Learning Analytics Tools

Professor Ramkumar Rajendran

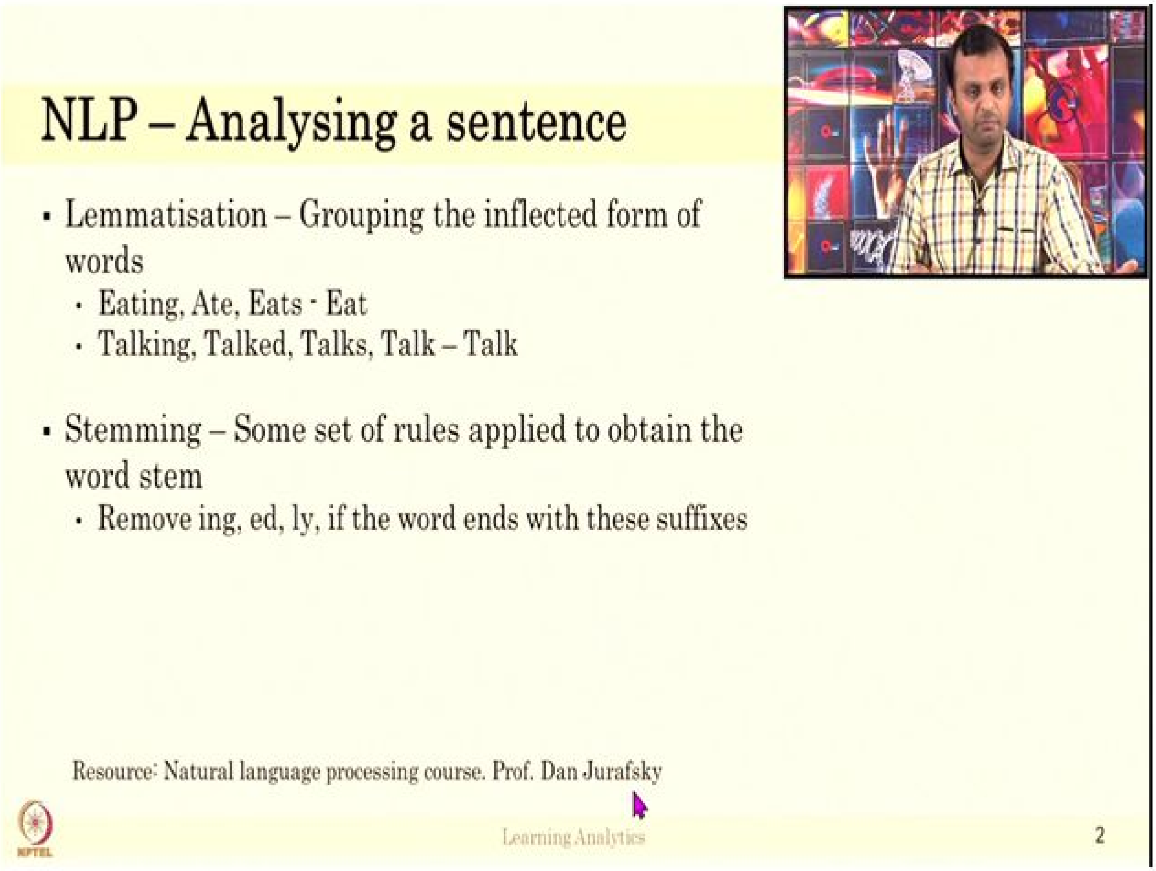
Educational Technology

Indian Institute of Technology, Bombay

Lecture 10.2: Intro to NLP

So, let us start with what is natural language processing. There will be a huge -one particular course of introduction to natural language processing itself. So, I will try to introduce a few basic concepts needed to create an automated grading system. That is what we will talk in the next video. I will introduce the concept needed for understanding how to create a system to automatically grade SA.

(Refer Slide Time: 0:45)



NLP – Analysing a sentence

- Lemmatisation – Grouping the inflected form of words
  - Eating, Ate, Eats - Eat
  - Talking, Talked, Talks, Talk – Talk
- Stemming – Some set of rules applied to obtain the word stem
  - Remove ing, ed, ly, if the word ends with these suffixes

Resource: Natural language processing course. Prof. Dan Jurafsky

Learning Analytics 2

So, let us talk about the first two concepts like lemma and stem. Lemma is, given a word, we need to identify the basic form. So, grouping the inflected form of words, for example, eating, ate, eat so eating, ate past tense or eats all these forms to a root word called "eat". It is a verb. The verb eat has been used as ate, eat, eating, eats all these things. Like talking, talked, talks, talk everything's root is talk. So, this is a lemma.

So, you have to find the root word, group those words based on that all the inflated form of these words and group into one particular form that is the lemma form. Lemma is in English it is easy because the dictionary is there. a lot of people worked on that, we have all the forms of words extensions, words inflated forms available. So, we have a huge dictionary. So, whenever word come we just use the dictionary and to find the root word and copy and paste it. So, the system is well developed to do this.

The simple thing is, in lemma is we might lose the meaning, so, we might lose the meaning of some words. So, not to do that, they use some other called stemming. Stemming is very simple rules the set of logical rules applied, the rules like remove "ing", remove "ed" or for example remove "s" or something like that. So, "ate" will be "ate" only, ate will not be converted to eat.

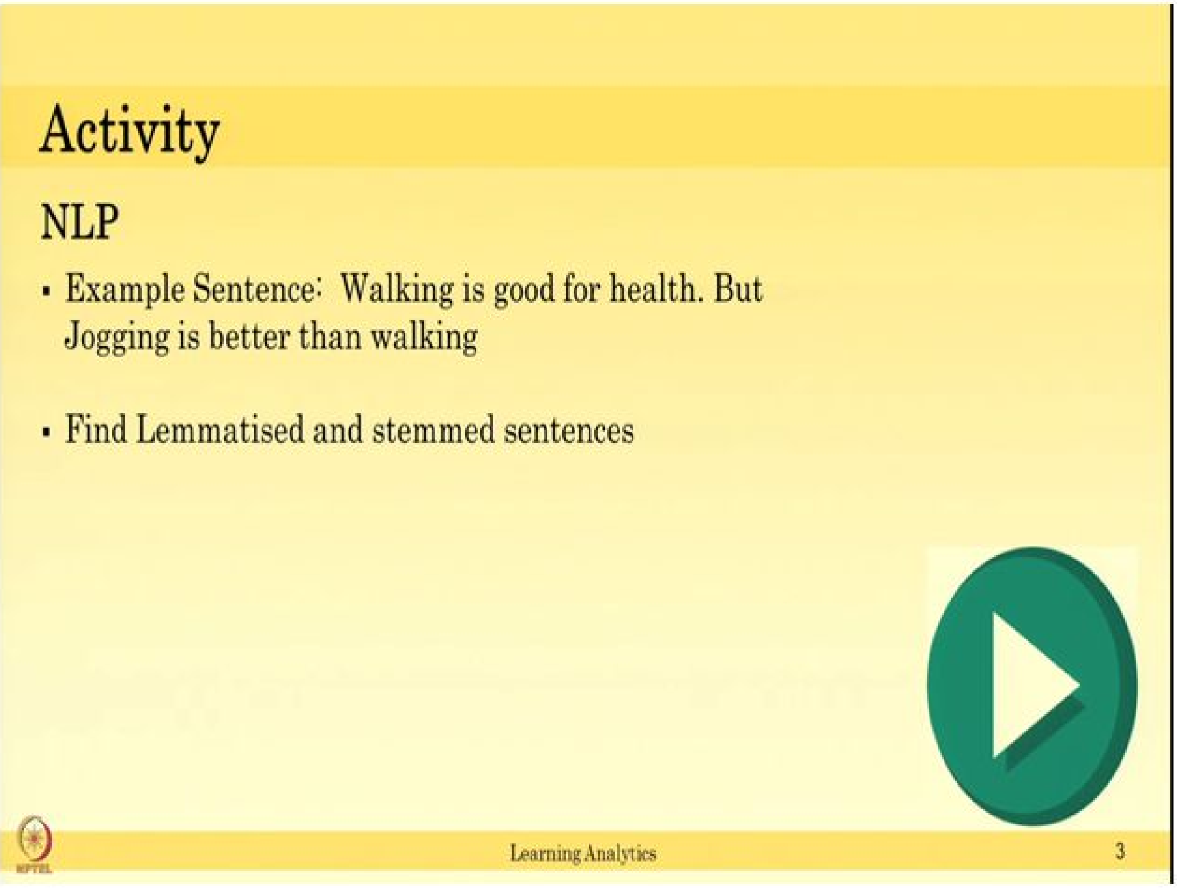
Eating, eat will be eat

talk, talked will be talk.

So, remove ing, ed, ly (if the word ends with these suffixes) that is the stem. So, you will take the stem of this, not the extra suffixes, you will be removing it. That is the thousand feet view of what is lemma and stemming. I am not talking about any mathematical forms or the algorithm behind it, that is not needed.

If you want to know more about, and find this particular video interesting, I request you to go and check the video called Natural Language Processing course by Professor Dan Jurafsky in a YouTube, it is all available freely. If you like the course, you can also just watch. But they explain completely the natural language processing traditions, how these words are to extracted and what is the information extractions, everything will be discussed in that particular course, it is an interesting course, check that out.

(Refer Slide Time: 3:14)



Activity

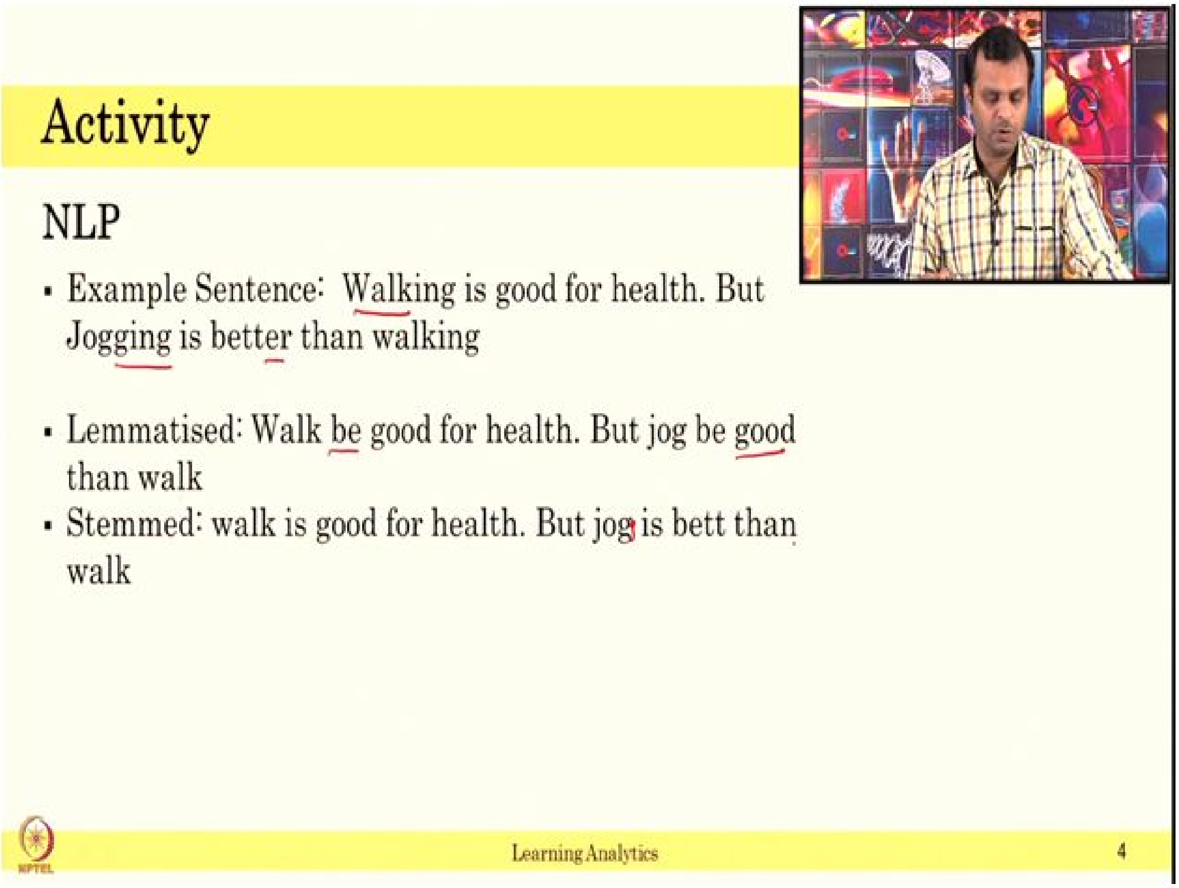
NLP

- Example Sentence: Walking is good for health. But Jogging is better than walking
- Find Lemmatised and stemmed sentences

So, since you saw what is lemma and stem in a basic form, I just want you to give a small activity. Consider sentence “walking is good for health, but jogging is better than walking”

Can you create a lemma, lemmatised form of this sentence also stemmed sentences from this sentence, stemmed words from the sentences. So, pause this video, write down your answers then show the video to continue.

(Refer Slide Time: 3:40)



Activity

NLP

- Example Sentence: Walking is good for health. But Jogging is better than walking
- Lemmatised: Walk be good for health. But jog be good than walk
- Stemmed: walk is good for health. But jog is bett than walk

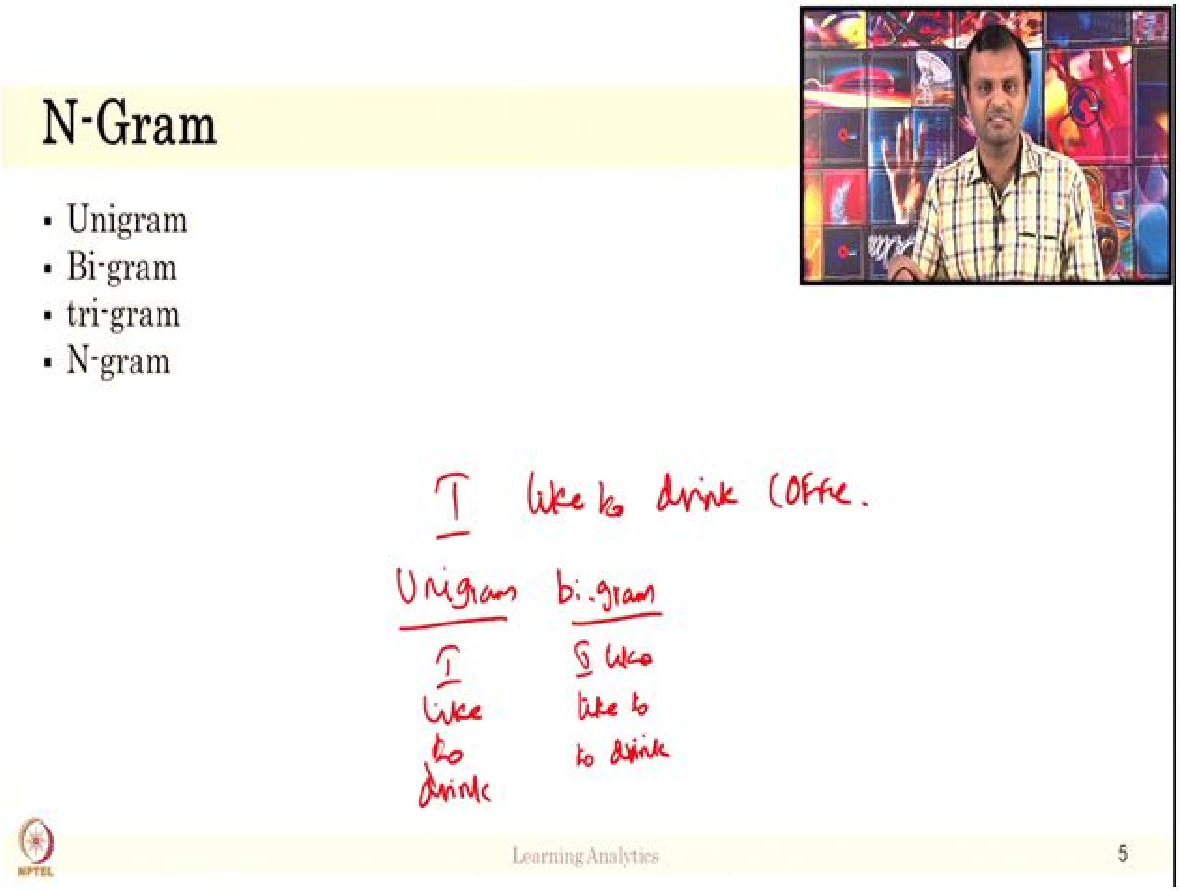
So, the lemmatised form is, it is simple, we know lemma is root form. So walking, the root form is walk. "Is" - root form is "be". Good for health.

But "jogging" will be "jog".

Be is again be,

better is actually a form of good. Then walk. This is a lemmatised form. Stem form is a bit different. It is good for health. But jog it, remove the jogging because you know jogging means there is a g which means you are removing not "ing" but ging. Sometimes if you do not have that rule, it will have a jogg(j -o- g- g), that is it. So, that is a problem. We do not know how to remove the words for better, then walk. So, that is a difference between lemma and stem.

(Refer Slide Time: 4:42)



N-Gram

- Unigram
- Bi-gram
- tri-gram
- N-gram

I like to drink coffe.

Unigram	bi-gram
I	I like
like	like to
to	to drink
drink	

Learning Analytics

5

But you know, the concept in NLP is very famous, is called the N-gram. It is, it is basically a unigram, bi-gram, tri-gram and N-gram. Given a sentence say “I liked to drink coffee”, something like that. So, if I say unigram, I want to create a dictionary with all the single words in it. So the dictionary will be unigram.

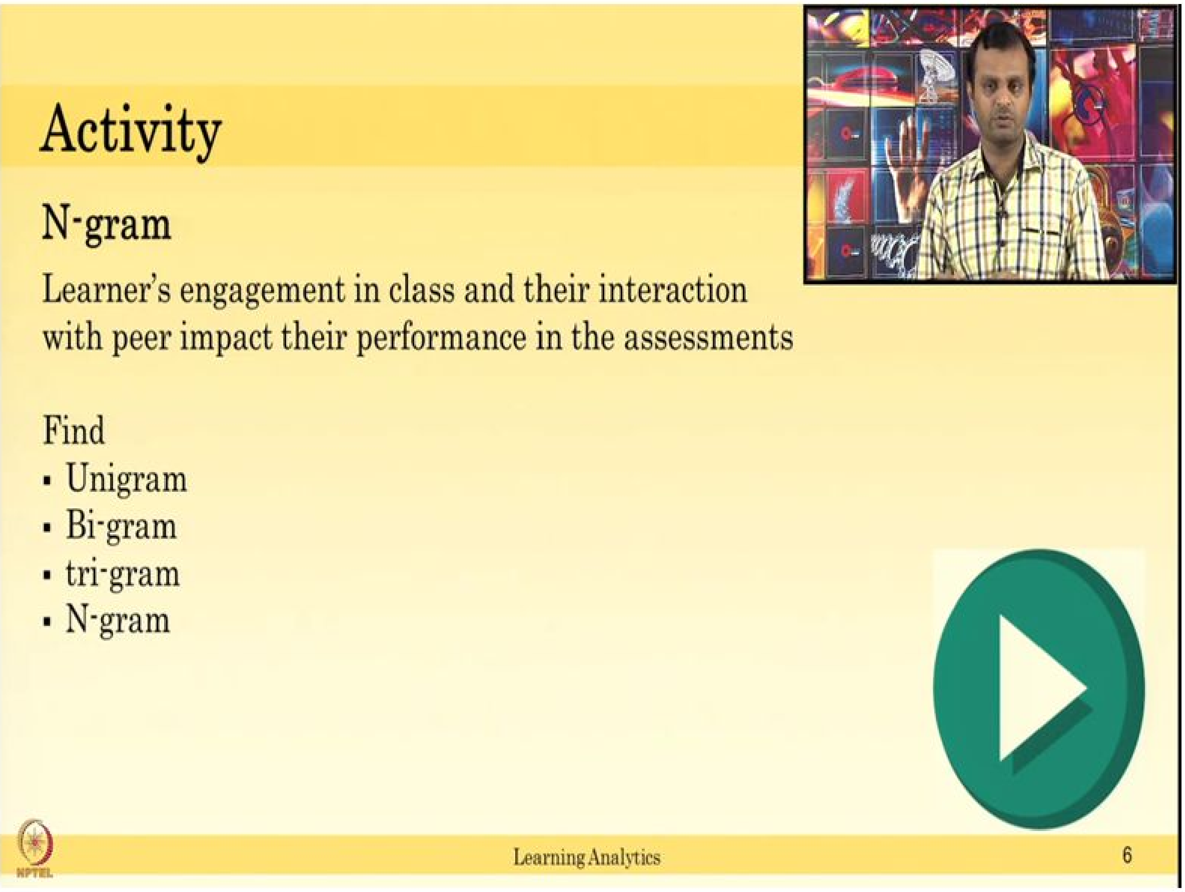
“I like to drink coffee”

Something like that. In bigram, I want all the two words together in the dictionary, like - “I like” is the one. “Like to” is the other. “To drink” is the third combination of the bi-gram in my dictionary, “drink coffee” is the one one more. If it is tri-grams, I like to, like to drink, to drink coffee, something comes in. So, unigram, bi-gram is used for applications like Google. I tell you where exactly.

But let us see, let us see how this can be used. So what is unigram, bi-gram tells us, like, given a lot of words, lot of content exists in the world, you can go call all the content in Wikipedia, Wikipedia database is available free. You can download and use it. Google News database or something like that. Can we create a dictionary of what is a combination of words that are occurring together? That is the idea of bigram. Tri-gram, it is not needed necessarily, but sometimes there is a word which is three-sentence, computer science and engineering.

So, computer, science and engineering is different, but computer science and engineering mean three words are needed to get that word. It is not necessary, but let us see how it works. And there is something called n-gram. So when anything is n-word collection, we call it as N-gram, like four-gram, five gram, six gram, something like that.

(Refer Slide Time: 6:53)



Activity

N-gram

Learner's engagement in class and their interaction with peer impact their performance in the assessments

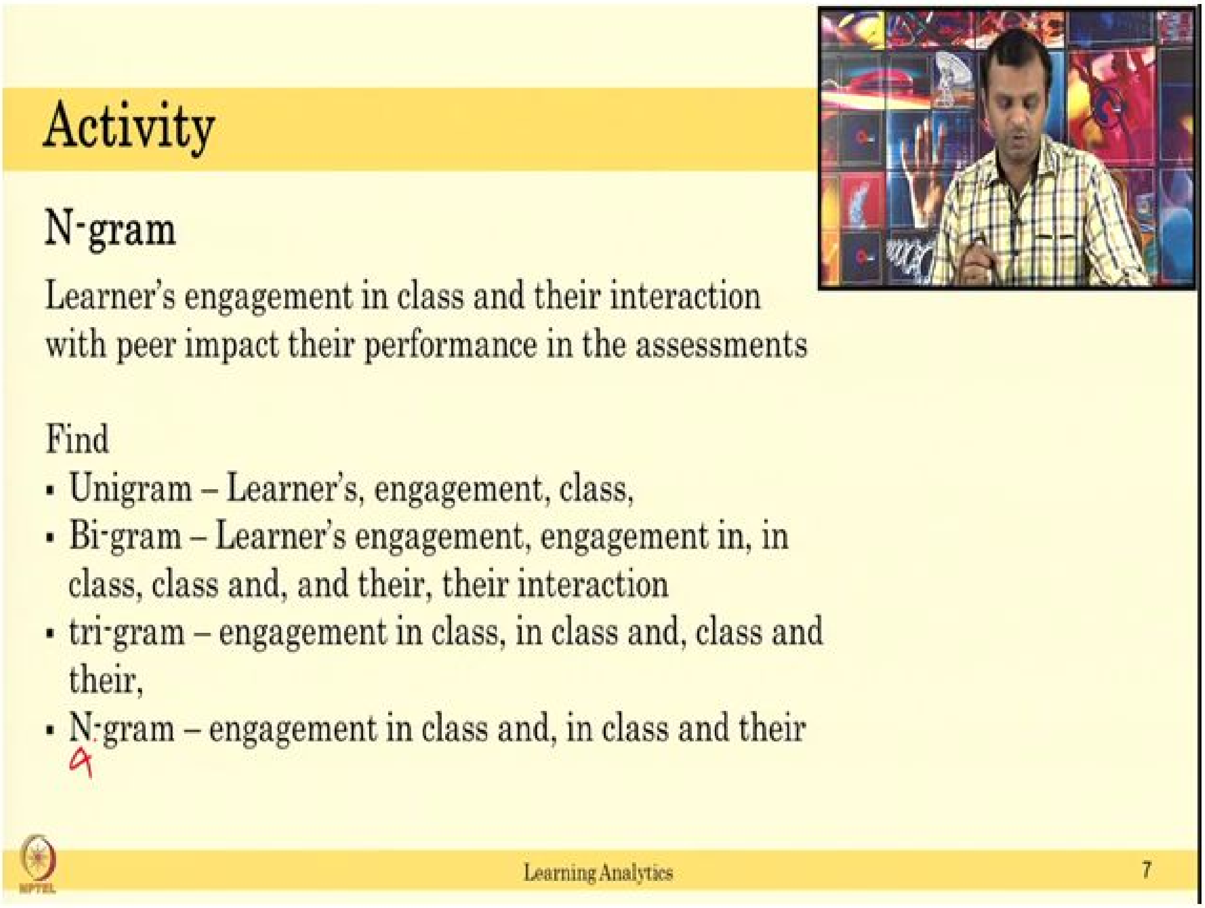
Find
- Unigram
- Bi-gram
- tri-gram
- N-gram

Learning Analytics 6

So, in this activity, check this sentence, “learner’s engagement in class and their interaction with peer is, peer impacts the performance in the assessment.”

The sentence, maybe not right, but consider learner’s engagement in the, in class and their interaction with peer, is peer impacts their performance in the assessments, some sentence like this. Can you find out the unigram, bi-gram tri-gram of this? Not N-gram. And if you want to do it, just go for four-gram, but not beyond that. Can you find this unigram, bi-gram, tri-gram for this sentence? So, after you do it, please touch in the video to continue.

(Refer Slide Time: 7:35)



Activity

N-gram

Learner’s engagement in class and their interaction with peer impact their performance in the assessments

Find

- Unigram – Learner’s, engagement, class,
- Bi-gram – Learner’s engagement, engagement in, in class, class and, and their, their interaction
- tri-gram – engagement in class, in class and, class and their,
- N-gram – engagement in class and, in class and their

4

Learning Analytics 7

So, the unigram will be

learner's

engagement

the

class etc…

Bi-gram will be

learner’s engagement,

engagement in,

in class, etc

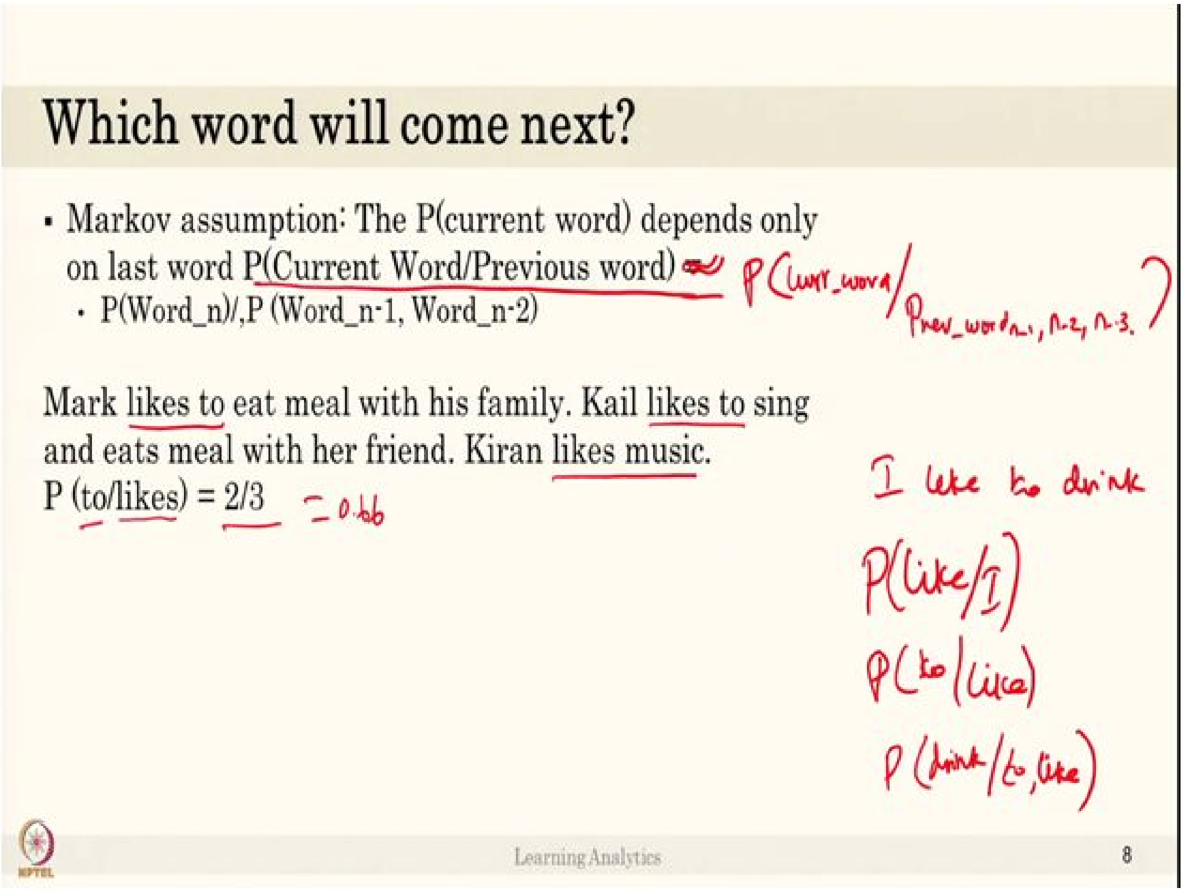
everything will be a bi-gram. So, tri-gram will be

learner’s engagement in,

engagement in the etc.

So we are actually moving, as such we, and shifting the windows like sliding the window slowly. So, the window is sliding. So, you can take the first three words, the next three words, the next three words. It is like that. So, tri-gram is there. Why we are doing this, it is very important. I will show you why we are doing it, it is very simple. Let us see.

(Refer Slide Time: 8:40)



Which word will come next?

- Markov assumption: The P(current word) depends only on last word P(Current Word/Previous word) = $P(\text{curr\_word}/\text{Prev\_word}_{n-1, n-2, n-3.})$
  - P(Word_n)/,P (Word_n-1, Word_n-2)

Mark likes to eat meal with his family. Kail likes to sing and eats meal with her friend. Kiran likes music.

P (to/likes) = 2/3 = 0.66

I like to drink

$P(\text{like}/I)$

$P(\text{to}|\text{like})$

$P(\text{drink}/\text{to}, \text{like})$

Learning Analytics 8

So, if you want to know what word will come next, how you can do that?

Have you wondered when you type in Google, you type the first word, it picks up and shows the second word. How the second word automatically appears? In Google Search engine is not just NLP. It based on the context where you are from and what is your personal profile, what kind of words you searched already, there is a lot of modelling of you happening in the search engine itself.

But the basic idea is NLP. So, all it tells you that the current word, that is a Markov assumption, okay. I told you about hidden Markov model last week, but that is Markov assumption is basically this. The current word depends upon the last word. So, I just go for the same example, I like to drink. What is the probability of like occurring if the given word is I? What is the probability of the second word would be like?

If you consider the English dictionary, a lot of sentences, the second word after I will be “am”, not like. But maybe in sentences if you have words like, I want, that kind may also occur. So that might have some, some probability. But the probability of to, given the current word “like” is high. Because most “like” fare followed by “to” because that is how all the sentences formed in the content like newspapers, Wikipedia.

So, from the content database, you can construct this probability So, you remember the text transition we talked about, we can create that kind of probability. That is what exactly is happening here. So, we are creating that. For example, probability of drink, given “to” is low. But consider if I have a probability of, the previous word is to, also that previous to the previous word is like to.

So, think about a sentence - “I like to go”, “I like to drink”, “I like to eat”. So, now there is an only certain form of things are coming, though, i.e. instead of “to drink” a lot of other words possible. Now, we have - eat, drink, go, all this thing. Now we can have a better probability. So, what happens is if you have more sentences given in the probability of the previous sentence, if you know about history like n, n -1, n - 2, n - 3, you might able to better predict it.

But Markov's assumption says that it is not needed. It mostly predict all these previous sentences by the previous word only. No need to give previous word n -1, n - 2, n - 3 words. The Markov assumption says that there is no need to do that. It is almost equal to considering the probability of only the previous word. So, considering the probability of current word, its previous word is enough and it is almost equal to the probability of a current word given previous word, n -1, n - 2, n - 3.

He says that it is no need to do that. So, that is what Markov’s assumption says. So, the current event depends mostly on the previous event. That is an assumption. We have seen this, I talked about bi-gram everything in a previous slide. This is a special form of that if you have done a bi-gram. You have a dictionary of all the words occurring together.

You have to compute the probability, just once you compute it, automatically the probability of which word will come next if you have the first word is “to”. So, what happens is when you type in a Google, so when you type the first word, it automatically picks up based on the probability of these words occurred in the content. But in Google search is different because based on the current trend, what are the people searching in the particular context and what is the latest news in the particular the IP address location, all these things is considered.

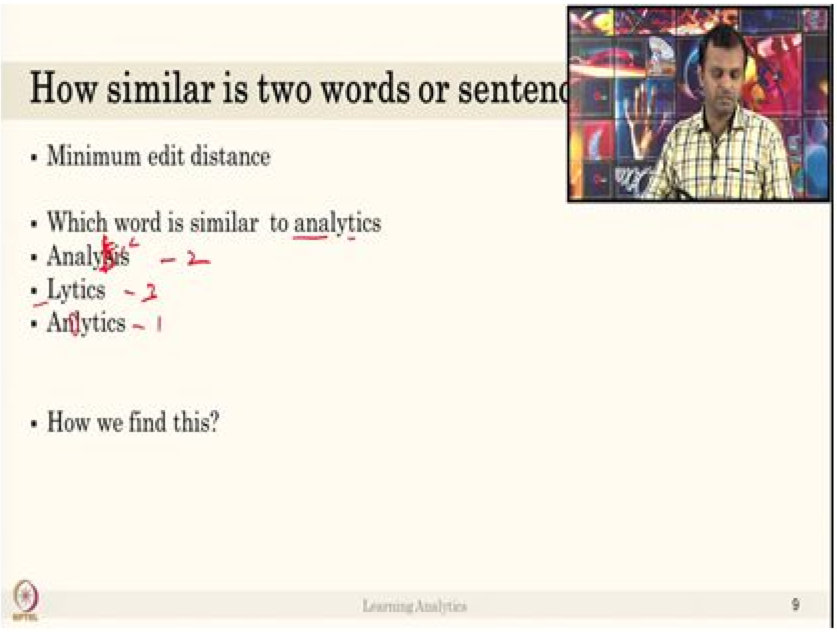
But in general, the idea is the finding the probability of current word, given the previous word,. So, let us look at one simple example to understand that more clearly. So, Mark likes to eat a meal with his family. Kail likes to sing and eat, and eats meal with her friends. Kiran likes music.

In your bi-gram dictionary, you will have - likes to, likes to and likes music.. How many words start with likes, three words. But in how many the second word is “to”? Two, i.e. “likes to”, “likes to”. So, two by three. The probability is 0.66, very simple. So, what happens now is when you type the word likes it automatically suggest you the word to, keyword was fixed up. So, that is a basic, very basic form of what word comes next. So, hope you understand this.

So, this is also a kind of a probability and finding the state transition things. But given a huge dictionary, a lot of content in English language ist available on the internet compared to most of the other languages. So, identifying this dictionary is not easy. So, you can go ahead and take all the content in Wikipedia or Google News and create a bi-gram. It is easy because you might use system and support.

Tri-gram and four-gram is needed sometime because that might give you better next word. It is not easy. It needs a lot of computational power, but Google already did it. Actually, they have four-gram database with them. But actually, Markov’s assumption says it bi-gram is enough.

(Refer Slide Time: 15:47)



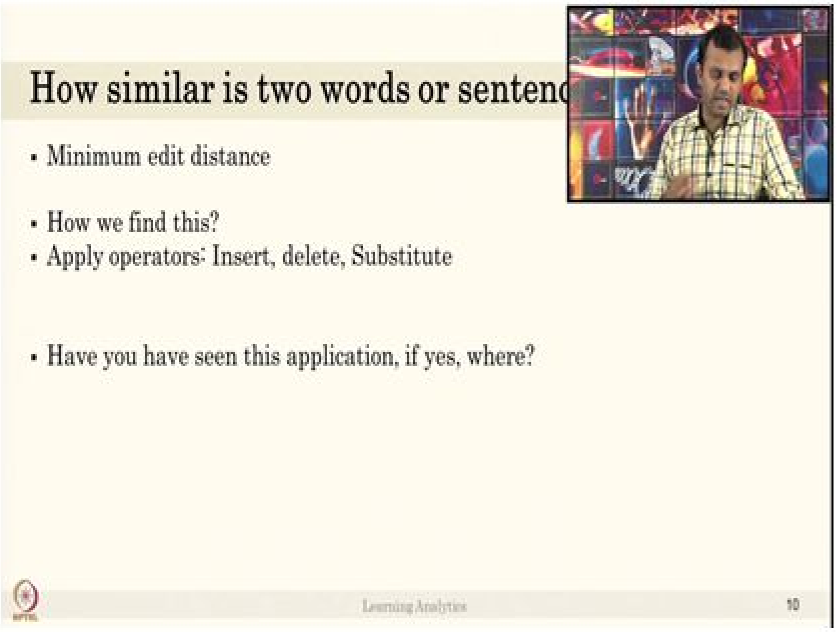
How similar is two words or senten

- Minimum edit distance

- Which word is similar to analytics
- Analysis – 2
- Lytics – 2
- Anlytics – 1

- How we find this?

Learning Analytics 9



How similar is two words or senten

- Minimum edit distance

- How we find this?
- Apply operators: Insert, delete, Substitute

- Have you have seen this application, if yes, where?

Learning Analytics 10

Let us look at the other concept. How similar is two words or sentences, is the next concept. So, let us see, there are some words. Analytics and other words, how similar these words given to this particular thing. So the similarity is computed by minimum edit distance. So, how much minimum edits you have to do to get this particular word to this word? So, "analysis" is almost same. So, t should be replaced with s and c should be inserted.

So, two edits you have to do. So this is two edits. Lytics the three edits are needed, “anlytics” that is one edit is needed. Let us see. How do you find this? There is a simple rule that is called you have to apply operators, operators like insert, delete, substitute and, and that is how the similarities are identified. Let us look at the example again. So, here “analysis”, I am inserting this word, I am replacing this word, substituting s with t, okay that is called substitute.

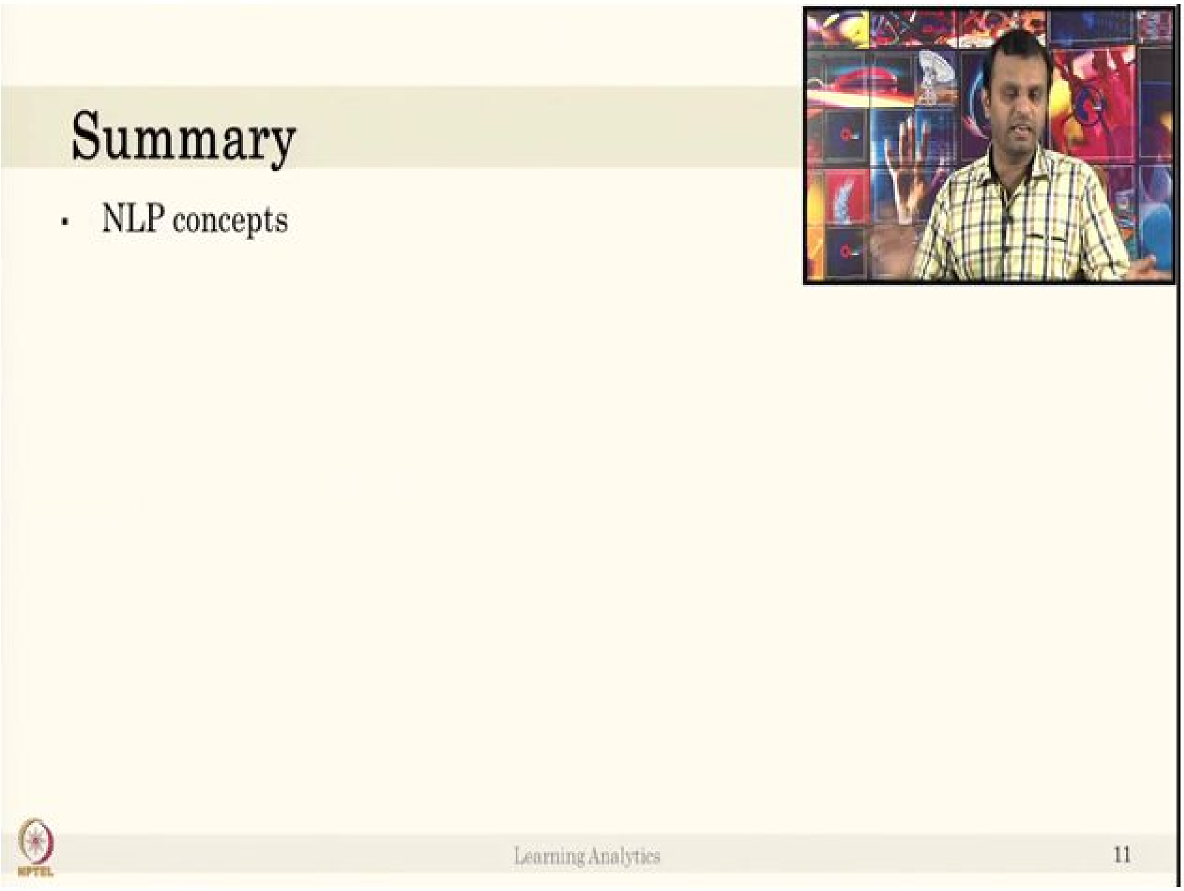
And I am inserting the word c, so I did one substitute one insert. So, there are two operations needed so two minimum edit distance. Here “lytics” is there, so I might need to insert three words, so insert three actions. And here only one word is missing. So, I am inserting only one word.. It is like I need to understand “lytics” is actually occurring in the last six words.

So, which means there should be a big matrix and comparing each and everything where it is matching. It is, it is kind of looks complex. But if you have the proper algorithm, it is not that much complex. It is easy to do it. So that is how the minimum edit distance is computed. So, how similar two words, it is basically based on this? This is very, very useful in bioinformatics.

There a lot of work on this particular exactly using the minimum edit distance. A lot of work has been going on and there is innovation happening in that field. If you are interested, go and check minimum edit distance and by informatics, you might find the answer. So, have you seen this application? Have you ever seen this application? If yes, where? Think about it. Again in Google search, you can type anything. It actually picks up the right word and shows out.

It also depends not just a minimum edit distance. Also the current trend in the particular location area. But in general, if you type the word it, it automatically corrects the sentence. But even in Power Point, I am using PowerPoint. It will show the wrong sentence in the red colour underlined. That is exactly right. So, in a word, corrections in PowerPoint at this has been used. So, you have seen this everywhere. This is the basic concepts I want you to understand in NLP.

(Refer Slide Time: 19:11)



Summary

- NLP concepts

Learning Analytics 11

Why I am talking about these concepts. I want you to use these concepts in order to create some automated grading system in the next talk, next video. Thank you.