

A Methodology to Identify Topic of Video via N-Gram Approach

Ramsha Pervaiz¹, Khalid Aloufi², Syed Shabbar Raza Zaidi¹ and Kaleem Razzaq Malik¹

¹ Department of Computer Science, Air University, Multan Campus, Multan, 60000 Pakistan;

² School of Computer Science and Engineering, Taibah University, Saudi Arabia;

Summary

Keyword helps in extracting the main idea from any document. It plays an important role in information retrieval from the content of the document. Keyword extraction is the process of detecting a keyword from any document that is easily understood by the users about the content of the documents. Keyword extraction is of vital importance in natural language processing. The keyword extraction is used for information retrieval, visualization, text summarization, classification, clustering, and web searching and topic detection. There are two main classification of keyword extraction, first one is supervised learning technique in which data is trained through dataset. Second is unsupervised learning in which no need to train the data and the data is collected from statistical approach. In this research, topic is generated from the video lectures according to the content of the videos. The topic is generated from the videos in which course code is mentioned instead of topic. Those videos cause problem in understanding main idea and content of the video lecture. The user have to listen all the videos without knowing the content of the video which is the wastage of time. Using unsupervised learning, frequency of words and combination of words is counted by N-Grams. The keyword extracted from these N-grams are compared with the data set of computer terms and the topic is generated of the video.

Key words:

Natural Language Processing (NLP), NLTK, N-grams, Keyword, Extraction, video lecture

1. Introduction

According to the International Encyclopedia of Information and Library Sciences [1] key word is defined as a word that concisely and exactly explains the topic or a part of the topic that is being discussed in a paper. Key term can not only be the single words i.e. key words but they can also be group of words i.e. key phrases. According to the book Foundation of statistical Natural Language processing, the phrases can be described as: Words do not occur in just any old order. Languages have constraints on word order. But it is also the case that the words in a sentence are not just strung together as a sequence of parts of speech, like beads on a necklace. Instead, words are organized into phrases, groupings of words that are clumped as a unit. One fundamental idea is that certain groupings of words behave as constituents. [2]

The keywords that are described to be series of one or more than one words, give a very dense portrayal of the material that is present in the document. Preferably, the keywords that are present in more dense or concise form give a very important part of the document. Keywords are more convenient to be explained, shared, memorized and revised. Thus they are being used extensively to explain the problems that involve the Information retrieval system. When compared with the mathematical systems, we realize that the keywords do not contain any extra written texts and they can be implemented around different corpora and Information retrieval systems.

Many articles, magazines and newspapers are being published online during these days. They make it very complicated to look through every document that is present. Hence, in order to eliminate this difficulty we use the keyword extraction methods as they give the important component that explain the whole topic of a document. By using it we can conclude the whole summary of the document by using these keywords. But As the keywords explain the meaning of the whole document, they can be taken advantage of and can be used by the applications for many purposes like retrieval of information, generating the report, visualization of the information, classification clustering, detection of topic, filtering, summarizing of the text etc.[3]

Key word generally contains one or more than one words that can explain the reader what is present in the whole content. We can use the keywords to show the main point of the article. Thus, when the readers read the keywords they can know if the article they're reading contains the content that they desire or not. By using this explanatory quality of the keywords we can tell if the keywords can help in the understanding of material in machine and help in the betterment of the standard of the clustering of the document categorization task in comparison to the other methods that are traditionally used and are involve the usage of every word that exists in the document. As talked about before, the keywords are very practical in the summarizing of the document doing it manually can consume a lot of time and also, there are more chances of errors in it. The language we use is very flexible, we can elaborate same thing by using distinct words. Plus it gets difficult for the reader to explore the similar articles that

have the same writers. It is not different when we talk about document clustering. In its scenario, it is difficult and time taking for the readers to understand that the article is related to which cluster. Thus, in order to resolve this difficulty, text mining system will be designed and used which will spontaneously give keywords to the articles and the cluster information. Not just this, it will also give the human feedback loop that will help in the betterment of the standards of the results of the system[4]. Both, the readers and the editors are helped by the text mining methods. The writers can be helped in their jobs as it helps them to work more quickly and more easily as it gives them tools for the exploratory analysis of the topics that are present in the document.

This helps the writers to investigate the way in which the topics become different with time and can conclude if it is beneficial to further explore the topic or not. There is also a need for the writers to put the data in the form of such an organization that the structure formed is in such a way that the topics with related information come together. The methods of text visualization and ontology organization will also be explored by us. [4]

In the automatic keyword extraction the key terms, words or the key phrases are located from the article or the document. These keywords can easily represent the whole idea of everything that is present in the document. [5].

The web constitutes of a very vast range of knowledge and it is still continuing to grow making the amount of the digital document that are present also grow and thus resulting in making the manual keywords extraction an impossible thing to be done. The key word extraction also plays a very great role in the mining of text, in the processing of the natural language and in the retrieval of information. A lot of applications including automatic indexing, automatic classification, automatic summarization etc. can take advantage of the keyword extraction process as it gives a very dense presentation of the document [6]. Keyword assignment and the keyword extraction are the two classes in which the automatic keyword generation procedure can be divided [2].

When talking about the keyword assignment, a group of practice able keywords are chosen from the limited words that are present. While the keyword extraction recognized the best related words that are present in the document under study.[5]

We can broadly divide the keyword extraction methods into four different approaches, namely linguistically approaches, machine Learning approaches, statistical approaches and other approaches [7]

An important subfield of text mining is text classification. It helps in assigning the text document to one or more than one groups that are already defined.

There are many ways in which the text is collected, including the articles, digital libraries, and the web pages. These are very significant ways of providing the information [7]. Therefore, we can say that the text classification technique is very beneficial in the field of research when talking about the library science, computer science and the information science[8]. We can model a lot of applications as the text classification problems. There are many things in these applications including the filtering of news, organization of the documents, sentiment analysis and filtering of the spam [9]. High dimensional feature space is very common challenge faced by the text classification application.[10]. The system of the classification of the texts turns into a computational intensive task while each and every word of the feature documents is used as the features[11]. Hence we can say that the keywords of the text collection are the major and related words of the text collection that can easily explain the content of the document and can be used as the candidates to explain the properties in the classification model construction [12].

The most important thing in the area of text mining is the keyword extraction technique. By using many different methods we can carry out the process of keyword extraction. These include the unsupervised and supervised machine learning, the statistical and the linguistic methods.

There are many ways for carrying out the process of keyword extraction and every process has its own advantage and disadvantage. There are four methods for carrying out this process including:

1.1 Rule Based Linguistic Approaches:

In these types of approaches, the rules of linguistic knowledge or features are used. These don't only require the domain knowledge but also the expertise in the knowledge. These are more correct but unfortunately they are also computationally insensitive.

Linguistic features of the words, mainly the sentences and documents are used by these approaches. The lexical analysis, syntactic analysis discourse analysis are the examples of these approaches.

Keyword extraction approaches usually use the linguistic information that is present in the texts for example all the words that are present in them. All the pieces of information that have been discussed till today cannot be described, however here are a few of them:

Sometimes, morphological or syntactic information, such as the part-of-speech of words or the relations between words in a dependency grammar representation of sentences, is used to determine what keywords should be extracted. In some cases, certain PoS are given higher

scores (e.g. nouns and noun phrases) since they usually contain more information about texts than other categories.

1.2 Statistical approaches

The statistical method approaches are simple. They do not require the training data. In these types of approaches, the keywords are identified by using statistical methodology. N-Gram is used by Cohen for the indexation of the document automatically. N-Gram is not dependent on the language or the domain. Word frequency, TF*IDF, word co-occurrence, [6] are also statistical approaches for the selection of keywords.

The statistical approaches are commonly built on linguistic corpus and statistical features that are derived from the corpus. These approaches are not difficult and they do not require any training data. We can use statistical information of the words to locate any keywords that are present in the document. The simplest way for locating the major keywords and key phrases is by using statistical methods.

Word frequency, word collocations and co-occurrences, TF-IDF (short for term frequency-inverse document frequency), and RAKE (Rapid Automatic Keyword Extraction) are the different examples for the statistical approaches.

It is not necessary to use training data for extracting the major keywords from a text. But as they rely on the stats, they can ignore the words that are not used more than once but still play a key role in the document. The extremely important benefit of the statistical approach is that they are not dependent on the language in which they are applied and thus, we can use this technique for the other languages as well. Although these approaches might not show the perfect results in comparison to the linguistic ones, but it has also made it possible to carry out the statistical analysis and get good results by the availability of large amount of dataset.

1.3 Machine Learning approaches:

When talking about the machine learning approaches, the supervised methods are generally applied. The keywords are taken out in these approaches from the training document in order to understand a model. The model can be tested by using the testing module. When the reasonable model is developed, it can be used to locate the keywords present in the document. In this method, the Naïve Bayes, Support Vector Machine, the Naïve Bayes are used. There is a need for the tagged document corpus for the supervised learning methods which is very complicated to be generated. When this type of corpus is

absent, the semi-supervised or the unsupervised methods are used as an alternate.

In many tasks that have the text analysis based learning like text analysis, the machine learning based approaches can be used. The answer to the question that what the machine learning approach is, is that it is a sub-division of the artificial intelligence. It involve the generation of the algorithms that have the ability to understand the examples and give the conclusions from them. The machine learning systems change the data into an understandable form instead of processing the data that is not structured. From this, a question arises that what can make a machine learning model do this? The answer to the question is that the data is first converted into the vectors which are actually the group of numbers that have encoded data. They have different features that are involved in the representation of the text.[13]

Many machine learning algorithm and methods are used for the extraction of keywords that are most related with each other. For example, the deep learning and the support vector machines. Following are the most usual ways for the extraction of keywords by using the machine learning method.

1.4 Domain specific approaches:

We can use many different approaches to a particular corpus in order to use the background knowledge that is connected to the domain for example ontology and thus, resulting in that specific corpus to recognize and separate the keywords. [13]

1.5 Other approaches:

In the remaining approaches of the keyword extraction the approaches that have been discussed before are combined or the other way is that the useful information from the task of keyword extraction is used. These include the size, the html tags, and positions and so on. Many extraction methods that have been talked about are for just one document, still there are some methods that can be used for more than one documents according to their requirement. We can extract the key words by locating the noun phrases. This is majorly because the noun phrases have very decisive details about the text document. The keywords are chosen according to their lingual properties[14]. And also the knowledge they provide [15]. These include the words that are highlighted. The keywords can also include the words that are present in the abstract, or the titles of the document. Methods which include co-occurrence [16] and the machine learning [17] can also be used for taking out a keyword from one document. By used the key word clustering, the topics can

be recognized. Plus, the keywords related to the clustering and extracting are usually. Based on the history of the query frequency [18] can also be a method that can be selected.

2. Literature Review

Natural language processing is very important for the mining of the text for the retrieval of information and hence it has been continuously spreading in these days. As the main aim of the text mining is to locate the important knowledge that is present in the document, thus a lot of the researchers are working very hard for proposing good techniques that will help in the elimination of the problems in the tasks of natural language processing. Keyword search is a very strong tool as it has the ability of scanning of big documents in an efficient way. Even though it does not have any idea about the semantics of the language nor does it have any information or any knowledge about it. The major aim is to keep the information accurate and efficient so that it is beneficial for the user. Thus, to achieve this goal, taking out keywords from a big document will be very beneficial for the detection of the topic and analyzing it in an effective way.[19]

In the methodology of the keyword extraction by using statistics, the steps used for the identification of the key phrases or the keywords are the statistical methods. These methods can include gram-statistics, word frequency, etc. There might not be a requirement of the training data for carrying out these methods and they can also be independent of domain[5]. The methodology for taking out keyword statistically was introduced [16]. In the start, many words were taken out. After that the occurring of the same term many times and the repeated term were observed. Then the determination of the document is done on the basis of the dispersal based on co-occurrence and the significance of the term that the text document comprises. There is no requirement of the training of the entity in this method. Also the results of this method are good enough to be compared with the TF-IDF measure. Another advanced method of key phrase extraction algorithm was introduced by Turney [20]. In these methods, there are statistical links between the key phrases. They help in the improvement of the quality of being logical for the keywords that are obtained. Web mining is used for the measurement of the links that are present between the key phrases. The text document can also be presented as an indirect graph by the other statistical keyword method [21]. Meanwhile upright part of graph has the phrases from text document, while edges have values that can be allocated on the basis of the

statistical measurement of the difference between multiple words. The methods that are linked to the linguistics, the qualities that are used for the location of keywords are based on languages. These examinations comprise of syntactic, discourse analysis and the semantic analysis [6]. The methods that involve linguistics depend upon the domain [2]. The inspection of the absorbance of the language based information including syntactical methods and the keyword extraction were carried out by [17]. The result of the experiment showed that there can be advancement by only using the statistical methods like the ngram or the term frequency. By only using the statistical methods, HaCohen-Kerner [22] gave a model for the keyword extraction in which the words are taken out only from the headings and the abstract. The methods like unigram, bigrams and trigrams are used for text presentation in this model. They keywords extraction algorithm was set forth by Nguyen and Kan [23] by using the scientific publications. In these procedures, language based properties like the location of the phrases in the document are considered.

A method of natural language processing was given [24] in which, the natural language processing method was changed into automatic key word extraction by using the scientific papers so that the execution of the machine learning algorithms can be enhanced. These include the vector machines and the Random Forests. The results that are gained by experiments are via the ACM dataset. The assessment is carried out by using the key phrases and the key word extraction algorithm that are assigned by the expert (KEA) The methodologies of machine learning like the support vector machines use Naïve Bayes, decision tree. It is important for the construction and the classifying model. The disadvantages of using the features that are based on the feature extraction model is that the model has to take a set of document that is tagged. An easy and systematic way of extracting the key phrase algorithm which uses Naïve Bayes algorithm for extracting the key phrase on the basis of domain was given [25]. In the following procedure, the keywords are selected by using lexical approaches and the better quality key phrases are acquired with the help of machine learning algorithms. An examination was done by HaCohen-Kerner, Gross, and Masa [22] that showed the results of what happened when the methodology of key phrase extraction and learning methodology were applied in the science based articles in the English language. The evaluation of the key phrase extracting methodology is done by using different methodologies of machine learning. The results of the experiments showed that C4.5 algorithm gives the best results as compared to the others for the domain. KEA++ is a method given by witten

[26] was an improvement in the key phrase extraction models.

The knowledge of phrases and terms that is taken from the thesaurus that is specific to the domain and is related to semantic is utilized for extracting the key words in automatic way. [6] made a model of the queries related to the key word extraction to a string labeling task. The conditional random field methods were used in this for the labeling purposes. The results given by the experiment showed that the conditional random field methods give good results in comparison methods like the conventional machine learning algorithms like the support vector machines and linear regression. The approaches that are based on graph are both the supervised keyword extracting methodologies. Text Rank is a text processing keyword as well as sentence extraction algorithm. It was introduced by [27]. In this method, the undirected/directed weighted co-occurring networks with changing window sizes are used. For extracting the keywords by this method the text of the document are compartmented into tokens. After that, the tokens are allocated with the tags of the parts of speech. For representing every token or a few, the nodes are made in correspondence to a specific part of speech. An association is made in between two nodes when there is an occurring of a word in multiple times [28]. An examination of the activities of supervised and unsupervised approaches that were based on graph summarizing the text was done by Litvak and Last [29].

For representing every token or a few, the nodes are made in correspondence to a specific part of speech. An association is made in between two nodes when there is an occurring of a word in multiple times [28]. An examination of the activities of supervised and unsupervised approaches that were based on graph summarizing the text was done by Litvak and Last [29]. For representing the text documents the syntactic representation method is used and it is graph based. A model for extracting the key terms was given by Grineva, Grinev, and Lizorkin [30]. It does the modeling of the text document in the form of a graph that shows relationship between the terms that are present. In the model that is made, the best related terms that give an idea for the heading of the document give thick components that are connected with each other. Graph community approaches are used for getting the thematic parts from the structure of the graph. In these techniques, the data taken from Wikipedia is used for weighing the terms and examining the semantic links in between them Huan, Tian, Zhou, Ling, and Huang. [31] did a study in which they showed a method for automatic key phrase extraction algorithm. It was helpful not only for supervised but also for unsupervised learning tasks. The basis on which the key

phrases are extracted are the structural dynamics of the semantic network. On a study that was presently carried out on the keyword extraction, a model was put forth that relied on the patterns of fraction [32]. It was shown from the results that the terms that were most related to the topic of the document had a fractional dimension that varied from one, while the terms that were not important had the fractional dimensional value of one. From this observation a conclusion can be drawn about the importance of words is calculated on the basis of fractional dimensions.

Writers suggested automatic extraction for keywords to meet corpus under a more organized way and bigram expansion [33]. They derived "entity bigram" by using bigram expansion in contrast to the TFIDF selection which is not supervised in that paper which has a good performance with POS filtering. TFIDF framework based keyword extraction was brought forward, which uses procedures such as POS of words as well as analyses the value of a graph based method.

In [34], a single loop feedback methodology was suggested for extraction of keywords by authors. Term peculiarity characteristics, decisions that construct a sentence as well as the traditional frequency or hints basing upon locations and collection of characteristics obtained from sentence summary. A feedback loop mechanism was suggested by them to devise improved system summaries which worked in the light of a framework that is supervised for improving association linking the keywords and the summary sentences. An LDA model based short web documents of hidden topic based framework [35].

Two main problems were treated using LDA model through MaxEnt classifier that is synonyms problems and data sparseness. Researchers have put forward many different algorithms for carrying out the process of keyword extraction and it can be divided into 3 different types [36]. These include simple statistics, linguistics and machine based. The basic techniques that are based on statistical approaches need limited pre-conditions and they are simpler to be understood as compared to the other methods. These are not focused on the lingual properties of the document.

The techniques for the extraction of keywords are also divided into two types, i.e. The supervised method and the unsupervised method. There is a requirement of annotated data for the supervised method to be carried out. The major focus that goes along with the supervised method is GenEx. This is a typically known and a very popular system. The whole keyword extraction is based on this. The classification of the keywords that are provided into the parts of the keywords is a task of classifying according to the binary system and it tells if the word is a keyword or not.

As talked about earlier in [17] Hulth has used Noun Phrase chunks instead of using term frequency and n grams. He has done exploration of the information based on languages into the taking out of the keywords. When the POS tags are added, they help in improving the quality of the results as they are then free of term selection approach. The NP-chunks yields good results when compared with the results of the n gram.

An algorithm for keyword extraction by using language based information from the scientific papers was proposed by Nguyen and Kan [23]. Few properties were introduced that could take out the salient morphological phenomena that are present in the scientific key phrases for example if the key phrase of the candidate that has been selected is an acronym or a particular terminology productive suffixes have been used.

In [24] Krapivin et al. Used the NLP methods for considering the machine learning approaches and improving them so that the queries related to the automatic keyword extraction from the scientific papers can be resolved. Interpretation exhibited that the outcomes which outperformed state of the art Bayesian learning system KEA on the same dataset without the use of controlled vocabularies.

In [27] Ha Cohen Kerner presented a very simple model that used the unigrams, 2 grams and 3 grams and stop word record and extracted keywords from the abstract and the titles. The model gave the weighing of the words and

the words with the highest weight of the group of words are called as the keywords.

The comparison of the supervised and unsupervised method was done Litvak and in [29]. This helped in identifying the method of extracting keywords from the document. The methods were on the basis of graph and the presentation on the document and the web document on the basis of the syntax. HITS algorithm gave outcomes of the document that was summarized gave results in comparison to the methods that were supervised (Naïve Bayes, J48, and SVM). The ranking based on the simple degree from the initial iteration of HITS, instead of doing it to the convergence must be taken into consideration.

In [37] a research conducted by Yang et al. put a focus on the difference of entropy in-between the intrinsic as well as the extrinsic ways for the extraction of keywords. It stated that the keywords that were extracted gave the mindset of the writer who had written that document. The entropy difference of Shannon was utilized between the intrinsic and the extrinsic mode in their methodology. It referred to the fact that the occurrence of the words can be adjusted according to the aim of the author whereas the words that are not relevant can be scattered in the text in a random way. Thus, it can be concluded that a document that has words natural language with the words that can be recognized in a clear way can be applied with these aims and do not need any information of the preceding syntax.

Table 1: Automatic key phrase extraction algorithms

| <i>Work</i> | <i>Year</i> | <i>Domain</i> | <i>Training Data</i> | <i>Method</i> |
|---|-------------|---------------------|----------------------|-----------------------------|
| Keyword and Key phrase extraction | 2015[2] | Dependent | No | Linguistic approach |
| An overview of graph-based keyword extraction method and approach | 2015[5] | Independent | No | Graph based method |
| Automatic keyword extraction from documents using conditional random fields | 2008[6] | Specific | Yes | Conditional random field |
| Keyword extraction from single document using co-occurrence statistical information | 2004[16] | Single document | No | Word co-occurrence |
| Improved automatic keyword extraction given more linguistic knowledge | 2003[17] | independent | Yes | NP-chunk |
| Coherent key phrase extraction via web mining | 2017[20] | independent | No | Web mining |
| Keyword extraction from single document using centrality measures | 2007[21] | independent | No | In directed graph |
| Automatic keyword extraction from abstract | 2003[22] | Abstracted document | No | Unigram, Bigram and Trigram |

| | | | | |
|---|----------|---------------------|--------|--|
| Key phrase extraction in scientific publications | 2007[23] | Scientific document | No | KEA $p < 0.5$ level |
| Keyword extraction from scientific document | 2010[24] | Scientific document | yes | Machine learning approach with NLP |
| Practical machine learning tool & techniques | 2016[25] | independent | Yes | Lexical |
| Thesaurus based automatic key phrase indexing | 2006[26] | specified | Yes | KEA ++ |
| Text Rank bringing order into text | 2004[27] | independent | Yes | Text Rank |
| Word clouds for efficient document labeling | 2011[28] | independent | No | Directed/In directed co-occurrence |
| Graph based keyword extraction from single document summarization | 2008[29] | Single document | Yes | Graph based |
| Extracting key terms from noisy documents | 2009[30] | independent | No | Graph based |
| Arabic key phrase extraction using hybrid approach | 2014[31] | independent | Hybrid | Statistical machine learning |
| The factorial pattern of words in a text | 2015[32] | Independent | Yes | Automatic Keyword Extraction Algorithm |
| Automatic keyword extraction using meeting transcript | 2009[33] | Transcript | No | Unsupervised Learning |
| Automatic keyword extraction using meeting transcript | 2019[34] | Transcript | Yes | Unsupervised Learning |
| A hidden topic based framework for keyword extraction | 2010[35] | Web Document | Yes | LDA Model |
| A vector space model for automatic indexing | 2018[36] | Independent | Yes | GenEx |

3. Methodology to Identify Topic of Video via N-Gram Approach

In this section, procedure is discussed through which key word is extracted for title generation. Keyword extraction is used for different purposes but in this research, keyword extraction is used to generate the topic of the video.

Mostly the videos have no title, so by using this approach title is generated according to the video content. By doing immense research, we have found out different approach in natural language field related to keywords extraction technique. From various methods, semantic relation caught our eyes and we developed a complete different approach to extract keywords from the documents (as shown in Fig. 1).

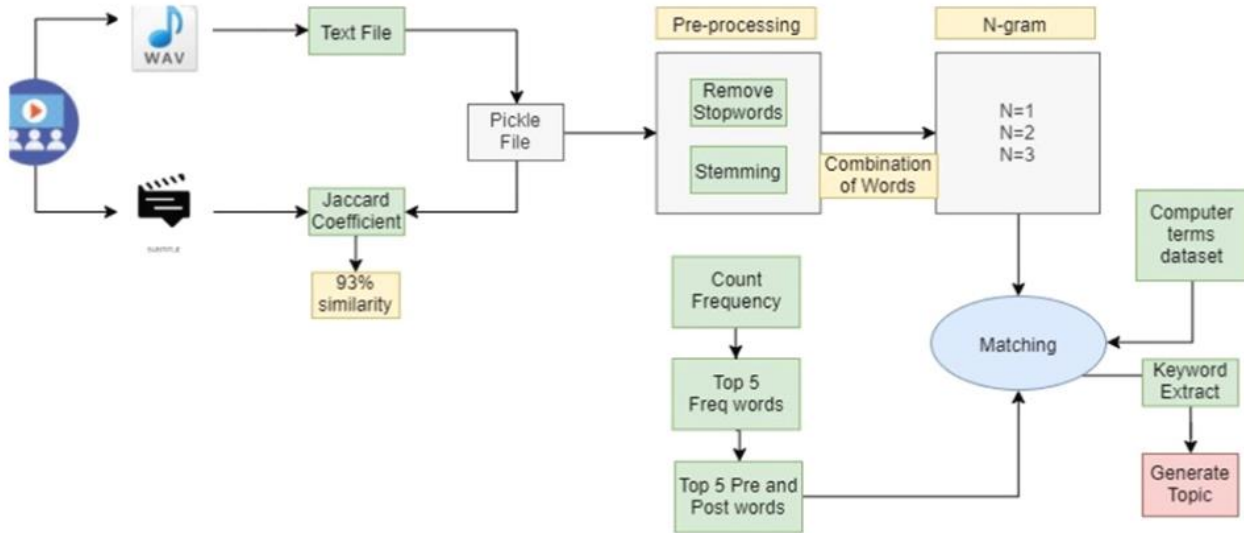


Fig. 1 Procedure of topic generation by using unsupervised approach

The procedure has the following steps:

1. Downloading the video lectures
2. Conversion of video into audio.wav
3. Conversion of audio.wav into text
4. Store the text into pickle
5. Preprocessing on text
6. Count frequency
7. Extract combination of words
8. Matching keywords
9. Keyword match with computer terms
10. Title /Topic is generated

3.1 Downloading the video lectures:

Download the video lectures from the internet with subtitles. Select the videos without topic but have mentioned course code on it

3.2 Conversion of video into audio

First a specified video is selected from internet. It is converted into audio by using google. In this way wav file is generated of the specified video (as shown in Fig. 2).

3.3 Conversion of audio into text

Now the duration of audio wav file is calculated and whole duration of audio is divided into 10 seconds. In this way every 10 second audio chunk is converted into the text (as shown in Fig. 3). Conversion results can be seen in Fig. 4 and Fig. 5.

Audio to text = Duration of audio in seconds / 10 seconds



Fig. 2 Audio into text

Conversion start with audio at 0 second



Fig. 3 Video into audio


```

loop iteration 0
Translation : 0 to 10 = in this lesson we will introduce you to linked list data structure in RPS lesson we tried to impleme
nt a dynamic list using arrays
loop iteration 1
Translation : 10 to 20 = and we had some issues there it was not most efficient in terms of memory usage in terms of memory
consumption when we use Aries
loop iteration 2
Translation : 20 to 30 = I have some limitations to be able to understand linked list well we need to understand this limita
tions so I am going to tell you a simple story
loop iteration 3
Translation : 30 to 40 = to help you understand this latest this is Computer memory and each partition here is 1 byte of mem
ory now as we know each bite of
loop iteration 4
Translation : 40 to 50 = has a dress we are showing only section of the memory that's why it is extending towards the bottom
and the top left side address increases
loop iteration 5
Translation : 50 to 60 = important to top so if this by his address 200 the next bite would be address 2019 next fight would
be address 202 and so on
loop iteration 6
Translation : 60 to 70 = I want to do is I want to do this memrv from left to right horizontally instead of trvine it from

```

Fig. 4 Snap of output at start of translation(Total Time of an Audio Clip in Seconds: 1032.08After finding the duration of audio clip, conversion of audio into text is started. Loop iteration started from 0 and every iteration contains 10 seconds conversion text data.)

Conversion end

```

Translation : 900 to 970 = the note with value for not to insert also we will have to traverse the list and go to that parti
cular position and show this will be big open again in terms
loop iteration 97
Translation : 970 to 980 = complexity the only thing is that the insertion will be a simple operation will not have to do al
l the ships as we have to do in an array to insert something in
loop iteration 98
Translation : 980 to 990 = Bihar to shift all the elements by one position at what I Indus similarly to delete something fro
m this list will also be o n so we can see some
loop iteration 99
Translation : 990 to 1000 = things about linked list there is no extra use of memory in the sense that some memory is unused
we are using some extra memory we are using some extra memory to store the
loop iteration 100
Translation : 1000 to 1010 = but we have the benefit that we create notes as and when we want and we can also free the notes
as and when we want we do not have to guess
loop iteration 101
Translation : 1010 to 1020 = size of the list before hand like in the case of Paris will discuss all the operations on linke
d list and the cost of these operations as well as
loop iteration 102
Translation : 1020 to 1030 = with our in our next lessons we will also be implementing linked list in C O C plus plus so thi
s is all for a basic introduction to link

```

Fig. 5 Snap of output at end of translation(Total audio clip duration divided by 10.so total iteration is 102All the audio data is converted into text but in the form of chunks.)

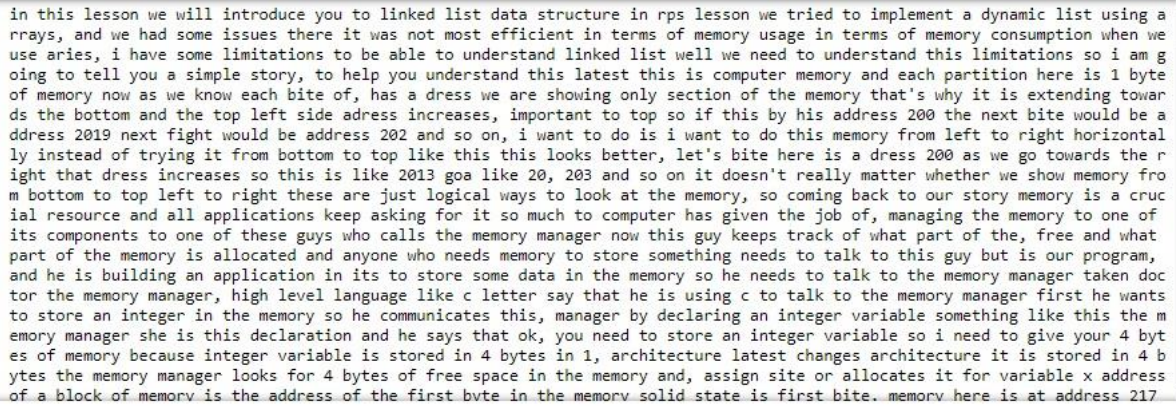
3.4 Store the text into pickle

After conversation text is stored in the pickle file. Pickle in Python is primarily used in serializing and desterializing a Python object structure. In other words, it's the process of converting a Python object into a byte stream to store it in a file/database, maintain program state across sessions, or transport data over the network (as shown in Fig. 6).



Fig. 6 Text stored as pickle file

Chunks are merged and separated by commas form a paragraph



in this lesson we will introduce you to linked list data structure in rps lesson we tried to implement a dynamic list using a arrays, and we had some issues there it was not most efficient in terms of memory usage in terms of memory consumption when we use aries, i have some limitations to be able to understand linked list well we need to understand this limitations so i am going to tell you a simple story, to help you understand this latest this is computer memory and each partition here is 1 byte of memory now as we know each bite of, has a dress we are showing only section of the memory that's why it is extending towards the bottom and the top left side address increases, important to top so if this by his address 200 the next bite would be address 2019 next fight would be address 202 and so on, i want to do is i want to do this memory from left to right horizontally instead of trying it from bottom to top like this this looks better, let's bite here is a dress 200 as we go towards the right that dress increases so this is like 2013 goa like 20, 203 and so on it doesn't really matter whether we show memory from bottom to top left to right these are just logical ways to look at the memory, so coming back to our story memory is a crucial resource and all applications keep asking for it so much to computer has given the job of, managing the memory to one of its components to one of these guys who calls the memory manager now this guy keeps track of what part of the, free and what part of the memory is allocated and anyone who needs memory to store something needs to talk to this guy but is our program, and he is building an application in its to store some data in the memory so he needs to talk to the memory manager taken doctor the memory manager, high level language like c letter say that he is using c to talk to the memory manager first he wants to store an integer in the memory so he communicates this, manager by declaring an integer variable something like this the memory manager she is this declaration and he says that ok, you need to store an integer variable so i need to give your 4 bytes of memory because integer variable is stored in 4 bytes in 1, architecture latest changes architecture it is stored in 4 bytes the memory manager looks for 4 bytes of free space in the memory and, assign site or allocates it for variable x address of a block of memory is the address of the first byte in the memory solid state is first bite. memory here is at address 217

Fig. 7 Chunks are merged and separated by commas form a paragraph

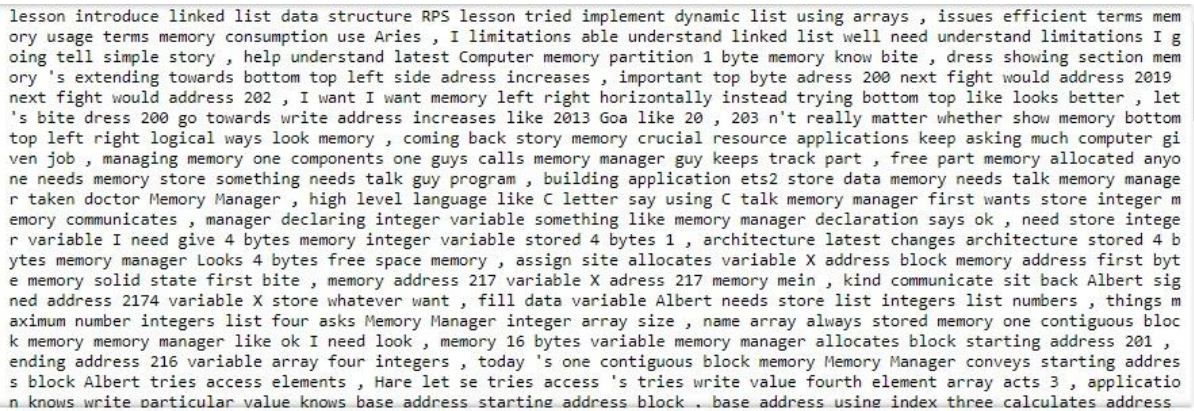
3.5 Preprocessing on text

Refine the data (as shown in Fig. 7) before processing and remove the data that is not useful. First remove the stop words and then perform stemming.

Stop words:

Documents contains grammatical words and other supporting words or POS to complete the sentences which doesn't contains any core information of the documents. In order to increase the processing speed, the system will

remove the Stop Words from the documents at the start of the procedure. In the selected topic, we collected a collection of stop words which contains almost 300 stop words. During the process of a document, stop words has been removed for further implementation procedure. Remove the stop words from the text in this way clean the data otherwise stop words used frequently in the videos and the main keywords are not extracted due to stop words (as shown in Fig. 8).



lesson introduce linked list data structure RPS lesson tried implement dynamic list using arrays , issues efficient terms memory usage terms memory consumption use Aries , I limitations able understand linked list well need understand limitations I going tell simple story , help understand latest Computer memory partition 1 byte memory know bite , dress showing section memory 's extending towards bottom top left side address increases , important top byte address 200 next fight would address 2019 next fight would address 202 , I want I want memory left right horizontally instead trying bottom top like looks better , let 's bite dress 200 go towards write address increases like 2013 Goa like 20 , 203 n't really matter whether show memory bottom top left right logical ways look memory , coming back story memory crucial resource applications keep asking much computer given job , managing memory one components one guys calls memory manager guy keeps track part , free part memory allocated anyone needs memory store something needs talk guy program , building application ets2 store data memory needs talk memory manager taken doctor Memory Manager , high level language like C letter say using C talk memory manager first wants store integer memory communicates , manager declaring integer variable something like memory manager declaration says ok , need store integer variable I need give 4 bytes memory integer variable stored 4 bytes 1 , architecture latest changes architecture stored 4 bytes memory manager Looks 4 bytes free space memory , assign site allocates variable X address block memory address first byte memory solid state first bite , memory address 217 variable X address 217 memory mein , kind communicate sit back Albert signed address 2174 variable X store whatever want , fill data variable Albert needs store list integers list numbers , things maximum number integers list four asks Memory Manager integer array size , name array always stored memory one contiguous block memory memory manager like ok I need look , memory 16 bytes variable memory manager allocates block starting address 201 , ending address 216 variable array four integers , today 's one contiguous block memory Memory Manager conveys starting address block Albert tries access elements , Here let se tries access 's tries write value fourth element array acts 3 , application knows write particular value knows base address starting address block . base address using index three calculates address

Fig. 8 Stop words remove, and content words just left

Stemming:

In the documents, every word needs to be processed to find out the keyword. Because of the words formation, the system could miss out same word. By stemming the words which means removing the suffixes and plural to singular words, the system could determine the same word and their relation among the sentences and occurrence in the documents.

It is better to use stemming that every word convert into their base words and content words easily extracted

Natural Language Toolkit:

NLTK is a leading platform for building Python programs to work with human language data. It provides easy interfaces over 50 corpora and lexical resources such as WordNet and text processing libraries for classification, **stemming**, **remove stop words**, tagging, parsing and semantic reasoning etc. In this research, we used different

modules from NLTK toolkit. NLTK libraries need to be imported in order to use its different modules

3.6 Count frequency

It is a statistical approach through which we calculate the number of words appears in the text. Term frequency is

calculating the occurrence of words in the text file. After processing the documents, we detect the term frequency of every words that plays a big role in keyword extraction technique. In this research, we used NLTK toolkit for calculating term (as shown in Fig. 9).

```
'name': 1, 'ending': 1, 'today': 1, 'conveys': 1, 'Halat': 1, 'fourth': 1, 'acts': 1, 'index': 1, 'calculates': 1, 'tree': 1,
'takes': 1, 'active': 1, 'race': 1, 'inconstant': 1, 'uses': 1, '"ll": 1, 'values': 1, 'physicians': 1, 'lecture': 1, 'declare
d': 1, 'fifth': 1, 'possible': 1, 'expect': 1, 'available': 1, 'cases': 1, 'stand': 1, 'options': 1, 'recreate': 1, 'solve': 1,
'costly': 1, 'login': 1, 'kept': 1, 'early': 1, 'option': 1, 'internal': 1, 'entirely': 1, 'feeling': 1, 'bad': 1, 'small': 1,
'wasted': 1, 'gross': 1, 'desperately': 1, 'seeking': 1, 'try': 1, 'solves': 1, 'cache': 1, 'unit': 1, 'cleaning': 1, 'request
s': 1, 'fast': 1, 'letters': 1, 'phone': 1, 'higher': 1, 'probability': 1, 'respectively': 1, 'disjoint': 1, 'non-contiguous':
1, 'Shobhit': 1, 'somehow': 1, 'calculating': 1, 'blog': 1, 'sexy': 1, 'parts': 1, 'restore': 1, 'example': 1, 'Lok': 1, 'II':
1, 'Birbal': 1, 'invalid': 1, 'used': 1, 'Mark': 1, 'That': 1, 'This': 1, 'End': 1, 'actually': 1, 'variables': 1, 'define': 1,
'fields': 1, 'typical': 1, 'star': 1, 'non': 1, 'nodes': 1, 'connected': 1, 'Data': 1, 'Structure': 1, 'points': 1, 'called':
1, 'complete': 1, 'Null': 1, 'zero': 1, 'travel': 1, 'start': 1, 'car': 1, 'end': 1, 'independently': 1, 'separately': 1, 'crea
ted': 1, 'Value': 1, 'property': 1, 'justice': 1, 'pattern': 1, 'null': 1, 'returned': 1, 'brown': 1, "ve": 1, 'identified':
1, 'could': 1, 'calculate': 1, 'started': 1, 'playing': 1, 'Treasure': 1, 'Hunt': 1, 'emotional': 1, 'lesser': 1, 'worst': 1,
'proportional': 1, 'words': 1, 'Big': 1, 'anywhere': 1, 'path': 1, 'save': 1, 'Nutan': 1, 'big': 1, 'open': 1, 'thing': 1, 'shi
ps': 1, 'Bihar': 1, 'shift': 1, 'Indus': 1, 'delete': 1, 'sense': 1, 'unused': 1, 'benefit': 1, 'guess': 1, 'hand': 1, 'Paris':
1, 'discuss': 1, 'lessons': 1, 'implementing': 1, 'basic': 1, 'introduction': 1, 'lesson': 2, 'limitations': 2, 'tell': 2, 'lat
est': 2, 'byte': 2, 'towards': 2, 'address': 2, 'increases': 2, '"nt": 2, 'really': 2, 'back': 2, 'keep': 2, 'asking': 2, 'muc
h': 2, 'allocated': 2, 'says': 2, 'space': 2, 'allocates': 2, 'solid': 2, 'whatever': 2, 'asks': 2, 'always': 2, 'write': 2, 'b
ase': 2, 'three': 2, 'constant': 2, 'erase': 2, 'Is': 2, 'extend': 2, 'extension': 2, 'make': 2, 'still': 2, 'large': 2, 'getti
ng': 2, 'solution': 2, 'named': 2, 'type': 2, 'another': 2, 'may': 2, 'location': 2, 'similarly': 2, 'gets': 2, 'together': 2,
'Note': 2, 'gives': 2, 'means': 2, 'view': 2, 'Notun': 2, 'us': 2, 'asked': 2, 'links': 2, 'properly': 2, 'filled': 2, 'third':
2, 'traverse': 2, 'cost': 2, 'complexity': 2, 'O': 2, 'insertion': 2, 'adjust': 2, 'operations': 2, 'plus': 2, 'terms': 3, 'us
e': 3, 'well': 3, 'simple': 3, 'story': 3, 'bottom': 3, 'left': 3, 'right': 3, 'instead': 3, 'logical': 3, 'look': 3, 'talk':
3, 'taken': 3, 'wants': 3, 'architecture': 3, 'tries': 3, 'se': 3, 'knows': 3, '.', 3, 'add': 3, 'way': 3, 'adjacent': 3, 'cop
y': 3, 'problem': 3, 'makes': 3, 'n': 3, 'operation': 3, 'understand': 4, 'bite': 4, 'would': 4, 'free': 4, 'say': 4, 'give':
4, 'X': 4, 'fill': 4, 'things': 4, 'previous': 4, 'see': 4, 'blocks': 4, 'stores': 4, 'extra': 4, 'pointer': 4, 'insert': 4, 's
tructure': 5, 'dress': 5, 'top': 5, 'application': 5, 'C': 5, 'stored': 5, 'integers': 5, 'numbers': 5, 'four': 5, 'particula
r': 5, 'separate': 5, 'two': 5, 'second': 5, 'link': 5, 'position': 5, 'last': 5, 'notes': 5, 'head': 5, 'show': 6, 'guy': 6,
'needs': 6, 'ok': 6, 'contiguous': 6, 'ask': 6, 'case': 6, 'information': 6, 'field': 6, 'let': 7, 'go': 7, 'starting': 7, 'cre
ate': 7, 'request': 7, 'get': 7, 'also': 7, 'using': 8, 'part': 8, 'something': 8, 'Albert': 8, 'number': 8, 'value': 8, 'Memor
y': 9, 'Manager': 9, 'size': 9, 'access': 9, 'integer': 10, 'bytes': 10, '"s": 11, 'time': 11, 'data': 12, 'elements': 12, 'wa
nt': 13, 'new': 13, 'note': 13, 'linked': 14, 'manager': 14, 'need': 15, 'variable': 15, 'first': 16, 'array': 16, 'one': 18,
'like': 19, 'element': 19, 'I': 22, 'next': 22, 'store': 26, 'node': 26, 'list': 45, 'block': 49, 'address': 54, 'memory': 60}
```

Fig. 9 Frequency of every words.

In this approach, count the occurrence of the every word exist in the video lectures. In this way words show with their frequencies. The word frequently used in the video helps to generate the title of the video. Extract top most 5 words that have highest frequency. It means these words are most repeated used in the video and also find the pre and post word of the top 5 words. For example linked word occur 45 times and list word also occur 45 times. Linked post word is list and list pre word is linked.so it helps to make the words or title. In this research, this methods helps to identify the keywords that is used in title generation (as shown in Fig. 10).

```
: Node 26
Next (22) node list (45)
Named (2) node like (19)
Address (54) node list (45)
Ask (12) node Albert (8)
Memory (60) node memory (60)
Block (53) node field (7)
Next (22) node store (36)
Notes (5) node store (36)
First (16) node also (7)

: list 45
Linked (14) list data (12)
Dynamic (1) list using (8)
Linked (14) list well (8)
Store (36) list integers (5)
Integers (5) list numbers (5)
Integers (5) list four (6)
Store (36) list I (27)
Element (31) list declared (1)
Element (31) list kept (1)
Numbers (5) list option (2)
Bad (1) list small (1)

: block 49
Address (54) block memory (60)
Contiguous (7) block memory (60)
Allocates (2) block starting (7)
Contiguous (7) block memory (60)
Address (54) block Albert (8)
Address (54) block base (2)
Extend (3) block memory (60)
Adjacent (3) block cases (1)
Sound (5) block case (7)
Next (22) block give (7)
Recreate (1) block new (13)
Previous (4) block new (13)

Address 54
Top (5) address next (22)
Would (4) address next (22)
Next (4) address block (53)
X (4) address block (53)
Memory (60) address first (16)
Memory (60) address variable (16)
Signed (1) address variable (16)
Starting (7) address ending (2)
Ending (2) address variable (16)
Starting (7) address block (53)
Base (2) address starting (7)
Starting (7) address block (53)
Base (2) address using (8)
Calculates (1) address tree (1)
Knows (3) address access (9)
New (13) address copy (3)
Starting (7) address allocated (2)

: Memory 60
Terms (3) memory usage (1)
Terms (3) memory consumption (1)
Computer (1) memory partition (1)
Byte (12) memory know (4)
Section (1) memory's (11)
Want (16) memory left (3)
Show (7) memory bottom (3)
Look (4) memory coming (1)
Story (3) memory crucial (1)
Managing (1) memory one (21)
Calls (1) memory manager (14)
Part (15) memory allocated (2)
Needs (6) memory store (36)
```

Fig. 10 Combination of word extraction

3.7 Extract Combination of Words

N-grams of texts are being used frequently in text mining and natural language processing tasks. It is basically a set of co-occurring words within a given document and when computing N-grams, we move one or more words forward. It depends on the number given in the N-gram calculation. Extract combination of words from N-grams. First, extract the single word, then bi-gram (combination of two words) then tri gram (combination of three words) and so on till the frequency of n gram words less than five

N-grams

```

n : 2, allocated : 2, says : 2, space : 2, allocates : 2, build : 2, whatever : 2, asks : 2, always : 2, write : 2, u
ase': 2, 'three': 2, 'constant': 2, 'erase': 2, 'is': 2, 'extend': 2, 'extension': 2, 'make': 2, 'still': 2, 'large': 2, 'getti
ng': 2, 'solution': 2, 'named': 2, 'type': 2, 'another': 2, 'may': 2, 'location': 2, 'similarly': 2, 'gets': 2, 'together': 2,
'Note': 2, 'gives': 2, 'means': 2, 'view': 2, 'Notun': 2, 'us': 2, 'asked': 2, 'links': 2, 'properly': 2, 'filled': 2, 'third':
2, 'traverse': 2, 'cost': 2, 'complexity': 2, '0': 2, 'insertion': 2, 'adjust': 2, 'operations': 2, 'plus': 2, 'terms': 3, 'us
e': 3, 'well': 3, 'simple': 3, 'story': 3, 'bottom': 3, 'left': 3, 'right': 3, 'instead': 3, 'logical': 3, 'look': 3, 'talk':
3, 'taken': 3, 'wants': 3, 'architecture': 3, 'tries': 3, 'se': 3, 'knows': 3, '': 3, 'add': 3, 'way': 3, 'adjacent': 3, 'cop
y': 3, 'problem': 3, 'makes': 3, 'n': 3, 'operation': 3, 'understand': 4, 'bite': 4, 'would': 4, 'free': 4, 'say': 4, 'give':
4, 'X': 4, 'fill': 4, 'things': 4, 'previous': 4, 'see': 4, 'blocks': 4, 'stores': 4, 'extra': 4, 'pointer': 4, 'insert': 4, 's
tructure': 5, 'dress': 5, 'top': 5, 'application': 5, 'C': 5, 'stored': 5, 'integers': 5, 'numbers': 5, 'four': 5, 'particula
r': 5, 'separate': 5, 'two': 5, 'second': 5, 'link': 5, 'position': 5, 'last': 5, 'notes': 5, 'head': 5, 'show': 6, 'guy': 6,
'needs': 6, 'ok': 6, 'contiguous': 6, 'ask': 6, 'case': 6, 'information': 6, 'field': 6, 'let': 7, 'go': 7, 'starting': 7, 'cre
ate': 7, 'request': 7, 'get': 7, 'also': 7, 'using': 8, 'part': 8, 'something': 8, 'Albert': 8, 'number': 8, 'value': 8, 'Memor
y': 9, 'Manager': 9, 'size': 9, 'access': 9, 'integer': 10, 'bytes': 10, 's': 11, 'time': 11, 'data': 12, 'elements': 12, 'wa
nt': 13, 'new': 13, 'note': 13, 'linked': 14, 'manager': 14, 'need': 15, 'variable': 15, 'first': 16, 'array': 16, 'one': 18,
'like': 19, 'element': 19, 'I': 22, 'next': 22, 'store': 26, node': 26, list': 45, block': 49, address': 54, memory': 60
memory
address
block
list
node
    
```

Fig. 11 N 1 based without word combination top 5 words

N=2

Combination of two words with top 5 highest frequency

```

memori', 'manag
link', 'list
block', 'memori
address', 'next
new', 'block
    
```

```

['one contiguous block
memory', 'time taken
access elements',
'pointer variable store
address', 'address next
block stores', 'block
address part would']
    
```

N=3

Combination of three words with top 5 highest frequency

```

contiguous', 'block', 'memory
address', 'next', 'block
starting', 'address', 'block
memory', 'manager', 'like
one', 'contiguous', 'block
    
```

N=4

Combination of four words with top 5 highest frequency

N-grams of texts are being used frequently in text mining and natural language processing tasks. It is basically a set of co-occurring words within a given document and when computing N-grams, we move one or more words forward. It depends on the number given in the N-gram calculation. In this research, after finding the top 5 words that have maximum frequency and check the combination of two words or three words and so on.

N=1

Without combination top 5 single word is displayed that have highest frequency (as shown in Fig. 11).

3.8 Matching keywords

In this research, after finding the top 5 words that have maximum frequency and check the combination of two words or three words and so on. After matching top 5 words, highest frequency of keyword is extracted After matching the keywords the keyword extract from 4th gram and top five words of count frequency

Linked list data structure

3.9 Keyword match with computer terms

The keyword extracted after matching the n-grams and count frequency keyword is extracted and this keyword

extracted compare with the computer terms dataset either it is a computer term or not?

Linked list data structure

- Checked linked words is exist? Yes
 - Checked list words is exist? Yes
 - Checked data words is exist? Yes
 - Checked structure words is exist? Yes
 - Checked linked list words is exist? Yes
 - Checked data structure is exist? Yes
 - Checked linked list data words is exist? Yes
 - Checked linked list data structure words is exist? Yes
- So proved these words are exist in the computer terms data set.

3.10 Title /Topic is generated

If the keyword is matched with the computer term then it is generated as the topic/title of the video.

Keyword is matched and the title is **Linked list data structure**

Verification:

Video lecture with subtitles to check the proposed model is working efficiently or not. How much it gives correct results. Therefore subtitles is matched with the text that is stored in pickle through jaccard coefficient (as shown in Fig. 12).

Similarity is achieved in this procedure.

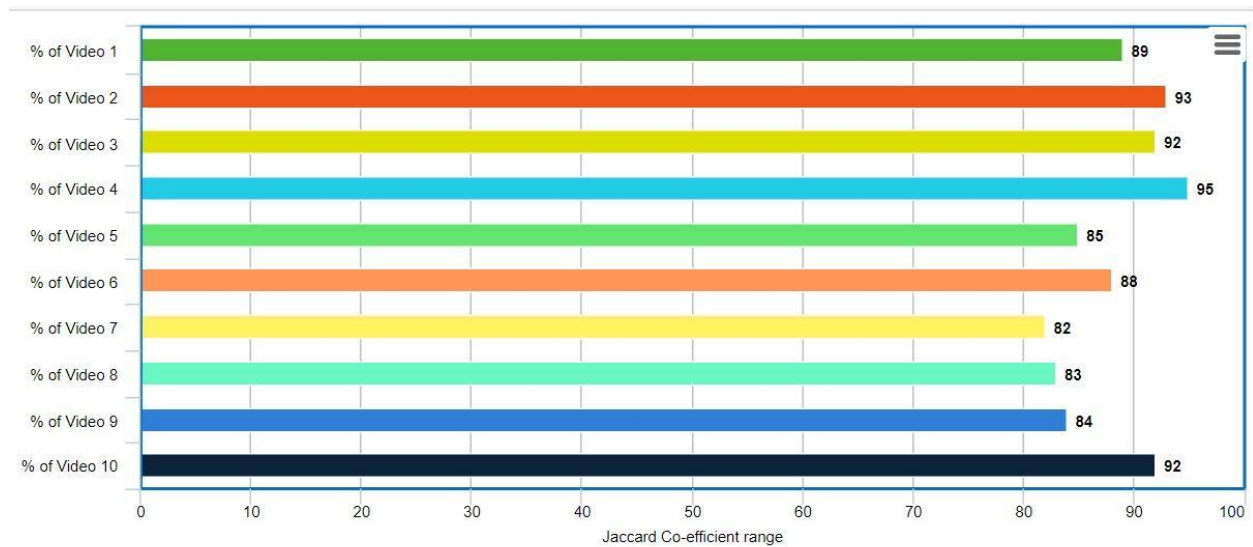


Fig. 12 Jaccard similarity coefficient

The Jaccard Index is also called as the Jaccard similarity coefficient. It is a statistics which is used to understand the similarities between the sample sets. The measurement focuses similarity between finite sample sets. It is formally defined as the size of the intersection divided by the size of the union of the sample sets.

$$(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$= \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

4. Results and Discussion

By doing immense research, we have found out different approach in natural language field related to keywords extraction technique. From various methods, semantic relation caught our eyes and we developed a complete different approach to extract keywords from the documents.

Through video lecture, extract audio from video then audio is converted in to text in 10 seconds chunk. Every 10 second chunk form paragraph separated by comma and the paragraph is saved as pickle file for further procedure. Check similarity between subtitles of the video and the text extract from the video by using jaccard similarity coefficient. In the result 93% data similar.

Need to refine the data, and text converted in to form that is useful for this approach. For refining the data, remove the stop words from text and apply stemming on them. After refining the data, only content words are left. Tokenization process perform on the content words. Each content word is converted in to tokens. Count the frequency of all the content words in the list. After count all the words along with their frequencies extract top 5 words which have the highest frequency among all. Extract the combinations of words, extract one word, then bigram (combination of two words) trigram (combination of three words and so on

Now extract the pre and post words of top 5 content words which have high frequency and also find the frequency of pre and post words. By using n-gram, check the combination of two words, then check combination of three words and so on with highest frequency and match with each other and extract the keywords. Then the given keyword matched with the computer term dataset. If the keyword exist in the computer science corpus the keyword used as title. In this way title and topic of the video lecture is generated.

Thirty videos with subtitles with subtitles were taken for the purpose of generation of the topic. The video were processed by taking combination of words by using N-grams, statistical approach. There was no need to train the data as each video content is different from one another.

The developed approach is able to generate the topic of the video. The actual subtitles and the text generated by the researcher were 93% similar

N-gram technique was used to extract keywords. Single word, combination of two words, combination of three words, combination of four words. After these combinations, check if any of the combinations match with top 5 highest frequency words with their pre and post words. The matched keyword is extracted. Extracted keyword is matched either the keyword exist in computer term data set. After confirmation that keyword is in data set then the keyword was declared as the topic or title of the video.

5. Conclusion and Future work

The paper has attempted to extract keywords from the text of the video lectures. In the existing approach, the extraction of keywords according to the content of the video is the main purpose. Find frequently used keywords used in the video and extract top 5 frequently used keywords. Top 5 keywords with high frequency matched with computer term and generate the title of the video. This approach can facilitate the user to get the title of the video and saved their time by not watching the whole video.

Finally, this method provides the best result by extracting keywords, and generate the title of the video by using n-gram. N-gram based extraction is also used for bigram and trigram expansion of keywords to extract the keywords from text of video and matching these keywords with computer terms dataset and detect/generate the topic of the video

This research was conducted on videos having subtitles. There was high accuracy in generating titles from videos

with subtitles. In future using this approach titles may be generated from videos without subtitles.

References

- [1] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory*, vol. 1, pp. 1-20, 2010.
- [2] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: a literature review," *International Journal of Computer Applications*, vol. 109, 2015.
- [3] S. Luthra, D. Arora, K. Mittal, and A. Chhabra, "A Statistical Approach of Keyword Extraction for Efficient Retrieval," *International Journal of Computer Applications*, vol. 168, 2017.
- [4] S. Stamenov, "Porozumění dokumentům pomocí metod text miningu," 2017.
- [5] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An overview of graph-based keyword extraction methods and approaches," *Journal of information and organizational sciences*, vol. 39, pp. 1-20, 2015.
- [6] C. Zhang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems*, vol. 4, pp. 1169-1180, 2008.
- [7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques (ed.).(J. Gray, Ed.) San Francisco," ed: CA, USA: Morgan Kaufmann Publishers, Elsevier, 2006.
- [8] A. Jain, A. Raghuvanshi, and G. Shrivastava, "Analysis of query based text classification approach," *International Journal*, vol. 2, 2012.
- [9] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science*, vol. 43, pp. 25-38, 2017.
- [10] T. Joachims, *Learning to classify text using support vector machines* vol. 668: Springer Science & Business Media, 2002.
- [11] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232-247, 2016.
- [12] Z. Liu, J. Liu, W. Yao, and C. Wang, "Keyword extraction using PageRank on synonym networks," in *2010 International Conference on E-Product E-Service and E-Entertainment*, 2010, pp. 1-4.
- [13] J. Kaur and V. Gupta, "Effective approaches for extraction of keywords," *International Journal of Computer Science Issues (IJCSI)*, vol. 7, p. 144, 2010.
- [14] X. Hu and B. Wu, "Automatic keyword extraction using linguistic features," in *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, 2006, pp. 19-23.
- [15] R. M. Alguliev and R. M. Aliguliyev, "Effective summarization method of text documents," in *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 2005, pp. 264-271.
- [16] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical

- information," *International Journal on Artificial Intelligence Tools*, vol. 13, pp. 157-169, 2004.
- [17] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 216-223.
- [18] T. Onoda, T. Yumoto, and K. Sumiya, "Extracting and Clustering Related Keywords based on History of Query Frequency," in *2008 Second International Symposium on Universal Communication*, 2008, pp. 162-166.
- [19] H. M. Lynn, E. Lee, C. Choi, and P. Kim, "Swifrank: an unsupervised statistical approach of keyword and salient sentence extraction for individual documents," *Procedia computer science*, vol. 113, pp. 472-477, 2017.
- [20] P. D. Turney, "Coherent keyphrase extraction via web mining," *arXiv preprint cs/0308033*, 2003.
- [21] G. K. Palshikar, "Keyword extraction from a single document using centrality measures," in *International conference on pattern recognition and machine intelligence*, 2007, pp. 503-510.
- [22] H. Cohen-Kerner, "Automatic extraction of keyword from abstracts," *Automatic extraction of keyword from abstracts, lecture notes in computer science*, vol. 2773, pp. 843-849, 2003.
- [23] T. D. Nguyen and M.-Y. Kan, "Keyphrase extraction in scientific publications," in *International conference on Asian digital libraries*, 2007, pp. 317-326.
- [24] M. Krapivin, A. Autayeu, M. Marchese, E. Blanzieri, and N. Segata, "Keyphrases extraction from scientific documents: improving machine learning approaches with natural language processing," in *International Conference on Asian Digital Libraries*, 2010, pp. 102-111.
- [25] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.
- [26] I. H. Witten and O. Medelyan, "Thesaurus based automatic keyphrase indexing," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'06)*, 2006, pp. 296-297.
- [27] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404-411.
- [28] C. Seifert, E. Ulbrich, and M. Granitzer, "Word clouds for efficient document labeling," in *International Conference on Discovery Science*, 2011, pp. 292-306.
- [29] M. Litvak and M. Last, "Graph-based keyword extraction for single-document summarization," in *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*, 2008, pp. 17-24.
- [30] M. Grineva, M. Grinev, and D. Lizorkin, "Extracting key terms from noisy and multitheme documents," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 661-670.
- [31] N. G. Ali and N. Omar, "Arabic keyphrases extraction using a hybrid of statistical and machine learning methods," in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 2014, pp. 281-286.
- [32] E. Najafi and A. H. Darooneh, "The fractal patterns of words in a text: a method for automatic keyword extraction," *PloS one*, vol. 10, p. e0130617, 2015.
- [33] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 620-628.
- [34] F. Liu, F. Liu, and Y. Liu, "A supervised framework for keyword extraction from meeting transcripts," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 538-548, 2010.
- [35] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, and Q.-T. Ha, "A hidden topic-based framework toward building applications with short web documents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 961-976, 2010.
- [36] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613-620, 1975.
- [37] Z. Yang, J. Lei, K. Fan, and Y. Lai, "Keyword extraction by entropy difference between the intrinsic and extrinsic mode," *Physica A: Statistical Mechanics and its Applications*, vol. 392, pp. 4523-4531, 2013.



Ramsha Pervaiz is M.S. in Computer Science from Air University, Multan Campus, Multan, Pakistan in 1997 and 1999, respectively. During 2017-2020, she is working on her MS thesis under the supervision of Dr Kaleem Razzaq Malik.



Khalid Aloufi is an associate professor in the Department of Computer Engineering, Taibah University, Madinah, Saudi Arabia. He received his Ph.D. and M.Sc. degrees in informatics from Bradford University, UK, in 2002 and in 2006 respectively. His B.Sc. degree in computer engineering was received in 1999 from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. From 2002 to 2006, he was part of the networks and performance engineering research group at Bradford University. Aloufi has been the dean of the College of Computer Science and Engineering at Taibah University, Saudi Arabia.



Syed Shabbar Raza Zaidi has completed Master's Degree in Network Communication from University of Hertfordshire London. He is currently working as Lecturer in Air University, Multan Campus, Multan, Pakistan from September 2017 - date performing duties like Teaching and Research. and he has 6 years' extensive experience in the field of IT, particularly in Computer Networks. He is CCNA certified and have excellent skills of Python programming language. He has training in MCSE 2012 from TNA and studied Course of Linux Administration from University of Hertfordshire.



Kaleem Razzaq Malik was born in Multan, Punjab, Pakistan in 1984. He received the M.S. degrees in computer science from the National University of Computer and Emerging Sciences, Lahore, PAK in 2008 and the Ph.D. degree in computer science from University of Engineering and Technology, Lahore, PAK, in 2018.

He is now working as Associate Professor in Air University, Multan Campus, Multan, Pakistan from March 2018 - date performing duties like Teaching and Research. He has worked as Assistant Professor in COMSATS Institute of Information Technology, Sahiwal, Pakistan from December 2015 – March 2018 and as Lecturer in Department of Software Engineering, Government College University Faisalabad, Pakistan from June 2013 – November 2015 (2 year 5 months) performing duties like Teaching. He has also worked as instructor of computer science in Virtual University of Pakistan. University level more than 10 years of teaching experience.