Learning Analytics Tools

Professor Ramkumar Rajendran

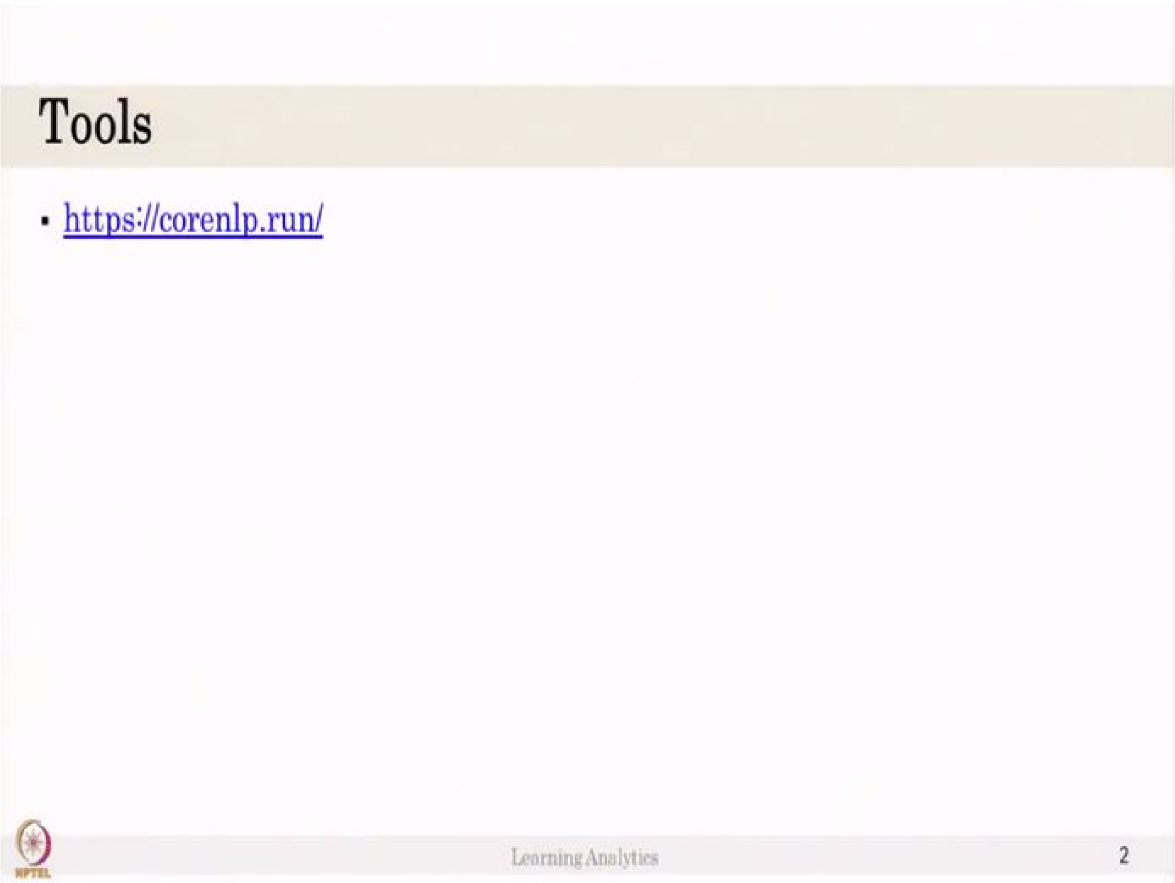
Education Technology

Indian Institute of Technology, Bombay

Lecture 10.4: NLP - Tools

In this video let us talk about NLP tools. There is one famous tool and I just want to introduce that, and also what is the latest in natural language processing, what I talked about Bag of Words is 10 years old.

(Refer Slide Time: 00:35)



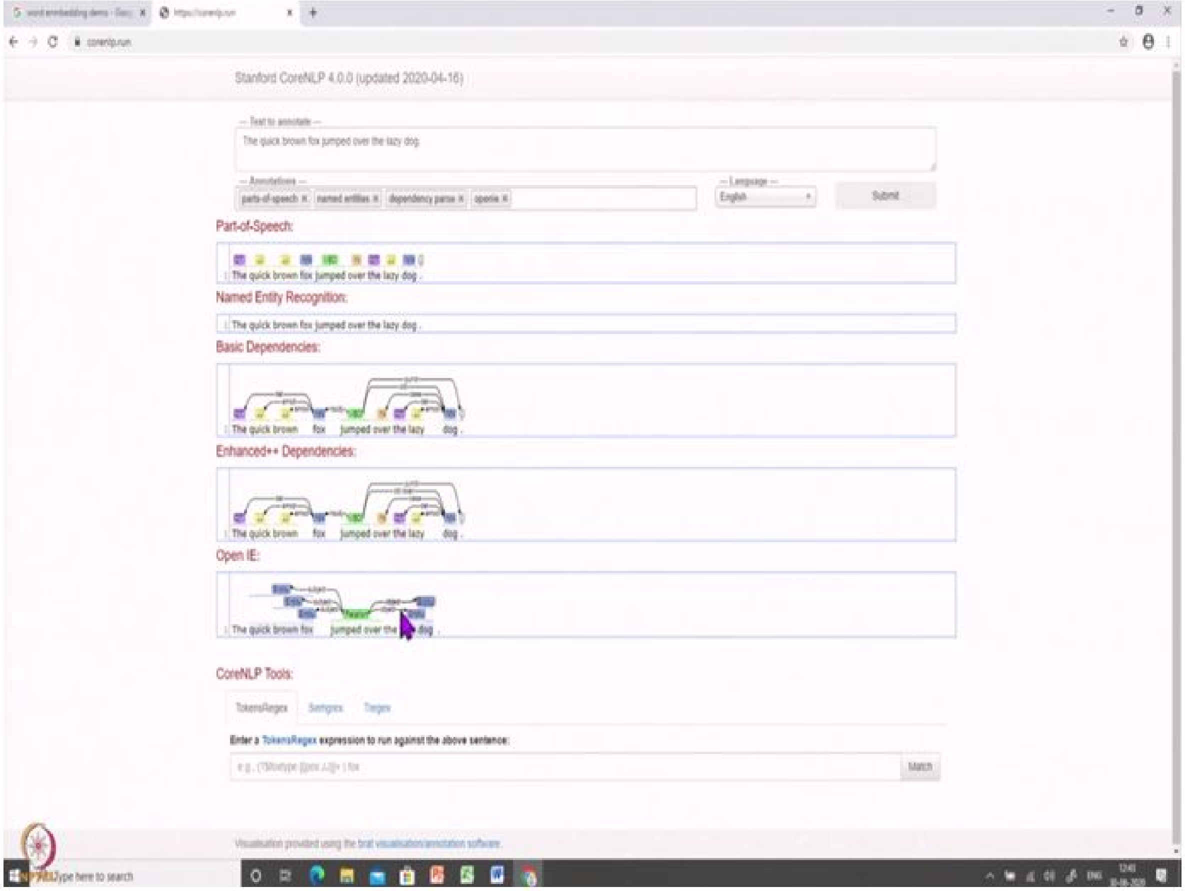
Tools

- https://corenlp.run/

Learning Analytics
2

So this is one tool "corenlp.run". This library is being created by Stanford University and is available for more than 15 years and they have started with the basic Bag of Words approach and a Java library. Now they have a sub-level and different script languages. So I was talking about how to compute get the dictionary from the given sentence. You just have to use this library it will give you automatically all these things. So this library is very, very interesting. NLP, what we saw is very basic, like the tip of the iceberg. It is not complete, what is NLP? NLP can do a lot of other things in it. I just want to show you a demo of this "corenlp.run".

(Refer Slide Time: 01:18)



Stanford CoreNLP 4.0.0 (updated 2020-04-16)

— Text to annotate —
The quick brown fox jumped over the lazy dog.

— Annotations —
parts-of-speech × named entities × dependency parse × openie ×

— Language —
English

Submit

The quick brown fox jumped over the lazy dog .

Enter a TokensRegex expression to run against the above sentence:

e.g., (?$foxtype [{pos:JJ}]+) fox

Visualisation provided using the brat visualisation/annotation software.

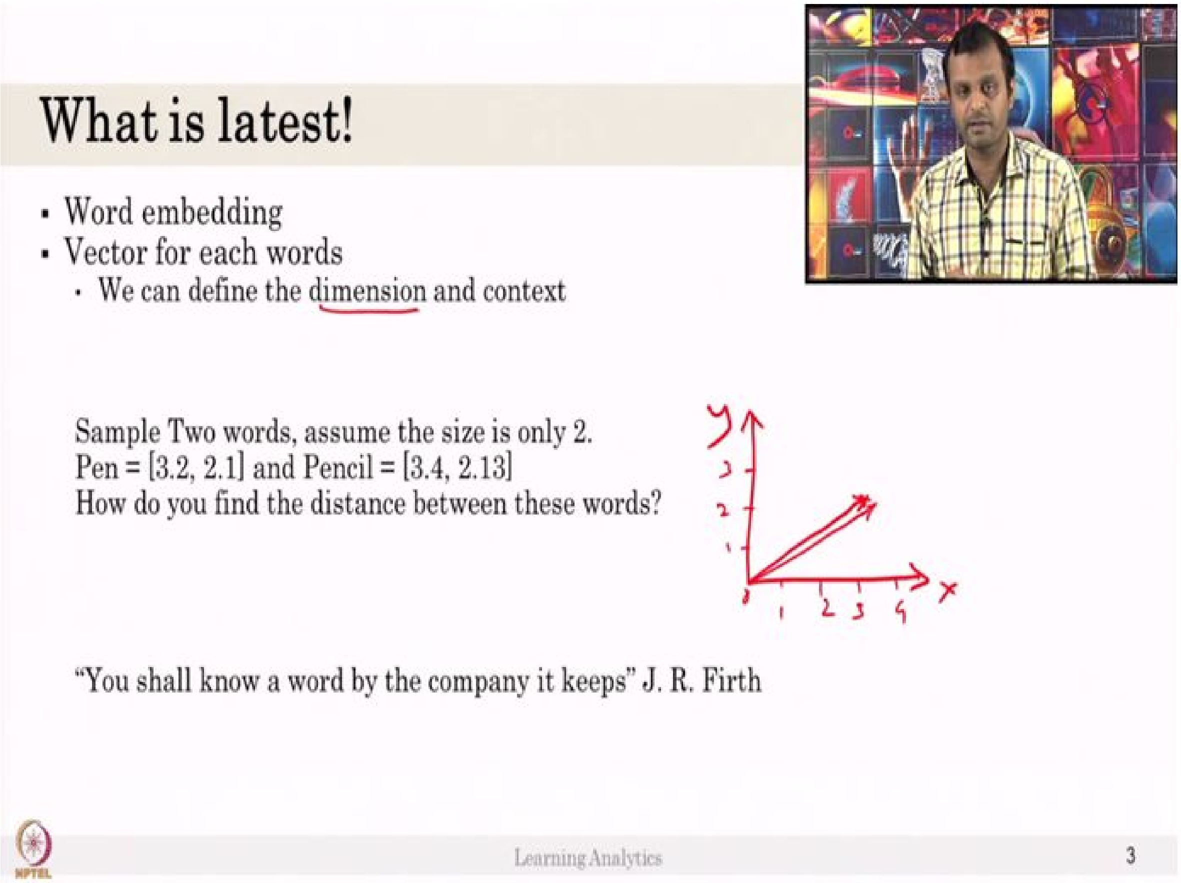
So if you go type corenlp.run, you will get this particular page you can type any text to annotate the version, you can download them. And you can do annotation based on parts of speech POS task, or named entities dependency pass, what is the dependency between these two sentences and everything. So I am going to put English. Let us put English because I just want to use the existing sentence I do not type anything.

The quick brown fox jumped over the lazy dog, you know this is a very famous sentence, it includes all the A to Z letters. It is just what we type when you try to learn typewriting you might have just started with this quick brown fox jumped over the lazy dog. So Part of Speech is kind of determine the nouns, pronouns etc. Like adjective (quick, brown) and the “fox” is a noun here and the verb is “jump”. And, so if you see it is, what I am trying to say is, this particular tool if you type the word it automatically identifies what is a noun, what is a verb, also object, or adjective.

Everything is identified automatically. So this tool has been created with a lot of training data. So you can use it directly. You just have to use this particular library and call this particular function called Part of Speech, it will give you these values. And also if you want a basic dependency like, fox is adjective on the quick brown is on the fox. What is the dependency of this “the”? Let us see, it is very interesting.

The fox jumped this is exactly the relation between the fox and the dog is jumped, that is what it exactly what it identifies. It extracts the information from this text automatically. So if you want more dependencies, enhanced dependencies are also there. And more importantly the information extraction I was talking about. So it extracts the information about these two things. So lazy dog is the entity, brown fox or quick brown fox or fox is the entity. If you consider fox is the entity, dog is the entity the relation between these two is jump. Let us go back to the next latest things in NLP.

(Refer Slide Time: 03:57)



What is latest!

- Word embedding
- Vector for each words
  - We can define the dimension and context

Sample Two words, assume the size is only 2.
Pen = [3.2, 2.1] and Pencil = [3.4, 2.13]
How do you find the distance between these words?

"You shall know a word by the company it keeps" J. R. Firth

Learning Analytics

3

So the latest thing latest in the sense latest from 2013 or 2014. The latest thing is Word embedding. So here each word is represented as the Vectors. The vector dimension can be 50, 100, 150, 200 or 300. In general 300 seems to be good, it works well for many other domains. Also in the education domain, there is a paper on that, in 2018 EDM conference which dimension is good for educational data set, it seems to be 300.

But you can fix your dimension to which dimension you want it and the context also, you want to create a word vector based on only the education data, Or you want to create a word vector based on the news dataset. Or you want to create a word vectors based on Wikipedia datasets, all are possible. So what is word vector means, they give this huge data, the data occurred in the news article, into a neural network a deep learning neural network.

And it tries to combine and get the relationship between each word based on its location and everything relation with other words, it gives you the number. The output layer in the neural network defines the dimension. So if more the dimension more the complexity, more the detailed values, less the dimension of abstracting the words, 50 seems to be good or 100 is good, so you can try different words.

If you want to create your own word vectors with your own sentence, the search program tools, everything available, just you have to use that particular program by Tomas Mikolov, I will give the link in next slide.

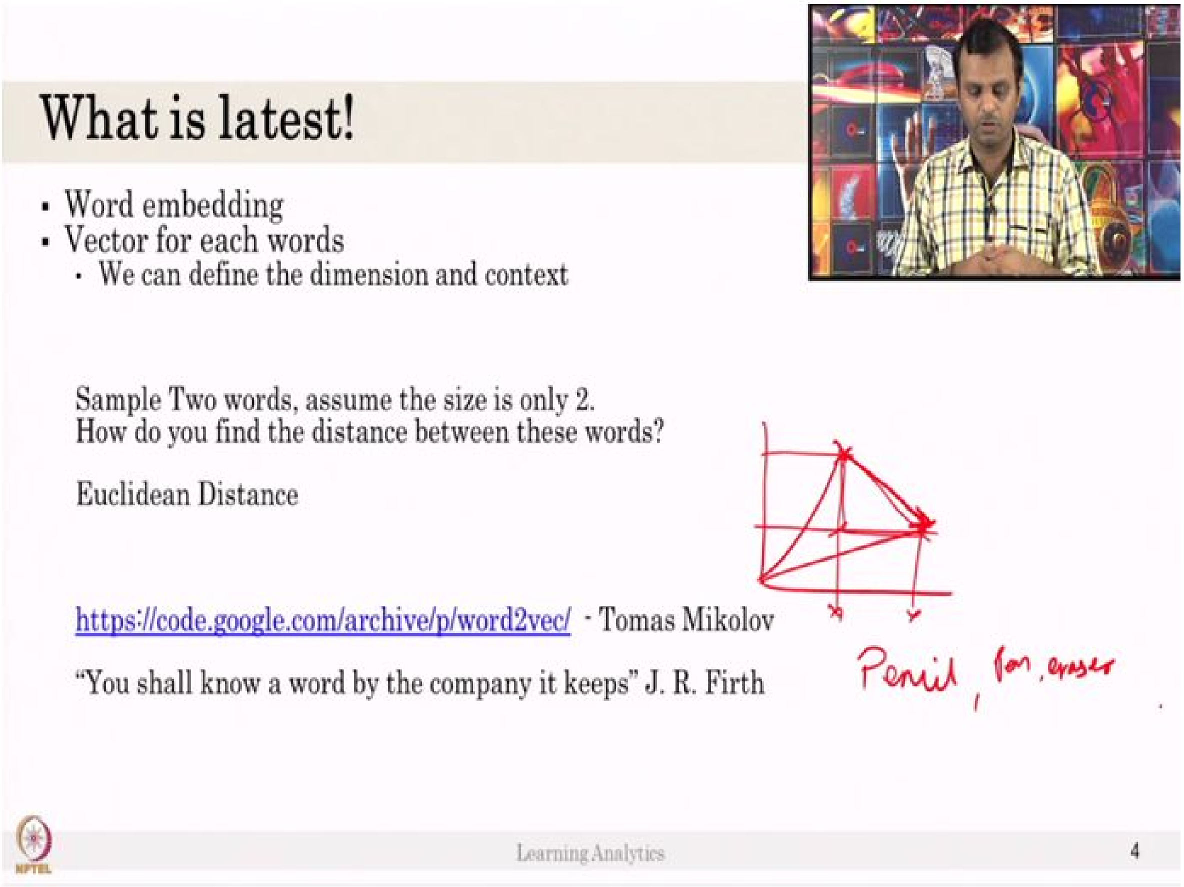
"You shall know a word by the company it keeps"

It is basically based on the words it is co-occurring with, that is the exact basic core of this. Let us see what is word vector, in a very simple example. I have two vectors. The vector size is 2 and then the dimension of the vector is 2.

I have two vectors called "pen" and "pencil". So, if the vector size is 2 hence the dimension is only 2. So the pen is (3.2,2.1), so the pen is something here, this is pen. Pen vector is from the origin. And the pencil is (3.4,2.13). Now, how closely pen and pencil related, maybe pen and pencil are stationary, it is a part of a geometry box or part of a stationary, people use it in schools and everything.

So these words might be occurring together too much in a given content or given article. So that is why it is also said like all coming together, something like that. Now the pen and pencil are very close to each other, so can you know, what is the distance between pen and pencil? Can you identify the distance? We did once to identify the distance between two vectors. That is how do you do that?

(Refer Slide Time: 07:32)



What is latest!

- Word embedding
- Vector for each words
  - We can define the dimension and context

Sample Two words, assume the size is only 2.
How do you find the distance between these words?

Euclidean Distance

https://code.google.com/archive/p/word2vec/ - Tomas Mikolov

"You shall know a word by the company it keeps" J. R. Firth

Pencil, pen, eraser

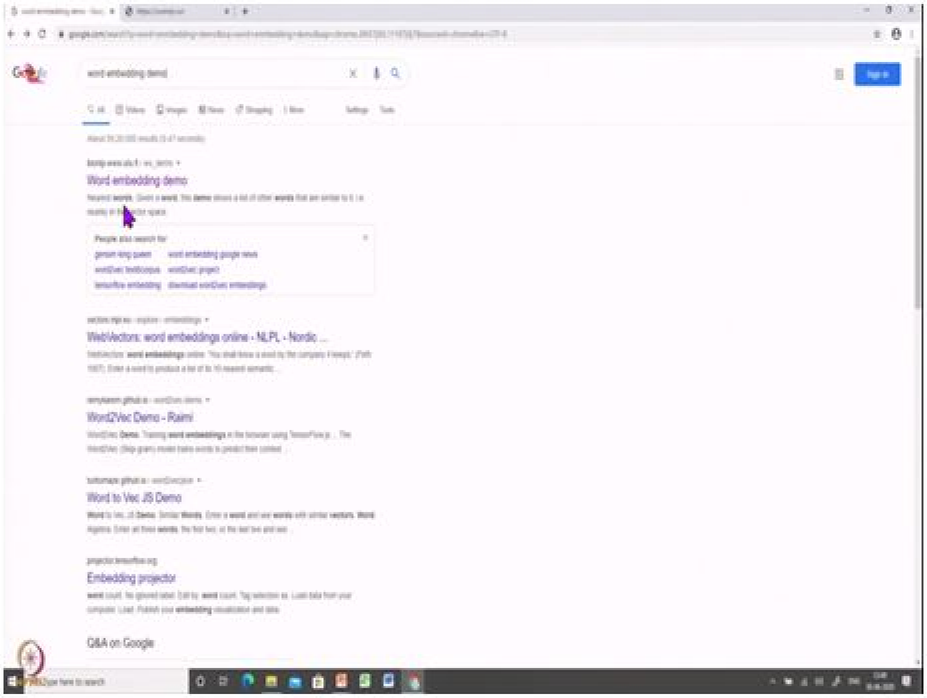
Learning Analytics 4

We did that using Euclidean distance. So this Euclidean distance actually does this. So what it does is, so now we have two vectors. Let us put it, let us put this complete graph. So this Euclidean distance between these two vectors is identified. How it identifies, if you have this origin. You know the X1 value., The difference between x-coordinates tells you this, this particular thing, the difference between y-coordinates tells you this.

If you know these value by Pythagoras Theorem then you know what is the value of this particular value distance between these two vectors, it is very simple. So, you can use the Euclidean vector to identify, to distance between these two words. So, let us say there is a word how closely is associated with pen and pencil. So, I want to give the word called pencil and I ask the system, can you bring up all the words related to the pencil? It can be a pen, eraser.

It can find all the related, how it finds a related word, what it does, it takes the vector, it finds the nearest vectors, nearest vectors computes the distance between all the other vectors, finds the nearest vector. Let us look at one tool, if you want to know more about that, please check the Tomas Mikolov's code here, code is on Google.com. Since I have a code for you to use it automatically. Let us look at that.

(Refer Slide Time: 09:13)



word embedding demo

Word embedding demo

People also search for

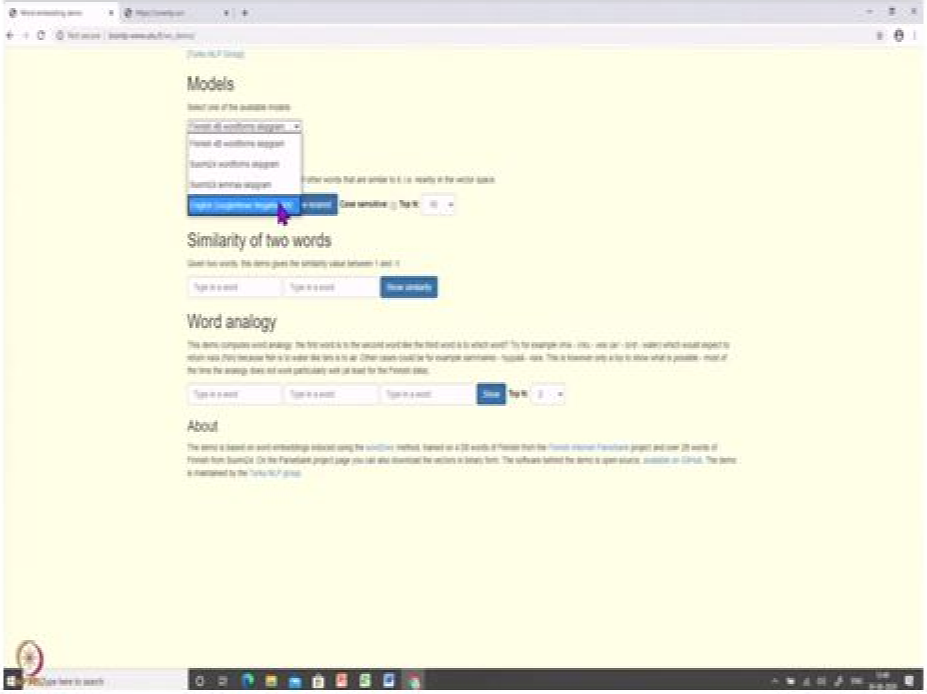
WebVectors: word embeddings online - NLPL - Nordic ...

Word2Vec Demo - Raimi

Word to Vec JS Demo

Embedding projector

Q&A on Google



Models

Similarity of two words

Word analogy

About

Let us go back to, just go to Google and type -“word embedding demo”. And just take the first one. There are a lot of demos check it. When you go there, check the English one. So lets use English Google News. Do not worry about the negative. It is a 300 dimension word. Okay, so the dimension is 300. They used Google news articles to create this statement.

So I want to use the word pencil, so let us see if I want to use pencil, show nearest words to this pencil and I want top N. So if you give this, it shows-

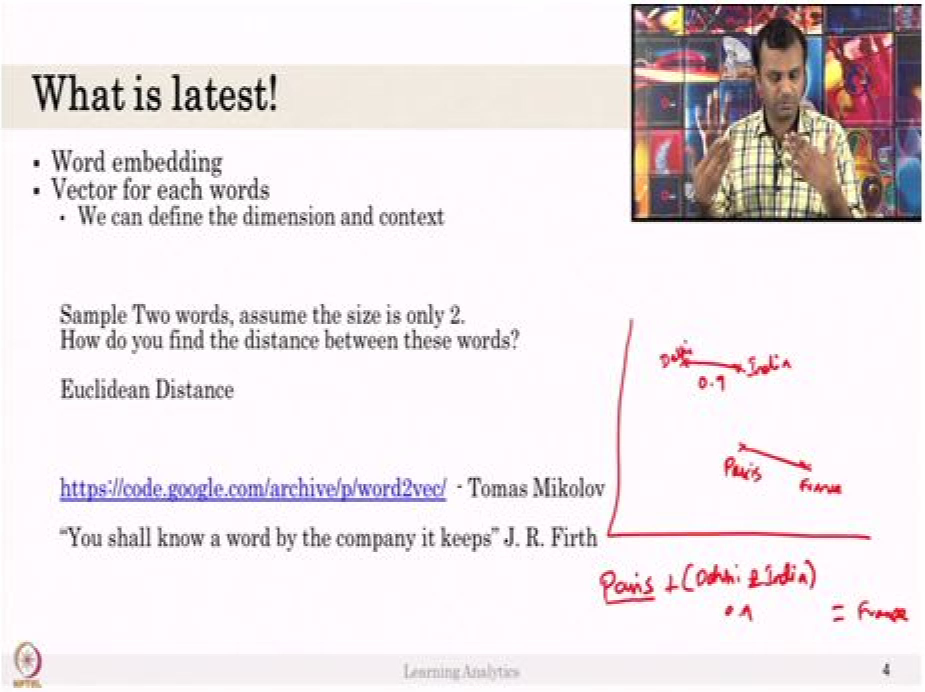
"pencils, crayon eraser, crayons, notepad, pen, scribbling, scribble paintbrush, Scribbles"

So you see if you give a pencil all the related words are coming together. You can see that users of this slide. So some people are writing works essays, they might be using related words.

So, if we use the similarity we are not able to find it. Now by using this someone might write a crayon, someone write a pencil, someone might use a pen. So by using these words, you are able to identify similar words and you might give a better grading or something like that. The most, more important thing here even not just finding the equal word. Let us see that. If you want to type some other words, type India, Delhi or any other word, an academic word, a technology word.

So I want to find Delhi is for India, what Paris is for? Delhi is capital of India, what is Paris capital of? I just want to show top words, France. How it identified, this again simple word vector. Since you have the word vector, you can use it. I will show you how it identified. Morocco is also spoken French, so Morocco is considered because a lot of people travel to Morocco, but I am not sure why Morocco, but the right word is France. If Delhi is the capital of India (what) what country is Paris capital of, it is France. It is very simple.

(Refer Slide Time: 11:54)



What is latest!

- Word embedding
- Vector for each words
  - We can define the dimension and context

Sample Two words, assume the size is only 2.
How do you find the distance between these words?

Euclidean Distance

https://code.google.com/archive/p/word2vec/ - Tomas Mikolov

"You shall know a word by the company it keeps" J. R. Firth

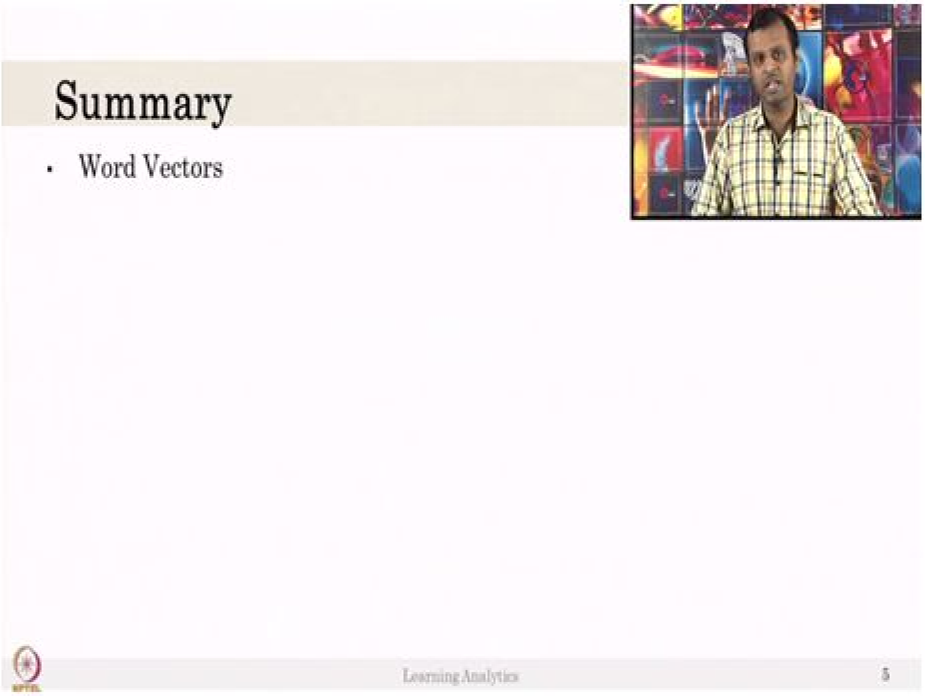
Let us go and look at how it identified. What happened here is very simple once you have the logic once you have the word vector, it is very, very simple to do it. First, what he did, it is a 300 dimension vector, it is not a two-dimensional vector, it is too big. I am just not able, I believe nobody can imagine it. So let us see there is a Delhi, there is something called India. There is a word called India.

So there is this word India or there is a word Delhi, there are two vectors, just simple words these are 300 dimension vectors. Now I identify the distance between these two. The distance can be say 0.9 something like that, very closely associated, something like that. So, now, what happens is now, if I apply Paris, the relationship between New Delhi and India this relation this 0.9 tells you whether it is a capital, whether it is enemy, or this is another country or something like that, this number tells you that.

What I have to do, I have to identify the vector Paris plus the distance between Delhi and, if I apply the distance that is 0.9, the next word will be France. The idea here is, if I have another somewhere not near Delhi in India is, another two vectors Paris and France, the relationship between these two will be the same as Delhi and India. That is it. So, check this particular tool and explore it. This is the latest.

That is since last seven-eight years, in NLP this has changed after 2014, everything we looked at NLP as a Bag of Words all these words concepts, all are gone. Now it is only vectors, no instead of million vectors, no need of million, million features, just we need a few features from the 300 words. You have to combine, add, average. You can compare to two words, you can compare two sentences if you think of how to do it. So this is the latest in NLP. I just wanted to inform you the one tool of NLP, also the latest in NLP in this video.

(Refer Slide Time: 14:18)



Summary

- Word Vectors

So that is all about Word Vectors and NLP tools. Thank you