

Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture 51
Text Mining & Sentiment Analytics

Hello everybody, welcome to Marketing Analytics course. This is Dr. Swagato Chatterjee from VGSOM IIT, Kharagpur who is taking this course. We have come a long path and we are now in week 10 and on this weekend and in the next week actually in two weeks we will be discussing about text mining and various applications of text mining and natural language processing in the context of analytics in the context of Marketing Analytics.

So, Text Mining in the context of marketing analytics is applied in a majorly applied because, that with the I would say internet availability of people and information and digital technologies being sprayed all over the world every person can express their feelings, express their information. So that is why, information is getting generated from various corners of the world and often those information is unstructured. So what is structured data and what is unstructured data that is something that we have to understand from analytics perspective.

Structure data are those kind of data which are kept in a tabular form, which is well, well put and I would say the various kind of statistical analysis can be done readymade that, the data is readymade for any kind of statistical analysis. On the other hand unstructured data is something which is not been given a form that can be used for data analytics. Yet so, that is something that is a problem for analytics person but, he has to do is he has to convert the unstructured data to some form of structural format and then with that structural format he can do statistical analysis, one of the examples of unstructured data is this text data.

So, consumers or individuals create text data in the context of marketing quite a lot. For example, customers write reviews, they can write review in let us

say [yelp.com](https://www.yelp.com) or [tripadvisor.com](https://www.tripadvisor.com) or various e-commerce websites they write reviews. Now, these reviews are text they are out structured data it do not have a proper tabular format for this particular data. To analyse this data you have to convert these unstructured data in some form of structural format and then have to analyse another example will be pictures.

For example, again various influencer marketing it is very well use that people post pictures. And you have to understand that whether, the picture what the picture is saying or what kind

of brands are more prominent of those pictures on what is the level of prominence of the brand?

So, all of these things will contribute towards the brand image when influencer marketers are actually posting their pictures. So, that is something that that excites marketing professionals and that is also one kind of unstructured data. Another type of unstructured data can be videos or audios, so video and audio both are unstructured data why are people take a decision that how I will actually what about this particular video is saying, whether this video is saying some action or video is saying some product, whether I can find out what kind of action is happening from this particular video and so on.

So, there can be so many different types of insights that you can generate from this unstructured data and the whole world is filled with more unstructured data and structured data. So, structure data will only be created when we forcefully captured the data and put it in the tabular form. But, there if you capture 10 percent of that world's data you are not capturing probably 90 percent of the world's data and that data is still lying ideal in various parts all you have to do is you to capture them, change them in some way and create a structure data. So, that is something which even in the marketing context we have to do.

So, customer reviews is one such unstructured data text, where text is involved and we will do natural language processing and text mining to do that, another places where the text comes of is blogs, consumer writes blogs, travel logs are created blogs are created where are these blogs are working with or blogs are good or blogs are bad that kind of thing makes sense.

Another classic application of marketing of text mining in the context of marketing and text mining is let us say click bait if, you have heard about click bait. What is a click bait? Have you heard about this kind of posts where people say that five ways to kill mosquitoes without using any kind of mosquito repellent or 10 ways to become successful in the world without doing anything.

So this kind of posts are actually called click baits, where they are baits they are driving you towards them so, that you click on them, what happens is when you click on them you go to the website and then that website is filled with various kinds of advertisements the only purpose of this click bait is to attract traffic towards them and the moment traffic comes to the particular website they make some money, that is all is the purpose.

Now, various platforms like Facebook and etc, is trying to stop this click baits, that how I can identify the click baits ahead of time and then I can block them so that people do not click them or my the content I would say the content that is being shared in Facebook the quality of the content does not go bad or often times violence identification if it is a picture or let us say nudity identification if it is a picture, all of these things can be done using various kinds of unstructured data mining, text mining is done for click bait text mining is done for sentiment mining for spam detection and various other stuff.

So here, in the next few hours we will discuss about how we can used natural language processing and text mining techniques do to implement it in the real world situation. Now, there is difference between text mining and natural language processing. What is natural language processing?

It is a area of computer science and linguistics you can say, where people try to find out how we talk, how we write based on that information they will process some text. In general, text mining is the broad area any text you mine that whether that has been made by human being, whether that has been created by machines, whether that has been created by some website writers any kind of mining of any text is text mining and they have certain process for that.

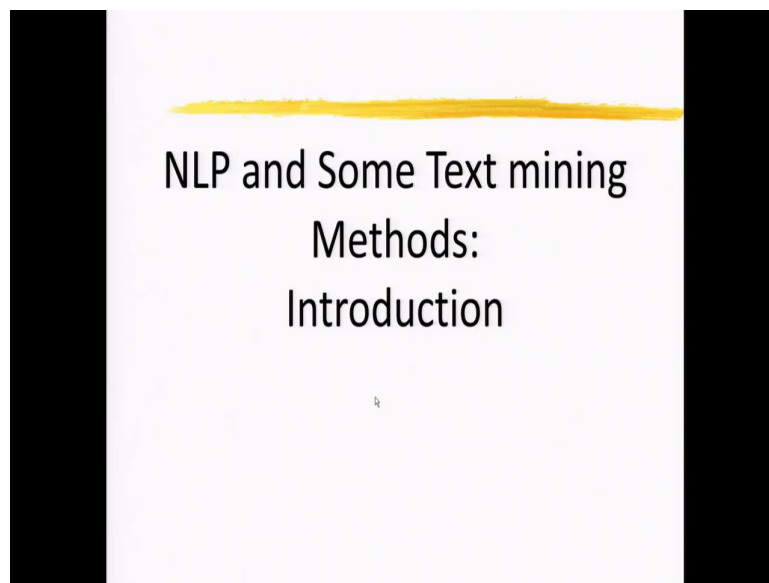
When you do the same thing text mining keeping in the mind of linguistic characteristics of the text that whether, the guy is for example, when I write, I write might written text you will see that they are often grammatically correct, when I am speaking I am not even completing a sentence sometimes I stay that okay, this, this and then after some time I break the sentence and go on in a different context all together. So, the sentences are also not full when we go and say something even for this particular course if you go and see the transcript of the course.

I think there are transcripts being created by NPTEL office. So, you go and see the transcripts of the tapes and not only my transcripts you can see any other professors transcript, you will see that the transcripts are not full sentences. Now, that is a text that is getting created from the linguistic characteristics of verbal communication.

The same thing will be absolutely different when you just analyse my emails my formal emails, my SMS my informal chats the type of linguistic characteristics that can will come up from this kind of text will be very different.

Now, when you do text mining keeping this kind of characteristics in mind, it is actually natural language processing. So, how naturally we create language or we use language how that knowledge can be used while you are doing text mining to create insights about the text that particular domain is natural language processing. Now, we have come up with a long path in natural language processing but, there are lots of other things that has to be done in natural language processing also so, we will in this particular section we will discuss about that.

(Refer Slide Time 09:35)

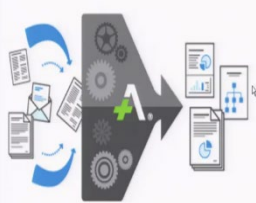


So, NLP and Some text mining methods an introduction is something that I will discuss in this particular class first.

(Refer Slide Time 09:44)

NLP & it's History

- **Natural Language Processing:**
 - *Natural Language Processing* is a sub field of Artificial Intelligence used to aid computers to programmatically understand the human's natural language to process and analyse large amounts of natural language data and develop insights out of that.
 - Data is question can be any unstructured data . E.g. Voice , Chat, social media data, feedback on any product, news etc.
- **History:**
 - Way back in seventeenth century, philosophers such as Leibniz and Descartes placed the proposal of codes to relate different languages, but those thought remained theoretical
 - In 1930s Georges Artsrouni, using paper tape, created the first bilingual dictionary
 - The other proposal, by Peter Troyanski, a Russian, was more detailed. It included both the bilingual dictionary, and rationale of grammatical roles between languages
 - In 1950, Alan Turing published his famous article "Computing Machinery and Intelligence" had a mention of thinking capability of machine
 - In 1980 first Statistical machine translation systems were developed: Based on statistical model



So the, what is NLP? So, natural language processing is a soft field of artificial intelligence used to aid computers to programmatically understand the humans natural language. So, that is the major goal, what is the goal? To programmatically understand the humans natural language and process and analyse large amounts of natural language data.

So, ideally there was always qualitative studies for done at in qualitative studies, let us say focus group discussion is happening, 5 people are sitting in the room and they are discussing there, what we used to do is we used to take pointers on them we used to record their particular audio whatever they are saying transcribe it and then there was one person who is to read and do content analysis on them. So, what kind of things are coming up by its simple reading.

Now, see if there are 5 persons doing this job of creating information and there is 1 person who is getting the information from the text of 5 persons conversation, then that is doable. The moment it becomes 500 persons or 5,000 persons or 5 lakh persons, it becomes very difficult it becomes very difficult for one single human being to do this. Now, you can say that okay, why will I put one single human being I can put 10,000 human beings.

Then, the inter human being reliability will go down, so whatever he thinks person 1 the analyser 1 thinks as a positive sentiment and whatever analyser two things has positive sentiment their level of positivity might be different there are trace holds might be different and then the analysis will have some amount of buyers depending on who is analysing it.

So, then to solve that problem we need a program but, the program has to have human level understanding of whatever being the being whatever, is being talked in the text. So, a word very good it is a very good can be said in various terms, there can be sarcasm they can be really good and etcetera. So, let us say I pass I give an exam and in that exam I score very well 90 out of a 100 or more than that, I come and say mom I have scored 90 out of 100, mom says very good beta, very good.

So, that is actual very good pride and good sentiment positive sentiment comes up. The same thing if I says mom I got 40 out of 100 mom, says and my mom says very good 40 out of 100 very good very good. So, that is basically a sarcasm, that mom is saying. When I convert it into text, both are same no both are very good.

So, if both are very good both are the same text, then the text has to the program has to understand by listening to the context that whether it is a sarcasm or actually true which becomes very difficult for a human being it is possible a human being if you understand that 40 out of 100 and then it very good comes in, that means it is sarcasm a human when he reads the text you will understand but a program might not understand.

So, you have to make the program understand in such a way such that, the programmatically understands humans natural language and to process and analyse large amount of data. So these two thing large amount of data programs can handle but, program cannot handle humans natural language, when these two things you try to bring in into the picture both are done at a suitable level of analysis, then you are doing natural language processing well.

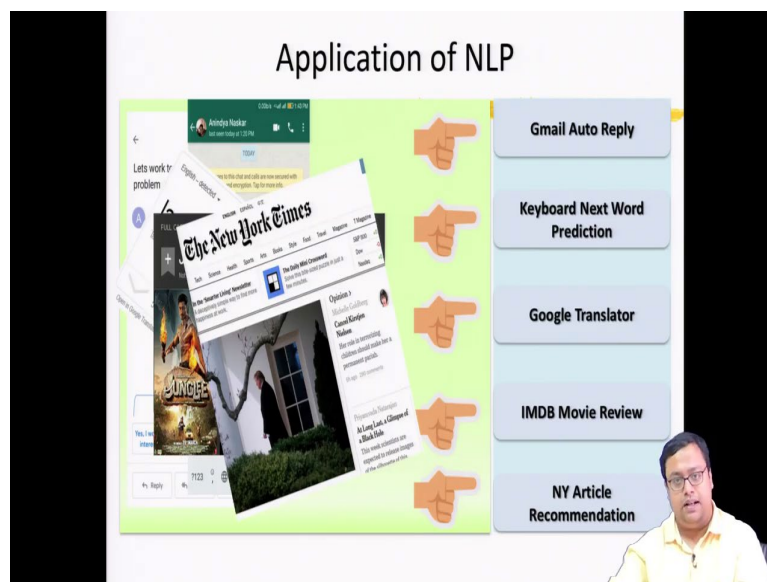
So, data in question in this case can be any unstructured data, it can be voice, it can be chat, it can be social media data Facebook or any product or news. So this kind of data can be analysed using natural language processing. The history is way back in 17 century, so it is being done for quite some time philosophers such as Leibniz and Descartes place the proposal of codes to relate different language.

So, this is something that has been happening till from the 17 century. But, those then remain theoretical, they were theoretical understanding. In 1930s Georges Artsrouni using paper tape created the first bilingual dictionary. So, dictionary is also one type of natural language processing.

The other proposal by Peter Troyanskii, a Russian was more detailed it included both the bilingual dictionary and rationally of grammatical roles between languages. So, slowly it evolved from 17 century then 1930 it was some bit of empirical work and slowly revolving 1950 Alan Turing published the famous article computing machinery and intelligence and had a mention of thinking about capability of machine. So now, we are bringing in AI a little bit.

And in 1980s first statistical machines translation systems were developed based on statistical model and probably in the 1990 and etcetera of we have used that in Google also the translation of language from one language to another language and so on.

(Refer Slide Time 15:36)



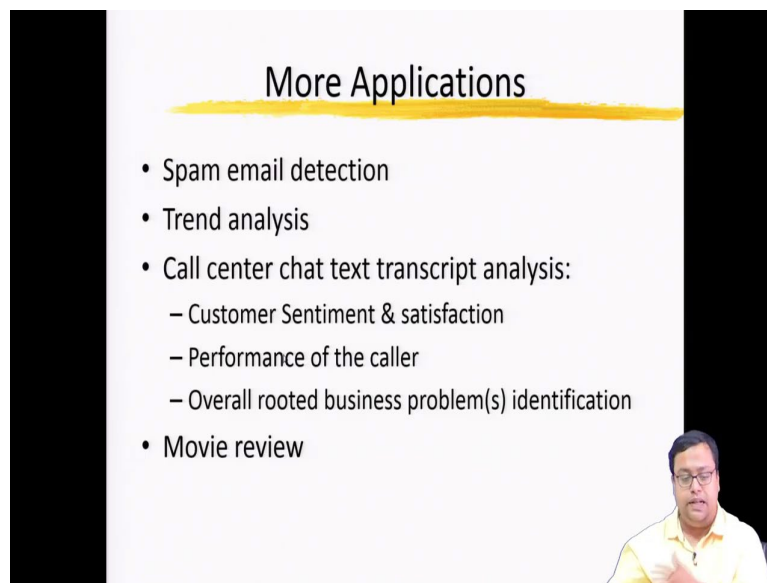
So, what is the Application of NLP? NLP can be applied in Gmail's Auto Reply for example, let us say, somebody says Saptrashi, we came across a business problem there is a huge scope of drive value research, let me know if you are interested, the moment somebody writes that it they see in Google at the bottom, there are some text comes up automatically. So you can just click on the reply goes up. So yes I would be interested I am interested or not interested so, these are auto replies. So, how I can get this auto replies you get that in chats you get that in emails? So, that is for natural language processing comes into the picture.

The same thing comes into the picture when you are typing, so I am working for Tata Consultancy and then the services is automatically written there. So, basically keyword, keyboard next word prediction is also what natural language processing comes in the picture. Then,

Google translator, so from language to bhasha natural language processing will come to the picture.

Then, IMDB movie review. If, I want to know that what kind of genre it is or what kind of movie reviews it is coming up what I should do what I should not do it comes to the picture. Again, article recommendation we have talked about recommendation engine, but we have not talked about how best of whatever, text to your reading how other articles which are of similar topic or we were of your similar interest level can be populated, that we have not discussed. That is also something that can be done using NLP.

(Refer Slide Time 17:25)



More Applications

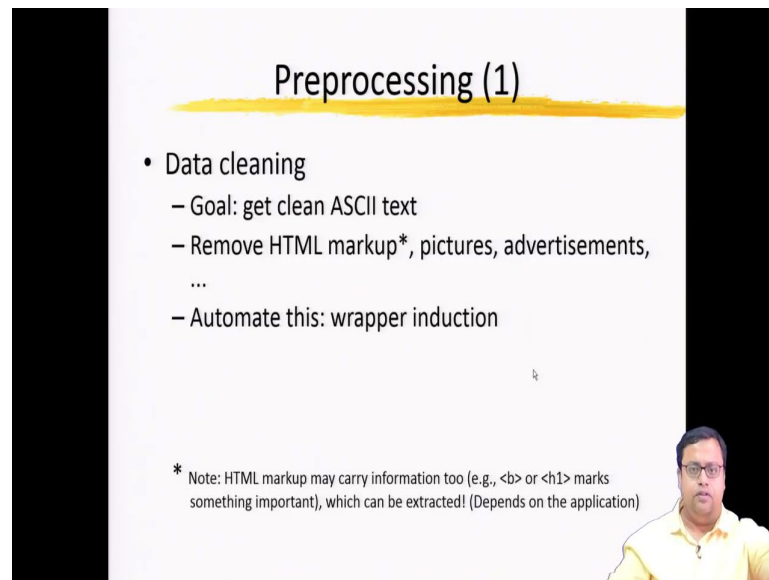
- Spam email detection
- Trend analysis
- Call center chat text transcript analysis:
 - Customer Sentiment & satisfaction
 - Performance of the caller
 - Overall rooted business problem(s) identification
- Movie review

Another kinds of applications are spam email detection this is something that we will do a case study on that, trend analysis we can do probably or I will at least suggest that how to do it. Call centred chat, chat text transcript analysis that is also something you can do to find out that whether people are working well or not what is the customer sentiment and satisfaction level, what is the performance of the caller, what is the overall robot business problem identification?

All of these things basically customer sentiment analytics, what customer wants, performance of the sales caller all of these things can be done. And then the movie review, whether the movie has been positively reviewed or negatively reviewed, what are the key themes were there giving positive reviews, what are the key things why they are given negative reviews, all of these things can be done using NLP. So, there are huge application of natural language

processing the context of marketing. So, when we do natural language processing, there are certain steps.

(Refer Slide Time 18:31)



Preprocessing (1)

- Data cleaning
 - Goal: get clean ASCII text
 - Remove HTML markup*, pictures, advertisements, ...
 - Automate this: wrapper induction

* Note: HTML markup may carry information too (e.g., `` or `<h1>` marks something important), which can be extracted! (Depends on the application)

The first step is pre-processing so, before you do any kind of text mining you have to process the data so, that it becomes usable, what are the processes various types of processes? The first job is clean the data. So, it is unstructured I would say raw data and the data dump probably and you have to clean the data first that is the first basic job that you have to do. So, how to clean the data?

Either, you can get clean ASCII text that is the best available, otherwise you remove HTML table mark-up pictures and advertisement. So, HTML mark-up may carry information to which can be extracted. For example, whether it is a heading or whether, it is a new paragraph or whether, it is it is a link email link or is a URL that kind that is actually given by this kind of signals ``, `<h1>` those we have done a little bit of with HTML will know.

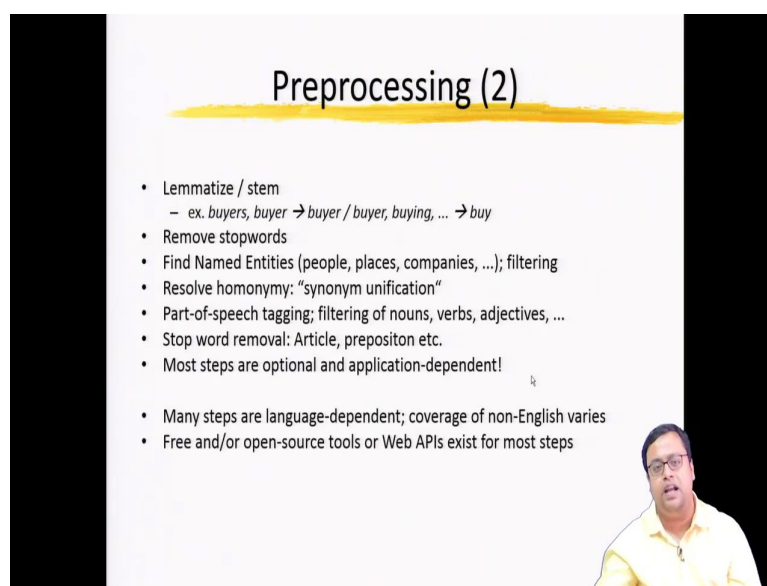
So, you might not want to remove that also if you want to remove, so you have to take a I would say very very informed call that what kind of HTML mark-ups I will remove and what kind of HTML mark-ups so I will keep in my data set so, that I can create certain information.

For example, let us say one of these things will be let us say in a any review there is a title of the review and then you write the whole text. So, it is a very good movie and then you give a explanation about how good it is.

Now, very good movie is the header that will come under delta H1 mark-up and then, the text is probably p paragraph. So, that H1 mark-up if you can copy then analyse the title of the review have separately and analyse the text of the review separately. Then, you might get the newer, newer information. So, this is something that you have to understand.

Then, automate this with wrapper induction so, you might not want to do all of these manually data cleaning you want to have a wrapper which will I will show you how to do it in the data mining process.

(Refer Slide Time 20:48)



Preprocessing (2)

- Lemmatize / stem
 - ex. *buyers, buyer* → *buyer / buyer, buying, ...* → *buy*
- Remove stopwords
- Find Named Entities (people, places, companies, ...); filtering
- Resolve homonymy: "synonym unification"
- Part-of-speech tagging; filtering of nouns, verbs, adjectives, ...
- Stop word removal: Article, preposition etc.
- Most steps are optional and application-dependent!

- Many steps are language-dependent; coverage of non-English varies
- Free and/or open-source tools or Web APIs exist for most steps

What else you can lemmatize or stem? For example, many words have the same key for example, let us say if I say talk about, I am eating, I eat rice and I have eaten rice. So, eating, eat, eaten all are eat basically, a verb the basic form of verb is so, when we transform the data to the basic code form of the word, that is called stemming or lemmatizing so, that is, that can be done that is something that we generally do to know that which one is most occurring.

So if I put eating, eat and eaten to be different words and we try to find out word frequency or something, we want to know that how much time people are talking about eating when they are giving a review of a restaurant, then if I keep these three words different then they will be calculated differently.

But, if I lemmatize them, stem them and make all of this what to be eat, then I will know that the frequency is much higher for what people are saying. So that kind of processing has to be

done. For example, buyers, buyer or buying all of these things are related to buy. So, if I can lemmatize that then we will know that people are talking about buying a lot.

Remove stopwords, so there are lots of words which are commonly used such as I, you, he, she, am, now, not, is, not is also stopword but, sometimes we should not remove that or let us say hi, hello and there can be so many different kind of stopwords so there is a stopwords dictionary that has been created by many various I would say linguistic professional or linguistic researches. Basically, these are the words, which does not contribute to any meaningful information in that text. So, we can remove the stopwords while doing the text mining, than find the Named Entities.

For example, proper nouns name of place name of company name of people sometimes we want to filter that also depending on the situation we want to filter. So, if I say that Rahul and I have went to this hotel and this hotel was very good and we have enjoyed a lot and this and that but, this name Rahul has nothing to do with this whole text.

So, I would rather not want Rahul to be in the text because, this a proper noun and I have nothing to do with that. So on the other hand when I am doing in other contexts this Rahul name might be useful, let us say I am talking about the chat transcripts or I am analysing the conversation of three 3, 4. People So, when we are converse when 3, 4 people conversing with each other, whom I am talking to can be found out from that Rahul name?

So, sometimes it will be useful sometimes it might not be useful you have to take a infirm decision that when I will keep it and when I will remove this particular thing from the text. I can reserve hominem also so, synonym unification all the synonyms can be joined together part of speech tagging sometimes we want to filter out nouns adjectives for separately.

For example, a classic example will be in the sentiment analytics, let us say we often I ask my students to do parts of speech tagging because, let us say if I ask you that how will you rate this particular course?

You rate this particular course based on the learning, based on the delivery quality, based on the probably the attentiveness or the responsiveness of the team and various other things till now, think about this thing the learning ability the learning the up the learning the delivery of the of this particular course or the responsiveness all of these things are noun, what is responsiveness?

It is a noun, it is a characteristics or let us say quality, qualities of noun product quality is a noun product quality is good, good is an adjective. So, you responsiveness is very, very, very good or is very highly, very high responsiveness high is an adjective responsiveness is the noun.

So, if I say that this book is very good, the book is the noun and good is my adjectives. So, often times we actually give sentiment of on a noun and the adjective gives me whether this element is positive or negative and etcetera. So, if I can tag them based on the parts of speech of the lines, then I will be knowing that whether, people are saying positive or negative about this text. So, that is something very important and we have to understand that.

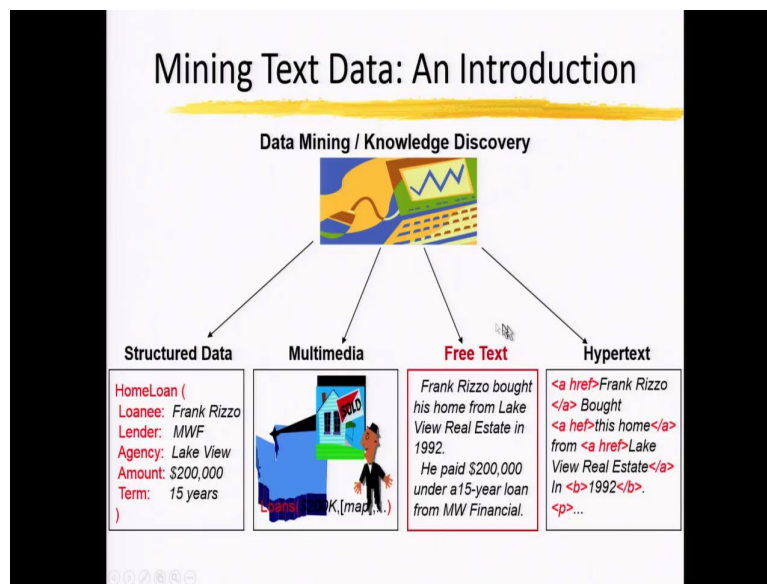
So, stop word removal, we have already talked about most steps are optional and application dependent. So as I was telling you again and again that you have to understand and you will not apply all these steps, you will choose which step to apply which step not to apply depending on the situation and I will show you in some situations some things will be applicable in some other situations something will not be applicable. And many sentences steps and languages-dependent coverage of non-English fairies.

For example, if by chance if there are certain things which are written in non-English language, then the stop words will be different the natural language processing of the same thing when I speak in Hindi and the same thing when I speak in English might be very different and the stop words might not exist in it in the or let us say or this part of speech tagging way of parts of speech tagging might be very different in Hindi and English.

So, we have to find out that, it is a very language-dependent activity the moment you shift the language the process of yours the lemmatization and etcetera might be different. So, that you have to think about when you do language. And free and open source tool and Web APIs exist for most steps.

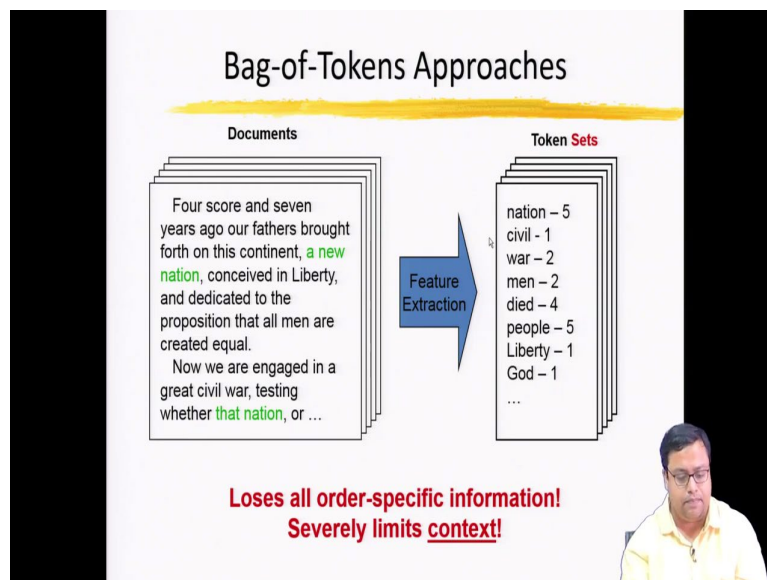
So, you do not have to quote for most of these things there are already somebody has done it at libraries we have been using libraries for various things. So, we will be using libraries here also. You have to remember the market analytics course is not to create algorithms, it is to apply already readymade algorithms in the context of marketing so that marketing decision making becomes easier we are not algorithm creators, we are users of an already created algorithms.

(Refer Slide Time 28:00)



Then, what is the text binding? So, this is how the structure data looks like and multimedia, free text and hypertext each of them can be the source of information, which is basically unstructured data.

(Refer Slide Time 28:19)



So, what we do here, the first thing that we do when we do text binding is called creating Bag-of-Tokens. So, this is what, while I will stop today, after this particular slide and we will continue in the next video.

So, what we do first thing is we convert documents in Tokens Sets. For example, you see that, in this particular thing nation has been used quite some time and then there is another

word call civil another call war. So, each of the unique words after removing stop word and punctuations and these and that whatever words are left, we convert them to unique words and then the corresponding who needs words frequency will also write.

So, losses all order specific information and severely limits context. So, this is one way of analysing but this generally limits the context of the data set. And still, this is though it limits the context of the data set, this is one of the most used version of Text Mining and we will discuss about this in the next video.

Thank you for being with me, I will see you in the next video with term frequency and inverse document frequency and how tokenizing can help in Text Mining. Thank you.