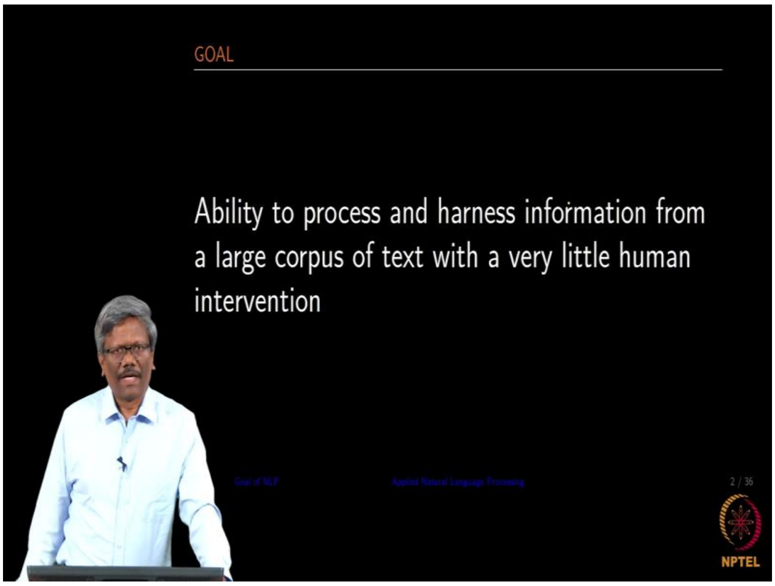
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 08
Statistical Properties of Words Part 01

Hello, again this is Ramaseshan continuing the lecture that I left last time. Last time I gave some introduction about natural language processing, artificial intelligence, how would natural language processing can be processed using some of the artificial intelligence techniques using some neural networks, and so on.

So, this time, we will get we will start diving deeper into what can be done with a corpus of text and then we can talk about how those texts can be converted into various forms, so that it becomes amenable for processing and the getting more inference from the text and so on and so forth ok. I will quickly go through the slides that I have, so that we will get to know what we are going to be covering in this class ok.

(Refer Slide Time: 01:11)

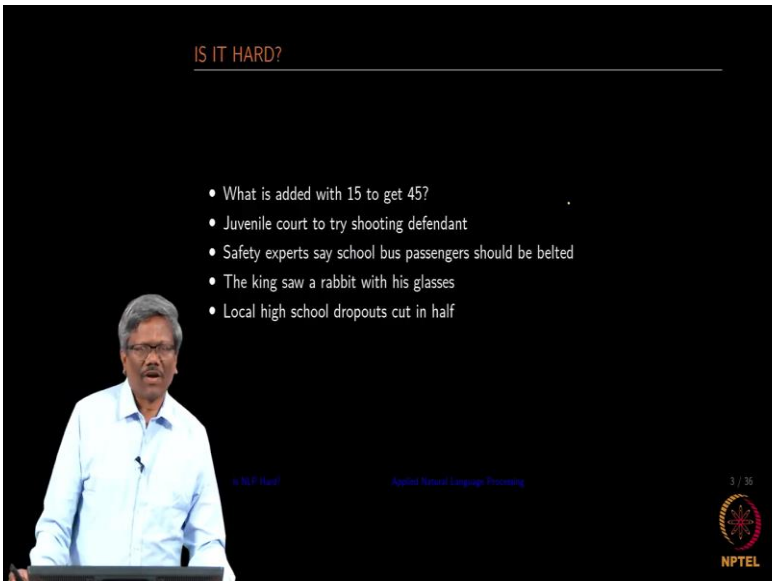


GOAL

Ability to process and harness information from a large corpus of text with a very little human intervention

So, I will be talking about the goal, this we will keep revisiting every now and then what is the goal of the natural language processing.

(Refer Slide Time: 01:20)

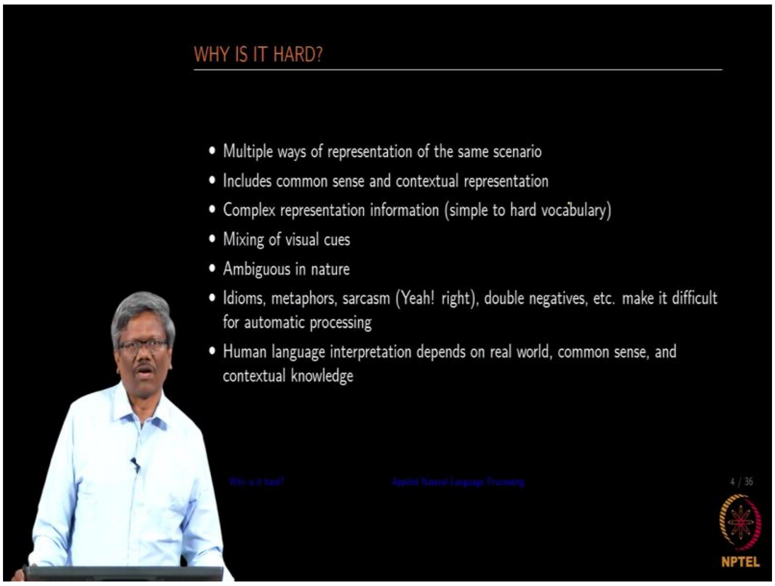


IS IT HARD?

- What is added with 15 to get 45?
- Juvenile court to try shooting defendant
- Safety experts say school bus passengers should be belted
- The king saw a rabbit with his glasses
- Local high school dropouts cut in half

And then we will talk about whether it is really hard to do some processing or not with the natural language text.

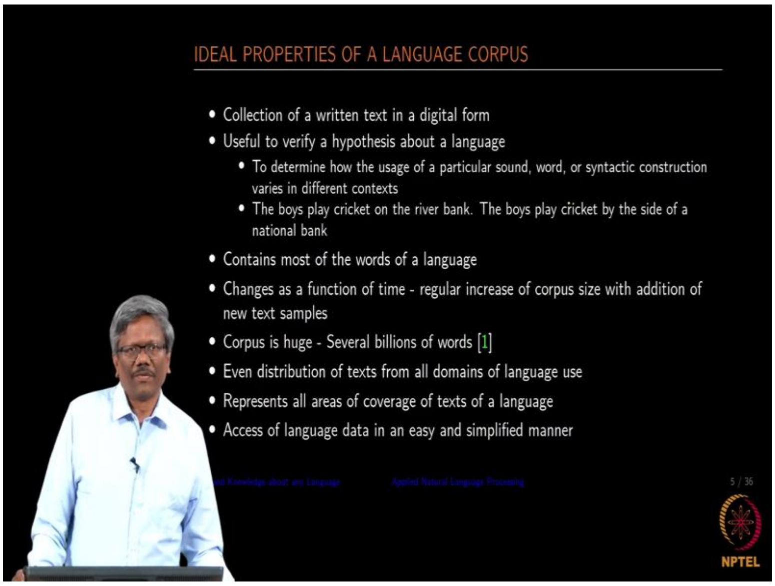
(Refer Slide Time: 01:30)



WHY IS IT HARD?

- Multiple ways of representation of the same scenario
- Includes common sense and contextual representation
- Complex representation information (simple to hard vocabulary)
- Mixing of visual cues
- Ambiguous in nature
- Idioms, metaphors, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing
- Human language interpretation depends on real world, common sense, and contextual knowledge

(Refer Slide Time: 01:33)



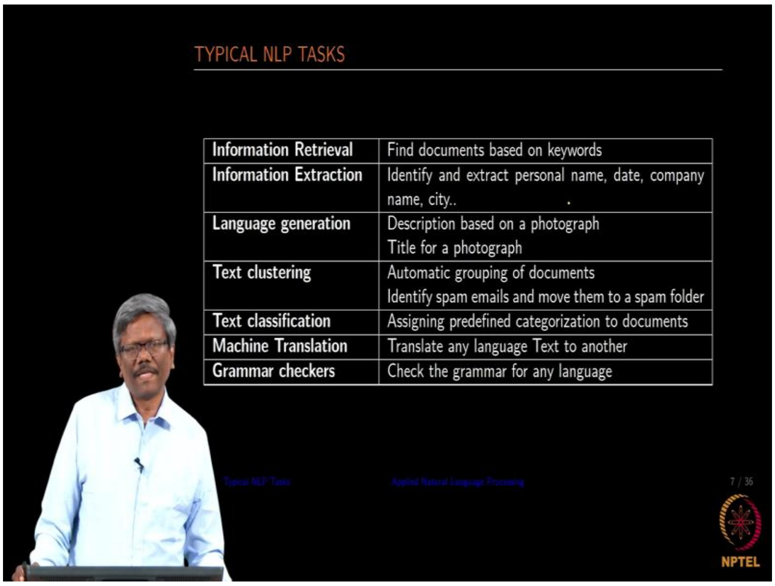
IDEAL PROPERTIES OF A LANGUAGE CORPUS

- Collection of a written text in a digital form
- Useful to verify a hypothesis about a language
  - To determine how the usage of a particular sound, word, or syntactic construction varies in different contexts
  - The boys play cricket on the river bank. The boys play cricket by the side of a national bank
- Contains most of the words of a language
- Changes as a function of time - regular increase of corpus size with addition of new text samples
- Corpus is huge - Several billions of words [1]
- Even distribution of texts from all domains of language use
- Represents all areas of coverage of texts of a language
- Access of language data in an easy and simplified manner

al Knowledge about any Language
Applied Natural Language Processing
5 / 36
NPTEL

And then we will also talk about why is it really hard, and then later we get into some the corpus which are going to help us in terms of identifying or the structures within the language of finding some patterns within the language that we have, and then figure out how we can really get the meaning of the word within a given context and so on and so forthright.

(Refer Slide Time: 01:56)

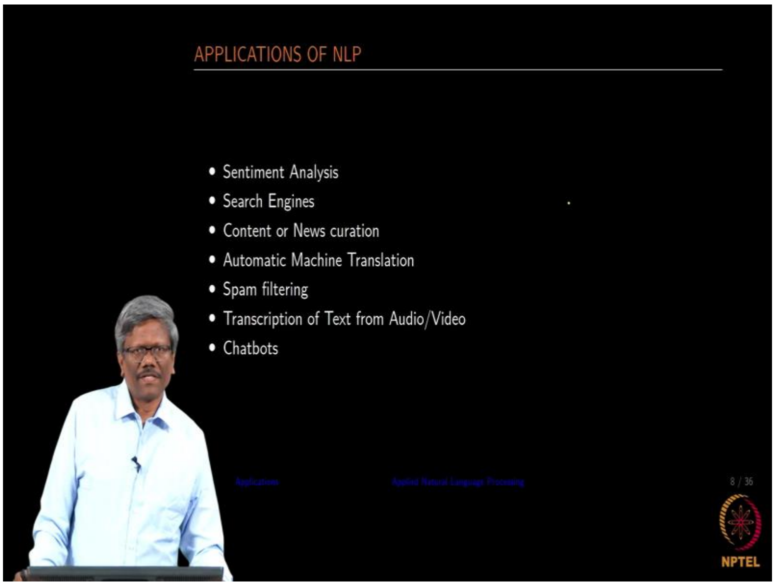


TYPICAL NLP TASKS

Information Retrieval	Find documents based on keywords
Information Extraction	Identify and extract personal name, date, company name, city..
Language generation	Description based on a photograph Title for a photograph
Text clustering	Automatic grouping of documents Identify spam emails and move them to a spam folder
Text classification	Assigning predefined categorization to documents
Machine Translation	Translate any language Text to another
Grammar checkers	Check the grammar for any language

Then we will talk about the typical NLP task that we will perform using the operation that we have performed using some statistical methods and so on.

(Refer Slide Time: 02:07)



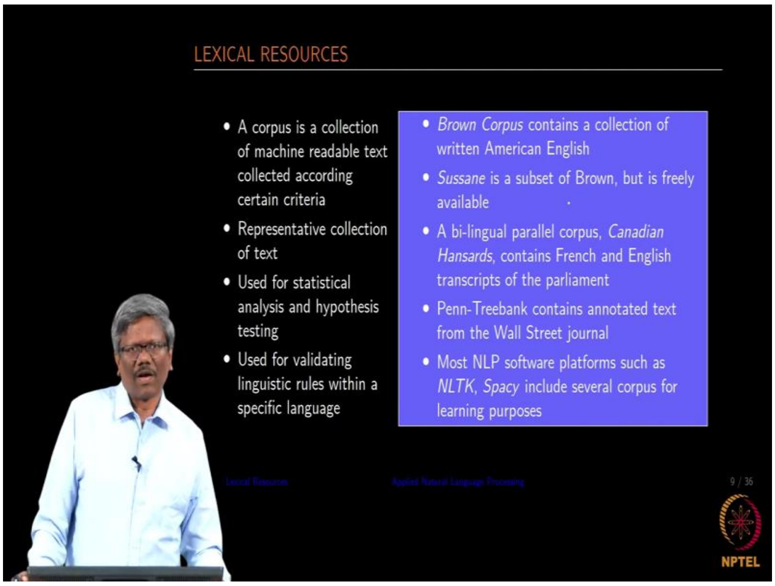
APPLICATIONS OF NLP

- Sentiment Analysis
- Search Engines
- Content or News curation
- Automatic Machine Translation
- Spam filtering
- Transcription of Text from Audio/Video
- Chatbots

Applications Applied Natural Language Processing 8 / 36

Later we will talk about some simple applications that can be created using the processes that we spoke about you earlier.

(Refer Slide Time: 02:18)



LEXICAL RESOURCES

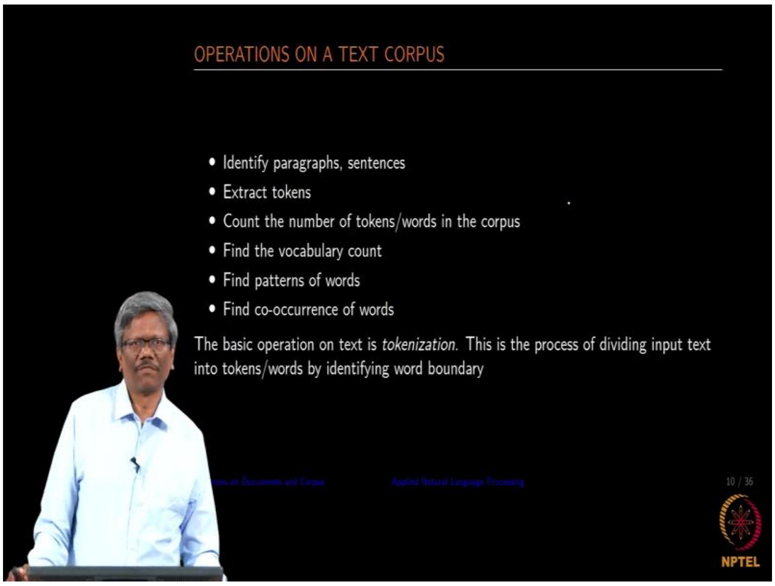
- A corpus is a collection of machine readable text collected according certain criteria
- Representative collection of text
- Used for statistical analysis and hypothesis testing
- Used for validating linguistic rules within a specific language

- *Brown Corpus* contains a collection of written American English
- *Sussane* is a subset of Brown, but is freely available
- A bi-lingual parallel corpus, *Canadian Hansards*, contains French and English transcripts of the parliament
- Penn-Treebank contains annotated text from the Wall Street journal
- Most NLP software platforms such as *NLTK*, *Spacy* include several corpus for learning purposes

Lexical Resources Applied Natural Language Processing 9 / 36

And then we will talk about the real corpus what should be the ideal corpus that we should have, we also talk about some are the corpora that are available right now for us to do some processing.

(Refer Slide Time: 02:29)



OPERATIONS ON A TEXT CORPUS

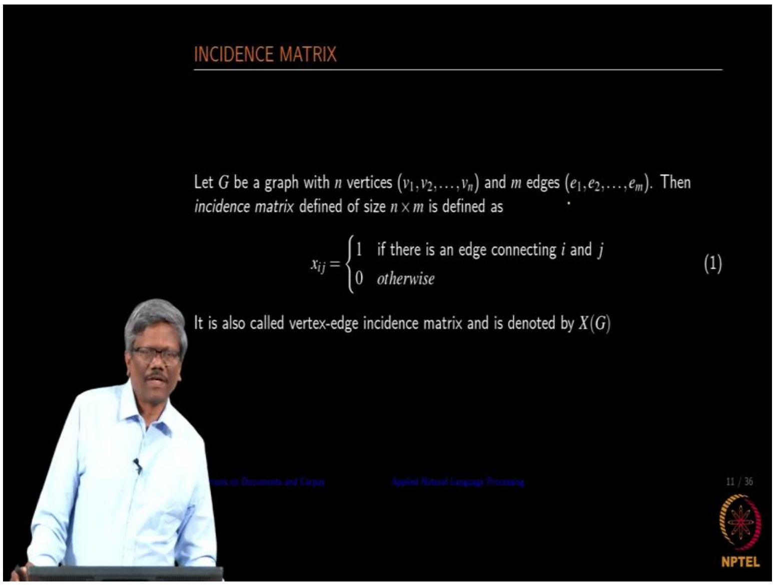
- Identify paragraphs, sentences
- Extract tokens
- Count the number of tokens/words in the corpus
- Find the vocabulary count
- Find patterns of words
- Find co-occurrence of words

The basic operation on text is *tokenization.* This is the process of dividing input text into tokens/words by identifying word boundary

tions on Documents and Corpus Applied Natural Language Processing 10 / 36

Then we will talk about a what is the operation that we can perform on a text corpus.

(Refer Slide Time: 02:39)



INCIDENCE MATRIX

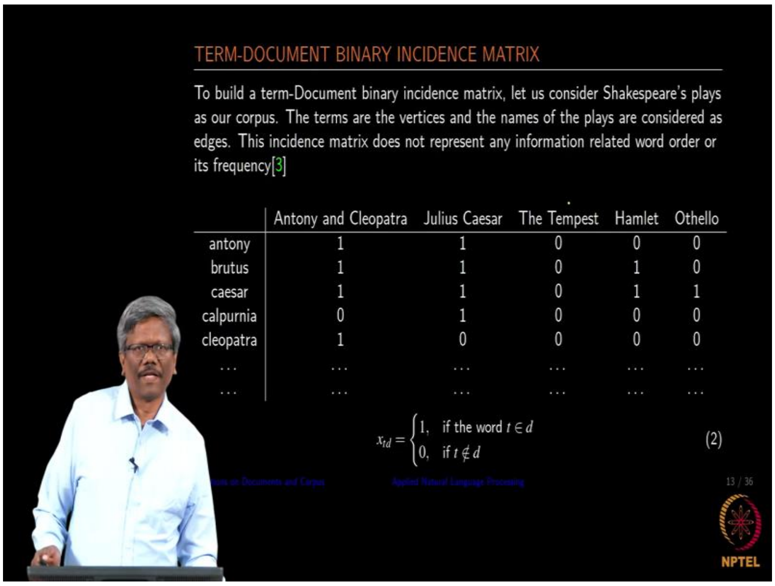
Let G be a graph with n vertices $(v_1, v_2, \ldots, v_n)$ and m edges $(e_1, e_2, \ldots, e_m)$. Then *incidence matrix* defined of size $n \times m$ is defined as

$$x_{ij} = \begin{cases} 1 & \text{if there is an edge connecting } i \text{ and } j \\ 0 & otherwise \end{cases} \quad (1)$$

It is also called vertex-edge incidence matrix and is denoted by $X(G)$

tions on Documents and Corpus Applied Natural Language Processing 11 / 36

(Refer Slide Time: 02:43)



TERM-DOCUMENT BINARY INCIDENCE MATRIX

To build a term-Document binary incidence matrix, let us consider Shakespeare's plays as our corpus. The terms are the vertices and the names of the plays are considered as edges. This incidence matrix does not represent any information related word order or its frequency[3]

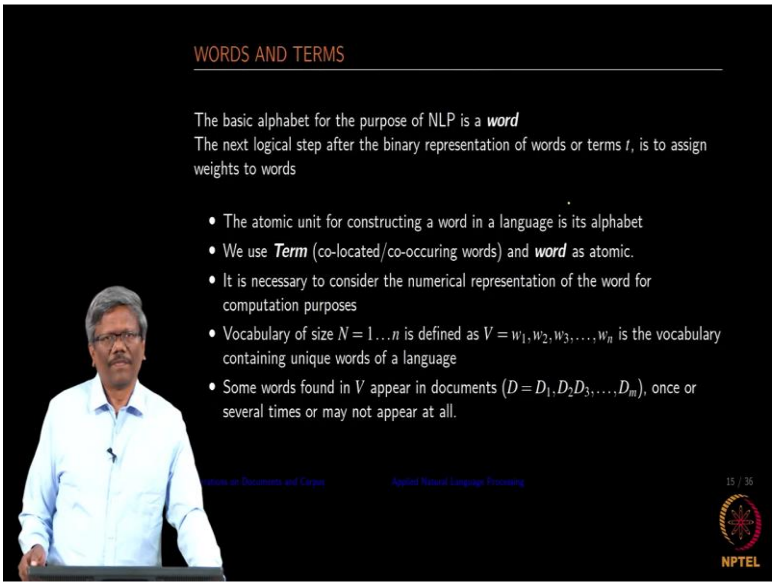
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello
antony	1	1	0	0	0
brutus	1	1	0	1	0
caesar	1	1	0	1	1
calpurnia	0	1	0	0	0
cleopatra	1	0	0	0	0
…	…	…	…	…	…
…	…	…	…	…	…

$$x_{td} = \begin{cases} 1, & \text{if the word } t \in d \\ 0, & \text{if } t \notin d \end{cases} \qquad (2)$$

…tions on Documents and Corpus Applied Natural Language Processing 13 / 36

And then later we talk about the incidence matrix, we will talk about the term document binary incidence matrix.

(Refer Slide Time: 02:47)



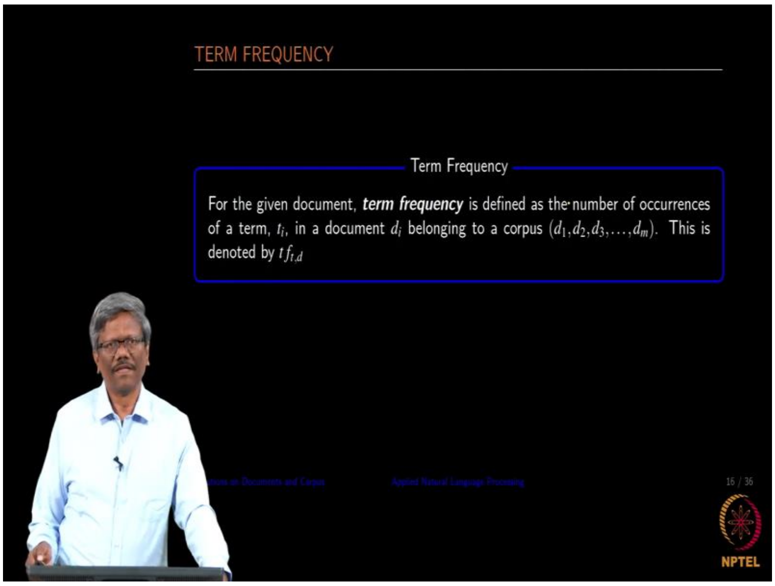
WORDS AND TERMS

The basic alphabet for the purpose of NLP is a ***word***
The next logical step after the binary representation of words or terms t, is to assign weights to words

- The atomic unit for constructing a word in a language is its alphabet
- We use ***Term*** (co-located/co-occuring words) and ***word*** as atomic.
- It is necessary to consider the numerical representation of the word for computation purposes
- Vocabulary of size $N = 1 \dots n$ is defined as $V = w_1, w_2, w_3, \dots, w_n$ is the vocabulary containing unique words of a language
- Some words found in V appear in documents $(D = D_1, D_2 D_3, \dots, D_m)$, once or several times or may not appear at all.

…tions on Documents and Corpus Applied Natural Language Processing 15 / 36

(Refer Slide Time: 02:50)



TERM FREQUENCY

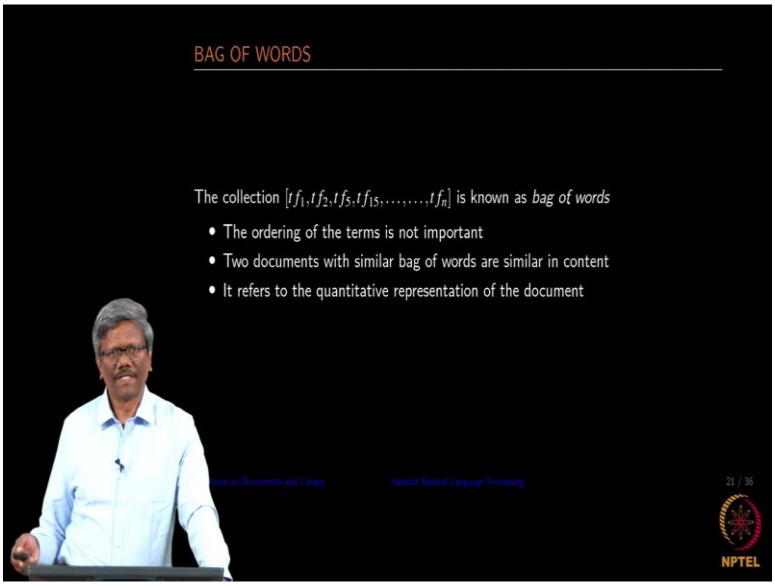
Term Frequency

For the given document, ***term frequency*** is defined as the number of occurrences of a term, t_i, in a document d_j belonging to a corpus $(d_1, d_2, d_3, \ldots, d_m)$. This is denoted by $tf_{t,d}$

...ations on Documents and Corpus
Applied Natural Language Processing
16 / 36
NPTEL

We will talk about the words and terms and the frequencies for each of the terms. We will also do some small demo to identify what can be done with the terms or words.

(Refer Slide Time: 03:03)



BAG OF WORDS

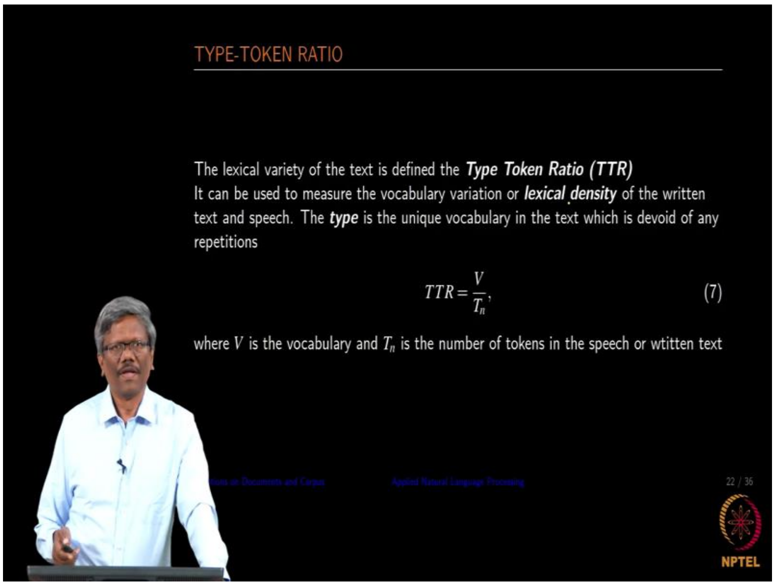
The collection $[tf_1, tf_2, tf_5, tf_{15}, \ldots, \ldots, tf_n]$ is known as *bag of words*

- The ordering of the terms is not important
- Two documents with similar bag of words are similar in content
- It refers to the quantitative representation of the document

...ations on Documents and Corpus
Applied Natural Language Processing
21 / 36
NPTEL

Then we will talk about the bag of words as another example of collecting the word frequencies and so on.

(Refer Slide Time: 03:09)



TYPE-TOKEN RATIO

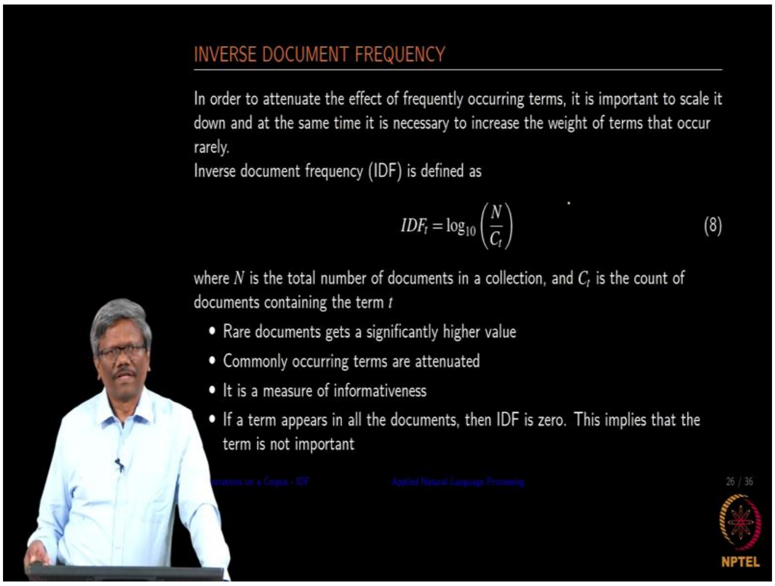
The lexical variety of the text is defined the ***Type Token Ratio (TTR)***
It can be used to measure the vocabulary variation or ***lexical density*** of the written text and speech. The ***type*** is the unique vocabulary in the text which is devoid of any repetitions

$$TTR = \frac{V}{T_n}, \qquad (7)$$

where V is the vocabulary and T_n is the number of tokens in the speech or wtitten text

tions on Documents and Corpus
Applied Natural Language Processing
22 / 36
NPTEL

(Refer Slide Time: 03:19)



INVERSE DOCUMENT FREQUENCY

In order to attenuate the effect of frequently occurring terms, it is important to scale it down and at the same time it is necessary to increase the weight of terms that occur rarely.
Inverse document frequency (IDF) is defined as

$$IDF_t = \log_{10}\left(\frac{N}{C_t}\right) \qquad (8)$$

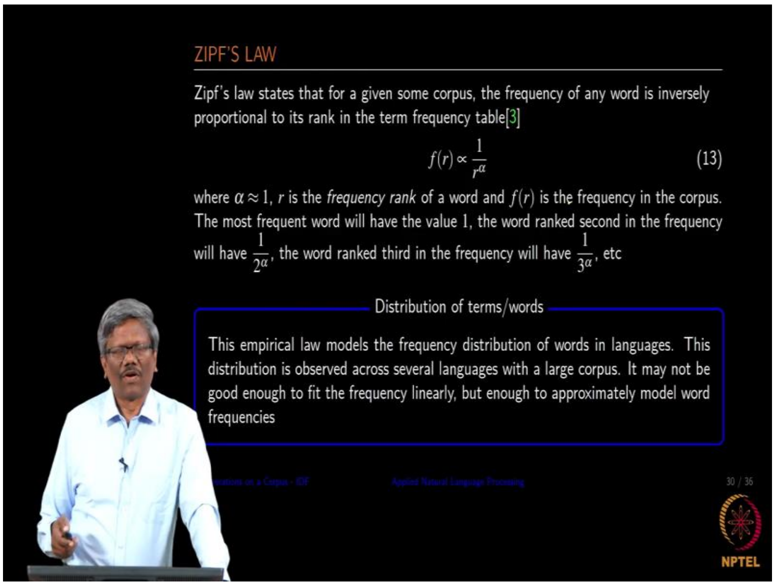
where N is the total number of documents in a collection, and C_t is the count of documents containing the term t

- Rare documents gets a significantly higher value
- Commonly occurring terms are attenuated
- It is a measure of informativeness
- If a term appears in all the documents, then IDF is zero. This implies that the term is not important

erations on a Corpus - IDF
Applied Natural Language Processing
26 / 36
NPTEL

We will talk about some empirical relationships like the type-token ratio. We will talk about the inverse document frequency.

(Refer Slide Time: 03:25)



ZIPF'S LAW

Zipf's law states that for a given some corpus, the frequency of any word is inversely proportional to its rank in the term frequency table[3]

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{13}$$

where $\alpha \approx 1$, r is the *frequency rank* of a word and $f(r)$ is the frequency in the corpus. The most frequent word will have the value 1, the word ranked second in the frequency will have $\frac{1}{2^{\alpha}}$, the word ranked third in the frequency will have $\frac{1}{3^{\alpha}}$, etc

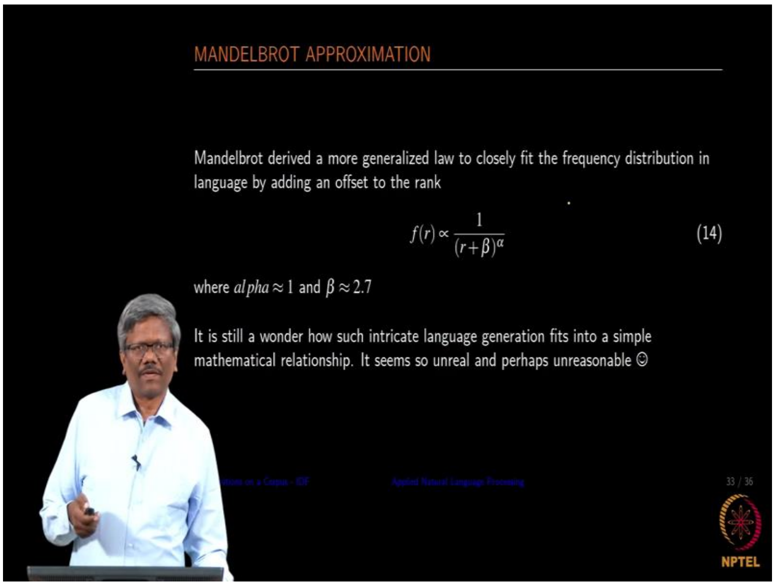
Distribution of terms/words

This empirical law models the frequency distribution of words in languages. This distribution is observed across several languages with a large corpus. It may not be good enough to fit the frequency linearly, but enough to approximately model word frequencies

...erations on a Corpus - IDF
Applied Natural Language Processing
30 / 36
NPTEL

We will talk about, we will talk about will we will some empirical lost alike ZIPF’s law.

(Refer Slide Time: 03:32)



MANDELBROT APPROXIMATION

Mandelbrot derived a more generalized law to closely fit the frequency distribution in language by adding an offset to the rank

$$f(r) \propto \frac{1}{(r+\beta)^{\alpha}} \tag{14}$$

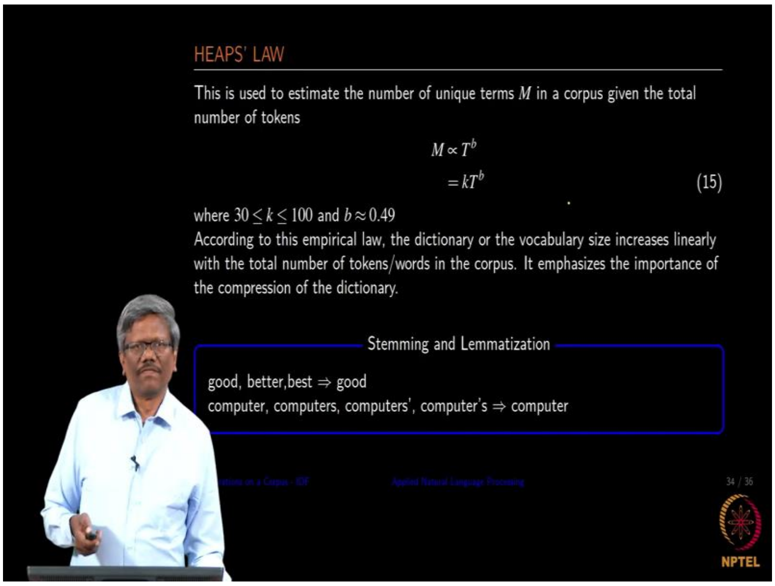
where $alpha \approx 1$ and $\beta \approx 2.7$

It is still a wonder how such intricate language generation fits into a simple mathematical relationship. It seems so unreal and perhaps unreasonable ☺

...erations on a Corpus - IDF
Applied Natural Language Processing
33 / 36
NPTEL

We will talk about the extension of these ZIPF’S law using Mandelbrot approximation.

(Refer Slide Time: 03:37)



HEAPS' LAW

This is used to estimate the number of unique terms M in a corpus given the total number of tokens

$$\begin{aligned} M &\propto T^b \\ &= kT^b \end{aligned} \qquad (15)$$

where $30 \leq k \leq 100$ and $b \approx 0.49$

According to this empirical law, the dictionary or the vocabulary size increases linearly with the total number of tokens/words in the corpus. It emphasizes the importance of the compression of the dictionary.

Stemming and Lemmatization

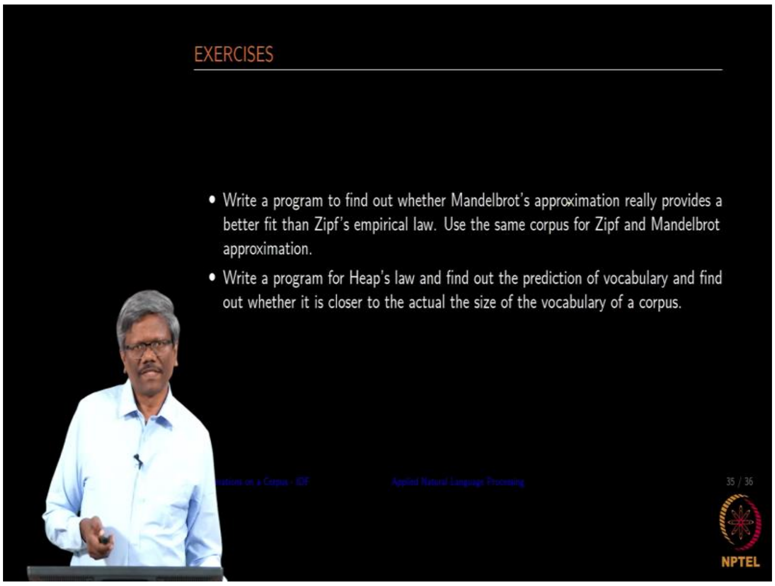
good, better,best ⇒ good

computer, computers, computers', computer's ⇒ computer

...tations on a Corpus - IDF
Applied Natural Language Processing
34 / 36
NPTEL

Then we will speak about another estimation process using Heaps' law.

(Refer Slide Time: 03:43)



EXERCISES

- Write a program to find out whether Mandelbrot's approximation really provides a better fit than Zipf's empirical law. Use the same corpus for Zipf and Mandelbrot approximation.
- Write a program for Heap's law and find out the prediction of vocabulary and find out whether it is closer to the actual the size of the vocabulary of a corpus.

...tations on a Corpus - IDF
Applied Natural Language Processing
35 / 36
NPTEL

And then finally, there would be two exercises that you can go back, and then perform on your computer am I right ok. let us start with the goal ok. The most important aspect of any natural language processing is to process and harness information from a large corpus of text with very little human intervention, please underline the word very little human intervention. we want to be able to process. You have a huge corpus a then capture information related to how frequently certain terms occurred, how frequently certain combination words of a occurred ok. we will keep revisiting every time when we talk about an important concept in natural language processing. Language processing is really hard and requires some contextual information to each and every one of the sentences that we speak and write.

To give an example let us take the first a bullet point that is added with 15 to get 45. if I ask a fourth-grader who understands this problem, he will be immediately doing 45 minus 15 and give me the answer. But when you ask the machine to do this there is no contextual information related to subtraction in this. we have to really give some additional information, so that machine finds the answer by subtracting 15 from 45 to get the answer ok. we do not have a clear cut definition or idea in terms of what should be done with this unless you understand the contexts very well.

So, let us take the next example of a juvenile court to try shooting defendant. you can form the sentence in different ways one. You can mean that juvenile court is going to start the trial for the shooting defendant right that is one way of doing. The second way of looking at address juvenile court is going to shoot the defendant, so but we know that is since this is going to be a court it is only going to go for a trial, and we know only one meaning if we take only one meaning of that. We do not take the second meaning of that.

The third one that we are going to be looking at is a safety expert says school bus passengers should be belted. we know that this particular sentence means that every passenger on the bus should be wearing a seatbelt and not like beating with a belt all that right. there is some kind of confusion when you look at these sentences. how are you going to teach the machine to understand the meaning only in one way right?

And we take the fourth example that we have here the king saw a rabbit with his glasses. Again there are two ways of representing the same one is the kings saw the rabbit with his classes; the second one the kings saw the rabbit carrying his class right. again we should be able to give some contexts to the system, so that it only means one thing. The last one is again having two different meanings, one is the local high school dropouts cut in half. we know that by the experience of reading so many school-related information the drop out percentage is reduced is what we really mean from these sentences, it is not really in terms of cutting somebody in half.

So, the context and the kind of information that we know based on what we have learned so far come into the play when we want to interpret this sentence right. it is really hard, natural language processing is a hard problem. we should be able to provide the contextual information to the system in such a way that it is able to interpret these sentences as which I have given as examples in the right way ok.

Let us find out why it is really hard. We have seen in the previous slide that there are multiple ways of representation of the same scenario. In the previous case we showed we had shown at least two different meanings for the same sentences. It includes common sense and contextual representation. For example, in the first case of the example what is added with for 15 to get 45. we need to understand it is a mathematical problem, and we need to understand there is subtraction involved in this, then only you can solve the problem ok. we need to have some common sense and contextual representation to be made available, so that machine is able to solve that particular example.

Complex representation of information, so some people use very hard vocabulary in their sentences, it is every time you need to really look into the dictionary to understand the meaning of that. Some use a very simple common vocabulary, so it becomes easier for us. it should be possible for you to interpret the meaning in two different ways; one is you can look at the dictionary for the meaning to figure out what it is, or in most cases what we do is looking at the contexts and the surrounding words and common sense that we have we try to interpret the meaning of the word that we do not really know right.

And then in some cases the visual cues are in interpret used. For example, you remember there was the color of the balloon that was flown yesterday so that visual cues coming from what I saw you yesterday is part of this sentence, which is not known directly to the system. how do you provide the visual clues are cues to the system?

And sentences are ambiguous in nature we saw through the five examples in the previous slide. We use idioms, metaphors, touches of sarcasm double negatives, etcetera, they make the interpretation very hard for automatic processing. For example, if somebody says something that you do not approve of instead of saying I do not know approve yeah right kind of you know sarcasm that we used. it is very difficult to bring that kind of interpretation into the machine, because yeah and right both mean a positive thing, you are not really negative what they are percentage is the same. this is an again difficult problem to address.

We have seen so far that human language interpretation depends on the real-world, common sense, contextual knowledge. without any of this is going to be hard for the machine to interpret the, so where we get this information, contextual cues, knowledge about the language, where are they going a come from, so it is going to be coming from a large language corpus.

So, remember when you were a child, nobody taught you how to really form us sentence using some syntax of a language right. you have learned to speak a language based on what you heard from you are parents, your peers, other children, and so on and so forthright. In the same way is it possible for us to really learn the language from a large corpus that contains representative representations of what we used on a regular basis or a daily basis. we will now look at what could be the ideal properties of a language corpus, and then see what is available to us in the water slides.

In general, if you want to define what a corpuses it is a collection of a written text in a digital form. You can consider all the HTML files on the internet as a collection or as a corpus. And if you want you still narrow it down, you can bring in all the collections related to kinematics in physics could be one small subcorpus that you can look at. Or you want to look at only the mobile reviews from all the internet pages that could be one of your corpus where you want to perform some operations related to mobile finding, these are reviews, sentiments what mobile you want to buy and so on and so forth.

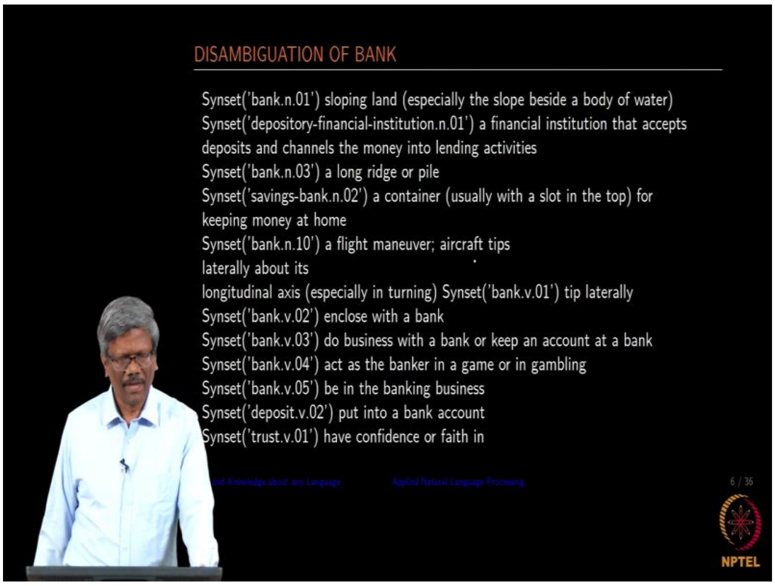
So, the corpus should be useful to verify the hypothesis about our language at two. Given examples to determine how the use of a particular sound, word or syntactic construction varies in a different context. I will give an example of that. The boys play cricket on the river bank. we know very well that the boys are really playing cricket by the side of the river; it is not really the bank does not mean here the place where you go and then do the financial transactions. The next sentence says the boys play cricket by the side of the national bank. we know very well that they are not playing cricket by the side of the river now, they have played cricket by the side of a bank.

So, these two contexts are very different, but playing cricket is our central context right, the central to this to these two sentences. by looking at the word surrounding play cricket, we are able to find out whether they are playing by the side of the river or near the nationalized bank and so on. it is very important know-how what we context is and what the word really means in every context. And so in this context the bank meant two different things right. this is called the polysemy in linguistic terminology. We want to have a corpus that contains large vocabulary, which means, it contains almost every possible word that we use in a given language.

Next one we also know that the language changes with respect to time, new vocabulary is added every year at least 20, 30, 100 depending on how new words are invented. this corpus will change with respect to time. we should be able to provide the ability to make changes stored. A corpus should be large if we need to have a really huge corpus so that all kinds of a representative word the sentence formations are represented within the given corpus.

We want to have a corpus where the text is distributed in a uniform fashion. We want to have a corpus that covers all areas in a given language right from the politics to news to science to technology and everything. Access to the corpus should be easy and in a simplified manner in some ways ok. this is the very ideal representation of a corpus that we are looking at, but in a real practical situation we may not have all of these. We may lack certain aspects of the real corpus that we going to be dealing with. when we do natural language processing on a corpus which is not really ideal, errors do occur ok.

(Refer Slide Time: 16:34)



DISAMBIGUATION OF BANK

Synset('bank.n.01') sloping land (especially the slope beside a body of water)
Synset('depository-financial-institution.n.01') a financial institution that accepts
deposits and channels the money into lending activities
Synset('bank.n.03') a long ridge or pile
Synset('savings-bank.n.02') a container (usually with a slot in the top) for
keeping money at home
Synset('bank.n.10') a flight maneuver; aircraft tips
laterally about its
longitudinal axis (especially in turning) Synset('bank.v.01') tip laterally
Synset('bank.v.02') enclose with a bank
Synset('bank.v.03') do business with a bank or keep an account at a bank
Synset('bank.v.04') act as the banker in a game or in gambling
Synset('bank.v.05') be in the banking business
Synset('deposit.v.02') put into a bank account
Synset('trust.v.01') have confidence or faith in

and Knowledge about any Language
Applied Natural Language Processing
6 / 36
NPTEL

So, to have a good corpus, so we need to be able to represent all the words including the synonyms and the polysemic words and so on and so forth in a very clear fashion. in this case, what you are seeing is the disambiguation of the word bank. if you go and then look at the word net which is a freely available thesaurus for digital dissemination of information, you will find various meanings represented for banks. what normally linguists do is look at the entire corpus and then find out how differently they are used, how differently the words are used, how these sentences are constructed using given word and so on, and then finally list them in some fashion.

So, what you are seeing in this particular slide is the listing of the word bank in the thesaurus. For example, in this case, if you look at the meaning of the word in the first case it is said sloping land, this one of the meanings, this slope beside other bodies of water. what has happened here is in a corpus the linguists would have found the use of banks in a different fashion. he will go and then look at, and analyze how that word is used and then create a synset for a given word, in this case it is bank ok. In the same fashion, he goes on the looks at you meaning of the word in a different context and then arrives at various synset.

So, the second one is a bank as we all know very well that it is a financial institution that accepts deposits and channels the money into lending activities, it is one of the meanings of the bank that was found to the corpus. And there are several other meanings see for the bank in the plane means of different things right. you latterly remove the plane it is flight maneuver, we have different other meanings like for example, have confidence is faith in is another representation of the same word. a corpus should be able to provide these kinds of contextual information, so that meaning can be obtained from every context.

So, what are the typical operations that we can perform using the corpus that we have in hand? I mentioned earlier that the corpus is a representative sample of the entire language for a given domain. Suppose, I want to find a document based on the keywords, and the keyword is very different from what is available in a document. For example, I want to search a document that does not contain the word computer but contains only the word machines.

And I want to still be able to get the documents where the machine is used instead of computers. when I use corpus where the machine is used in the context of computers, the machine would have got a sense that what I am trying to search are a computer and just the not alone machine. we will go and then bring all the documents that contain the word computer and machine. this is one example of finding the information from the set of documents.

And then I also have a need where I want to only extract the name of a person, date mentioned in a document, company name, the city name, and so on. I want to only extract that information. I would be able to go read, find out this info, and then retrieve the document and the related information along with that. Get me all the documents where cities are mentioned ok, I am not mentioning any name of the city here, only mentioning, in general, get me all the documents where cities are mentioned.

So, what the system should do is, it should be able to go through the entire collection and be able to find out a document that contains the names of the cities for example, Chennai, Bangalore, Trivandrum, Hyderabad, Kolkata and so on and so forth. all documents that contain the name of the cities or towns should be brought out through this. for that, you require corpus was that would have contained all the details that related to the cities and the cities are tagged in the document that is why you should that that is when you should be able to extract the information from that.

And then we want to generate the language this is one of the scenarios where you have been given a photograph, and you want to provide the title for the photograph. what you do you normally locate the photograph and then see what are all the objects that are available in that, and then string the words related to those objects, and then finally, provide a title to that. For example, if classroom photography is given you have a blackboard, you have a projector, you have a professor standing in front of the students, and so on right. By looking at the photographs you know that there are students, there are branches, there is a professor, there is a blackboard and so on. you would name the photograph as a classroom in general right.

So, for me to do that, I should be able to know what are all the objects that comprise a classroom. by looking at what I have extracted from the photograph, I string those words, and then finally figure out what would be the title for the photograph. This is one of the interesting examples that you will find in computer vision and natural language processing combined ok.

And then if you are given a huge collection of documents, you want to segregate them automatically in terms of the words that containing each of the documents. that is called a class string. This is an automatic process. You do not tell this system that gets me the documents related to physics, get me the document related to chemistry, get me the document related to maths like that. it automatically goes and then forms a cluster of documents and then gives it to you all that this particular set of this particular cluster contains a document wherein these words are found frequently. you go and then name that later.

So, the next one is text classification where you provide the label you tell that these are all the words that I want to want you to find. And when you find this document class provides the name of the class like physics, chemistry, maths and so on. the bucket names are known and using the words that are frequently found in the document, the documents are moved into respective buckets.

And then the last one is I am sorry then we have a machine translation, this is a very interesting task in the natural language processing where you want to be able to translate from one language to another automatically. we will talk about this in our later classes. And then grammar checkers I am sure some of you are most of you might be using all the web-based or the cloud base grammar checkers. what they normally do is, they just when you type a sentence it finds this error in your sentence of there is spelling mistake is identified and so on and so on forthright. again for you to do a good grammar checker, a very ideal English language text or any other language corpus should be made available ok.

So, let us move on to the application for that. we have spoken about what we can do with the corpus. now, we will find out what kind of application that we can build, you know this is a very crucial thing that is required for the human you know we want to build the application so that lot of things can be automatically picked up. the first one is going to be sentiment analysis.

Sentiment analysis, I am sure all of you would know like to find out whether the review of the movie is positive or not, whether it received 1 star or 5 stars, or is it a good movie or a bad movie, that actor is a good actor a bad actor, or is an average and so on. you want to be able to read the reviews and automatically find out whether the review provides a positive analysis or a negative analysis, whether it gave positive results or the negative result ok.

Then you know very familiar with the second one which is the search engine not going to be talking about this in detail. The third one is news curation is again something you have seen on you are mobile phones. If you keep reading certain news items that are available on the main page, it starts recommending certain news items automatically based on what you have read earlier.

So, the news curation is either it finds out what is your interest based on, what you have been reading over the last few months and starts recommending articles related to what you have read so far or it can also request you to say would you like to read articles related to cricket written by a certain author, or you want to related you want to read articles related to some political news written by a certain author. these are all the news curations or the content curation that you can develop based on natural language processing.

I spoke about the automatic machine translation, I am sure you also know about this well very well most of the mail that you received nowadays on over public email addresses are spammed only about 2 to 3 percent of the mails are very useful to us, rest of them are all spams. there is a mechanism by which the email providers create this spam filtering on automatically move the emails into the spam folder. If they are not able to do it, they also request you to mark certain emails as spam, so that next time when the mail of the type comes it automatically goes into the spam, spam folder.

Transcription of text from audio and video, for example, so I am now making a presentation and then speaking in English, is there a possibility of the machine reads the audio, and then finally transcribes the audio into the text format automatically? Again this requires a good amount of corpus in the language is also in the subject that I am talking, so that this sentence formation is legally correct.

The last one is the chatbots. The chatbots are something which is an emerging area where people would like to provide the experience of talking to a machine and provide the same experience that you might get from the human on the other side. this is again application that requires a lot of expertise and contextual information, so that it is able to provide meaningful transactions with the one who it is talking to right ok.

Let us now look at some of the lexical resources that are available, and you will also find a lot of them on the internet so that we can make use of that for your natural language processing. as I mentioned earlier that a corpus is a collection of machine-readable text collected according to certain criteria. I can collect the corpus for a given domain, maybe if I want to only talk about something in the natural language processing, I can collect all the documents related to that, and keep it as a natural language processing corpus.

By creating a corpus, we assume that it provides some representative collection of the text right. We want to perform statistical analysis and hypothesis testing on the corpus. We want to validate some linguistic rules within a specific language using the corpus. these lexical resources which are now available for us to do the natural language processing are given on the right side.

So, we have a brown corpus which contains a collection of written American English, this may not be suitable for arbitration gives as this is a very specific look at this you know. this is collected according to certain criteria. Other is Sussane in a subset of Brown which is freely available that we can make use of. A bilingual parallel corpus, Canadian Hansards, contains a French and English transcription of the Canadian parliament. There is a Penn-Treebank that contains annotated text from the Wall Street journal. You will also find many corpora for learning purposes from platforms like spacy and NLTK.

So, you have a collection of text, then it should be possible for us to do some statistical analysis on it, we can try some neural net-based model, build a language model. all those things are possible only if you have a good corpus. it is important for us to identify what we want to do and then get the corpus and start working on it ok.

So, we spoke a lot about what a corpus is, now we will talk about what kind of operation that it can perform on a corpus ah. that you know the corpus contains the text or which you can process using your normal application. For example, if we use the rep regular expression pattern identifier, you should be able to find out a word, and number, date, email address, and so on and so forth, so that is one of the fundamental requirements in the natural language processing.

So, I should be able to identify few patterns, for example, I want to capture two words at a time, three words of time, then I required some kind of a pattern-matching algorithm to get that. We require a very fundamental program that extracts tokens. A token is something that is separated by a boundary in the boundary sometimes space, a period, comma and so on and so forthright. I want to be able to extract that token based on some boundary conditions. I want to find out the number of tokens or words in a given document or in the entire corpus.

And you want to find out what is the vocabulary count in that. The vocabulary is nothing but the unique words that are represented in the corpus. You want to find patterns of words, for example, I want to find the twin words like in New York, New Delhi, hot dogs something like that, so you want to be able to find them out using the corpus that you have. You want to find the co-occurrence of words. I always want to find three words at a given time. if I have a small window where there are three holes, and then I use that window and move it across the text. Every time when we move from one word to the other, I capture three words. this is the co-occurrence of words that we are talking about. we will talk about this in detail as we move along.

So, the basic operation for all of this is an operation called tokenization. most of the applications of the platform provide these functionalities, so you do not have to really write a code to do the tokenization. this is the process where you divide the input text into tokens, words, by identifying the word boundary ok. these are the fundamental operation that you want to perform on a corpus.