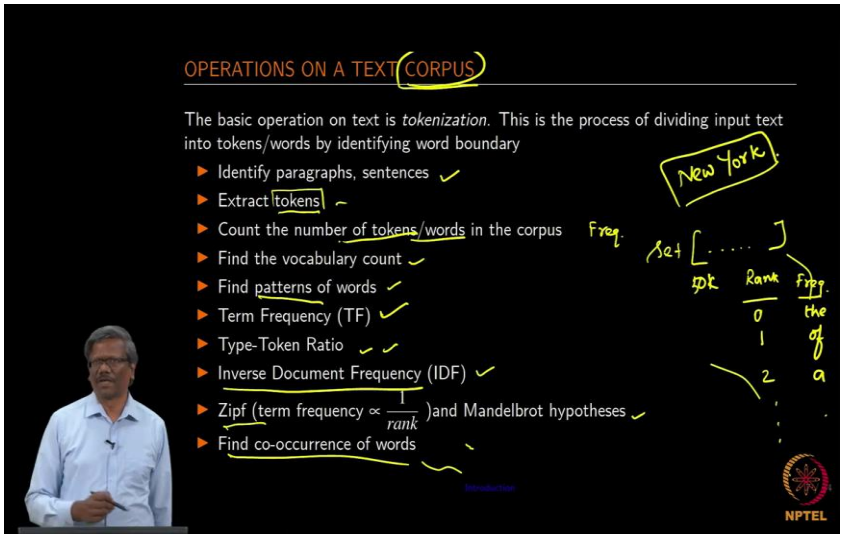
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 02
Operations on a Corpus

(Refer Slide Time: 00:15)



So, what I am going to be doing is I am going to be taking you through the 12 weeks in the next few minutes right. Whatever I am going to be doing in the next 12 weeks I will be given adjust of that in the next few minutes and then we will also see how all of these are connected and then the answer for why we are doing this would come into, then the answer for why we are doing this would be very clear at the end of this ok.

So, before that you know we will start looking at what kinds of things that we can do with other text that we have in hand. So, before that let me first describe what this corpus is. So, the corpus is nothing, but a collection of documents for example, you have been studying for certain topics let us say an operating system in computer science, you start collecting various information related to chapter the from the web some research papers and so on.

So, at the end of course, you know you might be having about you know 100 or 115 documents that you have in hand. So, that collection is called a corpus and you know you can name that corpus as an operating system corpus. So, that contains the relevant

information related to certain topics. So, in this natural language processing application we will always be dealing with certain corpus, we require a large collection of documents in order for us to deal with certain applications ok.

So, that is about the corpus ok, what else I can do with that corpus? So, when you are given a text you know for the first time what you will start looking at right. So, when you start looking at that you will say I am seeing a lot of keywords and then you will start underlining certain new words and then you will start underlining certain words that you do not understand or do not know the meaning of that or you will start underlining some important sentences as the next step and soon right.

So, we will start looking at the information from a very high level and then slowly start getting into a deeper and deeper level and at the end of that you would know what the document is all about right. So, the same thing that we are going to be doing in the entire course. So, we are going to be looking at the documents at a very high level, to begin with, and then start looking at it at a deeper level as we go further down the course.

So, for all of us to do certain operations on that corpus, we require certain tools and techniques right. So, the first thing that we do is we will just look at the documents from a very high level and since we are talking about the corpus it's going to be having lots of words in the corpus right. So, that is the first thing that we want to look at it contains about 1 billion words, it contains 500 million words, it contains about 50,000 vocabulary rest of the words are all repeated or repetitions of the vocabulary.

Certain words are useful in connecting these sentences, but not really useful in terms of understanding the content of the documents you know, for example, the of in you know we call them as stop words and we say they are not going to be really providing me some good information related to the content. So, I can probably remove them now we call them to stop words we will go through that one by one you know in this.

So, first what we do is we start identifying the paragraphs and sentences in the document. The second one that we look at is ok; so how many words are in the paragraph? We call them tokens here right. So, how many tokens are there? The tokens are nothing, but word separated by space and then we start counting the number of tokens. So, how many words are found in the corpus and you know we start looking at the frequency of the terms.

For example in this particular slide how many times tokens the word token has appeared? So, that is called the frequency of that see that is the number of tokens or words. So, in this slide we have about we have three. So, in this slide tokens have appeared three times. So, the entire collection of my 12 weeks PowerPoint presentation or PDF this word has occurred at least 300 times ok. So, that something that we start counting at a very high level without really and knowing what the meaning is right. So, we just start counting the words.

So, that is where we start looking at the frequency of the word, and then we start looking at the vocabulary count; vocabulary count is nothing, but the set of words right. So, right you just take the set of all the words that you have will give you only the unique words that it found in the entire corpus. So, most of the time the vocabulary for a very large corpus would be around 50 k and then find patterns of the words.

So, how many times certain words occur together? So, how many times and there is all two word that occur always together? Three words that always occur together. So, that is something that we want to look at. So, we will just look at this particular pattern that occurs a few times in this document.

So, as I mentioned we look at the term frequency and then we also look at you know based on the vocabulary and the total number of words in the document, we try to find out what could be the type-token ratio, is it going to be a very it's going to be the same for all corpus or it's going to be very different. So, there are the certain empirical formula that we start looking at by just doing the statistical counting of all these words. And then we start looking at how many times this particular word had occurred in the entire corpus.

So, we use some ratio to figure out this inverse document frequency we will talk about that in little later in the next as a session and so on. And then there are other empirical formulas or hypotheses that are available they are called Zipf and Mandelbrot. So, Zipf tells the term frequency is proportional to the rank. For example, when you start ranking them in terms of the frequency right the highest frequency will be let us say start with 0 the will have the highest frequency. This is the most commonly occurring word right and then 1 which will be let us say of and so on ok.

So, what Zipf does is it tries to find out the relationship between the term frequency and rank. Mandelbrot and also is very similar to what Zipf does and then we also tried to figure out what are the co-occurrences of words; that means, as I mentioned earlier. In terms of the patterns we try to find out these two words always occur together.

For example, you will find the words New York right. So, you can find the entire corpus how many times this is two words that occur together. There are since we are blindly doing it we do not know what exactly the meaning of all those we are just making the count of that. So, this is the very high-level task that we do when the corpus is given to you right. So, this is the first step.

Any, if you give any document the first and the foremost thing that we do is do the counting without really knowing what contents are and so on and we look at the number of tokens, we will look at the vocabulary, we try to compute the term frequency or idea or we also do the combination of term frequency ideas and so on, so forth, that is the first task that we do.