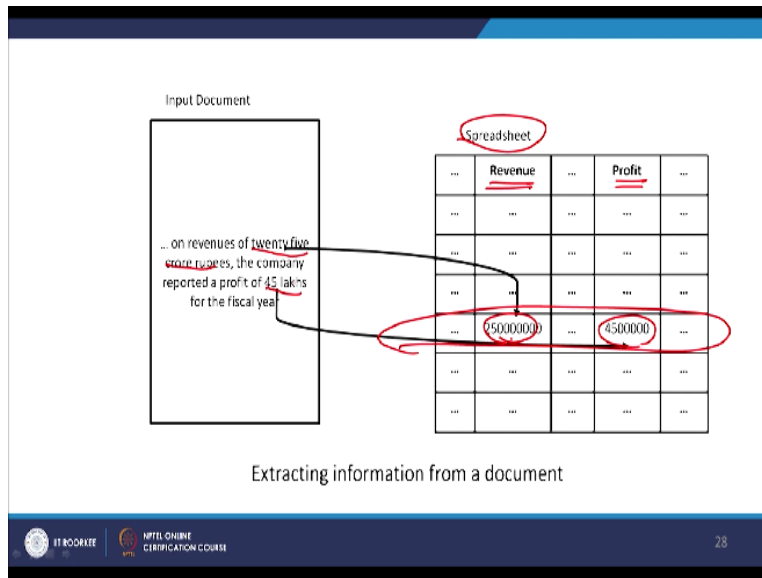
And their values you know related to that firm, so can we extract those values from the documents, so can we extract that information problem from the document and populate our variables in the database. So that is the whole idea, so we might have a given collection of documents and it is about automatically filling the relevant values associated with certain defined variables in a database.

So in our database we might have you know when we use industry code and etc. etc. and the idea is whether we can extract the relevant values and populate the you know store those values in the relevant row along that you know column. So the analytics component out of this business problem can be expressed like this. The analytics task is to extract data from an unstructured format based on words which can be higher level concepts or real valued variables.

So in the text document there would be a number of words there it would be consisting of a number of words and some of these words might be related to you know income or might be related to sales revenues various things. So can we extract you know data from there based on these words which might actually be real life or real valued variables. So another associated aspect that we need to understand is that variable that we are you know trying to measure here they might not have a fixed position in the document.

They might be anywhere, so the value might also be anywhere and they might be described in different ways. So that is the you know complexity that we might be dealing with in this problem, to give you an example extracting the sales volume and industry codes from company documents. So this could be one let us understand this you know using this figure.

(Refer Slide Time: 29:19)



So we have this input document and you can see a few you know lines here text lines here on the revenue twenty-five crores rupees the company reported a profit of forty-five lakhs for the fiscal year. So you can see that you know and then we have this spreadsheet where we have these variable revenue and profit. So this document you know might be along this row, might be associated with this row and whether we can extract the revenue number.

And the profit number, so you can see revenue number is written in words twenty-five crore rupees, so different format and the profit number is in numeric. So whether we can extract this information and you know fill our spreadsheet. So this task is about extracting this kind of in this fashion extracting the information and filling our database. Now few generic points about prediction and evaluation process that is performed in text mining.

(Refer Slide Time: 30:24)

INTRODUCTION

- Prediction and Evaluation
 - Text mining modeling process is similar to data mining modeling process
 - Process is about building models based on prior cases (from training partition)
 - Then the built model is used to predict the unseen cases (from test partition)
 - Evaluation of the model success is
 - Based on its performance on the test partition which is not part of the model building process
 - This mechanism works well for most of the text mining scenarios
 - However, there might be few special scenarios



So you know process text mining modeling process as such is similar to you know data mining modeling process that we have discussed in previous courses. So the process is about you know building models based on prior cases, so in these process prior cases are actually going to be forming our training partition. Then the model which is built on these prior cases using the training partition then this is used to predict the unseen cases from test partition.

Now how do we evaluate you know this the model success, so based on the performance on tests partition which is of course not part of the model building process, so we look at the performance numbers and how well the model is doing. So that is how the evaluation process is done. So if you look at this overall mechanism so this works for most of the text mining scenarios.

However, there might be few special scenarios, so that is something you know that also will discuss. So with this we would like to stop at this point and we will continue our discussion on prediction and evaluation in the next lecture, thank you.

Keywords: Structured Data, Unstructured Data, Text mining, Data mining, Classification, Categorization, Clustering, Partition, training etc.