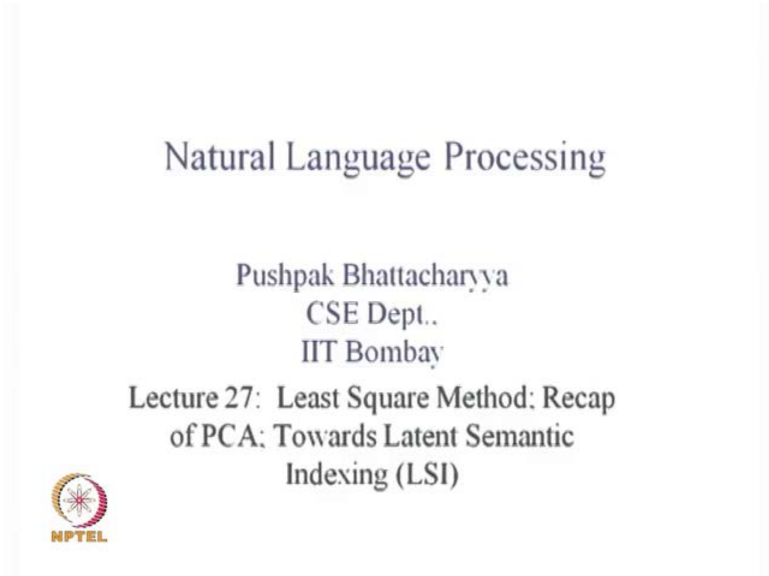
Natural Language processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 27
Least Square Method; Recap of PCA; Towards Latent Semantic Indexing (LSI)

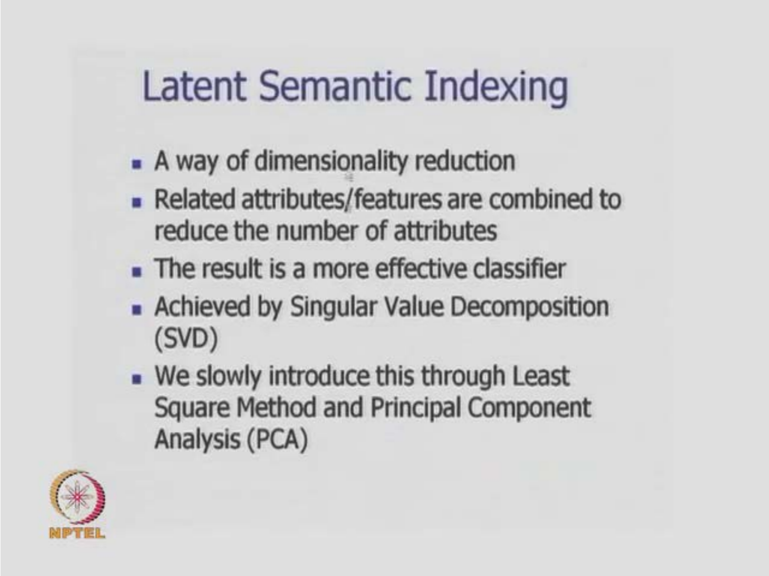
So in this lecture, we would like to discuss lexical semantic indexing or other latent semantic indexing. And this is a way of capturing co-occurrence amongst towards and there by doing more effective language processing and information retrieval tasks. So, as the title says, we would like to discuss least square method.

(Refer Slide Time: 00:47)



Then, we should like to have recap of the principal component analysis and then slowly move towards latent semantic indexing.

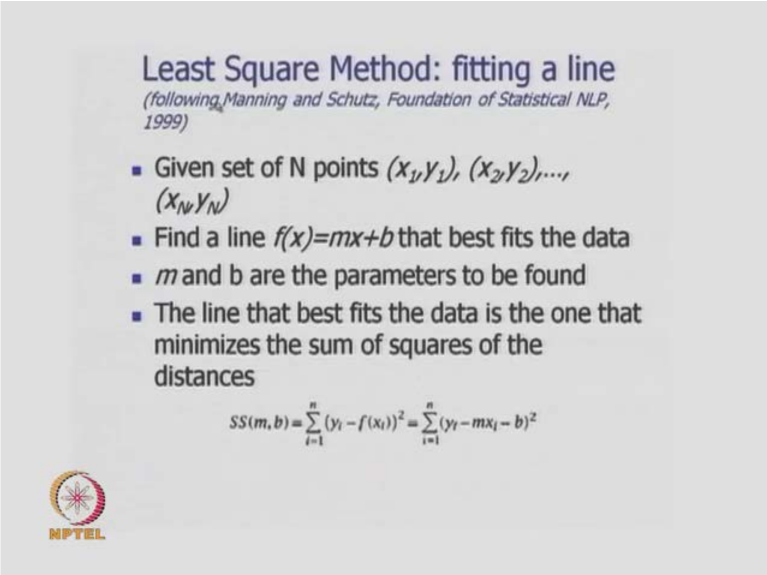
(Refer Slide Time: 00:58)



So, latent semantic indexing is a way of dimensionality reduction.So, this is a very important task, to find out which attributes actually are critical for a classification task, which features play a very important role and the idea is to take a combination of these attributes and features and thereby work with a smaller number of attributes and features.

We must note that, the attributes and features are not dropped. But, they are nearly combined to obtain a smaller number of changed attributes and feature. So, related attributes and features are combined, to reduce the number of attributes. The result is the more effective classifier. This is achieved by a technique which, we discussed called singular value decomposition and we will slowly introduce this, through the least square method and principal component analysis.

(Refer Slide Time: 02:08)

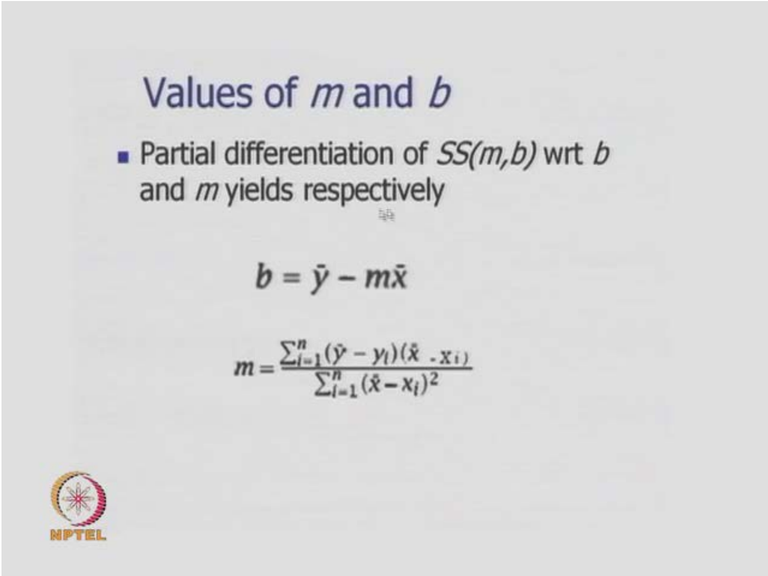


So, first we take a very simple problem. The problem of fitting a line, given a set of points and this is done by what is called the least square method. We follow here the discussion given in the Manning and Schutz foundation of statistical natural language processing, 1999. So, what is given is a set of n points X 1 Y 1, X2 Y2 up to X n Y n and our task is to find a line f X equal to m X plus b that best fits the data.

Now, why we would like to get this line and what does it have to do with dimensionality deduction, attributed deduction we will see very soon. But, we are saying that our task is to simply to fit a line, to a set of n points. Now, when we have these, this line f x equal to m x plus b what varies is this 2 values m and b.We try to find out m and b, such that the line best fits the data, and m and b are the parameters to be found. Now, what is the meaning of the line that best fits the data? The line that best fits the data is the one that minimizes the sum of squares of these distances.

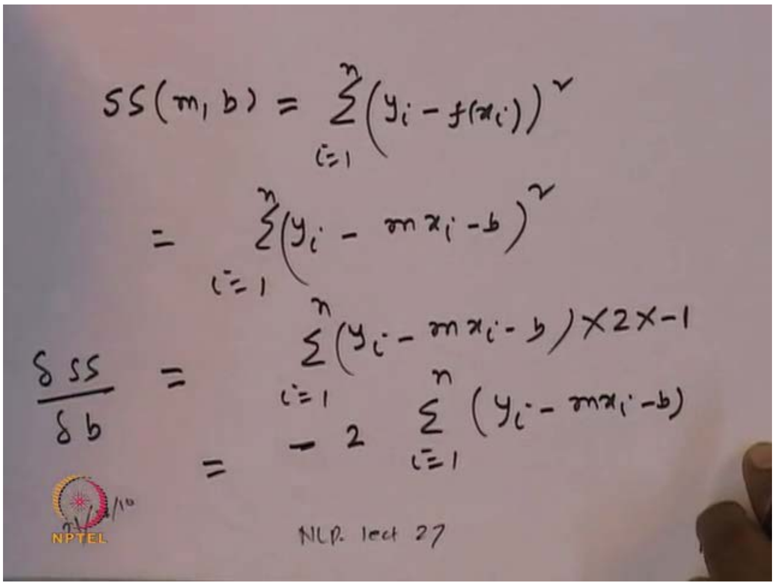
So, we, liked to find a line, such that the points which are given X 1 Y 1 up to X n and Y n; these point area least distance from the line. So, the sum square distance of each point from the line is taken and they are minimized. So, we can see that f X i is the value given by the line, f X i for a particular value of X f x i is the value given by the line. The actual value is Y ok, so Y minus f X i is the difference. If we take it square and sum it up then, there is no interference from change of sign. So, Y i minus f X i is a measure of error. The square of this error is summed up from 1 to n, for all these n points and the sum square distance is nothing but, Y i minus m Xi minus b whole square, i from going from 1 to n. So this is the quantity which should be minimized, so that the line best fits that data.

(Refer Slide Time: 05:15)



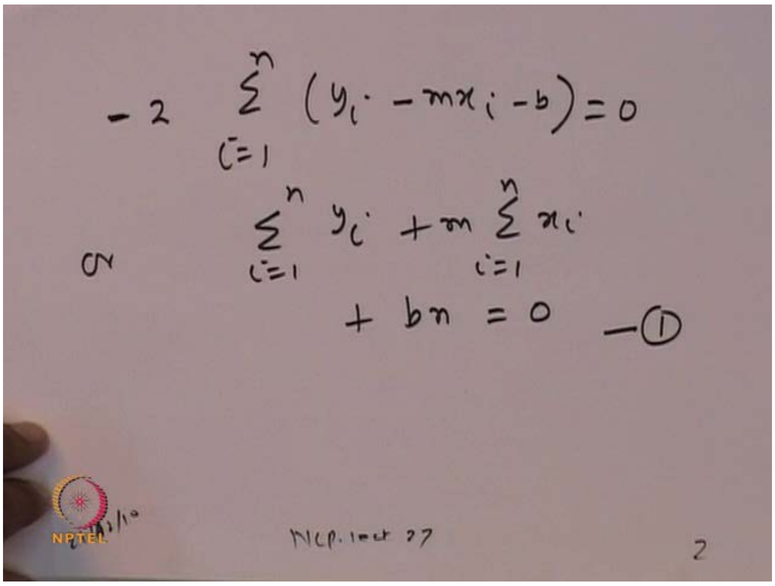
Now, how do we minimize? We take partial differentiation of these sum square distance with respect to b and m and if we do so, we get b equal to Y bar minus m X bar and m is this expression.So, for clarity let us do this derivation quickly.

(Refer Slide Time: 05:42)



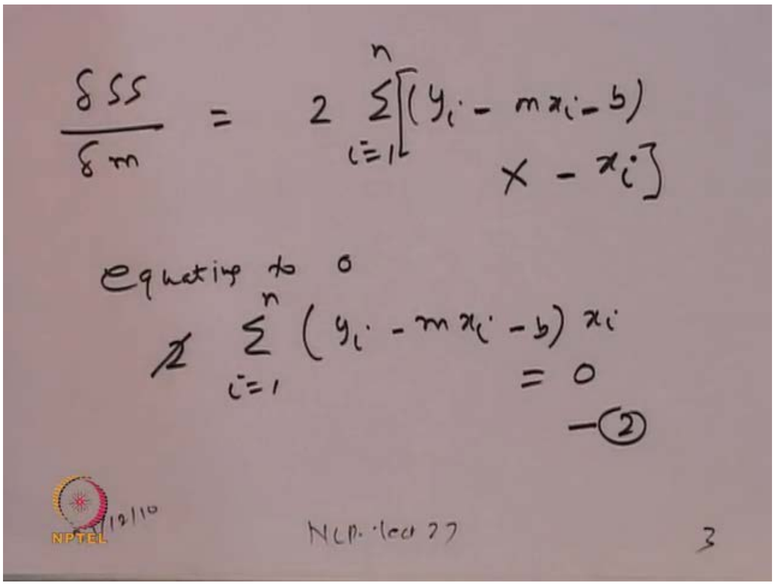
Now, we have S S m b is equal to sigma y i minus f x i square which is equal to y i minus m x i minus b square i equal to 1 to n. Now, delta S S with respect to delta b will be nothing but, sigma i equal to 1 to n, y i minus m x i minus b into 2 into a differentiation for minus b, which will be minus 1. So, this is equal to minus 2 into sigma i equal to 1 to n y i minus m x i minus b. So, this partial derivative should be equated to 0 for minimization.

(Refer Slide Time: 07:09)



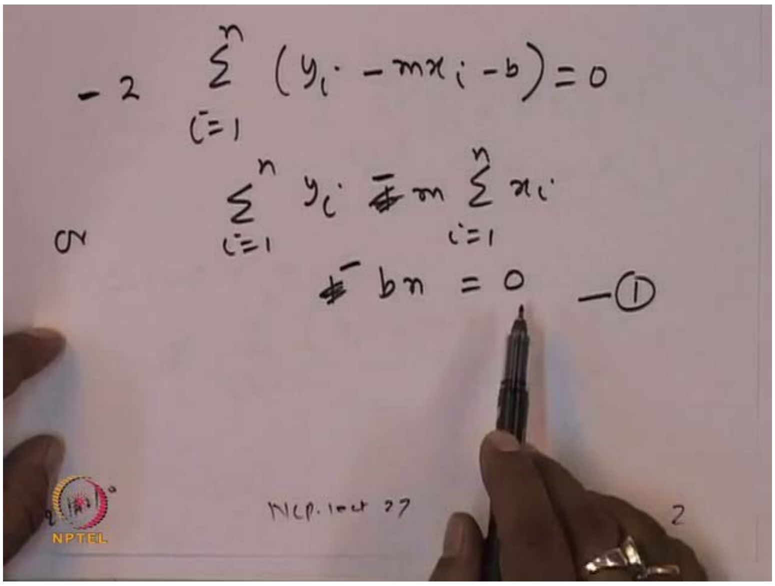
And therefore, we have minus 2 sigma i equal to 1to n, y i minus m x i minus b is equal to 0 or i equal to 1 to n, y i plus m x i i equal to n plus b n equal to 0. So, this is the 1st equation after taking partial derivation with respect to b.

(Refer Slide Time: 07:54)



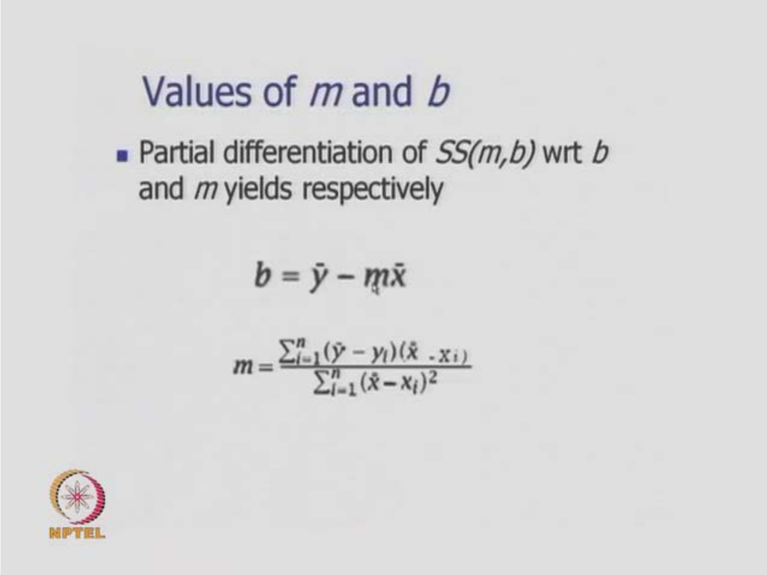
If I take partial derivative with respect to m, then delta S S delta m equal to 2 into i equal to n, y i minus m x i minus b into minus x i. So, equating to 0, we have i equal to 1 to n, y i minus m x i minus b into x i equal to 0. So, this is equation 2 so, equation 1 (Refer Slide Time: 07:09) was this sigma i equal to 1, 1 to n y i plus m sigma i equal to 1 to n, x x i plus b n equal 0.

(Refer Slide Time: 09:07)



Actually this will be minus, this will also be minus. So, this will be minus so, this is what it comes out to be as equation 1 and equation 2 is this.

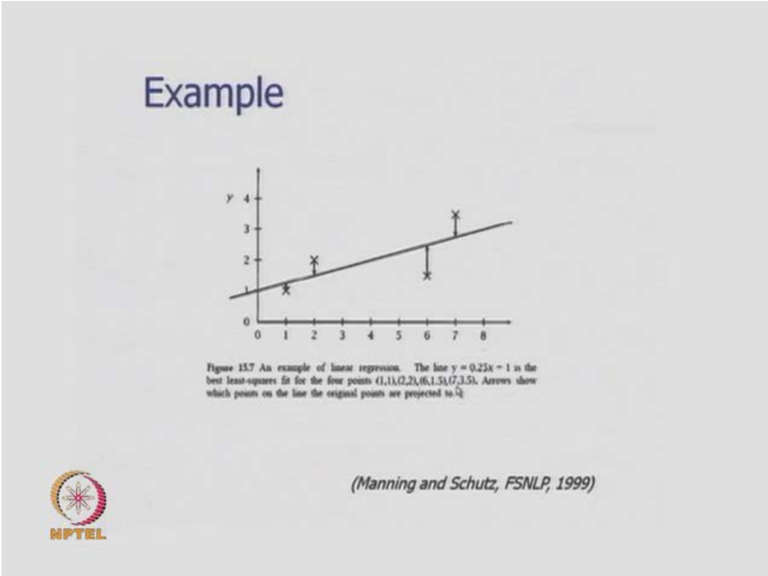
(Refer Slide Time: 09:26)



So, from these 2, we can get this expression which is b equal to y bar minus m x bar. So, b n was equal to sigma y i minus m sigma x i, if you divide by n then sigma y i by n is y bar sigma x i by n is x bar. So, b is equal to y bar minus m x bar and similarly, m can be processed and the value of b can put here and we obtained m to be equal to i equal to 1 to n y bar minus y i into x bar minus x i divided by i equal to 1 to n x bar minus x i whole square.

So, what is this quantity, this quantity is nothing but, the variance and this quantity is something like the covariance. So, this is like the covariance and this is the variance. So, this captures the difference, from the mean and this capture the difference from the mean for individual variables and takes their product and b is equal to y bar minus m x bar. So, given the data we can very easily find out m and having found m, we can find out the value of b because the mean of the values are known from the data ok.

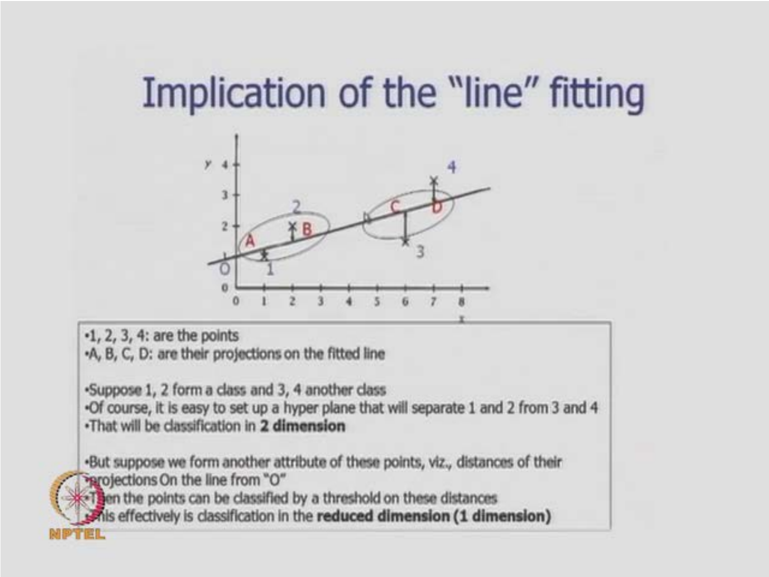
(Refer Slide Time: 11:11)



So, we take an example here, these points are given 1 1, 2 2, 6 1.5 and 7 3.5. This 4 points are given so, this is 1 1, this is 2 2, this is 6 and 1.5 and this is 7 and 3.5. So, these 4 points are given and we would like to find a line which best fits the data. So, by doing the analysis which we have shown before. We have, we can find out the individual mean of the x coordinate and the y coordinates. From the individual means, we can find out the m value as covariance and variance and then we can find out b.

So, the line y equal to 0.25 x plus1 is the best least square fit for the 4 points,which can be very easily find out. So this is the line which slope 0.25 and intersect on y axis as 1ok, so this line is found out. Now, the question n is as far as machine learning is concerned, dimension reduction is concerned, what does it bring to the plate? What do we gain by doing this least square fit?

(Refer Slide Time: 12:29)



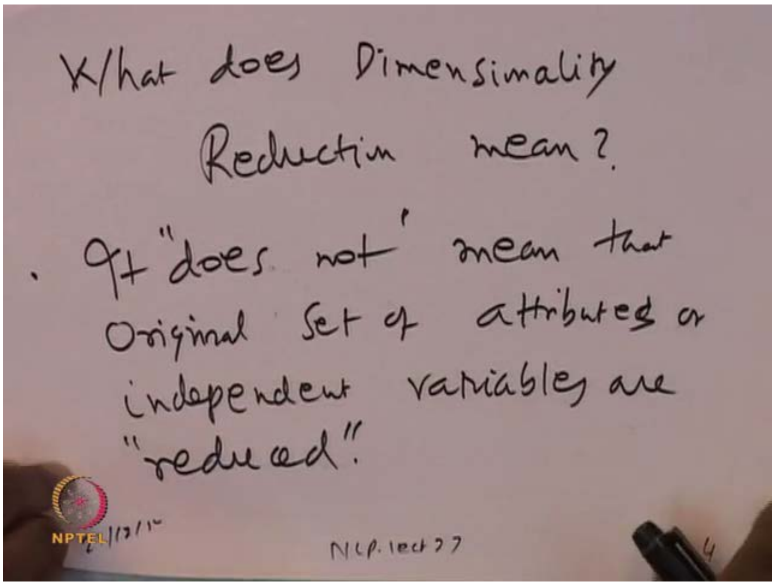
The next slide shows this the implication of the line fitting. So, if you look at this diagram, then this 4 points are 1,2,3 and 4. These are the 4 points and their projections on this best fitting line are A, B, C and D. So, these are the projections and this is the line. So, what we will gain from this fromclassification point of view? Now, imagine that these points 1 and 2 belong to one class, call it the positive class and 3 and 4, these 2 point belong to another class; call it the negative class. So, these 2 belong to 2 classes. So, there is no problem in setting up a hyper plane or a line in this case since, it is 2 dimensional data. One can put a line this way and separate the 4 points, 2 points in 1 class the other 2 points in the second class. Now, this decision is based on 2 attributes namely the x and y values of A B C D.

So, all the 4 attributes are involved in this classification. Suppose, we do it in a different way, what we say is that, we would do the same task in one dimension. So, what we will do is that, we will look at the projection of these points and we will form another attribute from these points. We will call this attributes, the distances of their projection on the line from O ok, this is the point O and the projection of 1 is A on the line. This distance of A from O is the attribute considered, this distance of B from O is the attribute consider; similarly, for C and D. Now, each point instead of having 2 point 2 specific pieces of information ok, have one specific information. So, which is a particular constricted information, which is that there is a point O which is special for this and distance for this A from this point, from B from this pointof C and D from this point. So, see how that dimensionality of the problem has been reduced ok.

So, right now the decision was with respect to 2 attributes of each point, after we decide that new attribute is distance from O of the projection of this line. We are dealing with only one pieces of information. There is only one dimension which is to be considered now and this is the heart of the problem; it is the heart of the point discussion. Now from multiple attributes we obtain a reduced set of attributes and based on decision on that.

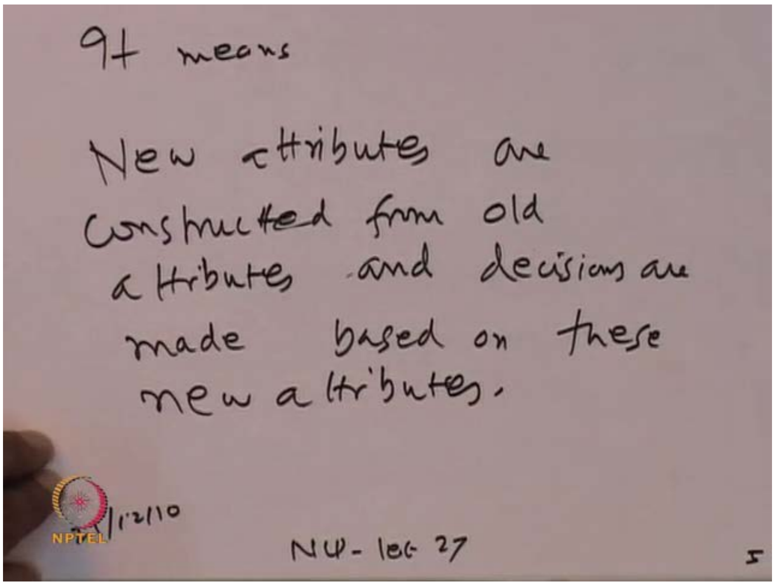
So now, when we talk about dimensionality reduction that is a misunderstanding. The misunderstanding is this, people think that we reduce the number of attributes. We do not reduce the number of attributes. What we do is that, weconstruct a new set of attributes.Based on the given set of attributes, just like from the attribute x and y we obtained a new attribute, called the distance from O of the projection.

(Refer Slide Time: 16:19)



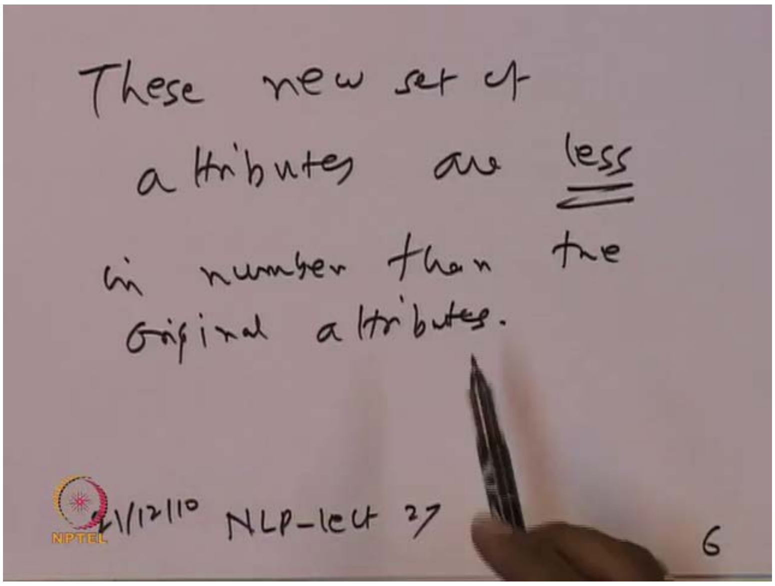
So, let us remember this point and let me write it down. What does dimensionality reduction mean? It does not mean that original set of attributes or independent variables are reduced. No, it does not mean that; it does not mean that original set of attributes or independent variables are reduced.

(Refer Slide Time: 16:56)



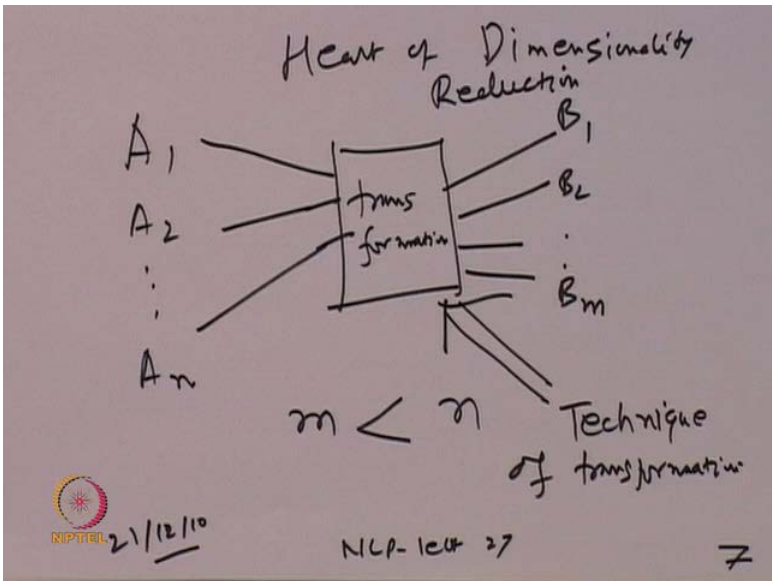
What it means? New attributes are constructed from old attributes and decisions are made, based on these new attributes ok.

(Refer Slide Time: 17:28)



More importantly, these new set of attributes are less in number, than the original attributes. So, this is the fundamental of dimensionality reduction and is at the heart of latent semantic indexing. And let me repeat this that, the original set of attributes transform into another set of attributes, which are functions of this original attributes.

(Refer Slide Time: 18:17)



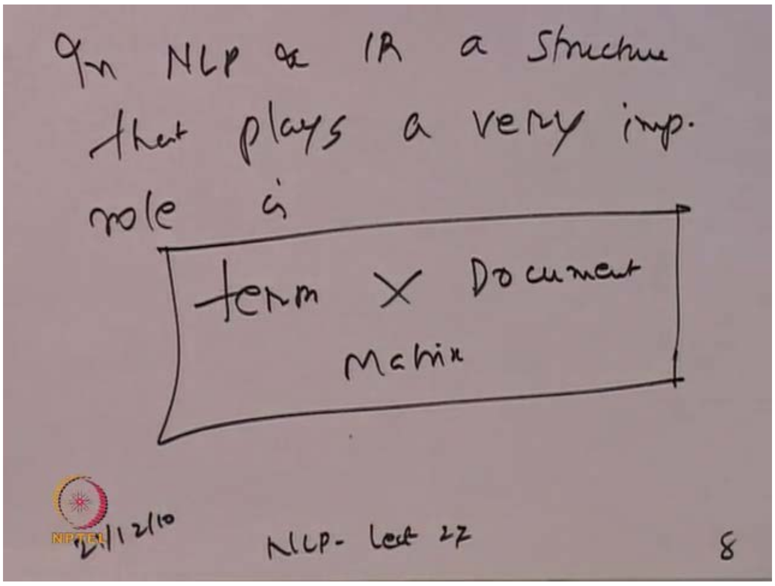
So, diagrammatically speaking, you have A 1, A2 up to A n. So, all these go through a transformation and produce B 1, B 2 up to B m, m is less than n and this is the heart of dimensionality reduction ok, fine. So, if this point is understood, let us look at this very important diagram again. We have set of attributes here, going through a transformation giving rise to another set of attributes B 1, B 2 to B m, m has to be less than n; otherwise there is no dimensionality reduction.

Now, comes the question of, what this box does, the technique of transformation? What is this technique of transformation? This is the question and there are different techniques. So, looking at the slide again, what was the technique of transformation here? The technique of transformation was, we took this point projected them to this line and the distance of the projection from this point O was found out. How was it found out? It was found outfrom the original attributes. Original attributes x and y, after projection they became something and those value then, gave rise to this distance attribute ok. So, this was one technique.

(Refer Slide Time: 20:19)

We go to another very well-known technique call the principal components analysis. Now, the difference between principal component analysis and singular value decomposition is that, both of them are for dimensionality reduction. The number of attributes get transformed, into another set of attributes and this new set of attributes is less in number. Now, principle component analysis is applicable only to square matrices ok, where as singular value decomposition is applicable to any general matrix.

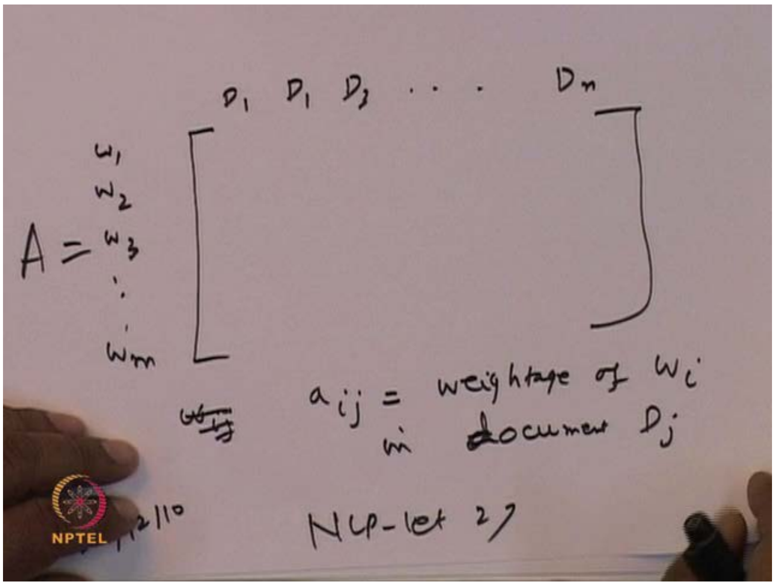
(Refer Slide Time: 21:04)



Now, what is the idea behind this? Let us first motivate this particular problem.

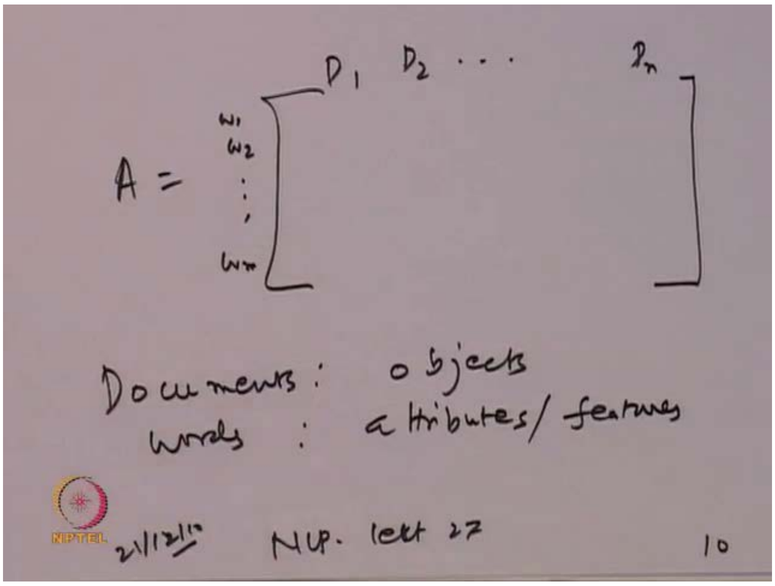
Now, what is the idea behind this? Let us first motivate this particular problem. So, in N L P and I R a structure that plays a very important role is term document matrix. So, this is a structure which plays a very important role.

(Refer Slide Time: 21:27)



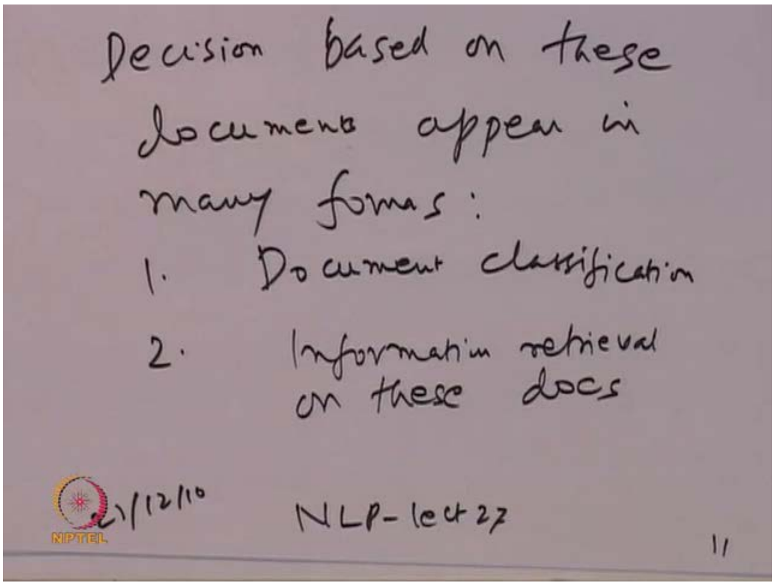
So, this term document matrix has the following shape. You have this document D 1, D2, D3 up to D n. And, there are words in the vocabulary of the language word 1, word 2, word 3 up to word m. So, any w i j, let us call this matrix A. So, any element a i j is equal to weightage of word i, in document d j. So, the weightage of the word w i in document d j. Now, if the document, if this matrix is already expressed in term presence or absence then, it will be filled with 1’s or 0’s. Otherwise there are different schemes for giving to weightage towards with respect to the document so, this we can see later.

(Refer Slide Time: 22:34)



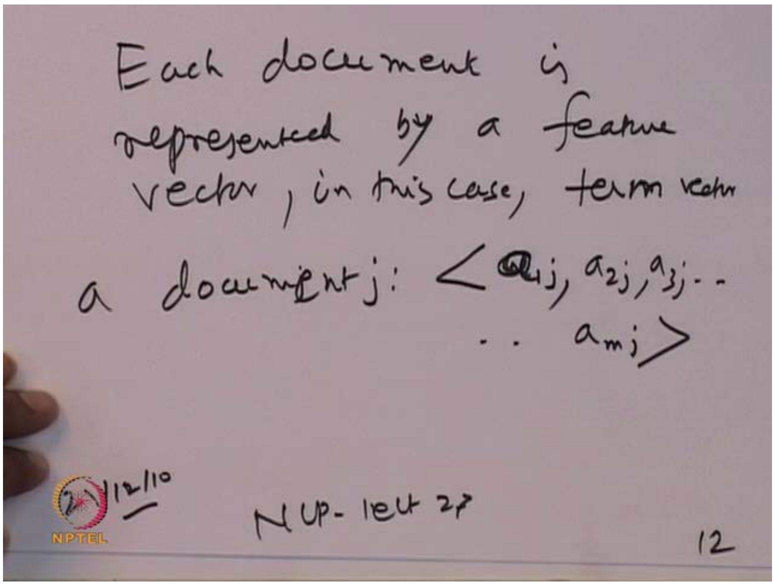
Now, this A matrix which is w 1, w 2 up to w m with columns D 1, D 2 up to D n. What has dimensionality reduction got to do with this particular matrix? So, in this matrix the objects are the documents D 1, D 2 up to D n and their attributes and feature these words w 1, w 2 up to w m. These words are the features and the documents are the objectives. Now so, we note this down, document are objects, words are attributes or features.

(Refer Slide Time: 23:27).



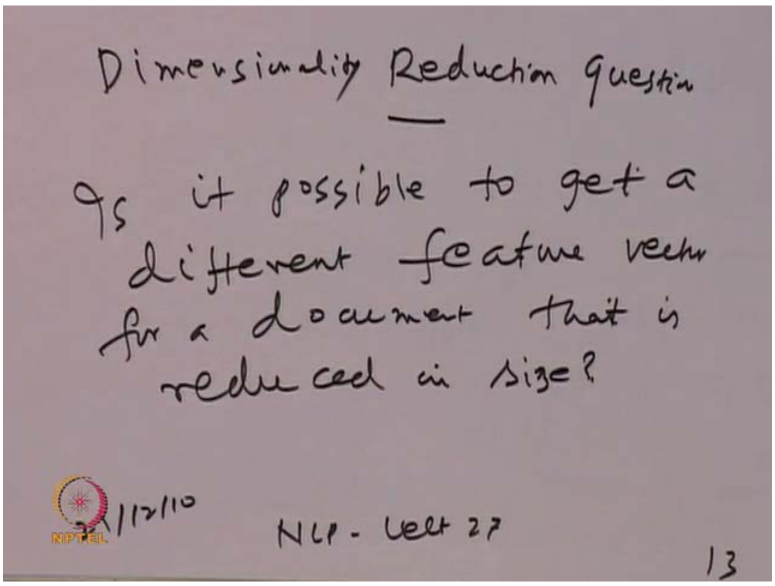
So, decision based on these documents, appear in many form; so, decision based on these document appear in many form. One could be document classification. Two could be information retrieval on these docs. So, these documents may be ranked after information retrieval and this ranking and scoring depends on these documents properties. So, classification and retrieval happen to be important tasks on this document.

(Refer Slide Time: 24:24)



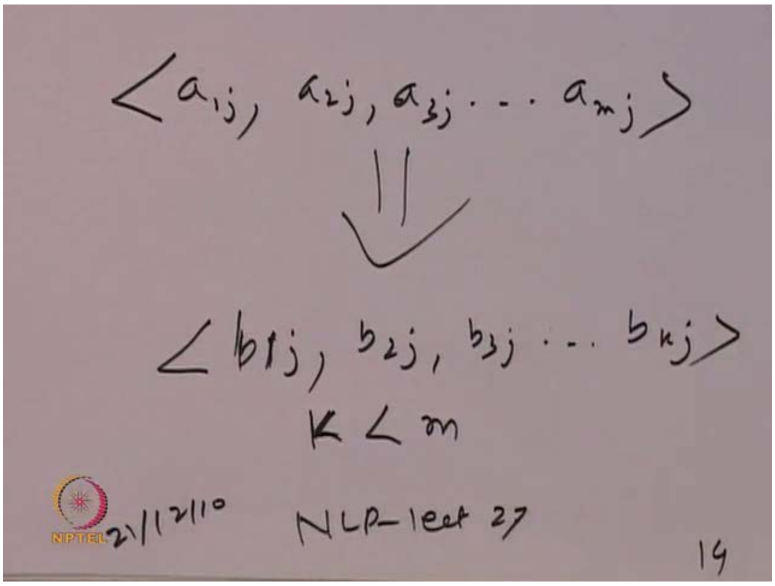
Now, each documentis represented by a feature vector and in this case the term vector. The document j is represented document a 1 j, a 2 j, a 3 j up to a m j. So, the weightage of these terms w 1, w 2, w 3 up to w m in the document d j. So, each document is became a feature vector and a decision is made based on this feature vector.

(Refer Slide Time: 25:10).



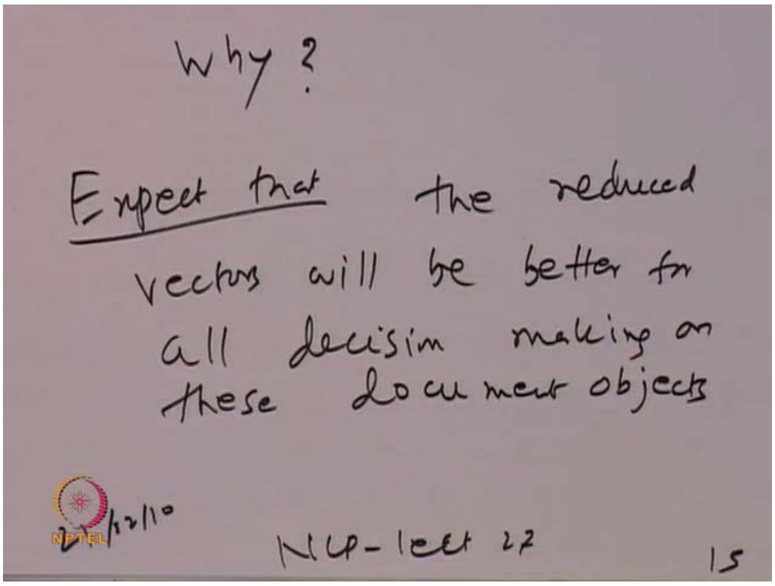
Now, the dimensionality reduction question is. Is it possible to get adifferent feature vector for a document that is reduced in size? That is reduced in size, this is the question. Is it possible to get a different feature vector for a document that is reduced in size?

(Refer Slide Time: 25:44).



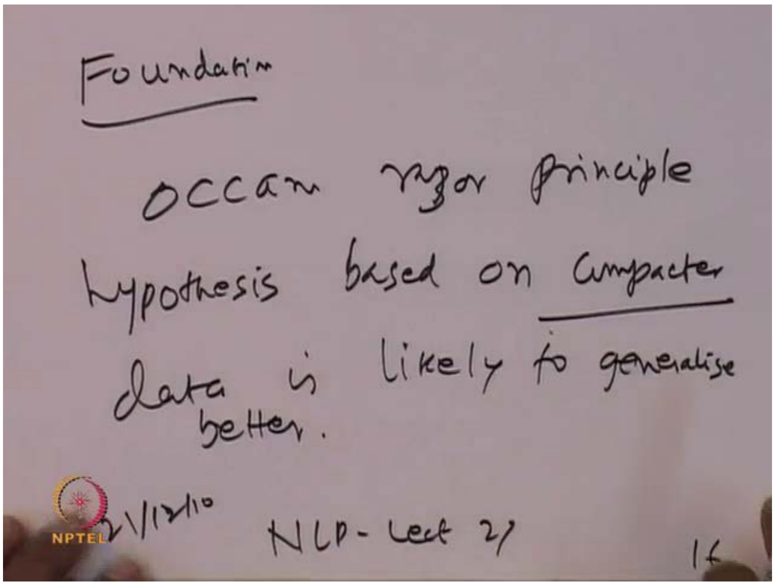
So this means, what it means is, that the document which has a 1 j, a 2 j, a 3 j up to a m j, can it be reduced to a reduce document b, b 1 j, b 2 j, b 3 j up to b k j where, k is less than m. So, this is the formulation of problem, can we reduce this document vector to a reduced vector going from the space of a’s to b’s? This is the dimensionality reduction question for the document.The question that actually that naturally arise is why?

(Refer Slide Time: 26:29).



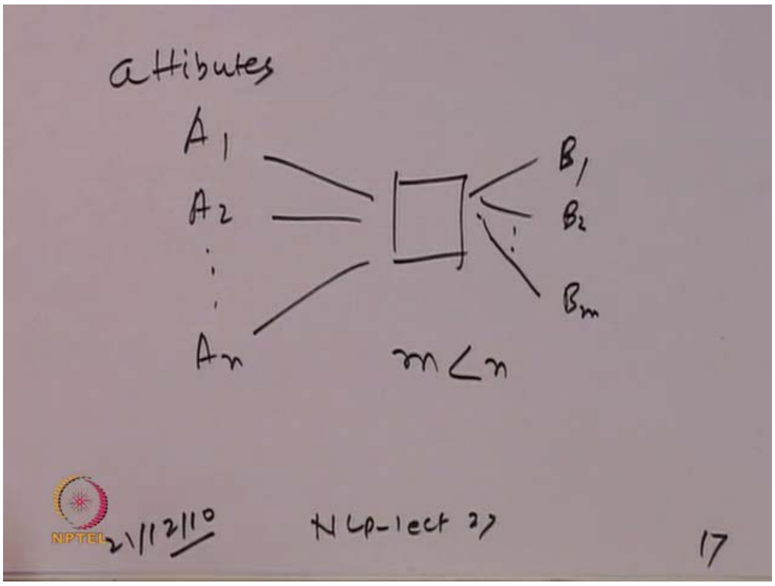
So, expect that the reduced vectors, reduced vectors will be better for all decision making on these document objects. We expect that the reduced vector will be better for decision making on these document objects ok; this is the purpose of dimensionality reduction.

(Refer Slide Time: 27:01).



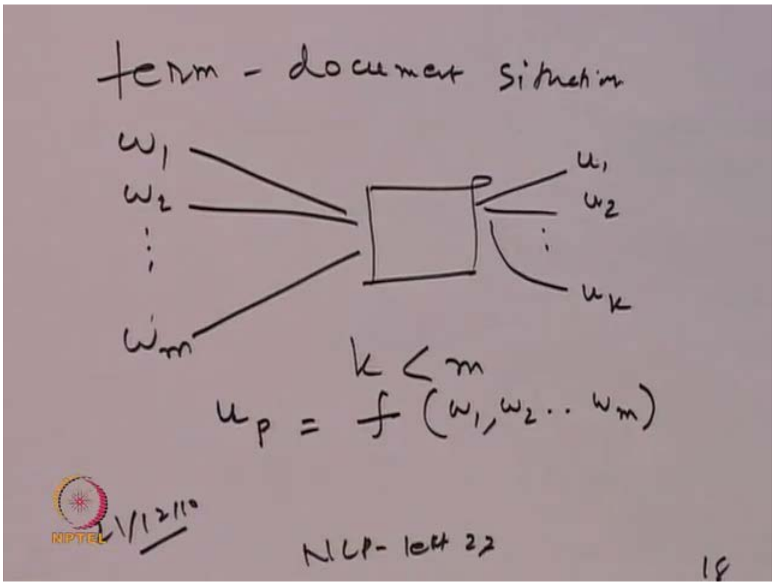
And the foundation is Occam razor principle ok. Hypothesis based on compacter data, is likely to generalize better. So, the hypothesis which is based on compacter representation data is likely to generalize better.

(Refer Slide Time: 27:37)



Now, what did you understand from our discussion, that the attributes A 1, A2 up to A n give arise to B 1, B 2 up to B m, m less than n. Now, in case of term document matrix. So, this is the picture for dimensionality reduction from set of attributes A 1, A 2 up to A n. We go to B 1, B 2 up to B m, m is less than n.

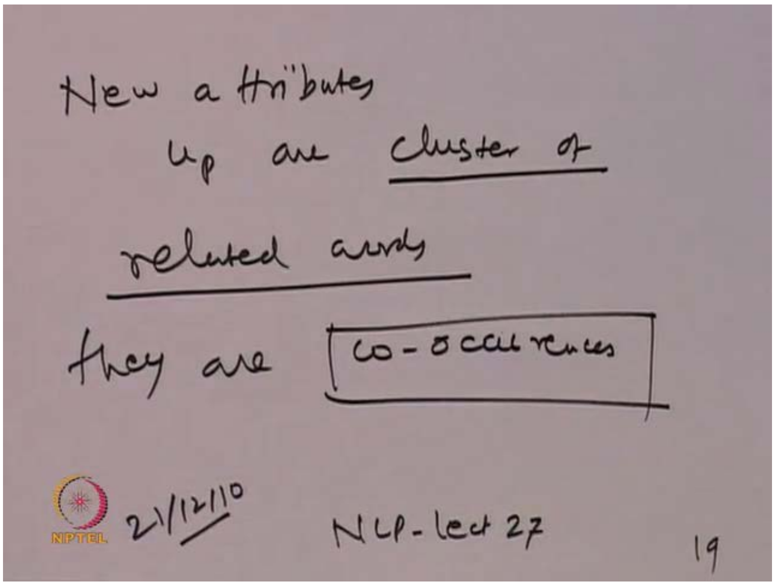
(Refer Slide Time: 28:20)



Now, for term document situation the words w 1, w 2 up to w m, after transformation. After dimensionality reduction give rise to entities u 1, u 2 up to u k ok u k, where k is less than m and this u 1, u 2 up to u k. So, each u p is a function of w 1, w 2up to w m.

So, each attribute u p is constructed from the terms w 1, w 2 up to w m. Now, the question is what is u p? What are this u 1, u 2 up to u k? So, that is the interesting question in language processing scenario ok. For other problems, all this new attributes which are constructed from old attributes have different meanings. But, in natural language processingsituation where the attributes are words and the new set of attributes have been obtained.

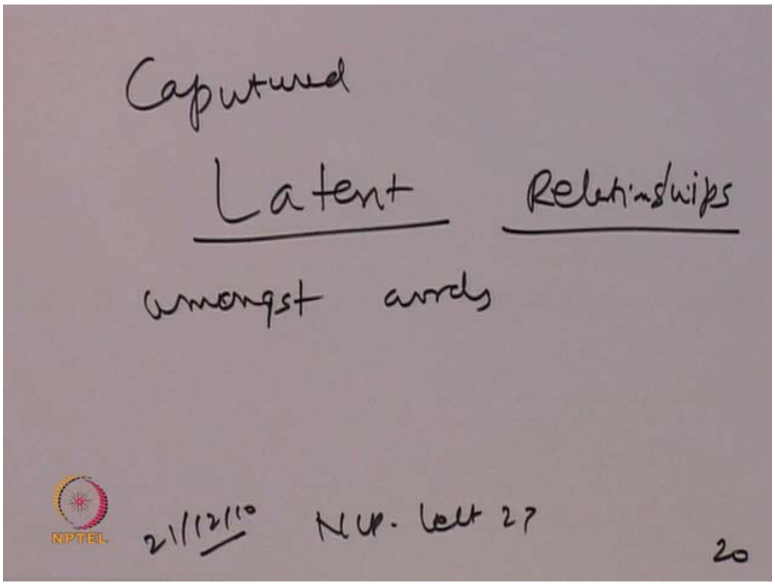
(Refer Slide Time: 29:42)



These, new attributes, new attributes u p are cluster of related words ok. So, this is the heart of dimensionality reduction again in natural language processing ok. This is the cluster of related words. In particular, they are co-occurrences so; we have hit up on a very important concept in Natural Language Processing. Dimensional reduction has been applied; the dimensionality reduction has been applied. We have got this words occurring in various document. We have applied dimensional reduction and what we have obtained is a set of new attributes. What this new attributes are, very strongly co-occurring words.

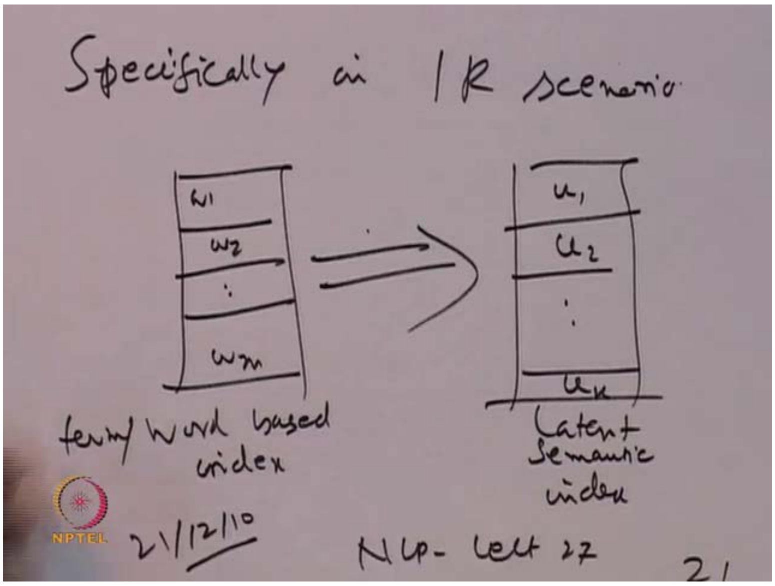
For example, if this words are Astronaut, cosmonaut, car, vehicle etcetera. Then we will see that related words like cosmonaut and astronaut they will be group together in terms of let say u 1, car, vehicle, auto, automobile they will be grouped together let us say u 2 with their specific weightages. And thus, we have after dimensionality reduction, a set of cluster of words and these cluster of words are better for decision making on the doc. So, that is the point. Now, instead of decision making on individual words, we capture this relationship between words and this related words are used for decision making.

(Refer Slide Time: 31:47)



So, we say that we have captured the Latent Relationships, we captured the latent relationships amongst words ok. That is how the term latent comes. They were latent in the words; latent relationships are captured.

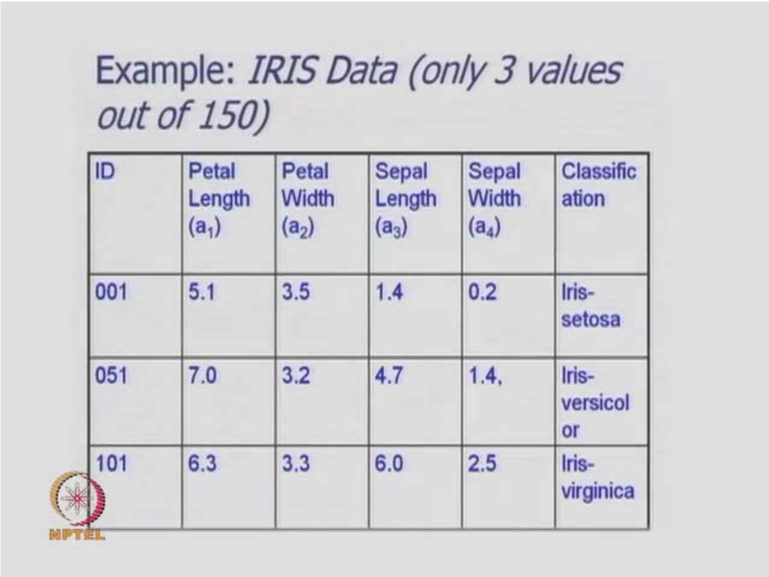
(Refer Slide Time: 32:05)



And in specific specifically in I R scenario,in scenario instead of this index table which is w 1, w 2 up to w m, we will have another index table which is u 1, u 2 up to u k. So, this is word based, this is term or word based index. So, this is on the other hand latent, this is latent index. But, this also capture some semantic in the form of word relationships, that is why it is called latent semantic index. So, this is the important of the word association, word occurrence giving rise to latent semantic index. So, instead of retrieving document based on this, we retrieved document based on this, and there by become more accurate.

So, this was in the context of natural language processing information retrieval. Now, we get an idea about what the exact mathematical techniques are and this facilitates by looking into the details of this important topic called (Refer Slide Time: 20:19) principle component analysis; P C A.

(Refer Slide Time: 33:33)



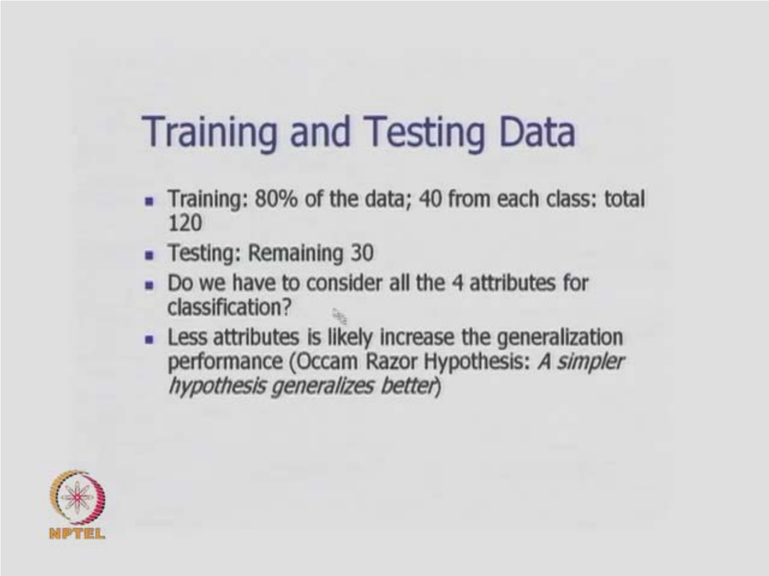
Example: *IRIS Data (only 3 values out of 150)*

ID	Petal Length (a_1)	Petal Width (a_2)	Sepal Length (a_3)	Sepal Width (a_4)	Classification
001	5.1	3.5	1.4	0.2	Iris-setosa
051	7.0	3.2	4.7	1.4,	Iris-versicolor
101	6.3	3.3	6.0	2.5	Iris-virginica

As remarked before, this is for square matrix. We take an example of a very famous machine learning problem known as the IRIS data and we show here only 3 out of the 150 data points. Now, there are 3 classes called iris setosa, iris versicolorand iris virginica. These are 3 classes of flowers and the idea is to classify a particular flower into one of these 3 classes. The deciding attributes of such classification problem are petal length, petal width, sepal length, sepal width. So,for example, the flowers with i d 001has a petal length of 5.1 units, petal width of 3.5 units, sepal length of 1.4 units, sepal width of 0.2 units. Then and it is classified as iris setosa.

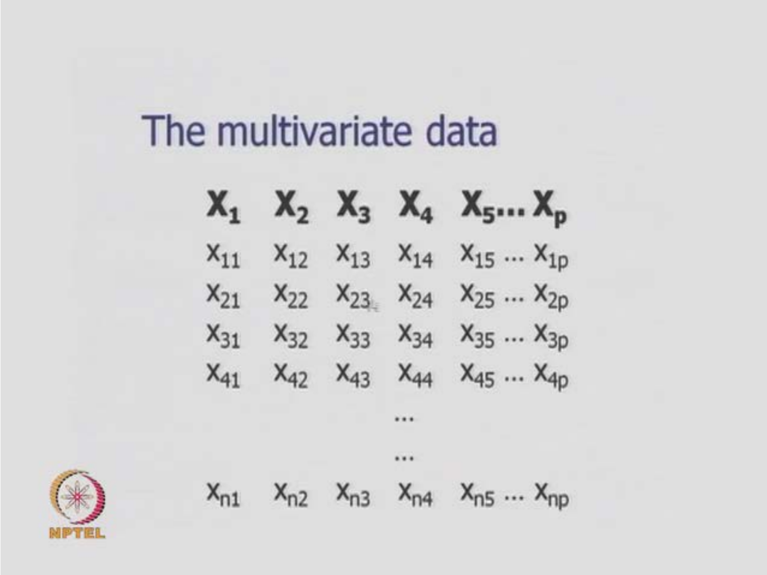
So, this way flowers with 150 i d’s is 001 to 150 are given with this values of 4 attributes. Now, this data is a very famous data and has served as a bench mark data for any machine learning problem. Whenever anybodydesign a new machine learningalgorithm, one of the important test for the algorithm is performance on the iris data. So, what is done, is that all this 150 data items are divided into the training data and testing data.

(Refer Slide Time: 35:18)



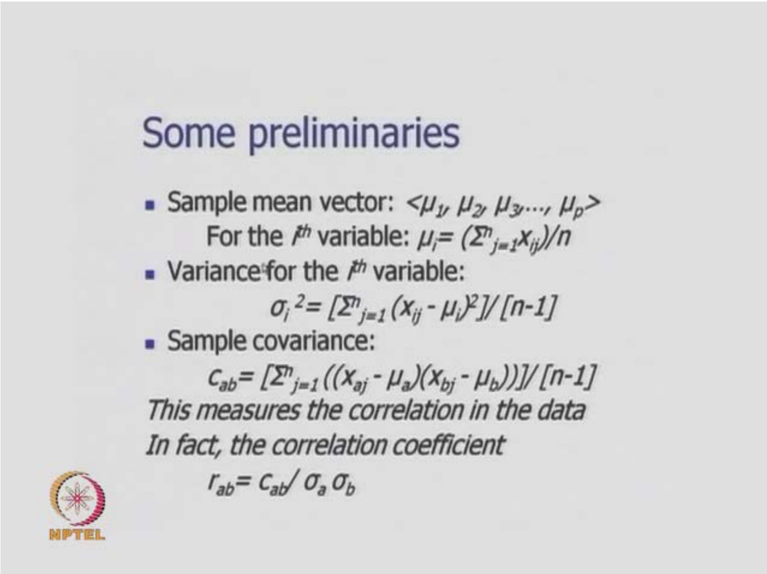
So, typically 80 percent of the data is used for training, 40 items from each class are taken for training. So, that makes totally 120 training points, testing is on remaining 30 data items. So, the question we ask is do we have to consider all the 4 attributes for classification? May be the attributes can be combined in some interesting way and then decision can be taken. So, less attributes is likely to increase the generalization performance. This is the Occam razor hypothesis ok.

(Refer Slide Time: 35:51)



Now, we discuss how to deal with multivariate data in the context of principle component analysis. So, we have this attributes X 1, X 2, X 3, X 4, X 5 up to X p and there are these n data points with value is X 1 1, X 1 2 , X 1 3 up to X 1 p, X 2 p. X 2 1 up to X 2 p and so on. So, each objects has these values under individual attributes.

(Refer Slide Time: 36:27)



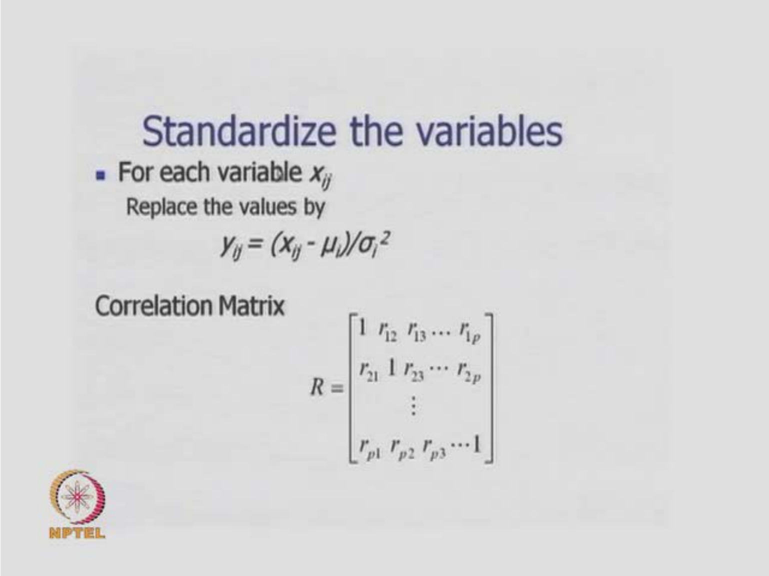
Now, some preliminaries are useful here in understanding the technique. We get the sample mean vector which is mu 1, mu 2, mu 3 up to mu p. For the i-th variable mu i is nothing but, sigma j equal to 1 to n, x i j divided by n ok. So, for i-th variable this is the mean. So, what this means is that, for these particular attributes let us say x 3, we will sum up all these values and divide by n to get the mean for the attribute x 3, which will be mu 3.

So, mu 1, mu 2 mu 3 up to mu p are obtained for each of this attributes. The variance is also obtained for each attributes sigma i square is nothing but, the departure from the mean. So, x i j is subtracted with mu i; the attribute mean the square is taken and summed up from j equal to 1 to n and is divided by n minus 1 to give the variance. So, again if you go to the table we have already calculated the mean and we see the difference of each value from the meaning ok. Take its square sum it up divided by m minus 1 and we get the variance.

The next important parameter is the sample covariance which is nothing but, the co-variance between 2 columns2, 2 different attributes. So, if we have X a and X b then what we do is that, we take the difference of the a-th attribute from the mean the b-th attribute from the mean. Take the product, take the sum from j equal to 1 to n divided by m minus 1 and this gives the sample covariance. (Refer Slide Time: 35:51) So, again coming to this multi variant data, if i want to find out the covariance between: X 1 and X 2, i have the mean from this column. I have the mean from this column. So,I have the departure also, from the mean for the each value. So, corresponding rows I take this difference multiply some of all the differences divided by n minus 1 and thereby get the covariance.

So, covariance between any 2 attributes is measured this way. This measures the correlation in the data. In fact the correlation coefficient is nothing but, r a b which is equal to the covariance between a and b divided by sigma a and sigma b. Product of sigma and sigma b ok C a b by sigma and sigma b, that is covariance divided by standard deviation of the 2 attributes. So, this is the correlation coefficient.

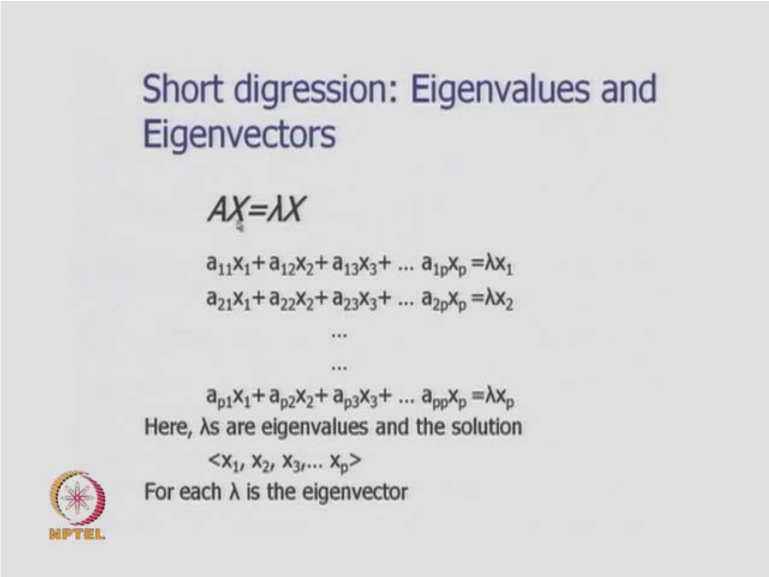
(Refer Slide Time: 39:19)



After that, we do what is called standardization on the variables. We, replace each value by this new quantity y i j which is x i j minus mu i divided by sigma square. So, if you go to this table once again, each value is subtracted from the main and is divided by the variance, each value is treated this way. So, this is a standardization of the variables then, we get the correlation matrix, which is one, for all the diagonal elements because, the correlation between a row and itself; between an attributes and itself is surely 1 and after that the correlation is expressed through r 12 to r 13 etcetera.

So, what is the meaning of r 12? The meaning of r 12 is the correlation coefficient between attribute 1 and attribute 2, r 13 is between attribute 1 and attribute 3, which you can find out very easily from the column values, their mean their variance and their product of and the covariance. So, this gives us the correlation matrix.

(Refer Slide Time: 40:43)



Now, we need 2 important concepts, which are Eigen values and Eigen vectors. So, A x equal to lambda X is the definition of the Eigen value. So, where lambdas are the Eigen values and for each lambda, we get the Eigen vector. So, this a 1 1 X 1 a 1 to X 2 up to a 1 p X 1 is equal to lambda X 1 a p 1 X 1 a p 2 x 2 up to a p p X p equal to lambda X p. So, these are the equations obtained from a X equal to lambda X the defining equation from Eigen value. So, these lambdas are solved and there from there the Eigen vectors are computed. So, from this Eigen values and the Eigen vectors, we will proceed to principle component analysis, then to singular value decomposition and then to latent semantic indexing. This is for the next class.