IT204 Project Report: Lecture Summariser
Group 34

Shlok Bhosale - 201IT258
Information Technology
National Institute of
Technology Karnataka
Surathkal, India 575025
shlok.201it258@nitk.edu.in

Annant Maheshwari - 201IT109
Information Technology
National Institute of
Technology Karnataka
Surathkal, India 575025
annant.201it109@nitk.edu.in

Kowshic V - 201IT231
Information Technology
National Institute of
Technology Karnataka
Surathkal, India 575025
kowshicv.201it231@nitk.edu.in

***Abstract*—In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most ef iciently. Lecture Summarizer is a tool that helps anyone to summarize an audio file and make the most out of their time. The lecture summarizer identifies the most important meaningful information in an audio file and compresses it into a shorter version preserving its overall meanings. The absence of sentence boundaries in the recognized text complicates the summarization process. This report compares the various approaches to make this possible and explains the algorithm of the best method out of the listed methods. This paper provides an abstract view of the present scenario of research work for audio summarization**

I. Introduction

Using the text summariser we can convert audio to text and then create a summary out of it. This is an excellent time saving tool and solves time management problems. It is quite challenging to summarise a long audio or piece of text, this tool does it excellently. There aren't any popular mainstream solutions to this problem. One might a few github repo's here and there but nothing as helpful as our implementation.

The core idea of our project is to first convert audio to text and then use extractive text summarisation.The existing solutions vary in the methodology. Our project can only be implemented for summarising English audios or lectures.

II. Literature Survey

Text summarization is a technique to get the most necessary information from a given input text. There are two possible approaches to this, extractive and abstractive. [1] talks about extractive summarization. It summarizes texts based on word frequency , where high weighted frequency sentences are extracted and printed .

. [4] talks about the idea of using latent semantic analysis in text summarization . inspired by the latent semantic indexing, applied the singular value decomposition (SVD) to generic text summarization.

Along with this , we proceeded with comparing our model algorithm and its results with those obtained by using other algorithms as well, such as text rank, cosine similarity, TF-IDF, etc. [7], [8].

Authors	Methodology	Merits	Limitations
J. N. Madhuri and R. Ganesh Kumar	Sentence ranking	Works on extractive summarization	Some important words might be overlooked due to low frequency
Moratanch, N. Gopalan, Chitra Kala	Uses LSA with SVD	Works on extractive summarization	No word knowledge. Some sentences might not make sense.
S. Bhattacharjee, A. Das, U. Bhattacharya, S. K. Parui and S. Roy	Sentiment analysis using cosine similarity measure	Compares every two sentences	Result might not be as accurate
M. R. Ramadhan, S. N. Endah and A. B. J. Mantau	Implementation of Text rank Algorithm	Ranks sentences according to their weightage	Important texts with lower frequencies might get missed out

Fig. 1

III. Problem Statement

Problem name lecture summariser where we need to take an audio file and return the summarized text of the content.This requires converting the audio to text and then summarizing the text.

A. Objectives

- Using python libraries for converting speech to text
- Analyzing the various methodologies of text summarization
- Giving the most optimal output (mostly using Latent Semantic Analysis-LSA)

IV. Methodolgy

There are various implementations for the text summarisation. We looked through and compared the following approaches/algorithms and improvised them to implement our project (ref:Fig.2):

A. TF-IDF

- Term Frequency (TF) - with the simplest being a raw count of instances a word appears in a document
- Inverse Document Frequency (IDF)- This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is

Fig. 2

- Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

B. TextRank (Graph Theory)

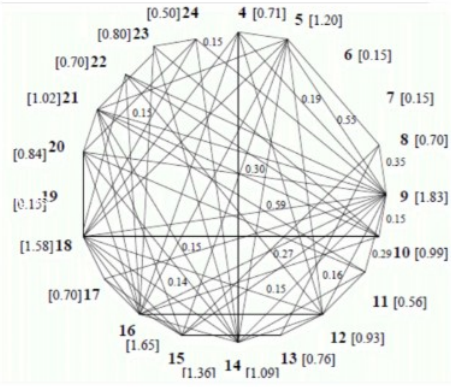

Fig. 3

- Unsupervised graph based
- Numbers in bold are number of sentences
- Sentences are vertices
- Numbers in brackets are weights
- Similarity as edges

C. Cosine Similarity

Cosine Similarity measures the similarity between two sentences or documents in terms of the value within the range of [-1, 1] whichever you want to measure. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space.

In this context, the two vectors are arrays containing the word counts of two documents. When plotted on a multi-dimensional space, each dimension corresponds to a word in the document, the cosine similarity captures the orientation (the angle) of the documents.Smaller the angle, higher the similarity

D. Latent Semantic Analysis (LSA)

Now last but not the least, the main algorithm that we are using for this project is Latent Semantic Analysis The algorithm for LSA consists of three major steps:

- Input matrix creation- The input document is represented as a matrix to understand and perform calculations on it.
- Singular Value decomposition(SVD)- SVD is an algebraic method that can model relationships among words/phrases and sentences.
- Sentence Selection- Here we have used Topic method to extract concepts and sub-concepts from the SVD calculations and are called topics of the input document.

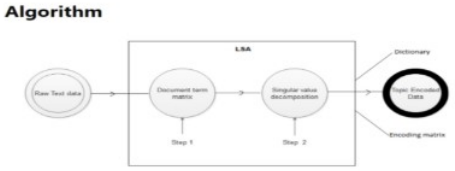

Fig. 4

We implemented our model using LSA and SVD using the following python libraries: Sumy and nltk.

V. Results and Analysis

Codebase (Google Collab): https://colab.research.google.com/drive/13V54uUsDZkWHxVJouM3wGPmmFo3vSj4R?usp=sharing#scrollTo=WDifICllGAlQ

We implemented text summarisation using all the above discussed algorithms. The outputs and the comparison are as follows:

ORIGINIAL TEXT

Junk foods taste good that's why it is mostly liked by everyone of any age group especially kids and school going children. They generally ask for the junk food daily because they have been trend so by their parents from the childhood. They never have been discussed by their parents about the harmful effects of junk foods over health. According to the research by scientists, it has been found that junk foods have negative effects on the health in many ways. They are generally fried food found in the market in the packets. They become high in calories, high in cholesterol, low in healthy nutrients, high in sodium mineral, high in sugar, starch, unhealthy fat, lack of protein and lack of dietary fibers. Processed and junk foods are the means of rapid and unhealthy weight gain and negatively impact the whole body throughout the life. It makes able a person to gain excessive weight which is called as obesity. Junk foods tastes good and looks good however do not fulfil the healthy calorie requirement of the body. Some of the foods like french fries, fried foods, pizza, burgers, candy, soft drinks, baked goods, ice cream, cookies, etc are the example of high-sugar and high-fat containing foods. It is found according to the Centres for Disease Control and Prevention that Kids and children eating junk food are more prone to the type-2 diabetes. In type-2 diabetes our body become unable to regulate blood sugar level. Risk of getting this disease is increasing as one become more obese or overweight. It increases the risk of kidney failure. Eating junk food daily lead us to the nutritional deficiencies in the body because it is lack of essential nutrients, vitamins, iron, minerals and dietary fibers. It increases risk of cardiovascular diseases because it is rich in saturated fat, sodium and bad cholesterol. High sodium and bad cholesterol diet increases blood pressure and overloads the heart functioning. One who like junk food develop more risk to put on extra weight and become fatter and unhealthier. Junk foods contain high level carbohydrate which spike blood sugar level and make person more lethargic, sleepy and less active and alert. Reflexes and senses of the people eating this food become dull day by day thus they live more sedentary life. Junk foods are the source of constipation and other disease like diabetes, heart ailments, clogged arteries, heart attack, strokes, etc because of being poor in nutrition. Junk food is the easiest way to gain unhealthy weight. The amount of fats and sugar in the food makes you gain weight rapidly. However, this is not a healthy weight. It is more of fats and cholesterol which will have a harmful impact on your health. Junk food is also one of the main reasons for the increase in obesity nowadays. This food only looks and tastes good, other than that, it has no positive points. The amount of calorie your body requires to stay fit is not fulfilled by this food. For instance, foods like French fries, burgers, candy, and cookies, all have high amounts of sugar and fats. Therefore, this can result in long-term illnesses like diabetes and high blood pressure. This may also result in kidney failure. Above all, you can get various nutritional deficiencies when you don't consume the essential nutrients, vitamins, minerals and more. You become prone to cardiovascular diseases due to the consumption of bad cholesterol and fat plus sodium. In other words, all this interferes with the functioning of your heart. Furthermore, junk food contains a higher level of carbohydrates. It will instantly spike your blood sugar levels. This will result in lethargy, inactiveness, and sleepiness. A person reflex becomes dull overtime and they lead an inactive life. To make things worse, junk food also clogs your arteries and increases the risk of a heart attack. Therefore, it must be avoided at the first instance to save your life from becoming ruined. The main problem with junk food is that people don't realize its ill effects now. When the time comes, it is too late. Most importantly, the issue is that it does not impact you instantly. It works on your overtime; you will face the consequences sooner or later. Thus, it is better to stop now. You can avoid junk food by encouraging your children from an early age to eat green vegetables. Their taste buds must be developed as such that they find healthy food tasty. Moreover, try to mix things up. Do not serve the same green vegetable daily in the same style. Incorporate different types of healthy food in their diet following different recipes. This will help them to try foods at home rather than being attracted to junk food. In short, do not deprive them completely of it as that will not help. Children will find one way or the other to have it. Make sure you give them junk food in limited quantities and at healthy periods of time.

Fig. 5

SVD

Junk foods taste good that's why it is mostly liked by everyone of any age group especially kids and school going children. To make things worse, junk food also clogs your arteries and increases the risk of a heart attack. Therefore, it must be avoided at the first instance to save your life from becoming ruined. The main problem with junk food is that people don't realize its ill effects now.

TF-IDF

They are generally fried food found in the market in the packets. It increases the risk of kidney failure. It is more of fats and cholesterol which will have a harmful impact on your health. This may also result in kidney failure. It will instantly spike your blood sugar levels. When the time comes, it is too late. Most importantly, the issue is that it does not impact you instantly. It works on your overtime; you will face the consequences sooner or later. Moreover, try to mix things up. Do not serve the same green vegetable daily in the same style. Incorporate different types of healthy food in their diet following different recipes. Children will find one way or the other to have it.

COSINE

Junk food is the easiest way to gain unhealthy weight. One who like junk food develop more risk to put on extra weight and become fatter and unhealthier. To make things worse, junk food also clogs your arteries and increases the risk of a heart attack. Junk food is also one of the main reasons for the increase in obesity nowadays. This food only looks and tastes good, other than that, it has no positive points. Eating junk food daily lead us to the nutritional deficiencies in the body because it is lack of essential nutrients, vitamins, iron, minerals and dietary fibers

TEXTRANK

['Eating junk food daily lead us to the nutritional deficiencies in the body because it is lack of essential nutrients, vitamins, iron, minerals and dietary fibers.', 'Junk food is also one of the main reasons for the increase in obesity nowadays. This food only looks and tastes good, other than that, it has no positive points.', 'To make things worse, junk food also clogs your arteries and increases the risk of a heart attack.', 'The amount of fats and sugar in the food makes you gain weight rapidly.', 'It is found according to the Centres for Disease Control and Prevention that Kids and children eating junk food are more prone to the type-2 diabetes.']

Fig. 6

VI. Conclusion

- LSA is efficient and easy to implement .
- LSA reduces noise.
- LSA gives a decent result that is much better as compared to other methods.
- LSA is faster compared to other algorithms as it involves document term matrix decomposition.

Although many different approaches can be used to solve the text summarisation problem we figured that LSA is one of the most efficient ones and gives the closest and most accurate summary.

Individual Contribution

Shlok Shivale	Anmol Maheshwari	Kowshic V
• Read various research papers and references • Implemented the TF IDF method for comparison • Compiled all the codes • Created the presentation	• Went through different research papers. • Built the problem statement. • Used the text rank algorithm for summarization. • Helped in creating slides	• Analyzed the various approaches to the problem • Came up with the implemented algorithm • Worked out the implementation with cosine similarity • Helped in creating slides

Fig. 7

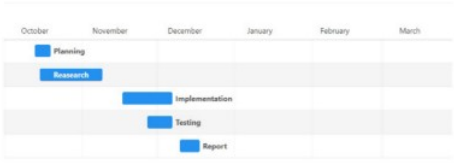

Fig. 8