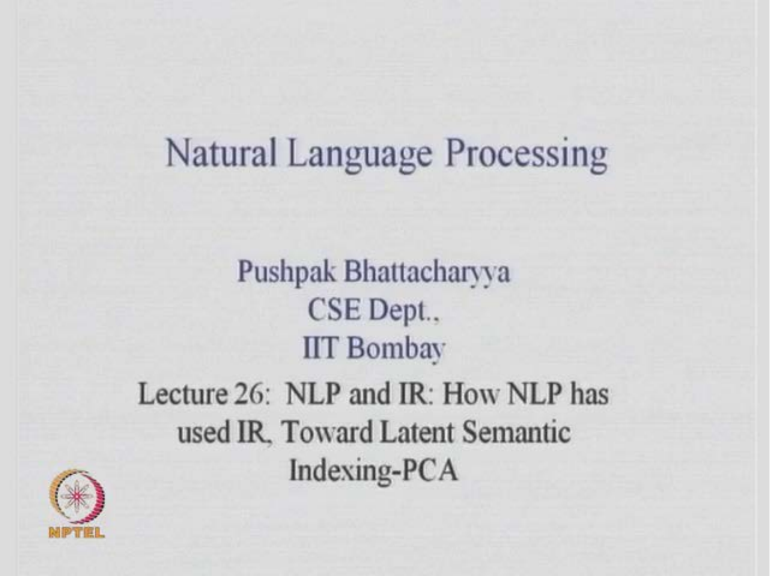
Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 26
NLP and IR: How NLP has used IR, toward Latent Semantic Indexing – PCA

Today, we will continue exploring the relationship between n l p and i r, in particular will move to a very powerful technique called the latent semantic indexing, which is bridge between n l p and i r.

(Refer Slide Time: 00:33)



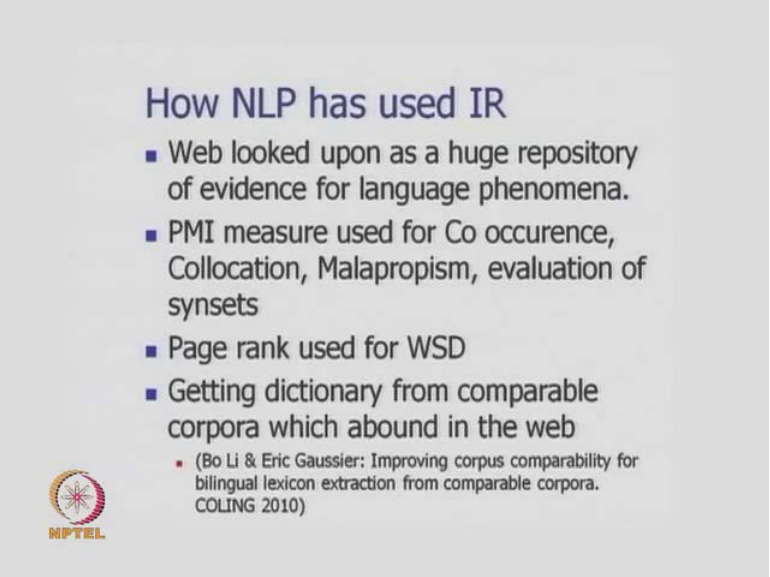
Natural Language Processing

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 26: NLP and IR: How NLP has used IR, Toward Latent Semantic Indexing-PCA

So, today is topic is n l p and i r, how n l p is used i r toward latent semantic indexing and these discussion will be facilitated by a discourse on principle components analysis, alright.

(Refer Slide Time: 00:53)



How NLP has used IR

- Web looked upon as a huge repository of evidence for language phenomena.
- PMI measure used for Co occurence, Collocation, Malapropism, evaluation of synsets
- Page rank used for WSD
- Getting dictionary from comparable corpora which abound in the web
  - (Bo Li & Eric Gaussier: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. COLING 2010)

So, we are discussing the relationship between n l p and i r, in particular we would like to see how natural language processing has used i r. We were discussing this in last class also and we mention that, web is looked up on has a huge repository of evidence for language phenomena. So, the point being measured is that, natural language is a very dynamic evolving entity and since, language always evolves and language has many different forms across culture, across country, across times evidences of language usage have to be found in various repository and web happens to be a very important repository for language usage.

This is where in the web itself, we find language in use in varied forms and all these language phenomena need to be proceed by the machine, we have to design techniques algorithm for processing these phenomena. What we find is that different language is just to take particular case, different languages of different delis in a different part of the country or even a state. If we take the case of English, has many different form in India, in U K, in South Africa, in Kenya, in Australia and so on and so forth and natural language processing technique have to be revised for all this.

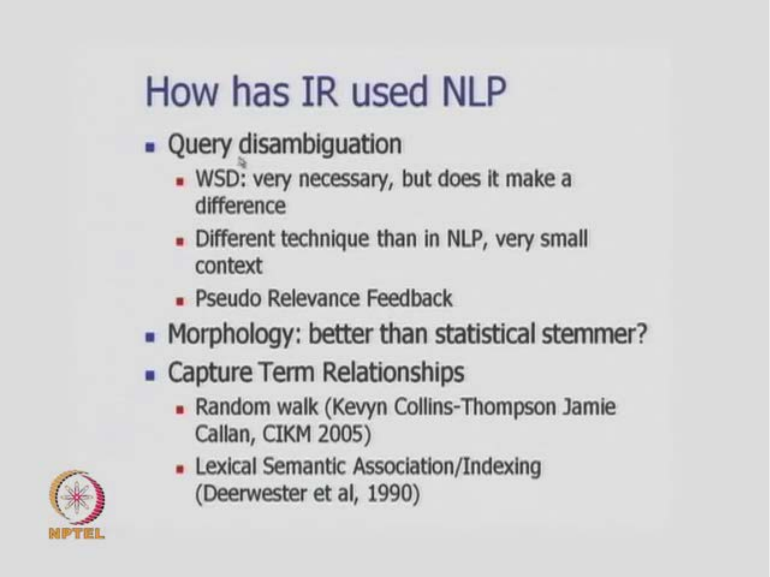
Now, this technique require training through machine learning algorithm, evidence have to be shown to this training algorithm and they can be found from the web. So, that is the main idea behind this 1st point web is looked upon as a, huge repository of a evidence for language phenomena. Now, what you find is that, there are many measures which are used for: co occurrence, co location, malapropism, evaluation of synsets which are typical natural language processing problems and one can make use of P M I, point vise mutual information measure for this n l p phenomena. We will see, what this measure and how it looks. Another important use of higher ideas has been application of page rank algorithm for words and disambiguation.

This we will discuss briefly and the very important of current topic is getting dictionary from comparable corpora, which abound in the web. So, we see papers even this year, which tries to get the dictionary from comparable corpora and comparable corpora abounds in the web for example, from news paper domain. Now, getting the dictionary which is probabilistic, a statistical dictionary is a very important task in natural language processing. It is used for many things and the most important amongst those is statistical machine transition.

So, getting the statistical dictionary from aligned copra is a very important task. But, that needs aligned copra aligned copra is by itself a very t d s task, it is man power intense, money intense, time intense. On the other hand, one can easily find copra which discuss the same topic but, are from different sources in different languages. For example, a an article on India’s foreign policy, India’s recent foreign policy can appear in times of India, which is a English daily and can also appear in our valid times which is the Hindi daily. And, they are likely to discuss the same topic but in a different languages also the sentence need not be exactly parallel to each.

So, in such case also it should be possible to extract the dictionary from this corpora. It has not taken any human labor to unite the comparable copra, it is the day to day activity of some two news papers or it is possibility to generate a dictionaries form this. So, getting dictionary from comparable copra is a very life problem and, this is where we see i r providing resource to natural language processing.

(Refer Slide Time: 05:43)



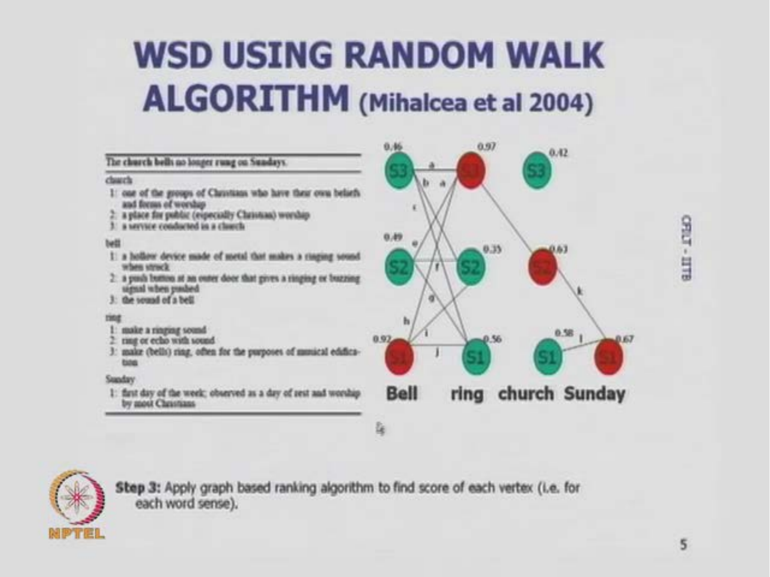
How has IR used NLP

- Query disambiguation
  - WSD: very necessary, but does it make a difference
  - Different technique than in NLP, very small context
  - Pseudo Relevance Feedback
- Morphology: better than statistical stemmer?
- Capture Term Relationships
  - Random walk (Kevyn Collins-Thompson Jamie Callan, CIKM 2005)
  - Lexical Semantic Association/Indexing (Deerwester et al, 1990)

And, how as i r used as natural language processing, query disambiguation surely has been a very important problem in i r and it is thought that, words in disambiguation is necessary here but, does it make really difference? We can discuss this later in some other discussion. The query disambiguation must have different technique than, in n l p because of the very small context. The query size is typically 3 words and here one can see the use of what is called pseudo relevance feedback, which helps disambiguate and expand the query, i r has also used statistic stemmer in place of morphology but, there are language whose information retrieval needs elaborate morphological analysis.

Because, the words are formed in a complex way then, in information retrieval it is important to capture term relationship. So, they are random work algorithm, semantic association based algorithms alright.

(Refer Slide Time: 06:48)



WSD USING RANDOM WALK ALGORITHM (Mihalcea et al 2004)

The church bells no longer rung on Sundays.

church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edification

Sunday
1: first day of the week; observed as a day of rest and worship by most Christians

Step 3: Apply graph based ranking algorithm to find score of each vertex (i.e. for each word sense).

NPTEL

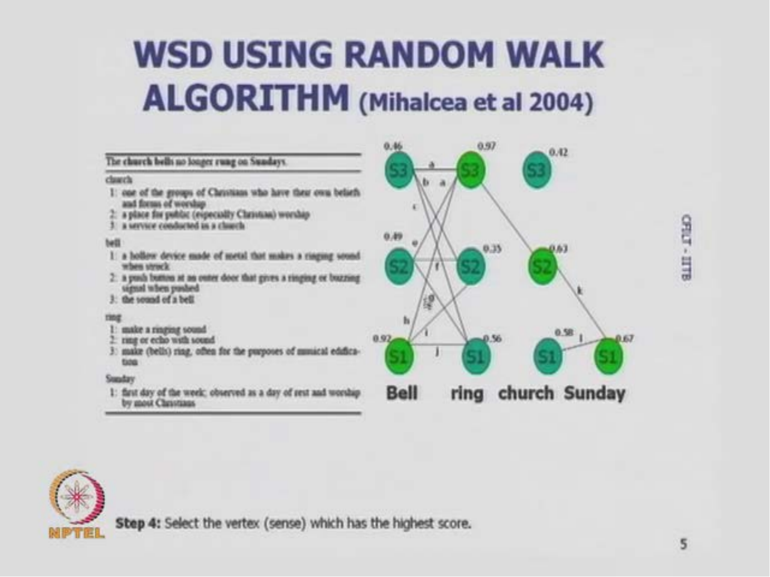
5

So, we now take a look at a what is the disambiguation using random work algorithm or page rank. So, here is a situation where there are four words: bell, ring, church and Sunday. So, we have removed from the sentence this was the sentence. So, may be the sentence is all bells ring in the church on a Sunday. So, words like all the a etcetera have been removed because, they are not ambiguities words but, what remains are: verbs, noun, adjective and adverbs and they need to be disambiguated.

Now, let us look at this 4 words and their senses, church as 3 senses, one of the group of christen, who have their own belief and forms of worship. 2nd meaning is the place for public especially christen worship, 3 is service conducted in the church. Similarly, bell has 3 sense a hollow device made of a metal that makes ringing sound when struck a push bottom at an outer door that, gives a ringing or busing signal when pushed or the sound of the bell. Ring again means make a ringing sound ring or eco with sound make bells ring often for the propose of musical edification. Sunday is not ambiguous, this is this has the meaning, 1st day of the week absorbed as a day of christ and worship by most christen alright.

So, we have this four words and on top of each word we erect the senses. So, bell has 4 senses: s 1, s 2 and s 3, ring also has 3 senses, church has 3 senses and Sunday has a single senses. So, after this senses are erected then, we create ages between the senses going one what to the next and these ages are embellished with weights, which uses definitions based semantic similarity. This is the least method and then, we apply a graph based ranking algorithm to find the score of each vertex that is for each word sense and we select the vertex of the sense, which has the highest score.

(Refer Slide Time: 09:12)



WSD USING RANDOM WALK ALGORITHM (Mihalcea et al 2004)

The church bells no longer rung on Sundays.

church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edification

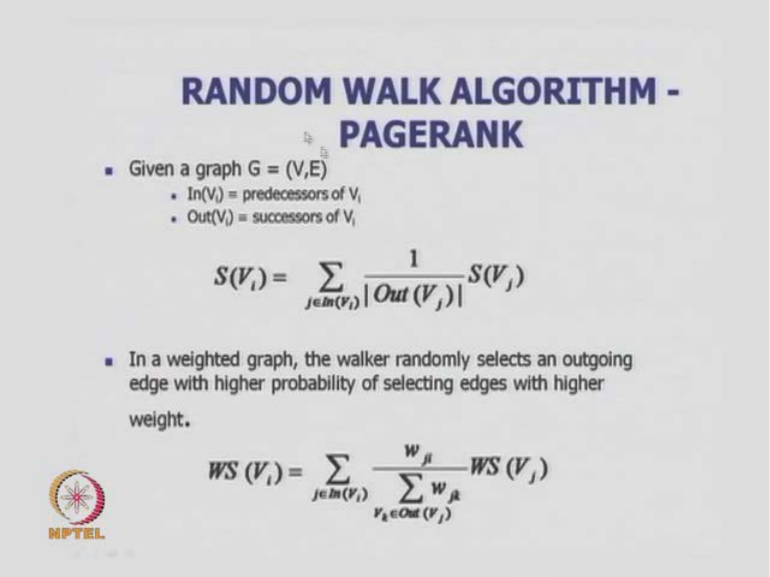
Sunday
1: first day of the week; observed as a day of rest and worship by most Christians

Step 4: Select the vertex (sense) which has the highest score.

5

So, when the algorithm stops, s 1 on bell which is hollow device made of metal that makes a ringing sound when struck. This is the sense which will become the winner, ring s 3 is the winner, s 3 is make bells ring often for purposes of musical edification and for church s 2 is the winner which is the place for public especially christen worship and for Sunday, which has the sense this is the winner. So, after the learning of the page rank algorithm, which is based on importance of a particular note and the important it transfers by it leakages to other notes, which is a very information retrieval each idea were nodes are pages and edges are links to other pages as been bored into words and disambiguation. This is Mihalcea et al work in 2004.

(Refer Slide Time: 10:15)



RANDOM WALK ALGORITHM - PAGERANK

- Given a graph G = (V,E)
  - $In(V_i)$ = predecessors of V_i
  - $Out(V_i)$ = successors of V_i

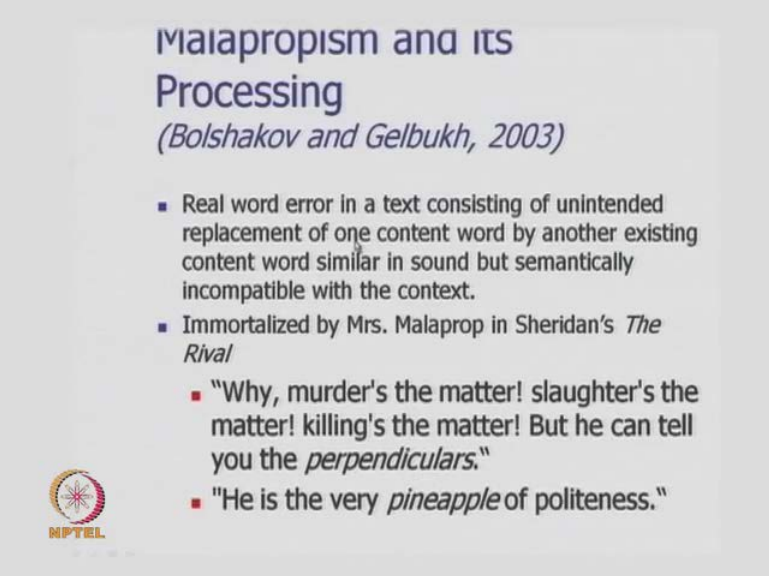
$$S(V_i) = \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

- In a weighted graph, the walker randomly selects an outgoing edge with higher probability of selecting edges with higher weight.

$$WS(V_i) = \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

And, these 2 formula show how nodes transfer there important to other nodes. So, this is a the mathematical expression which says that the important of the node V i, which is connected to other node V j were j is in the link of V i is equal to importance of V j divided by the outgoing links of V j. So, we take some of all this V j which are linked to V i and then we obtain the importance of V i. Similarly, a weighted some of the importance is used to get the weighted importance of the node V i alright. So, these we need not go into lot of details, what I would like to stress is that, these idea of important going from one note to another over the links is a very higher is idea.

(Refer Slide Time: 11:30)



Malapropism and its Processing

(Bolshakov and Gelbukh, 2003)

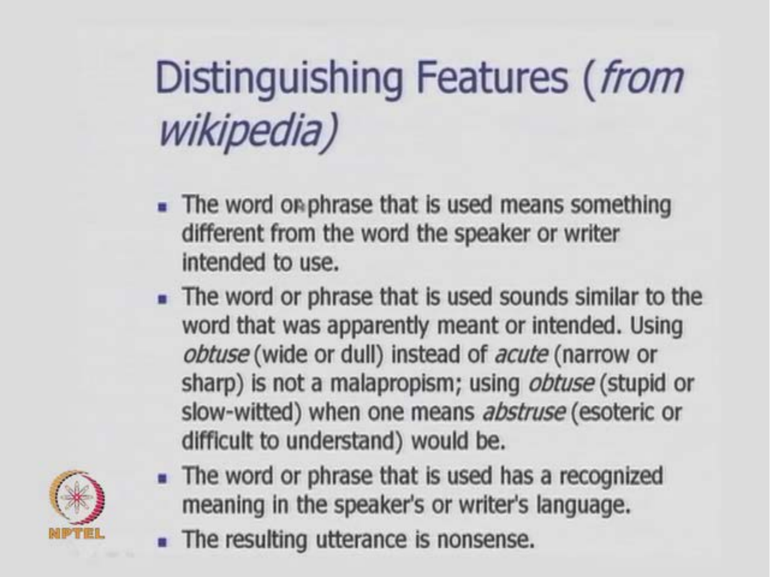
- Real word error in a text consisting of unintended replacement of one content word by another existing content word similar in sound but semantically incompatible with the context.
- Immortalized by Mrs. Malaprop in Sheridan's *The Rival*
  - "Why, murder's the matter! slaughter's the matter! killing's the matter! But he can tell you the *perpendiculars*."
  - "He is the very *pineapple* of politeness."

Now, we discuss a very interesting problem which is in the heart land of natural language processing. So, this there is phenomena called malapropism and there was work trying to show, how this problem could be solved. So, malapropism is real word error in a text consisting of unintended replacement of one content word by another existing content word similar in sound but, semantically incompatible with the context alright. So, in malapropism one replaces a word by another valid of the language but, the new word is completely miss fit semantically in the context. Malapropism has been immortalized by misses Malaprop in Sheridan's, the rival, is very famous novel.

But, more famous novel is this charter in the novel, misses Malaprop because of her odd speech because of her usage of words, which sound similar to an accurate word in the context but, the what she uses is completely misfit. So, here is an example: why, murder is the matter! slaughter is the matter! killing is the matter! but he can tell you the perpendicular. So, the word perpendicular here is a complete misfit what misses Malaprop meant was but he tell you the particulars. So, she meant particular instead of she uses perpendicular, which is complete misfit.

The 2nd example is, he is the very pineapple or politeness. The adiabatic expression in English, for this politeness which is very appropriate and which is in right amount, the actual sentence would, he is the very pineapple of politeness. Instead of pineapple she has used the word pineapple, which is the fruit. So, these 2 example show what malapropism is, it is essentially substituting a valid word by a accurate, what which is semantically what.

(Refer Slide Time: 13:50)

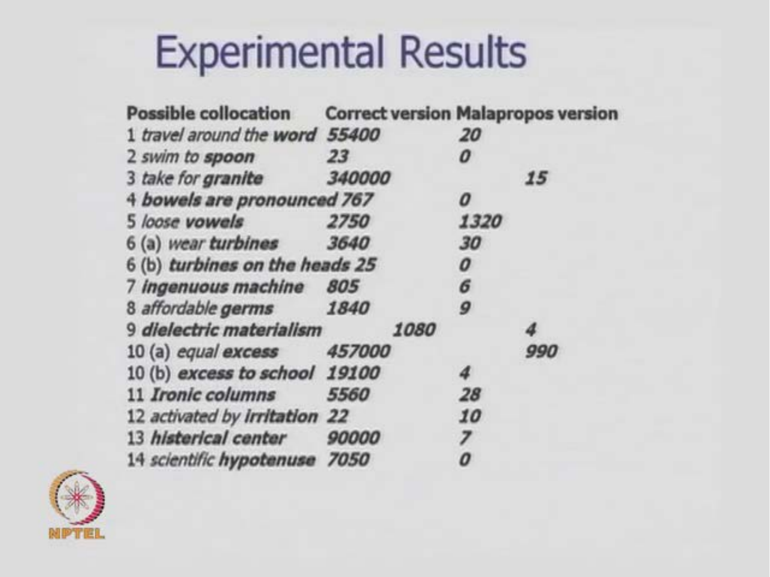


Distinguishing Features (*from wikipedia)*

- The word or phrase that is used means something different from the word the speaker or writer intended to use.
- The word or phrase that is used sounds similar to the word that was apparently meant or intended. Using *obtuse* (wide or dull) instead of *acute* (narrow or sharp) is not a malapropism; using *obtuse* (stupid or slow-witted) when one means *abstruse* (esoteric or difficult to understand) would be.
- The word or phrase that is used has a recognized meaning in the speaker's or writer's language.
- The resulting utterance is nonsense.

Now, the distinguish features of malapropism which is a reproduced from Wikipedia is that, the word of phase that is used means something different from the word, the speaker or writer intended to use. The word or phrase that is used sounds similar to the word that was apparently meant or intended. Using obtuse wide or dull instead of acute narrow or sharp is not a malapropism; using obtuse stupid or slow witted, when one means abstruse esoteric or difficult to understand would be. So, the point here is that, the word or the phrase must sound similar and it should be a completely misfit in the context, the word of phrase that is used as the recognized meaning in the speakers or writers language. The resulting is utterance is a non sense.

(Refer Slide Time: 14:39)

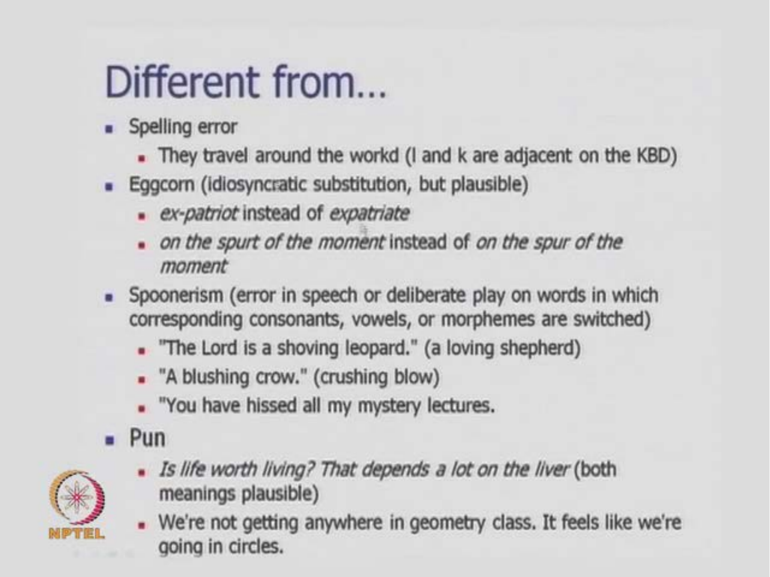


Experimental Results

Possible collocation	Correct version	Malapropos version
1 *travel around the* ***word***	55400	20
2 *swim to* ***spoon***	23	0
3 *take for* ***granite***	340000	15
4 ***bowels are pronounced***	767	0
5 *loose* ***vowels***	2750	1320
6 (a) *wear* ***turbines***	3640	30
6 (b) ***turbines on the heads***	25	0
7 ***ingenuous machine***	805	6
8 *affordable* ***germs***	1840	9
9 ***dielectric materialism***	1080	4
10 (a) *equal* ***excess***	457000	990
10 (b) ***excess to school***	19100	4
11 ***Ironic columns***	5560	28
12 *activated by* ***irritation***	22	10
13 ***histerical center***	90000	7
14 *scientific* ***hypotenuse***	7050	0

So, this particular problem was tackled (Refer Slide Time: 13:50) using point wise mutual information. And here, sound the experiments result are shown which will come to little later.

(Refer Slide Time: 14:59)



Different from…

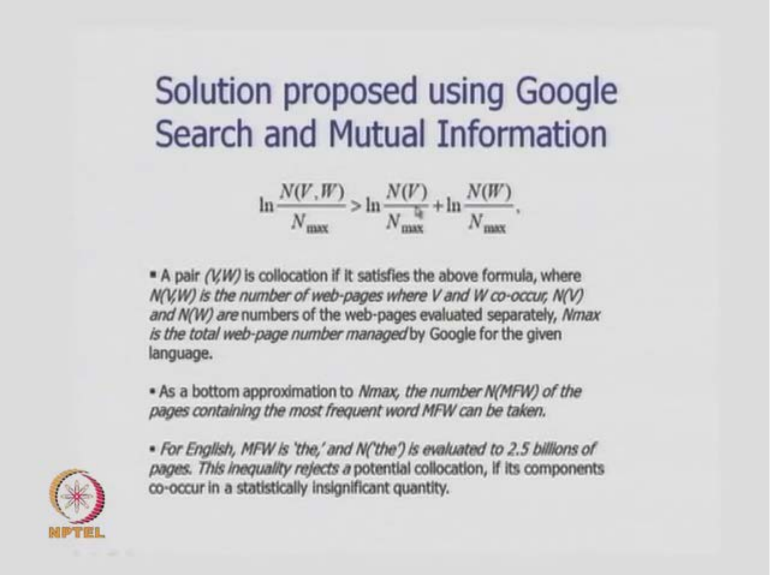
- Spelling error
  - They travel around the workd (l and k are adjacent on the KBD)
- Eggcorn (idiosyncratic substitution, but plausible)
  - *ex-patriot* instead of *expatriate*
  - *on the spurt of the moment* instead of *on the spur of the moment*
- Spoonerism (error in speech or deliberate play on words in which corresponding consonants, vowels, or morphemes are switched)
  - "The Lord is a shoving leopard." (a loving shepherd)
  - "A blushing crow." (crushing blow)
  - "You have hissed all my mystery lectures.
- Pun
  - *Is life worth living? That depends a lot on the liver* (both meanings plausible)
  - We're not getting anywhere in geometry class. It feels like we're going in circles.

But, malapropism could be differentiated from a few related one is spelling error. They travel around the what here, the workd actually is world the letter k came here, because l and k are adjusted on the key board and the person has type k instead of l. Now, the spelling error is the solve problem, there are many system which detects spelling error and make very sense able suggestion. There is another phenomena, which is related to malapropism that is called egg corn. This is idiosyncratic substitution but, plausible. For example, one uses the word ex-patriot instead of expatriate and the word means approximately same as the expatriate but is not exactly that.

On the spurt of the movement instead on the spur of the moment is a fixed expression. On the spurt of the movement is not idiomatic but still there is not much valuation of the meaning. Spoonerism is interesting phenomenon were there is error in speech or deliberate on words, in which corresponding consonants vowels or morphemes are switched. So for example, the expression here, a very famous duplicate expression, a loving shepherd is expressed has the lord is a shoving leopard. The actual sentences should be the lord is a loving shepherd.

A blushing crow instead of, a crushing blow, he have hissed all the mystery classes instead of you have missed mystery classes. Here also, one does not see much improvement in the area. The next point is pun, here are 2 example is life worth living that depends lot on the liver, there is pun on liver, can mean a body part or a person who is living and both meaning are possible. We are not getting any were in geometry class. It feels like going in circle here, also there is pun on the circles.

(Refer Slide Time: 16:59)



Solution proposed using Google Search and Mutual Information

$$\ln \frac{N(V,W)}{N_{\max}} > \ln \frac{N(V)}{N_{\max}} + \ln \frac{N(W)}{N_{\max}},$$

- A pair *(V,W)* is collocation if it satisfies the above formula, where *N(V,W) is the number of web-pages where V and W co-occur, N(V) and N(W) are* numbers of the web-pages evaluated separately, *Nmax is the total web-page number managed* by Google for the given language.
- As a bottom approximation to *Nmax, the number N(MFW) of the pages containing the most frequent word MFW can be taken.*
- *For English, MFW is 'the,' and N('the') is evaluated to 2.5 billions of pages. This inequality rejects a* potential collocation, if its components co-occur in a statistically insignificant quantity.

So, the solution which has proposed for doing malapropism deduction was through the Google search and the mutual information. Here is the mutual information formula, this quantity N V W is the number of web pages which contains v and w together. This N max is the normalizing factor which is the number of times or the number of pages containing the most frequent word in the language. Now, this algorithm of N V W divided by N max should be greater than individual algorithm of N V divided by N max plus algorithm of N W divided by N max. So, what it means is that, the proportion of pages containing V and W together should be greater than the sum of the number of pages containing V and those pages containing W individually.

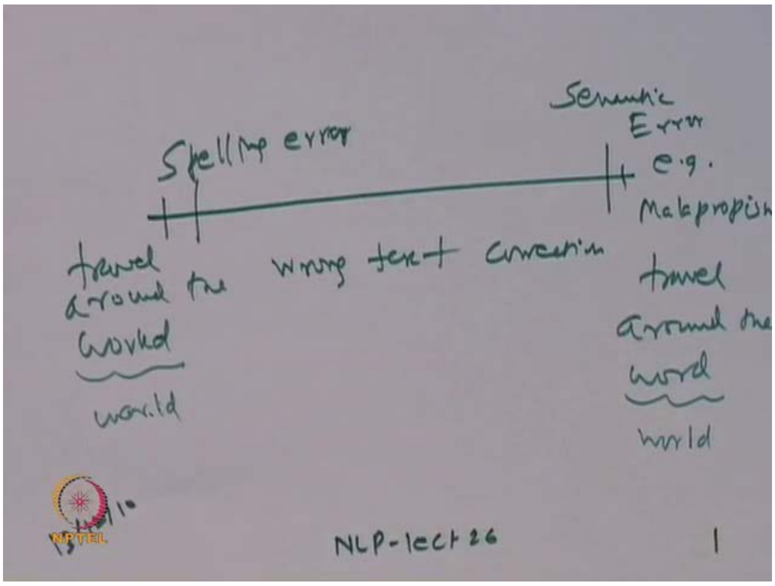
So, this and also logarithm of course, so this formula indicates that, if V and W is a strong collocation then, they are occurring together is much more than, they are accidental individual occurring together in a page. So, this is the point wise mutual information formula. So, the formula detects very strong collocations between 2 words: V and W. Now N max is found in the following way, we find the number of pages were the word the, appears. So, the number of page containing the word the, that should give me N max, that maximum number of pages containing the most frequent word in the language. So, what is the overall idea? The overall idea is that, V and W are checked for their being collocated.

Now, if a word is a malapropism word in a sentence then, it will not have strong collocation with any of the words in the language. For example, I travel around the world, travel and world have strong collocation but, travel and world will not have the strong collocation. So, if we try to see if travel and world are collocated we launch a search with travel and world together and identify the number of pages, were travel and world occur together. We also launch another search with travel and world occurring together. And there, we will find a very large number of pages because this is a strong collocation.

So here, i r or the web is helping us to detect collocation and there by solve the problem of malapropism. So, here are some experimental result which we had shown before. Travel around the world, the correct version which is world, travel around the world as 55,000 web pages showing this pattern. But, travel around the world this has already 20 pages, swim to spoon has 23 correct version and 0 malapropism version take for granite instead of take for granted. So, take for granted will have past 340000 web pages and take for granite has only 15 pages. Bowels are pronounced instead of bowels are pronounced the correct version is 767 pages and so on.

So, this number shows that words which are strong collocation have very strong web evidence of being together in a page and malapropism words when tested collocation with other words in the sentence, show very little evidence on the web. So, based on that we can identify, if a word is a malapropism word or not. The propose of this discussion was to show that natural language processing can make use of the resources of the web to solve its intricate and interesting problems. Now, what is the use of malapropism detection? Malapropism detection comes in the spectrum of correcting desktop publishing.

(Refer Slide Time: 22:02)



So, we write here a few points, if I look up on this spectrum of wrong text correction then, we find that at one end of the spectrum is spelling error and the other end of the spectrum is semantic error for example, malapropism. So, spelling error will detect a wrongly spelt word. So, travel around the workd instead of traveling around the world. And here, we have traveled around the word instead of world, here also instead of world ok. Now here, what has been miss spelled has what, l substituted by k, this is an easy error to detect. But, world replaced with world, where l has been dropped is the deletion error and this was the substitution error.

This is the former difficult to detect because all the words are correct word of the language but, travel around the world does not make any sense. So, that has to be detected through sophisticated techniques for example, malapropism detection ok. So, this is the motivation behind tackling this error alright. So, we have seen a an example of hard code natural language processing problem, the problem directly from the heart line of natural language processing which has been solved by making use of the information retrieval resource ok.

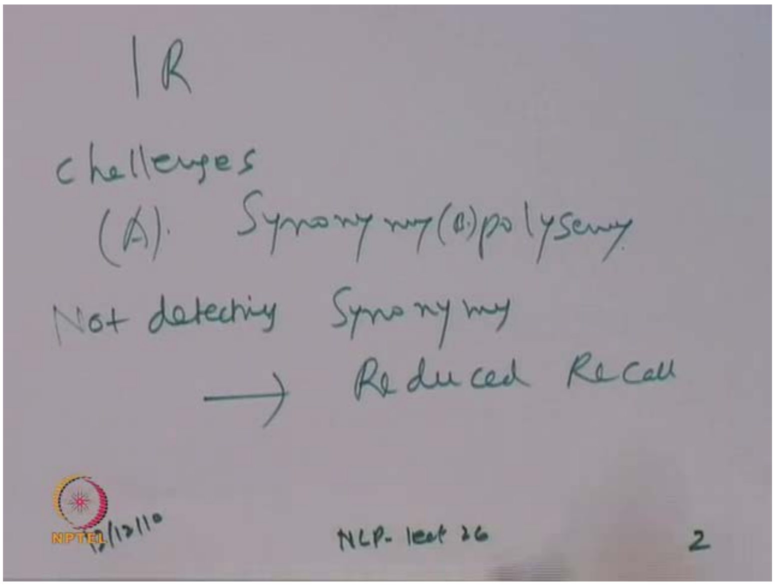
(Refer Slide Time: 23:57)

Towards Latent Semantic Indexing: Principal Component Analysis

So, we move ahead and then we bring up a very important topic namely: latent semantic indexing and for these technique, we need to understand a another technique called principle component analysis, p c a that is what we will do. Now, what is the important of latent semantic indexing l s i, as it called, l s i is very important because, it shows and association between words which are semantically related through the evidence of document which contain those words ok. So, we saw a collocation detection through the evidence of the web. But, that became possible when such engine technology mature and that, web contain lot corpora and it was possible to extract or being pages from the web at a fast speed.

But, what we find is that when before even before, i r became a very mature, very efficient. There was a need for detecting words which are semantically related based on the evidence of corpora were they occur ok. So, this is the propose and the main goal here is to capture semantic association between words. And, this is used for improving the recall in information retrieval that means the words which are semantically related will also be used for detecting for preaching pages from the corpora.

(Refer Slide Time: 25:58)



IR

challenges

(A). Synonymy (B) polysemy

Not detecting Synonymy

→ Reduced Recall

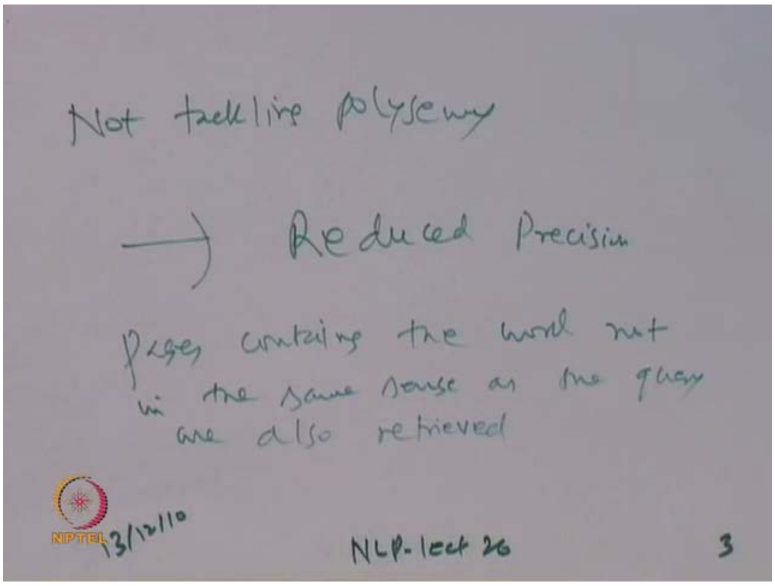
13/12/10

NLP- lect 26

2

So, now we will write a few points again, in i r, a very important challenge or there are two important challenge are: one Synonymy and B Polysemy. So, not detecting synonymy leads to reduced recall ok. So, that means a query has been launched and the query words are used to face a document but it is possible that, the same query words are synonymy and just because the words have been matched, the actual web page containing synonymy words will not be brought. So these bring down the recall of such.

(Refer Slide Time: 26:51)



Not tackling polysemy

→ Reduced Precision

pages containing the word not in the same sense as the query are also retrieved

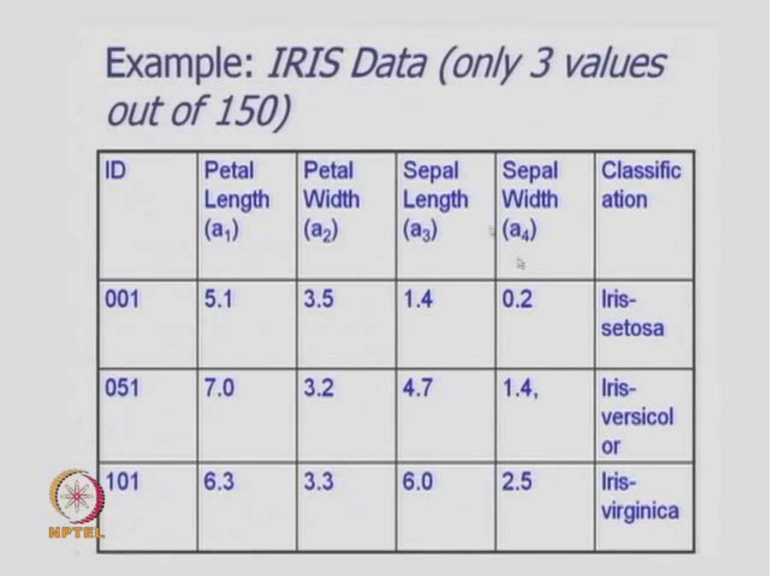
13/12/10

NLP- lect 26

3

On the other hand, not tackling Polysemy can leads to reduced precision. So, this means that pages containing the word not in the same sense has the query are also retrieved. So, this brings down the performs of the search engine by reducing precision. But, the previous problem were, if the synonymy is not detected then that brings down the recall of the system. So, both situations are undesirable and at least for synonymy one can make use of this powerful technique of latent semantic indexing and improve the recall of retrieval ok.

(Refer Slide Time: 27:49)



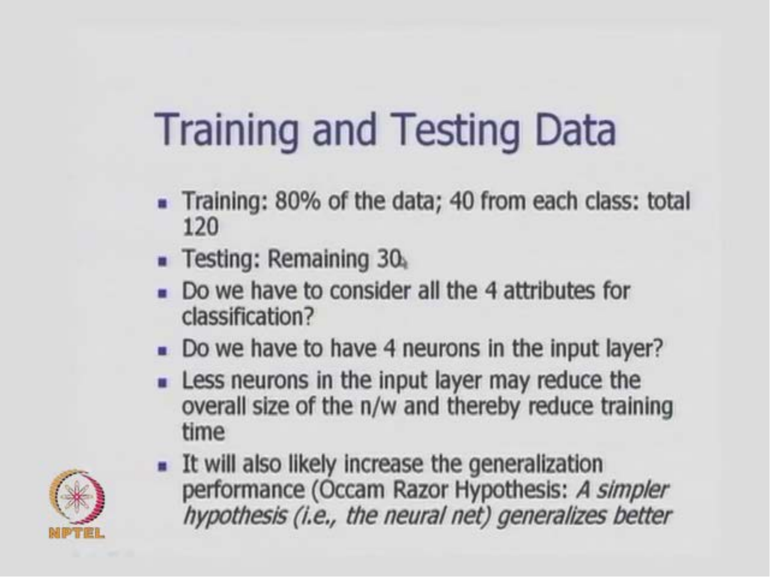
Example: *IRIS Data (only 3 values out of 150)*

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

So, we proceed ah from a basic discussion, which is on principle component analysis. Here is a example, there is the famous data which is used by machine learning researches, these data is called IRIS data. So, there are 150 such rows as it shown here, the 1st column is the i d of the data item, the 2nd column is the attribute which is the petal length, 3rd column is attribute which is petal width, 4th column is the attribute called sepal length followed by attribute sepal width. And, based on this 4 parameter namely: petal length, petal width, sepal length, sepal width, we identify the flowers as one of the 3 classes: Iris setosa, Iris versicolor and 3rd class is Iris virginica, iris setosa, Iris versicolor and Iris virginica.

So, here this table shows 3 such classes and there corresponding attribute. For example, if the petal length is 5.1 unit, petal width is 3.5 unit, sepal length is 1.4 units and sepal width is 0.2 units then, the classification of the flower is into the Iris setosa the class. So, abstractly speaking here, is the problem were there are 3 classes and there are 4 attributes and our goal is to create a learning system which, when given the petal length, sepal length and sepal width will produce the classification of the flower alright.

(Refer Slide Time: 29:37)



Training and Testing Data

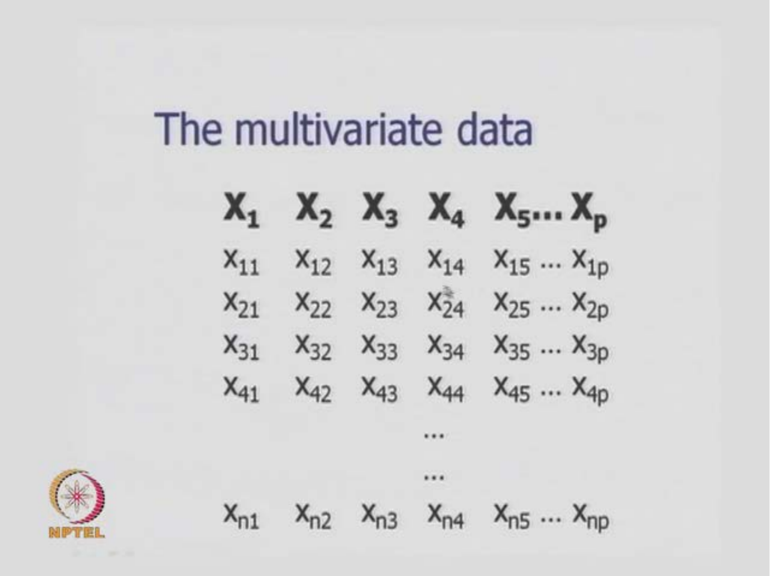
- Training: 80% of the data; 40 from each class: total 120
- Testing: Remaining 30
- Do we have to consider all the 4 attributes for classification?
- Do we have to have 4 neurons in the input layer?
- Less neurons in the input layer may reduce the overall size of the n/w and thereby reduce training time
- It will also likely increase the generalization performance (Occam Razor Hypothesis: *A simpler hypothesis (i.e., the neural net) generalizes better*

So, we discuss the learning situation once again. The training happens on 80% of the data for, there are 50 examples each from the 3 classes making it up has 150 example for the 3 classes all together. Out of this 50, we take up 40 for trading and lift 10 for testing. So, 40 for each class so, totally there are 120 class examples for training, testing is one remaining 30 examples ok. So after training the machine will given the 4 attributes of any of the 30 example and it is suppose to classify the example correctly.

Now, the question we ask is, do we have to consider all the 4 attributes for the classification? Do we have to have 4 neurons in the input let? This is with respected on neuron deter which is used for machine and less neuron on the input layer may reduce the overall size the network and thereby reduce training time. It will also likely increase the generalized performance which is the Occam Razor hypothesis. A simpler hypothesis the neural net for example, in this case analyses better. So, the main point behind this slide is that, we make a portion of the given data as training data and rest of the data is used for testing efficacy of the learned system. And, on the way we ask if all the attribute which are used for the deciding the class are needed ok.

So, these are the important point because if all attributes are not needed then, we have the much more compact machine learning system

(Refer Slide Time: 31:25)

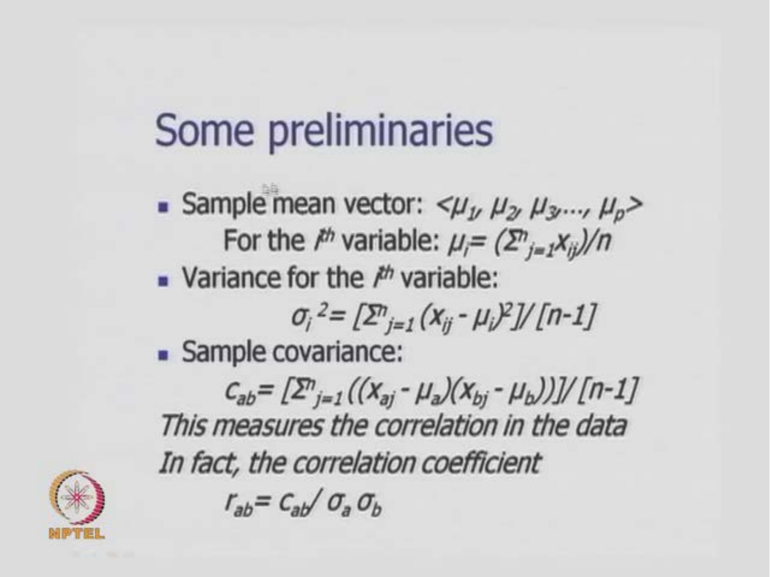


The multivariate data

X_1	X_2	X_3	X_4	X_5	…	X_p
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	…	x_{1p}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	…	x_{2p}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	…	x_{3p}
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	…	x_{4p}
			…			
			…			
x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	…	x_{np}

So, we generate the discussion and go into what is called the processing of multivariate data. Here, we have this p attributes and each of the values of the attribute are given for each example. So, the first row is for the first example were the attribute x 1 has the value x 1 1, x 2 has the value x 1 2, x 3 has the value x 1 3, x 4 has the value x 1 4 similarly, the p-th attribute has the value x 1 p. So, proceeding this way the last example has the value: x n 1, x n 2, x n 3, x n p for the attribute x 1 p x p ok. So, these are tables which is given along with its classification, there should be a y column which shows the classification and we discuss how to process the multivariate data.

(Refer Slide Time: 32:21)



Some preliminaries

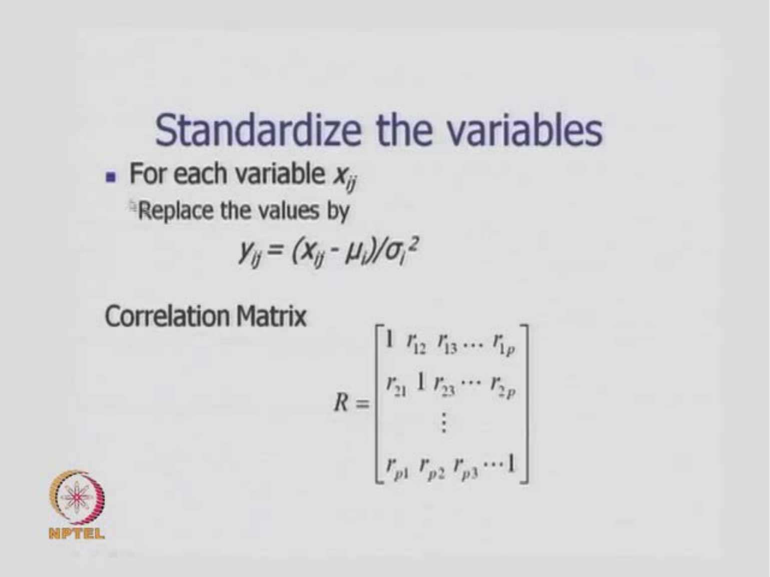
- Sample mean vector: $<\mu_1, \mu_2, \mu_3, \ldots, \mu_p>$
  For the i^{th} variable: $\mu_i = (\Sigma^n_{j=1} x_{ij})/n$
- Variance for the i^{th} variable:
  $\sigma_i^2 = [\Sigma^n_{j=1} (x_{ij} - \mu_i)^2] / [n-1]$
- Sample covariance:
  $c_{ab} = [\Sigma^n_{j=1} ((x_{aj} - \mu_a)(x_{bj} - \mu_b))] / [n-1]$

This measures the correlation in the data
In fact, the correlation coefficient
$r_{ab} = c_{ab} / \sigma_a \sigma_b$

So, we make use of some preliminary notions. 1st is the sample mean vector which is expressed as mu 1 to mu 2 up to mu p, for the p attributes and were mu i for the i-th attribute is sigma j equal to 1 2 n x i j divided by n. So, what it means is that, if we look at the multivariate data once again for example, the sample means is a vector of the mean values of the attributes and there values. So, mu 3 for example, will be by summing up all this value and dividing it by n. So, for an attribute what the values are and the mean of those values. Similarly, the variance for the i-th variable is the standard formula sigma i square is equal to j equal to 1 2 n x i j minus mu i whole square divided by n minus 1.

So, this shows the deviation from the mean of a value and sample co variants also can be found, which is c a b equal to x a j minus mu a into x b j minus mu b divided by n minus 1. So, do take this product of x a j minus mu a and x b j minus mu b, j varying from 1 to n and divided by n minus 1 and we have sample co variants matrix. These measured the correlation in the data in fact, the covariant co efficient r a v equal to c a v divided by sigma a into sigma b.

(Refer Slide Time: 34:14)



Standardize the variables

- For each variable x_{ij}

Replace the values by

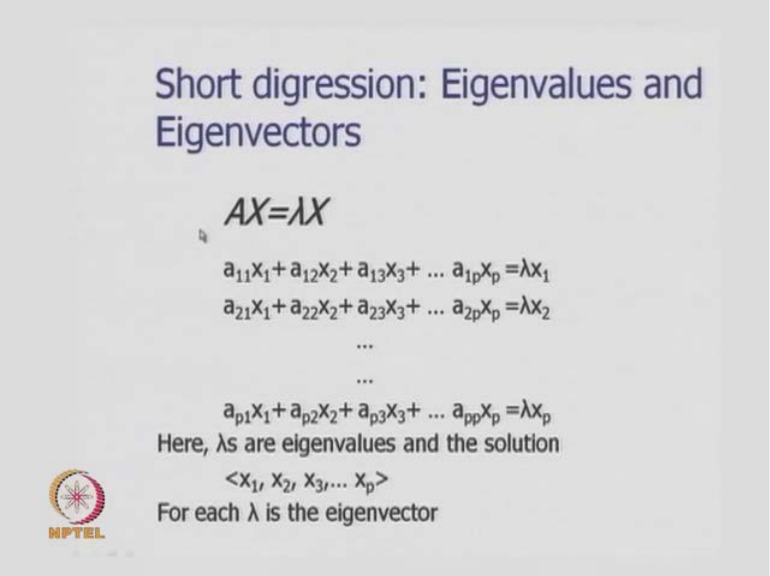
$$y_{ij} = (x_{ij} - \mu_i)/\sigma_i^2$$

Correlation Matrix

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} \cdots & r_{1p} \\ r_{21} & 1 & r_{23} \cdots & r_{2p} \\ & & \vdots & \\ r_{p1} & r_{p2} & r_{p3} \cdots & 1 \end{bmatrix}$$

Now, the first thing which is done for that each variable value x i j we normalize it. So, we get from x i j y i j which is nothing but, x i j minus mu y divided by sigma i square sigma j sigma i square s. So, sigma i square is the variants of i-th attribute and x i j minus mu i is the departure from the attribute mean. Similarly, the correlation matrix correlation is written here which is: 1, r 1 2, r 1 3 up to r 1 p. So, each r i j is the correlation co efficient as define in the previous slide by means of c a b and sigma a and sigma b. So, this is the correlation matrix all whose diagonal elements are 1 and other values are correlation co efficient

(Refer Slide Time: 35:16)



Short digression: Eigenvalues and Eigenvectors

$$AX=\lambda X$$

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots a_{1p}x_p = \lambda x_1$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots a_{2p}x_p = \lambda x_2$$
$$\ldots$$
$$\ldots$$
$$a_{p1}x_1 + a_{p2}x_2 + a_{p3}x_3 + \ldots a_{pp}x_p = \lambda x_p$$

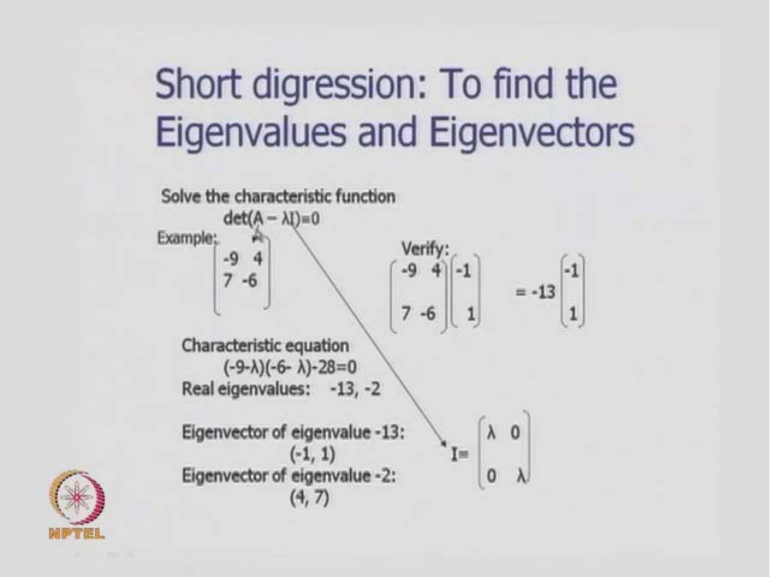
Here, λs are eigenvalues and the solution

$<x_1, x_2, x_3, \ldots x_p>$

For each λ is the eigenvector

Now, we make the short digression and discuss the important Eigen values and Eigen vectors in this whole topic. So, given a matrix A we can find out the Eigen values of the matrix by means of this fundamental equation, A X is equal to lambda X. And, this equation can be expanded to write this p equations, x a 1 1 x 1 plus a 1 to x 2 plus a 1 x 3 plus a 1 to x p is equal to lambda x 1. Similarly, a 2 x 1 x p is equal to lambda x 2 and so on. And finally, we have a p 1 x 1 a p to x 2 up to a p p x p equal to lambda x p. Here, these lambdas are Eigen values and the solution x 1 x 2 up to x p for each lambda is the Eigen vector.

(Refer Slide Time: 36:15)



Short digression: To find the Eigenvalues and Eigenvectors

Solve the characteristic function

$\det(A - \lambda I) = 0$

Example: A

$\begin{pmatrix} -9 & 4 \\ 7 & -6 \end{pmatrix}$

Verify:

$\begin{pmatrix} -9 & 4 \\ 7 & -6 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -13 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

Characteristic equation

$(-9-\lambda)(-6-\lambda)-28=0$

Real eigenvalues: -13, -2

Eigenvector of eigenvalue -13:

(-1, 1)

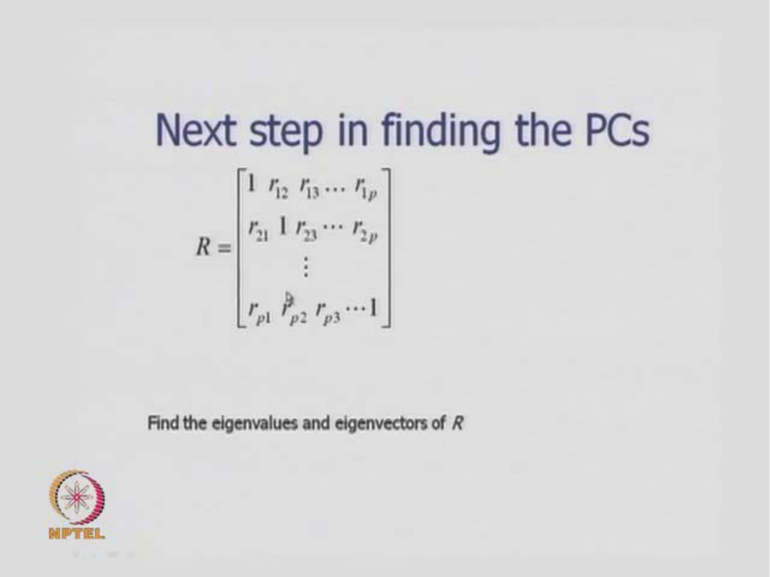
Eigenvector of eigenvalue -2:

(4, 7)

$I = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$

So, let us take an example suppose, the matrix is minus 9, 4, 7, 6. So, the 1st two element of minus 9 and 4 2nd row element is 7 and minus 6. So, we have to solve the determinant a minus lambda i equal to 0 and get the value of lambda. So, a minus lambda i were, i is identity matrix which is nothing but 1 0 0 1, lambda i is the matrix lambda 0 0 lambda. So, a minus lambda i will give us a minus lambda i, will give us this characteristic equation, minus 9 minus lambda into minus 6 minus lambda minus 28 equal to 0. Solving this equation we find that the lambda values are minus 13 and minus 2. So, the from these 2 Eigen values we get this 2 Eigen vector, minus 1 1 and 4 and 7. So, for Eigen values minus 13 the Eigenvector is minus 1 1 and for the Eigen values of minus 2, the Eigen vector is 4 7 ok.

(Refer Slide Time: 37:38)



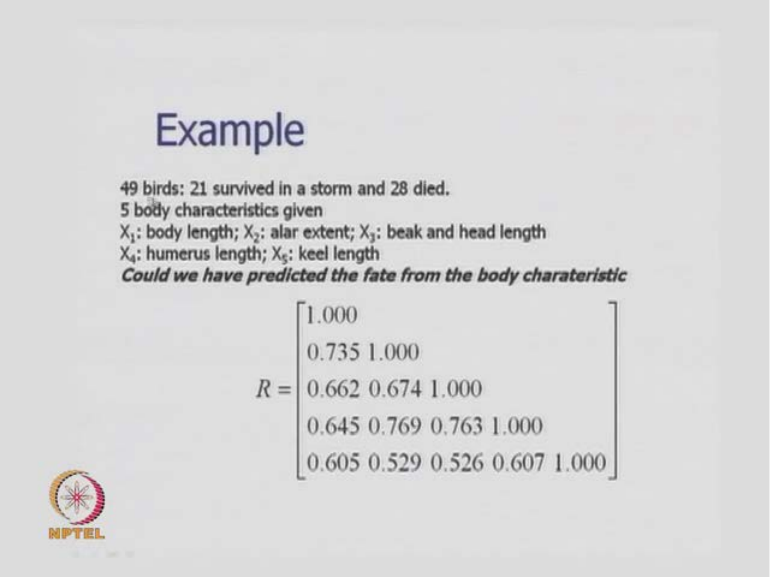
Next step in finding the PCs

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ & & \vdots & & \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix}$$

Find the eigenvalues and eigenvectors of R

So, after this deviation we now discuss, how we find the principle components for the multivariate data. Now, we see that from the multivariate data we have got this rank correlation matrix which is 1 r 1 3 up to r 1 p, each r i j is the coloration coefficient between the i-th attribute and the j-th attribute. So, for this multivariate data, from this rank matrix, we find the Eigen values and Eigen vectors of R.

(Refer Slide Time: 38:12)



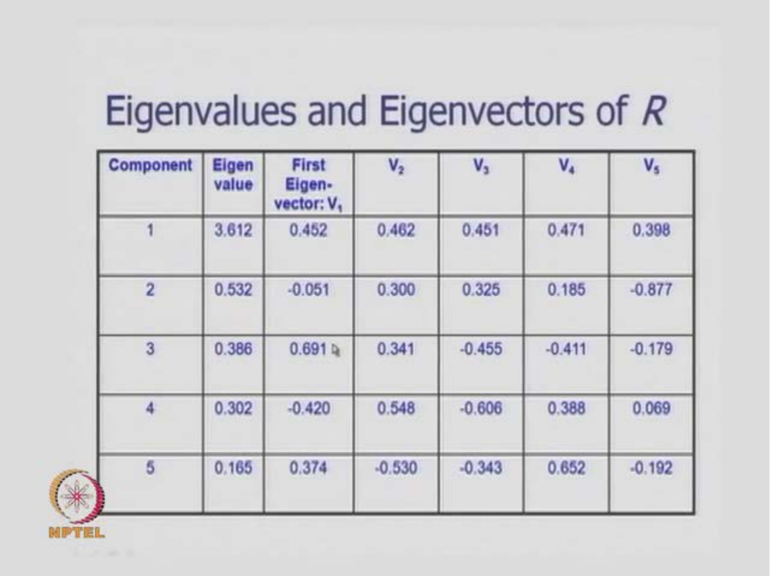
Example

49 birds: 21 survived in a storm and 28 died.
5 body characteristics given
X_1: body length; X_2: alar extent; X_3: beak and head length
X_4: humerus length; X_5: keel length
Could we have predicted the fate from the body charateristic

$$R = \begin{bmatrix} 1.000 & & & & \\ 0.735 & 1.000 & & & \\ 0.662 & 0.674 & 1.000 & & \\ 0.645 & 0.769 & 0.763 & 1.000 & \\ 0.605 & 0.529 & 0.526 & 0.607 & 1.000 \end{bmatrix}$$

So, we take here an example, there are 49 bars 21 which survived in a storms and 28 died and 5 body characteristic are given. 1st attribute is body length, 2nd attribute is a biological term called alar extent, x 3 is beak and head length, x 4 is humerus length and x 5 is keel length. So, depending on this 5 attribute were suppose to decide if the bar as to survive in a storm or not. So, could be predicate the fate from the body characteristics. So, from the multivariate data, from this 5 attributes for 49 data items we can make use of the rank, we can produce the rank correlation co efficient matrix.

(Refer Slide Time: 39:13)

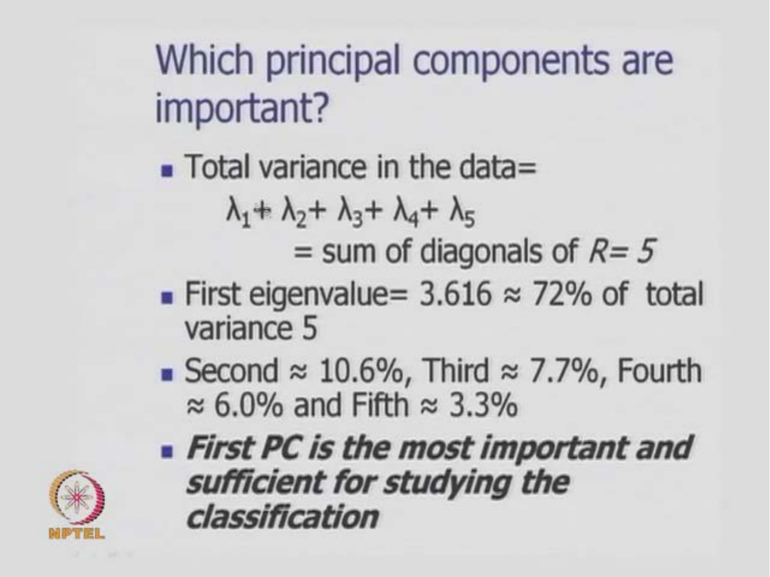


Eigenvalues and Eigenvectors of *R*

Component	Eigen value	First Eigen-vector: V_1	V_2	V_3	V_4	V_5
1	3.612	0.452	0.462	0.451	0.471	0.398
2	0.532	-0.051	0.300	0.325	0.185	-0.877
3	0.386	0.691	0.341	-0.455	-0.411	-0.179
4	0.302	-0.420	0.548	-0.606	0.388	0.069
5	0.165	0.374	-0.530	-0.343	0.652	-0.192

And, having found the r matrix, we can get the Eigen values and Eigen vectors. So, what we show here is for the 1st component the Eigen value is 3.612, 1st Eigen vector V 1 is 0.452, for the 2nd Eigen vector this is 0.462, V 3 is 0.451 we put this 0.471 and V 5 is 0.398, for all the 5 component.

(Refer Slide Time: 39:53)



Which principal components are important?

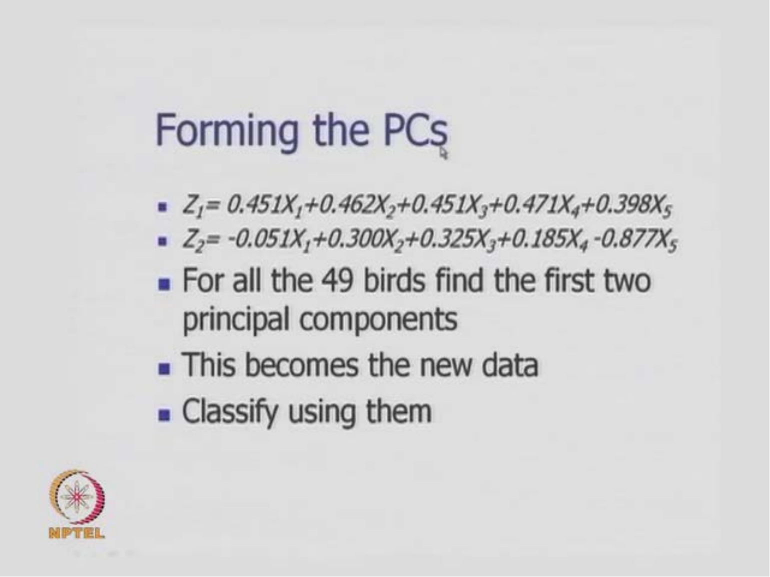
- Total variance in the data=

  $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5$

  = sum of diagonals of *R= 5*
- First eigenvalue= 3.616 ≈ 72% of total variance 5
- Second ≈ 10.6%, Third ≈ 7.7%, Fourth ≈ 6.0% and Fifth ≈ 3.3%
- ***First PC is the most important and sufficient for studying the classification***

Now we ask, which principle component are important. So, the total variance in the data can be measured as the sum of the diagonal of rank coloration coefficient matrix which comes to be 5. So, lambda 1 plus lambda 2 plus lambda 3 plus lambda 4 plus lambda 5, this is equal to sum of diagonals of R which is equal to 5. And, the 1st Eigen value is 3.616 which is 72 percent of the total variance 5 and 2nd Eigen value corresponds to 10.6 percent of the total variance. 3rd corresponds to 7.7 percent, 4th corresponds to 6.0 percent and 5th corresponds to 3.3 percent. So, 1st principle corresponds to you most important and sufficient for studying the classification because these cover 72 percent of the total variance.

(Refer Slide Time: 40:48)

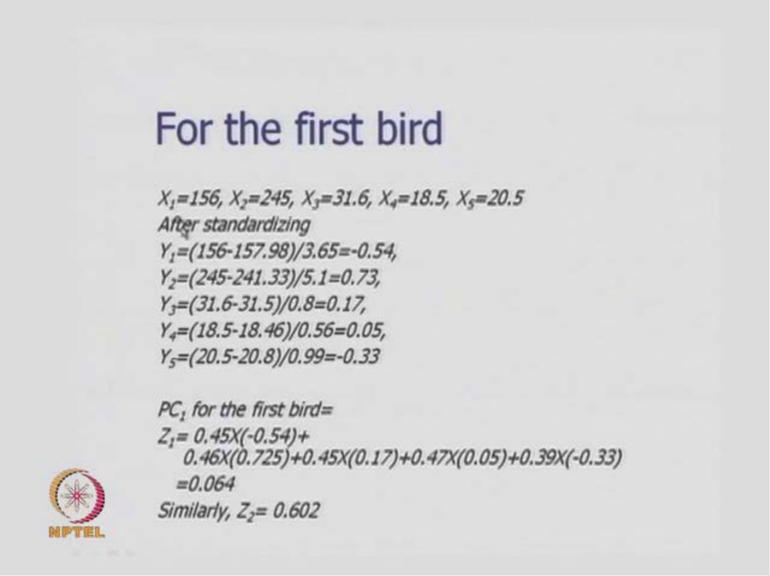


Forming the PCs

- $Z_1 = 0.451X_1+0.462X_2+0.451X_3+0.471X_4+0.398X_5$
- $Z_2 = -0.051X_1+0.300X_2+0.325X_3+0.185X_4 -0.877X_5$
- For all the 49 birds find the first two principal components
- This becomes the new data
- Classify using them

The principle components can be formed by making use of the 5 components and corresponding Eigen values. So, we can introduce two variables Z 1 and Z 2, Z 1 is found using this formula: 0.451 x 1 plus 0.462 x 2 0.451 x 3 0.471 x 4 and 0.398 x 5. Similarly, Z 2 can be founded so, for all the 49 words we can find out the 1st two principle components and this becomes the new data and we can make it classify by making use of these two components ok.

(Refer Slide Time: 41:43)



For the first bird

$X_1=156, X_2=245, X_3=31.6, X_4=18.5, X_5=20.5$

After standardizing

$Y_1=(156-157.98)/3.65=-0.54,$

$Y_2=(245-241.33)/5.1=0.73,$

$Y_3=(31.6-31.5)/0.8=0.17,$

$Y_4=(18.5-18.46)/0.56=0.05,$

$Y_5=(20.5-20.8)/0.99=-0.33$

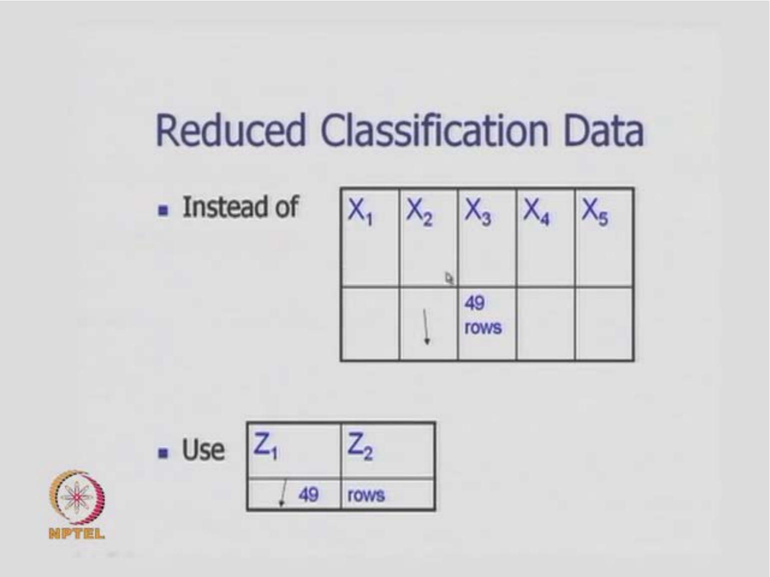
PC_1 for the first bird=

$Z_1= 0.45X(-0.54)+$
$0.46X(0.725)+0.45X(0.17)+0.47X(0.05)+0.39X(-0.33)$
$=0.064$

Similarly, $Z_2= 0.602$

So, for the 1st part x 1 is 156, x 2 is 245, x 3 is 31.6, x 4 is 18.5, x 5 is 20.5. So, these are the various value for: the beak length, body length, alar parameter and so on and so, after standardizing which is essentially saving the divagations from the attribute mean and dividing by the sigma square of the attribute, we get y 1, y 2, y 3, y 4 and y 5 as these values minus 0.54, 0.73, 0.17, 0.05 and minus 0.33. So, principle component 1 for the 1st part is z 1 equal to 0.45 into minus 0.54 plus 0.46 into 0.725 plus 0.45 into 0.17 0.47 into 0.05 0.39 into minus 0.33. Similarly, we can find the value of Z 2, as 0.602.

(Refer Slide Time: 42:51)



Reduced Classification Data

- Instead of

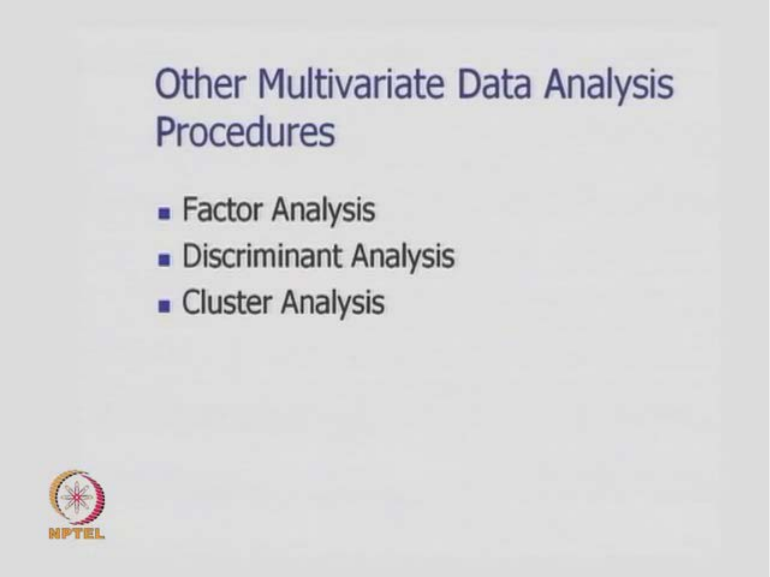
X_1	X_2	X_3	X_4	X_5
		49 rows		

- Use

Z_1	Z_2
49	rows

So, now the classification data gets a reduced form. Instead of this 49 rows with 5 attributes instead of the 5 by 49 column, we have a 2 by 49 column. So, these 5 attributes have been now reduce to two attributes ok. So data size is reduce, the f x of this 5 attributes have been distilled in to two attributes which is Z 1 and Z 2 through the technique of principle component analysis which rest on Eigen vectors and Eigen values.

(Refer Slide Time: 43:34)

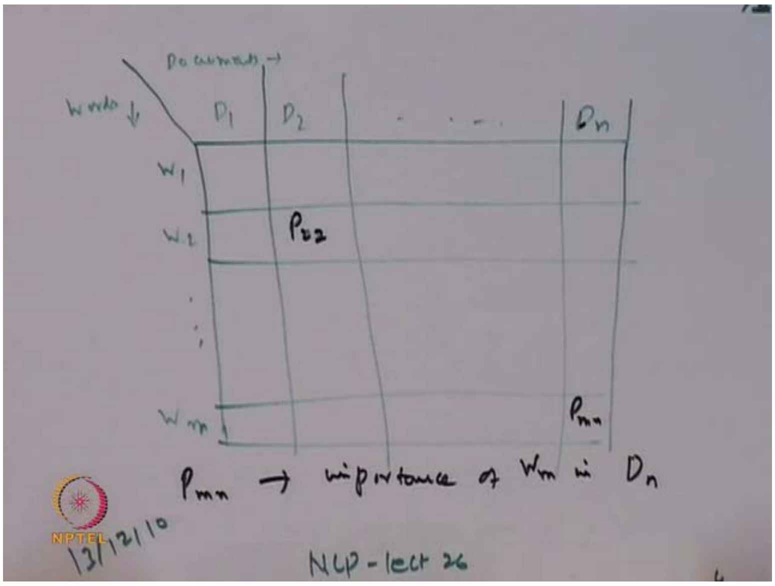


Other Multivariate Data Analysis Procedures

- Factor Analysis
- Discriminant Analysis
- Cluster Analysis

So, this is the essential idea behind principle component analysis and this forms a very important component in our discussion on latent semantic indexing. Then, other methods of multivariate data analysis like: factor analysis, discriminant analysis, cluster analysis and so on. We can take a look at them later but now we discuss a few points on latent semantic indexing which principle components analysis actually facilitates.

(Refer Slide Time: 44:20)



Words ↓ / Documents →	D_1	D_2		. . . .	D_n
W_1					
W_2		P_{22}			
.					
.					
W_m					P_{mn}

P_{mn} → importance of W_m in D_n

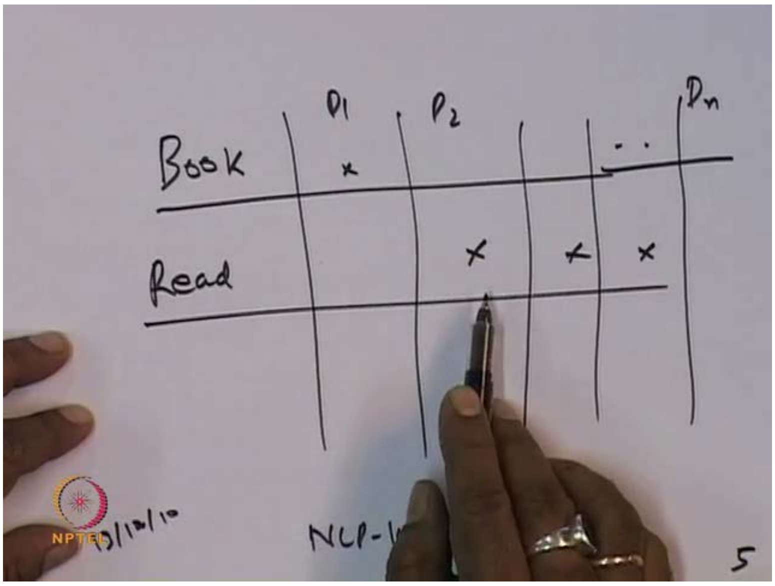
13/12/10

NLP - lect 26

In latent semantic indexing what we have is, a word versus document matrix. Suppose, there are D 1, D 2 up to D n documents and there are m words: W 1, W 2 up to W m alright. So, here are words and here are documents. So we create a large matrix in the form of words and documents. So, a number here which is let us say p 2 2, this will indicate the importance of word 2 in document 2. So, this can for example, p represented by the number of time the doc 2 appears in the document 2. Similarly, the number of times the word m appears in the document n. So, P m n is importance of W m in D n.

Now, this is a the very large matrix with all the documents and their corresponding vocabulary, the words and their appearance in this document. These matrixes also quite parts ok many of the entries are 0 in this whole large matrix and on this, if we apply principle component analysis we can reduce the size of the matrix ok only some of the words or some of the documents become important depending on whether the words are looked up in the attributes so, the documents are looked on as attributes. So, once this principle component analysis is done we find out the important words or important documents. And then, we can find out the words which are important ok, the attributes which are important that is words are the documents which are important and based on that we can find out the association between the words.

(Refer Slide Time: 46:55)



	D1	D2			Dn
Book	x			..	
Read		x	x	x	

So, assume we take two synonym words, book and very strongly associated word. Let us say read take a book and read now, the word book appears in many documents D 1, D 2 up to D n read also appears in many documents and this matrix will try to capture the as strong association between book read. So, we will go into more details of principle component analysis and the latent semantic indexing in the subsequent class and sure the strong connection between l p and i r through this.