Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 05
Lecture - 23

Lecture 23: Feature Engineering – II

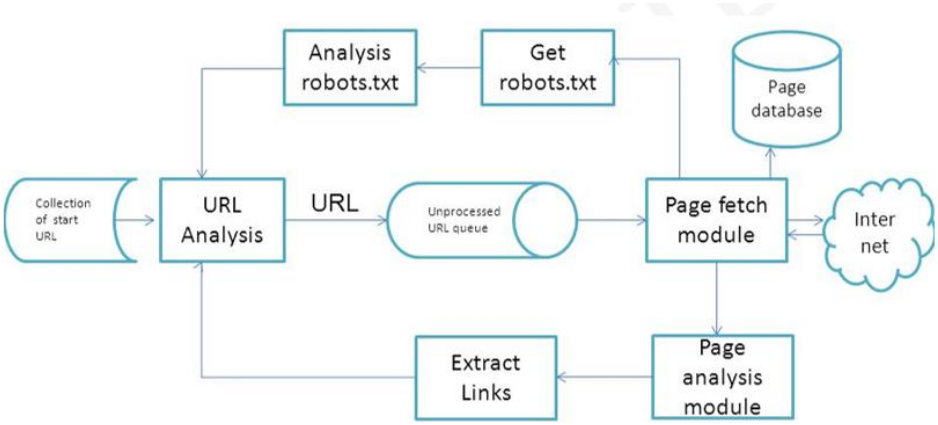
Hello everyone. Welcome to the second lecture on Feature Engineering. Which we are talking in the context context of content based recommender system. We are again going to talk continue our discussion on feature extraction and that too we will be talking about text features, then we will be talking about feature selection. So, if you remember last class we were talking about the first step towards text analytics is data collection and standardization. And we saw the example of a movie recommender system where the data ah in the context of movie recommender system the data may be coming from two different sources.

One you may get from the ah movie lens data set where the ratings are specified and maybe another you could be getting the item descriptions from IMDB database. Now, from IMDB if you do not have an API or something or even if you have API and you extract the data it may be coming in XML form where you will have to extract the extract the irrelevant tags and all. So, therefore, getting the data from the websites is a very important part. So, for this purpose typically we use some kind of web crawler.

These web crawlers go through the this is a typical structure of a web crawler and this web crawler try going through the ah web page going through various links and download the relevant pages. And after you get those relevant pages you will be extracting ah you will be removing the tags and all then from the text you will be extracting the features. Now, for this URL analysis you have to be really careful about the robot dot txt which specifies what kind of information you can scratch, you can scrap from the ah web page. But whatever may be the case you have to create a database of those pages and then from that database you have to extract the features. When we talk about the features this features can be word level features.

So, this ah word level features when we talk about the word level features this is the most common kind of thing. But however, while considering their these features we have to take care of the relationships that exist among this words like this homonymy, polysemy, synonymy, hyponymy etcetera. So, ah for example, if consider the case of homonymy. In this case if you are considering the word like that of bank then that bank can relate to river bank or it can be financial institution. Similarly, polysemy is about the getting the words of same form, but having related meaning.

For example, here again this blood bank and financial institution can take place, synonymy, synonymous words like singer and vocalist, hyponymy, the word ah denotes the sub class of another breakfast and meal. Then we have to look for those words and their frequencies. So, typically you will find that there are very small number of words small number of very frequent words and very big number of low frequency words. Basically if you plot them they will follow a power distribution. You have to help take help of certain tokenization software which can split the text into different words.

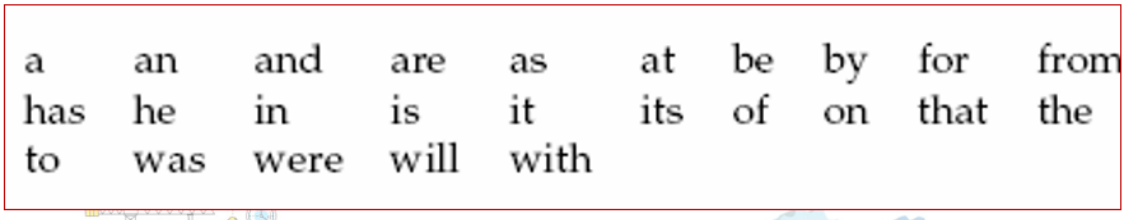


Now come to tokenization. This is about breaking the stream of the character into words and more precisely tokens. So, this sentence this first sentence has many commas and each of the words removing the commas are represented below there is a semicolon as well. Then this token is an instance of sequence of characters that are grouped together and is useful as a semantic unit. The class of all token as containing the same character sequence are called type.

The term that is includes in the dictionary is a normalized form of this. Now come to the characters like space tab new line etcetera which are called delimiters and they are not counted as tokens and they are collectively called white spaces. But while doing so, you have to be little careful. For example, here there is a white space here there is a white space here there is a white space as well, but still they are exceptions and they these two words together will be calling called a term. So, while tokenizing such exceptions should be taken care of.

During this process though these are typically thought of delimiters this may not be so in many example situations. For example, here dot here apostrophe dot ok they cannot be delimiters and all here for example, dash it it is not here comma and so on. So, many times while tokenizing you have to take care of this special situations and this tokenizations are also language specific. So, specific rules need to be language specific rules need to be understood. Now, coming to the next stage we are supposed to we you have now tokenized the data and you have got many terms and those terms you are now supposed to see as I told you they follow some kind of power distribution.

There are some terms which occur so frequently that they do not have any discriminatory power neither they have can describe any kind of relationship within among the terms. So, such terms are called stop words. So, they do not carry any additional information they mostly have functional roles they usually help the methods to perform I mean if you remove them they help the method to perform better. This stop word list is also language dependent for example, these are some typical stop words in case of English language besides that there can be domain specific stop words as well. For example, in a collection of details let us say you are talking about the movie recommend a movie reviews the word movie will be coming many times.



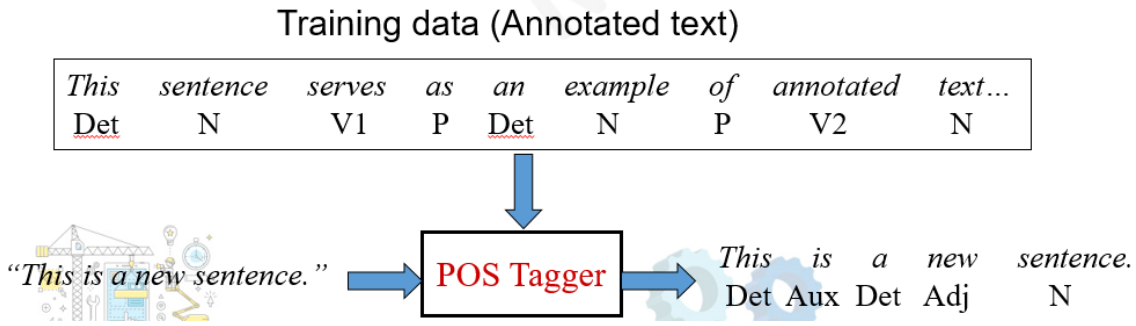
So, such highly frequent words which occur in almost every document will not have discriminatory power. So, those stops have to be removed. Then it is about finding the root word. So, how do you find the root word? Finding the root word there are two approaches stemming and limitization. In case of stemming you simply chop the word for example, you decide here you chop the word you take up to this many number of letters.

So, this is a stemmed word of this, but in case of limitization you have to find out some root word for example, here democratic democracy is the root word. Sang sing is the root word. Now the features can be obtained from the tokens if there is no special need else other steps are performed depending on the requirement. So, which means the next stage that we are going to look at a little bit advanced. So, if you are happy with your word level features you can stop here after this stemming and limitization otherwise you are supposed to go for others.

For example, here part parts of speech tagging. Now if the goal is more specific say recognizing names of people, place, organization it is usually desirable to perform certain additional linguistic analysis on the text to extract more features. Parts of speech analysis is one such step. So, there will be certain parts of speech tagger available in your NLP package whatever package you are trying to use and here this has to be trained this particular tagger has to be trained with some data. So, that this parts of speech like noun, verb, different types of verb, ah, determiner and so on has to be determined.

So, once this is ready when a new sentence you put the tagger will be attaching the tags auxiliary verb, determiner, adjective, noun and so on. So, through this process if you are interested to find out certain names and all you will be able to find out. Next is your word sense disambiguation. So, there are many English words which can be ambiguous and

when you isolate them from their parts of speech status they become ah meaningless without any reference. So, while finding out such words we take help of something called dictionaries.

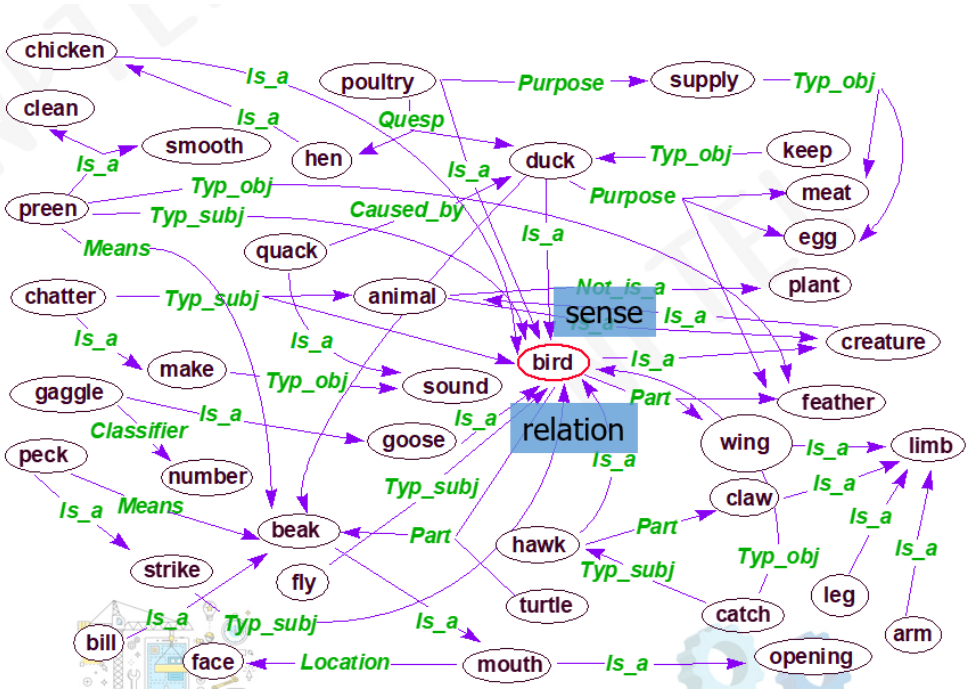


So, there are many large scale dictionaries like WordNet there are many projects which develop this develops these dictionaries which shows their words and their interrelationships. So, by using such taxonomies you can have ah I mean if you use taxonomies or thesaurus you can now find out the synonymous words and may converse them into a single word. In a typical this is just some details about the WordNet dictionary. So, here it is a database of lexical relationships. So, it has nouns, verbs, adjectives and adverbs and it can tell you if the words are synonymous or not.

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

This is just a small part of the word where it shows like a bird which is called a sense what are the relationships there are many relationships bird creature bird is a creature then ah goose is a bird so and so on. So, this is just a graphical view of a very small part, but the this will be actually related ah represented in terms of many relationships. What are various relationships? Hypernym which connects certain lower level concepts to higher level concept for example, breakfast is a type of meal. So, why do we do so? For example, if you like to reduce the features and if you have if it mentions that breakfast then lunch then dinner everything you can replace something by called meal ok. So,

similarly hyponeme from concepts to subordinate details if somebody has mentioned as meal and if you want to elaborate the feature you can think of whether it is a lunch or not.



Similarly has a member has a part antonym and synonym and so on. Then you can also recognize the phrase. Phrase in this sense here it is called ah set of a group of words which together makes some sense. So, when we talk about the togetherness of the word they can be sequentially they has to be they are sequentially and they can have certain meaning when they appear in sequence. And you when collecting them from the ah from the sentences you have to decide how many such words you will be taking together.

So, here for example, if you decide let us say our one of the sentence from the movie review I have taken the sound of music is an all time classic family musical that has always been one of my favorite. So, what are the n grams here? Suppose we decide n equal to 2 the sound of the sound sound of of music. So, these are n grams. Similarly taking 3 words at a time in an overlapping window makes the 3 grams. So, many of this widely accepted n grams, bi grams, tri grams etcetera they also are provided by ah industry leaders like that of Google and they can be directly used to consider as features in the text document.

In a document if there is some name even if it is in two different parts this you would like to combine them and treat them as one term. So, that is how you will have some kind of again automated process which will be helping you to get those names. So, they are to be trained based on their domain. So, they are domain specific names of drugs, medical conditions, names of ships, bibliographic references, movie details in our example

continue with our example ok. Now, we are in a position to represent the text in the form of features.

So, when we talk about the texts now what we did we do? We did tokenization to quad level feature tried removing some synonyms etcetera, try getting the root word ah. We tried getting the root word using this lemmatization etcetera then we tried keeping the word with comparing it with the synonymous word and keeping one word. So, words sense disambiguation we did and we also thought that why one word we can consider multiple words. Sometimes this multiple words can be with certain semantic connectivity like name of a person, name of some place combined together with a white space in between or they can be simple bigrams or trigrams without considering such relationships simply moving and window in the text and taking the bigrams and trigrams whatever may be the case right now we consider all of them as the features and terms or the terms. So, now, you will have such terms represented representing one document.

Relation	Definition	Example
Hypernym	From lower to higher concepts	breakfast -> meal
Hyponym	From concepts to subordinates	meal -> lunch
Has-Member	From groups to their members	faculty -> professor
Member-Of	From members to their groups	copilot -> crew
Has-Part	From wholes to parts	table -> leg
Part-Of	From parts to wholes	course -> meal
Antonym	Opposites	leader -> follower

Let us say these are my news articles. So, these will be my terms. So, whatever it may be word level feature it can be some kind of phrase or some kind of engram and so on. Moving on this is again not there is no a fixed method to do so, this requires the knowledge of little your domain knowledge as well as the underlying data set and as I told you this may not be that this this is an iterative task. In one go there is no rules or algorithms which can tell you that these are going to be the features.

Now next task is once we identified what are going to be the features we are going to put certain weights. So, what are the weights? So, first of all simply present or absent present or absent. So, such 0 1 level binary features you can adopt or you can simply adopt the frequency of the terms. So, you can simply use the term frequency. So, what is the term frequency? It is how many times the word is occurring within the sentence.

So, similarly you can have something called inverse document frequency which says in each document in how many documents they are referred. So, now, look at this we have something called term weights or the term frequency. So, more frequent the idea is more frequent the term in the document they are more important they are more indicative of the topic. And sometimes we would like to normalize it normal in the sense like the total frequency divided by the maximum frequency among all the terms. Then another way out of looking at getting this term weights is to get the inverse document frequency which says that the term that appears in many different documents are less indicative of overall topic.

For example, if we are talking about movie recommender system if movie is appearing in every document then it loses is distinguishing characteristics. So, this DFI is the document frequency of term I the number of documents containing that is number of documents containing the term I and IDF is the inverse of this inverse of this and log of this is used to dampen the effect of relative term frequency and relative to term frequency. Now it is observed that typically if you combine these two it gives a very good feature very good weight for the feature. So, this T f into IDF which is T f into log of n is the total number of documents total number of reviews etcetera that you have divided by the document frequency. Now once you have found this or you have decided the see T F IDF or only T f normalized T f or only binary factors binary ah values it is your choice and you are supposed to decide which works best with your predictor that you are going to use.

So, experimentally it has found that this T f IDF works very well, but it may not be the case may not be in your case. So, you have to decide, but whatever may be the case you have to do in depth study of the text analytics area, but now the idea is how we are representing the items. What was our original topic from which you have come here? We were trying to talk about content based recommendations where the recommendations were actually getting generated using some kind of predictive model and that predictive model requires certain input and the item features will be the input and the output will be the users preference. So, for that purpose we were trying to generate a data set. In that data set this features represents the features describe the items and the response variable which is collected from the from the rating matrix will be serving as the additional dependent variable.

So, sometimes it is also possible to combine suppose this and one more thing I forgot to tell you. In fact, we will be discussing them in detail coming to this intrinsic feature we have to now decrease the number of words will be very large number of terms. So, we have to decrease and take very few. So, considering all the extrinsic features and intrinsic features is also possible. So, getting them together we can make a complete data set.

Now, what we have done so far? We have now decided this structured representation of the item ok. So, now, based on this we have to now decide our training example. For getting this training example one additional data that is response variable need to be collected. Where from it will be collected? It will be collected from the rating matrix. Now, these users like dislike or in the form of implicit data has to be now collected.

Text opinions as well. So, these users like and dislikes can be collected in from 4 sources. Rating in this case the user specifies ratings indicating their preferences for the item. Rating can be binary interval based ordinal ordinal in rare cases rating can be real value valued as well. The nature of the rating has a significant impact on the model used for learning the user profile. Now, this can be collected from the implicit data which refers to the how the how the user behaves while browsing etcetera.

Now, it can come from the text opinion of the user who rates an item as well as give some additional feedback. It can also come from the example cases which the user has already seen. So, now combining this we have now training data set. These are the item features and this is our response variable ok.

So, now our training data set is ready. I am starting now ok. So, so with this we now stop this lecture and these are our references and as you can see here we have referred to text one text predictive text mining book and another book on information retrieval. So, these two book you can refer to get more on how to extract the text features and how to represent them in numeric form. So, these are our concluding remarks. For the purpose of content based recommendation the items are to be represented in the form of certain feature vector and this feature vector can contain text descriptions and this while having this text descriptions we have to convert them into numeric form to do so, we have to discover the terms which can be single word or multiple words and have some kind of numeric representation for this which can be used as the input and some such numeric representations are either it can be binary, it can be used certain kind of term frequencies rather normalized term frequency and TF-IDF rating.

Now, this item features can be combined with rating values to make the training data set. Thank you.