A PROJECT REPORT

ON

“EXTRACTIVE TEXT SUMMARISATION”

BY

DEEPTI CHAMOLI	MT2017039
DEVYANI BAJAJ	MT2017040
VEERAPURAJU DHANYA AKHILA	MT2017041

UNDER THE GUIDANCE OF
PROF. SRINIVASA RAGHAVAN

ABSTRACT

Text summarization is the process of automatically creating a shorter version of one or more text documents. It is an important way of finding relevant information in large text libraries or in the Internet. Essentially, text summarization techniques are classified as Extractive and Abstractive. Extractive techniques perform text summarization by selecting sentences of documents according to some criteria. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. Word and sentence scoring is the technique most used for Extractive text summarization. This project concentrates on Extractive summarization.

Keywords: Extractive, Abstractive, Scoring

Contents

Chapter 1

Introduction

1.0.1 Problem Statement

To Develop a prototype that automatically summarises (extracts major points) the given text.

1.0.2 Motivation

To take the appropriate action, we need latest information. But on the contrary, the amount of the information is more and more growing. There are many categories of information (economy, sports, health, technology...) and also there are many sources (news site, blog, SNS...).

1.0.3 Extractive Text Summarization

As part of the project we have implemented summarisation in an extractive way.

- Select relevant phrases of the input document and concatenate them to form a summary (like copy-and-paste”).
- PROS : They are quite robust since they use existing natural-language phrases that are taken straight from the input.
- CONS : But they lack in flexibility since they cannot use novel words or connectors. They also cannot paraphrase like people sometimes do.

Chapter 2

Technologies and Data

2.1 Technologies and Tools Used

- Python 3.6
- Scikit-learn
- Pycharm IDE
- GIT

2.2 Data and Source Code

2.2.1 Data Set

`https://github.com/saksham-singhal/IRE-Text-Summarizer/tree/master/DUC-2004/Cluster_of_Docs`

2.2.2 Source Code

`https://github.com/Jolig/TextSummarization`

Chapter 3

Underlying Concepts

3.1 Score Based Algorithms

The Project implements Score Based Algorithms to implement summariser. i.e

- Word Scoring Algorithms
- Graph Scoring Algorithms

3.2 Word Scoring Algorithms

The initial methods in sentence scoring were based on words. Each word receives a score and the weight of each sentence is the sum of all scores of its constituent words. The approaches are explained briefly.

3.2.1 Word Frequency

As the name of the method suggests, the more frequently a words occurs in the text, the higher its score. In other words, sentences containing the most frequent words in a document stand a higher chance of being selected for the final summary. The assumption is that the higher the frequency of a word in the text, the more likely that it indicates the subject of the text.

3.2.2 TF/IDF

The hypothesis assumed by this approach is that if there are more specific words in a given sentence, then the sentence is relatively more important. The target words are usually nouns except for temporal or adverbial nouns.This algorithm performs a comparison between the term frequency (tf) in a document (in this case each sentence is treated as a document) and the document frequency (df), which means the number of times that the word occurs along all documents.

The TF/IDF score is calculated as follows:

TF/IDF(w) = DN(log(1+tf)/log(1+df))

where DN is the number of documents

3.2.3 Upper case

This method assigns higher scores to words that contain one or more upper case letters. It can be a proper name, initials, highlighted words, among others.

3.2.4 Proper noun

Usually the sentences that contain a higher number of proper nouns are more important; thus, they are likely to be included in the document summary. This is a specialization of the Upper case method.

3.2.5 Word co-occurrence

Word co-occurrence measures the chance of two terms from a text appear alongside each other in a certain order. One way to implement this measure is using n-gram), which is a contiguous sequence of n items from a given sequence of text or speech. In short, it gives higher scores to sentences that co-occurrence words appear more often.

3.2.6 Lexical similarity

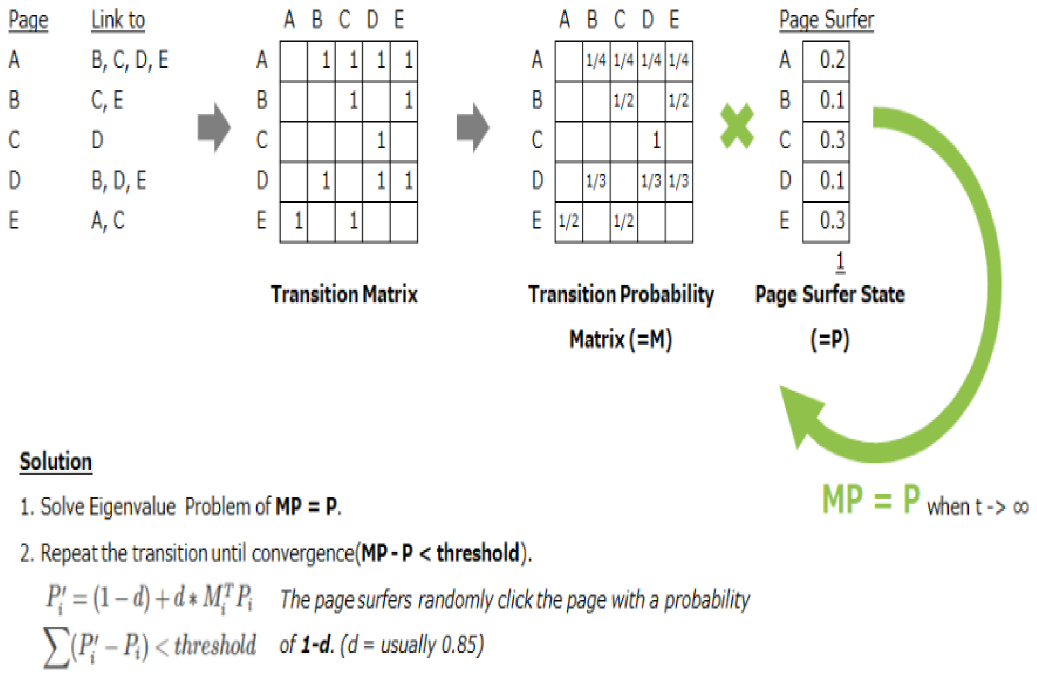
It is based on the assumption that important sentences are identified by strong chains . In other words, it relates sentences that employ words with the same meaning (synonyms) or other semantic relation.

3.3 Graph Scoring

3.3.1 TextRank

TextRank is a graph-based ranking model for text processing . It extracts important keywords from a text document and also to determine the weight of the importance of words within the entire document by using a graph-based model. Sentences with a larger quantity of keywords get higher scores.

It is based on PageRank algorithm that is used on Google Search Engine. Its base concept is ”The linked page is good, much more if it from many linked page”. The links between the pages are expressed by matrix (like Round-robin table). We can convert this matrix to transition probability matrix by dividing the sum of links in each page. And the page surfer moves the page according to this matrix.



Page	Link to
A	B, C, D, E
B	C, E
C	D
D	B, D, E
E	A, C

	A	B	C	D	E
A		1	1	1	1
B			1		1
C				1	
D		1		1	1
E	1		1		

Transition Matrix

	A	B	C	D	E
A		1/4	1/4	1/4	1/4
B			1/2		1/2
C				1	
D		1/3		1/3	1/3
E	1/2		1/2		

Transition Probability Matrix (=M)

Page Surfer	
A	0.2
B	0.1
C	0.3
D	0.1
E	0.3
	1

Page Surfer State (=P)

MP = P when t -> ∞

Solution

1. Solve Eigenvalue Problem of **MP = P**.
2. Repeat the transition until convergence(**MP - P < threshold**).

$P_i' = (1-d) + d * M_i^T P_i$ *The page surfers randomly click the page with a probability*

$\sum(P_i' - P_i) < threshold$ *of* ***1-d****. (d = usually 0.85)*

Figure 3.1: PageRank Example

Chapter 4

Implementation

4.1 Word Scoring

- Firstly we extract all the sentences from the given text
- For each sentence use each of the six above mentioned word-scoring algorithms and extract the scores
- Perform Linear Regression

4.1.1 Semi-supervised Learning

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.
*** Since there is no direct labelled data for our project we have chosen Semi-supervised learning, so that we can make use of a small labelled data that is created manually.

4.1.2 Linear Regression

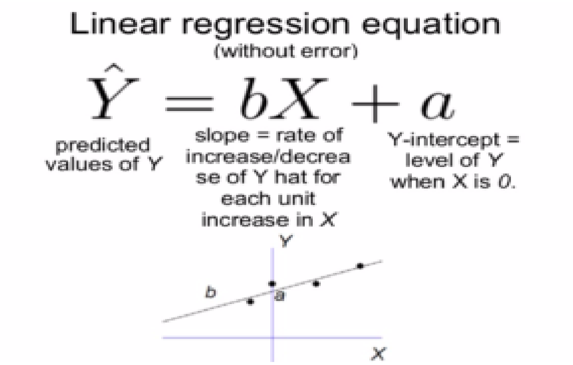

Linear regression equation
(without error)
$\hat{Y} = bX + a$
predicted values of Y
slope = rate of increase/decrease of Y hat for each unit increase in X
Y-intercept = level of Y when X is 0.
Y
b
a
X

Figure 4.1: Linear Regression Plot

Features

Score from each of the word-scoring algorithm acts as a feature.
i.e we use linear regression to attach weights to each of these scores, so that output will not be biased of any one.

Labels

As mentioned, scores would act as features for each and every instance of data set and the labels would be 0, 1.
i.e
0 - if the sentence is NOT present in summary
1 - if the sentence is present in summary

Training

We train the linear regression model with the features and labels(labels that are added manually) we had and get the weights.

Prediction

After estimating the weights, when a new sentence comes the model will predict the label close to 0 and 1.

Extracting Summaries

After predicting labels for each and every sentence we pick some of the sentences(with a threshold) based on the labels.
i.e the higher the label the more probable that it should be present in the summary as it is more close to 1.

4.1.3 Workflow

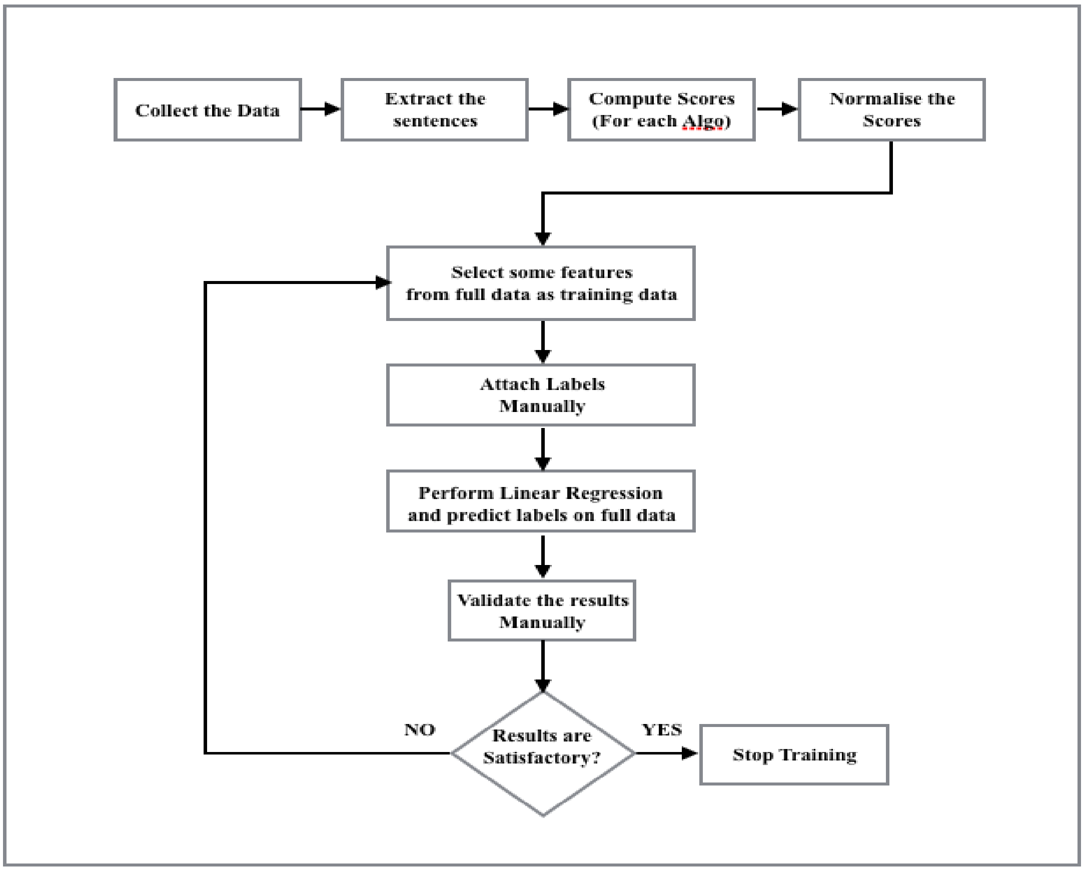

Collect the Data
Extract the sentences
Compute Scores (For each Algo)
Normalise the Scores
Select some features from full data as training data
Attach Labels Manually
Perform Linear Regression and predict labels on full data
Validate the results Manually
Results are Satisfactory?
NO
YES
Stop Training

Figure 4.2: Training Workflow

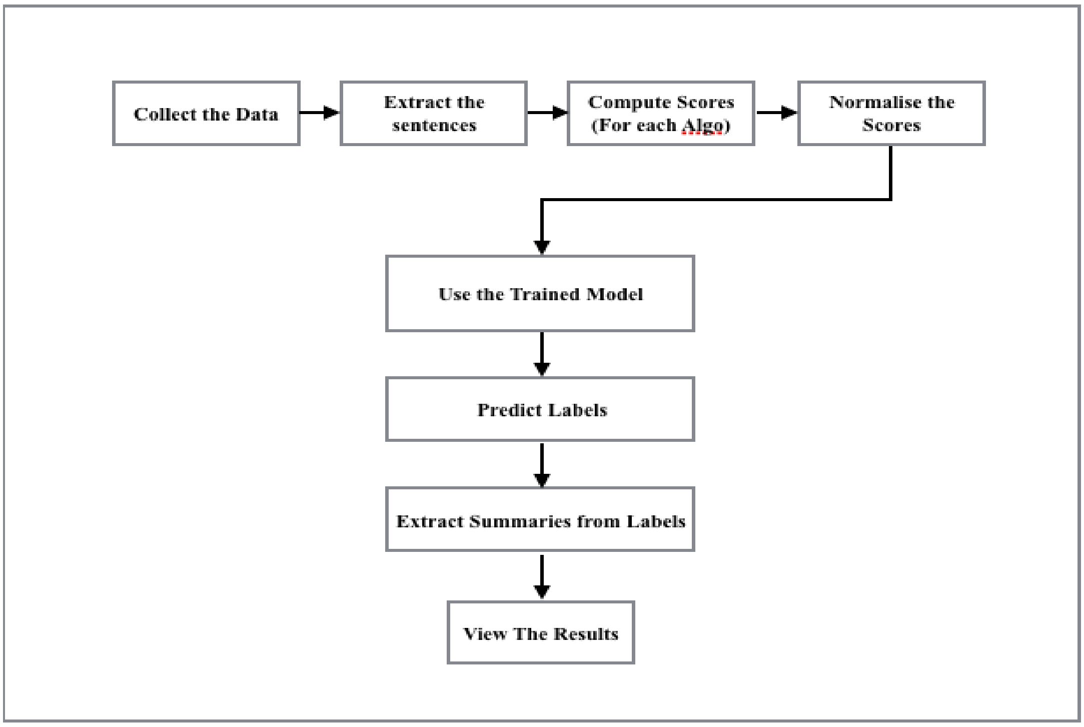


Figure 4.3: Prediction Workflow

4.2 Graph Scoring

4.2.1 About Page Rank

- It is used to compute the rank of web pages, but funny enough, its not named after its use (ranking pages) but rather after its creator: Larry Page, one of Googles founders.
- Important pages are linked by important pages (a recurrent definition).
- The PageRank value of a page is essentially the probability of a user visiting that page.

4.2.2 PageRank Algorithm

1. Let N be the number of pages with links between them.
2. P = PageRank vector (P[i] = the pagerank of page i).
3. A[i][j] is the probability of the user transitioning from page i to page j.
4. A[i][j] is initialized with 1.0 / outbound links count(from node=i) if theres a link between i and j and 0 otherwise.
5. If a page i has no outbound links, then A[i][j] = 1.0/N. Assign equal probability to transitioning to any page. Pages with no outbound links are called dangling pages.
6. Build an index of the webpages and assigns them a numeric index which serves as adjacency matrix.
7. Feed transition probability matrix to PageRank to get the scores.A[i][j] is the probability of transitioning from page i to page j.

4.2.3 From PageRank to TextRank

1. We consider sentences as equivalent to web pages.
2. The probability of going from sentence A to sentence B is equal to the similarity of the 2 sentences.
3. Apply the PageRank algorithm over this sentence graph.

A few observations:

1. In the case of TextRank, the graph is symmetrical.
2. It is up to us to come up with a similarity measure and the algorithms performance depends strongly upon this measure.
3. The algorithm is language agnostic. (It can become language dependent due to a language dependent implementation of the similarity measure).
4. It does not require any training, it is a totally unsupervised method.

Some conditions a good similarity measure has to obey

1. 0 less than or equal similarity(A, B) less than or equal to 1
2. similarity(A, A) = 1
3. similarity(A, B) =/= 1 if A =/= B

*** In our project we have used Cosine Similarity.

Chapter 5

Results and Conclusion

5.1 Sample Results

5.1.1 Word Scoring

Text-1

In his 2018 letter to Amazon shareholders, Jeff Bezos revealed, uncharacteristically, a key indicator of the companys impressive achievements: Amazon Prime subscriptions have topped 100M. This was followed by glowing quarterly numbers that proclaimed 51B in revenue, more than 10 percent of which was garnered by Amazon Web Services (AWS), a genuinely amazing service that saw 49 percent year-to-year growth. But whats the most impressive facet of this litany? Its this: Jeff Bezos is the too-rare CEO who writes to his shareholders every year. Wait what is that too-rare epithet supposed to mean? Dont company CEOs duly and regularly pay their respects to company owners in the cover letters affixed to their annual reports? Ah, yes, they want us to think they do. But our gut knows better. Blame attorneys, PR consiglieri, or a weak-spined CEO for yielding to societys offensive demand that we not offend anyone ever. Whatever the reason, when we listen to typical corpospeak there is no music, no soul, no human reaching out to us. Theres no such lack of soul in Amazons annual letters to shareholders. Founder CEO Jeff Bezos rejoices, occasionally apologizes, and always expounds his companys management philosophy and practices and he does so with wit and good grace. While the annual letters always take care of business, you can see Bezos growing more confident as he breaks from the traditional form. He devoted the 2004 letter to a hypothetical people transport machine company as an illustration of why Amazon is obsessed with free cash flow per share rather than the usual corporate fascination with earnings and profit. In 2010, he penned a tribute to Amazons engineers by explaining, and not just in laymans terms, what they do. The tone was just right, neither disingenuously geeky nor overtly tongue-in-cheek:The diversity of products demands that we employ modern regression techniques like trained random forests of decision trees to flexibly incorporate thousands of product attributes at rank time. Now, if the eyes of some shareowners dutifully reading this letter are by this point glazing over, I will awaken you by pointing out that, in my opinion, these techniques are not idly pursued they lead directly to free cash flow. What he was doing, although we may not have fully appreciated it at the time, was giving a brief tour of AWS, arguably Amazons most important technology.

Summary-1

This was followed by glowing quarterly numbers that proclaimed 51B in revenue, more than 10 percent of which was garnered by Amazon Web Services (AWS), a genuinely amazing service that saw 49 percent year-to-year growth. But whats the most impressive facet of this litany? Its this: Jeff Bezos is the too-rare CEO who writes to his shareholders every year.Founder CEO Jeff Bezos rejoices, occasionally apologizes, and always expounds his companys management philosophy and practices and he does so with wit and good grace.

Text - 2

Theres a famous quote by Albert Einstein:Logic will take you from A to B,Imagination will take you everywhere. I think this sums up the biggest challenge that we have today with innovation in India. Children are extremely imaginative, but as they grow, imagination dies as people struggle to conform. Growing up, my father always encouraged me to ask questions and challenge status quo. I believe that a healthy dialogue is crucial to build a culture of innovation. If we look at world history, the path breaking innovations we see today have been the direct result of people going against the tide and daring to ask questions. There are five barriers to innovation, perceptual, emotional, cultural, intellectual and environmental. Theres a saying in Kannada, which when loosely translated means a poets eyes can see what even the Sun cannot see; a drunkard can see what even the poet cannot see. Any situation doesnt just depend on the external factors but also on your own mental beliefs. Teachers need to create a safe environment where diversity of thought is encouraged. Teachers here include parents, bosses, relatives and neighbours, who play a big role in shaping a childs mind. Our pedagogy and culture in universities and schools needs to evolve accordingly at a disruptive pace.

Summary -2

Theres a famous quote by Albert Einstein:Logic will take you from A to B,Imagination will take you everywhere. If we look at world history, the path breaking innovations we see today have been the direct result of people going against the tide and daring to ask questions. There are five barriers to innovation, perceptual, emotional, cultural, intellectual and environmental. Theres a saying in Kannada, which when loosely translated means a poets eyes can see what even the Sun cannot see; a drunkard can see what even the poet cannot see.

5.1.2 Graph Scoring

Text-1

In his 2018 letter to Amazon shareholders, Jeff Bezos revealed, uncharacteristically, a key indicator of the companys impressive achievements: Amazon Prime subscriptions have topped 100M. This was followed by glowing quarterly numbers that proclaimed 51B in revenue, more than 10 percent of which was garnered by Amazon Web Services (AWS), a genuinely amazing service that saw 49 percent year-to-year growth. But whats the most impressive facet of this litany? Its this: Jeff Bezos is the too-rare CEO who writes to his shareholders every year. Wait what is that too-rare epithet supposed to mean? Dont company CEOs duly and regularly pay their respects to company owners in the cover letters affixed to their annual reports? Ah, yes, they want us to think they do. But our gut knows better. Blame attorneys, PR consiglieri, or a weak-spined CEO for yielding to societys offensive demand that we not offend anyone ever. Whatever the reason, when we listen to typical corpospeak there is no music, no soul, no human reaching out to us. Theres no such lack of soul in Amazons annual letters to shareholders. Founder CEO Jeff Bezos rejoices, occasionally apologizes, and always expounds his companys management philosophy and practices and he does so with wit and good grace. While the annual letters always take care of business, you can see Bezos growing more confident as he breaks from the traditional form. He devoted the 2004 letter to a hypothetical people transport machine company as an illustration of why Amazon is obsessed with free cash flow per share rather than the usual corporate fascination with earnings and profit. In 2010, he penned a tribute to Amazons engineers by explaining, and not just in laymans terms, what they do. The tone was just right, neither disingenuously geeky nor overtly tongue-in-cheek:The diversity of products demands that we employ modern regression techniques like trained random forests of decision trees to flexibly incorporate thousands of product attributes at rank time. Now, if the eyes of some shareowners dutifully reading this letter are by this point glazing over, I will awaken you by pointing out that, in my opinion, these techniques are not idly pursued they lead directly to free cash flow. What he was doing, although we may not have fully appreciated it at the time, was giving a brief tour of AWS, arguably Amazons most important technology.

Summary-1

In his 2018 letter to Amazon shareholders, Jeff Bezos revealed, uncharacteristically, a key indicator of the companys impressive achievements: Amazon Prime subscriptions have topped 100M. This was followed by glowing quarterly numbers that proclaimed 51B in revenue, more than 10 percent of which was garnered by Amazon Web Services (AWS), a genuinely amazing service that saw 49 percent year-to-year growth. But whats the most impressive facet of this litany? Its this: Jeff Bezos is the too-rare CEO who writes to his shareholders every year. Founder CEO Jeff Bezos rejoices, occasionally apologizes, and always expounds his companys management philosophy and practices and he does so with wit and good grace.

Text - 2

Theres a famous quote by Albert Einstein:Logic will take you from A to B,Imagination will take you everywhere. I think this sums up the biggest challenge that we have today with innovation in India. Children are extremely imaginative, but as they grow, imagination dies as people struggle to conform. Growing up, my father always encouraged me to ask questions and challenge status quo. I believe that a healthy dialogue is crucial to build a culture of innovation. If we look at world history, the path breaking innovations we see today have been the direct result of people going against the tide and daring to ask questions. There are five barriers to innovation, perceptual, emotional, cultural, intellectual and environmental. Theres a saying in Kannada, which when loosely translated means a poets eyes can see what even the Sun cannot see; a drunkard can see what even the poet cannot see. Any situation doesnt just depend on the external factors but also on your own mental beliefs. Teachers need to create a safe environment where diversity of thought is encouraged. Teachers here include parents, bosses, relatives and neighbours, who play a big role in shaping a childs mind. Our pedagogy and culture in universities and schools needs to evolve accordingly at a disruptive pace.

Summary -2

Theres a famous quote by Albert Einstein:Logic will take you from A to B,Imagination will take you everywhere. Children are extremely imaginative, but as they grow, imagination dies as people struggle to conform. There are five barriers to innovation, perceptual, emotional, cultural, intellectual and environmental. Theres a saying in Kannada, which when loosely translated means a poets eyes can see what even the Sun cannot see; a drunkard can see what even the poet cannot see.

5.2 Conclusion

This project implements the most important text summarization strategies of word scoring and graph scoring. The Results of the implementation are quite good, summaries are able to represent the original text to maximum extent. Also we found that there is around 70 to 80 percent similarity between the summaries generated by word scoring and graph scoring algorithms.

References

[1] *Assessing sentence scoring techniques for extractive text summarization*; Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, Luciano Favaro
https://www.researchgate.net/publication/257404974_Assessing_sentence_scoring_techniques_for_extractive_text_summarization

[2] https://github.com/icoxfog417/awesome-text-summarization

[3] https://nlpforhackers.io/textrank-text-summarization/