LECTURE 30 : Text Mining - Introduction

Hello learners, welcome to the course. So what we have done until now, we have extracted the data from different bibliographic data sources. Then we have seen the different type of analysis and visualization on these data. But we have not analyzed two of the things of the data set that is title and abstract. So if we see our data set, so this is our data set. So we have done analysis on these authors.

Then we have also done analysis here. We have seen the analysis on different like cited by. Then we have seen the analysis on affiliation. Then we have seen on the keywords, whether the author keywords or index keywords, or then different parameters like language of the documents or what exactly the publication and then what about that whether the publications are in open access or not.

So we have done this kind of analysis. So now one of the things which we have left with is that this is the abstract and then this title. And title and abstract provides the concise summaries of the content of the research publication. We can use these for different text mining techniques like topic modeling to identify latent topics or themes discussed in the documents. So this helps in understanding the key themes or the area discussed on those publications.

So the analysis of this text mining on this title and abstract will help us in knowing the key themes or the topics related to those field or the domain. So in this week, I will be discussing about what text data is, how to do the text mining, what are its some of the applications. Then we will also see how to do the pre-processing of the text data, which is one of the primary steps in the text mining. And in pre-processing, we will see what is regular expression, how regular expression helps in the pre-processing of the text data. And after that, we will see how to remove such other things like stop words, punctuation, transforming the data to lowercase and all.

After pre-processing of this data, we will apply the topic modeling approach. Basically, we will be using the LDA approach. So among various approaches of topic modeling, LDA is one of the most popular. So we will be using the LDA or doing the topic modeling on title and abstract data. And we will be doing this whole analysis on one sample data set of like 50 publications of one of the journals and we will see how exactly we can interpret the results.

So text data is, it is a kind of a data that is available in natural language, text format and meant for human for their communication purposes. We can see the text data in article, books, web pages, emails, tweets, maybe Facebook or LinkedIn ports, reviews and all. So with the rapid usage of many of the online tools and platform, the plethora of text data have been generated. So if we see this is our text data, we have this is the title and then we have this abstract. So this is one of the publication whose title and abstract we have.

In similar way, we have this another title and abstract. So this text data is in the natural language. So natural language is a primary mode of communication used by humans in their day to day linguistic expression. So this is an example of a natural language of the part of the admin Smith describing in the Egyptian script about a different surgery cases from 1600 BC. So we can have the text data in various natural languages.

If you see on my screen, this is an example of a text which is from Egyptian script. So if we see here, these are the text data in the English language. Here also we have the English language. So in this week when we will be doing the analysis of topic modeling, we will be taking off the English language text, we will be not taking other language text data. So the another thing about the text data is that it is unstructured in nature.

So what exactly this unstructuredness is there. So there is a two kinds of data generally we have. So one is that structured and another is the unstructured data. So if we see this data, if you see here on my left hand side, this is an example of a structured data where we have a kind of rows and columns and this is an example of unstructured data on my right hand side. So structured data typically organized in tabular format with rows and columns.

So this is a structured data and from here we can easily do the analysis like who is the most prolific author. If you see like there are different authors are there and there are two authors are there Mukesh and Prankav which have published two documents and or in terms of like if we have to see who has the more number of citation, maybe total citation or a given year of the citation. But in this right hand side we have this unstructured data and we can't extract this information from this kind of a text data. So in unstructured data it does not have any kind of predefined data model or format. It is typically not organized in tabular format.

So some of the unstructured data example is text, images, audio and video content, but we will be discussing about the text data. So our focus is on the unstructured data which is a text. So if we have to process this unstructured text that I will see, so this is like that. So this is like I have colored the different observation in this data. So there is one is here that whatever the publication by Mohit Garg, then this is about Mukesh Vira, then this is about Priyanka and then this is about Vijay.

You see the analysis on structured data is quite easy, but the analysis on unstructured data is not that kind of easy. But with the help of tools we can do easily find those analysis. So text data often have noises like typographical errors, misspellings, grammatical errors or maybe the informal languages. Handling these noises is crucial for accurate analysis and interpretation. For example, many of the text from online social media sources does not have proper words like morning, morning is written as like m r n g or maybe your is written as your.

So this kind of noises is there in many of the text data. But luckily what the data we are

handling in this particular course is about the text data which is peer reviewed and published in different sources. So we need not to worry about these kind of noises in our text data. But again we have to do some kind of pre-processing, but we will be discussing later in this part. So this is the complexity of unstructured data.

So how we can extract the knowledge from this text data, but which is exactly unstructured in nature. So with the help of text mining. So what exactly this text mining is? So text mining is the intersection of natural language processing, data mining, machine learning and statistics. It is used to extract actionable insight and knowledge from unstructured textual data. It is a process of transforming silos of text data from unstructured to a structured understandable form so that effective decision can be made.

So there are two types of text mining is done. So one is based on bag of word model and another is based on this syntactic parsing. So what we will be focusing in this course is bag of word model. So what exactly this bag of word model is? Say for example we have two sentences S1, S2.

So S1 is like Mr. X is a good boy and a cook. And for the sentence 2 we have like say for example Mr. Y is a good boy and a driver. So in bag of words model we consider each word or group of words as a unique feature and our focus is a note on the grammatical structure of the text. So the bag of word model is a computational cheaper compared to the syntactic parsing and it is quite easy to understand.

So in bag of word model what we do is we generally simply count the number of words and their occurrence. So say for example if we see here the Mr. is occurred two times in this both the sentences then X is 1, Y is 1 then we have is is is occurrence is 2 then we have occurrence of A is 1, 2, 3, 4 then we have this good occurrence of this good is 1 this good is 1 and then we have cook 1 driver 1. So in bag of word model just we simply see this whole as a bag. We don't take care of this grammatical like what exactly the syntax of there.

In bag of word model we generally take all those words what we call is in that language of text mining is that tokens. So each of these words are called as tokens. So we extract the tokens from the text and then we just see the frequency and we do our analysis what kind of words are coming together also in bag of word model we do like n gram kind of thing. So what exactly n gram is there so when we are considering this Mr. only as a word so it is a one gram word because one word is there.

Two gram will be so what will be Mr. and X so we will see the whether Mr. and X are coming together so in all the our documents so that will be our two gram. So it is generally called as bigram and this is called as unigram. And likewise it can go to n gram so we can just check that what exactly the words or group of words are coming together.

So syntactic parsing is different from the earlier bag of word model. So here the grammar of

the sentence is taken into consideration. It uses technique like part of speech tagging to understand the relationship between word based on syntax rules. So in this course we will be using the bag of word model. So let us now understand that what exactly the process of text mining.

So in the process of text mining first we need to identify our aim of the study. So let us say that our aim of the study whatever the text mining we are aiming to do. So what exactly our problem is and what we need to do. For example we want to understand the key topics of discussion on the online forums or maybe about the key trending topics are published in that particular journal or maybe related to that domain. So after defining our aim so the next step is to select the data source.

What exactly data source is? Data source is where we will be extracting our data. So if we are doing the analysis on online platform or online discussion forum like stack or flow or quora so our data source will be those platform. But in our case what exactly our data source is? The bibliographic data basis like Scopus or Web of Science or Permet. So whatever the text data we are getting like title and abstract we will be considering that for our text mining purposes.

So our data source is selected. So this is our second step and this is our first step. So after selecting the data sources we need to extract the data from those sources. So then we have to go for data extraction. So we have already discussed about how we can extract the data from these different bibliographic data sources like Scopus, Web of Science or Permet. But if we have to extract the data from different online discussion forums or different other platform we have to see the different kind of like we have to use different API's or different tools for extracting the data.

So until now these three steps we have already discussed a lot about in this course. So the next thing is that we have extracted the data and now we have decided that we will be taking care of title and abstract. So the fourth step in the text mining process is about the preprocessing of the text data. So preprocessing of the text data is the primary step in any of the text mining approaches. This process have multiple layers of operations like removal of stock words, text normalization, lametization, stamming and others.

So why preprocessing is done? So if you see our example here. So if you see these two sentences that these two goods they are conveying the same meaning but written in different way. One is in the lower case and this is in the upper case. So if we do not preprocess this text data and we directly move to analyze this so it will be considered as two different token. So this is one example where we have to do the preprocessing.

Another example can be say for example in this sentence Mr. is written here with dot but here the Mr. after Mr. there is no period sign is there.

This dot is not there. So it will also consider as two different token but they are the same but we have to remove this punctuation also like is, a and other thing like and these are the common words. These generally terms as a stock words. So we need to remove these also because they are just connecting the sentence they are as per grammar. So what exactly these stock words are and what is punctuation and what these lower casing will discuss when we will be doing the preprocessing but preprocessing is required for doing the text mining. And finally in the last step what we do, we do apply a technique for the analysis on this preprocess data.

For in our case we will be applying the topic modeling technique and using LD approach. So this is all about the process of text mining, how exactly the text mining works. We will not be discussing these three steps because we have already discussed in this initial weeks. We will be taking care of these two steps. We will be preprocess the data and then we will do the analysis using LDA.

So if we see the application of text mining, so there are like various applications of text mining in different areas. So the mining of the text data of discussion posted on various online and social media platforms can help us in to understand the sentiments and opinion about an event or a product or a thing. More personalized services may be provided based on the text responses received from the customer support. In health science we can use the text mining for the analysis of medical reports and other related documents. Then we can use text mining in the information retrieval from large number of documents.

Then we can do this text mining analysis on the responses we have received from this open ended questionnaire and surveys. And it has various applications. I have listed here some of the application of text mining but it has many applications. So there are like different techniques are there for text mining. So one is text classification, another is text clustering, then information extraction, topic modeling, sentiment analysis and text summarization.

So we will briefly discuss what exactly all these techniques are. So in text classification what we do? So in text classification we map the documents with the predefined categories. And the best example for understanding the text classification is what we do in the libraries. So what exactly we do in the library? We classify the content based on some scheme.

We have the predefined categories. Say for example we have this DDC. DDC is a DeWitt decimal classification where we have 10 categories like 004 computer science, then 100 for philosophy and psychology, then we have 200 for religion, then we have 300 for social sciences, 400 for languages, 500 for science, 600 for technology, 700 for arts and creation, 800 for literature and 900 for history and geography. So in text classification we have these predefined categories. So we have these 10 categories and whatever the text document comes, we label based on this category. So say for example we have a text on maybe chemistry or maybe on economics.

So chemistry is in under which category? Chemistry is here. So chemistry is basically 540 and then economics is here in social science. So it is 330. So whatever the text it comes, it will be labeled as 540 for chemistry and then 330 for economics. Now if we receive another text like say for example on inorganic chemistry or then for labor economics.

So inorganic chemistry is 546 and the labor economics is 331. So now this labor economics and economics they will be under the same rank and this inorganic chemistry and chemistry will be on the same rank. Now this labor economics and economics will be on the same rank and this chemistry and inorganic chemistry will be on the same rank. So what exactly it will help? So it will help the user to identify the collection easily. So this is what exactly text classification is. So what exactly we do in text classification? Basically we classify the text based on some predefined categories.

So in our example we have taken up the 10 categories but it is up to the analyst maybe some other one to classify the text maybe in five categories or six categories but those categories are predefined. So in machine learning terminology this kind of approach is known as the supervised approach. What exactly supervision is? Supervised is that we have defined the categories. These are the labels whatever the content is there it should be labeled with this.

So the next text mining technique is the clustering. So the major difference between the classification and the clustering is that in classification we have the predefined category but in clustering we don't have the predefined category. We give the data and let machine decide what could be the categories of this particular data set. What all these different categories are? These different categories are based on different perspective. Maybe we can have the text on different subject cluster or maybe like we have the clusters on polarity of the text. So what exactly the content is positive and what exactly the content is negative and that is why the clustering approach is also known as the unsupervised way because here we don't have predefined categories.

So the next text mining technique we have is information extraction. So in information extraction what we do? We take the unstructured text data and we extract some of the key information from that unstructured data. So if you see this example here so this was the unstructured data and from this particular data we have extracted these authors. So these were the author and these were the author.

We have extracted these years. Then we have extracted the citations. So these are the citations are there here. And we have extracted the title of the publications. So these are the title. So this is one title then this is one title. This is another text mining technique of information extraction from the unstructured text data.

So next we have this topic modeling approach but we will be discussing throughout this week. So in topic modeling we extract the key topics from the text. So say for example we

have these different documents. So these are the documents.

Document 1, document 2, document 3, document 4, document 5. Then say for example we have topic 1, topic 2, topic 3, topic 4 and topic 5. So what exactly topic modeling is that? So documents have topics and all topics have some words. We have now words here. So all documents have some portion of a topic and those topics are made up of a word.

So this is what exactly the topic modeling is based on. And these constituent words are selected based on each word's probability of association with the topic. So I will be discussing more about this topic modeling when I will be taking up the example of LDA and how we can do the topic modeling on our text data set that consists of title and abstract of 50 publications. So the next we have this sentiment analysis. So what exactly the sentiment analysis? So say for example we have this content 1, this content 2, content 3, content 4, content 5. Now within these five content we want to know that what exactly has been polarity of this particular content.

But whether it's positive, negative or neutral. So this is what we do in sentiment analysis that we want to know that if some post has been posted on any of the platform, whether it's sentiment is positive, whether the opinion about a product is positive of a customer or a negative or whether they are in neutral. So this is what sentiment analysis is. Then in the last we have another text mining technique which is for text summarization. So say for example we have a long papers.

So if we see here that this is our paper is there. So in this paper we have this title and then those introduction, methodology and then results. And we have different such papers. So this is paper 1, paper 2, paper 3. And now we want to summarize it.

So in text summarization what we do, we summarize the content. So maybe these are like 5000 words are there. Now we want to summarize it in maybe in like 200 words or 300 words because the whole publication is generally contains of many things. So what we want to know only the key information from that paper.

So like what we do in abstract. So abstract is a kind of a summary of a paper. So in one way you can say that we have a full publication and now we want to have abstract from that text data. So this is what we do in text summarization. So this is all about the text mining. We have just discussed very briefly about this text mining and then we have discussed about what exactly the process of text mining and what are the different techniques are therefore doing the text mining. And we will be focusing mainly on this topic modeling and in the upcoming lectures we will be discussing about the two of the steps of the text mining because we have already extracted the data.

Now we will be doing the pre-processing and then we will be doing the analysis. So in the next lecture I will be discussing about the regular expression which is a very important for

doing the text pre-processing because we have to match the strings whenever we have to remove something we have to remove unwanted things in a text data. So after studying the regular expression we will see the how we can use those regular expression in pre-processing. We will also see that how we can map the whole data into a lowercase or we can remove the punctuation or numbers what exactly the stroke parts are there and how we can remove the stroke parts. And then in the final step we will see how we can apply the topic modeling on this pre-process data. Okay so see you in the next lecture thank you.