International Journal of Core Engineering & Management
Volume-6, Issue-09, 2020 ISSN No: 2348-9510

TEXT SUMMARIZATION USING TEXT RANK AND LATENT SEMANTIC ANALYSIS (LSA) ALGORITHMS

Krishna Mohan
Department of Computer Science
University of Texas at Dallas
Dallas, TX
nagasaikrishnamohan.pitchikala@utdallas.edu

Naveen Kumar
Department of Computer Science
University of Texas at Dallas
Dallas, TX
naveenkumarreddy.inaganti@utdallas.edu

Sai Krishna
Department of Computer Science
University of Texas at Dallas
Dallas, TX
saikrishna.chirumamilla@utdallas.edu

Abstract

Text Summarization is used to brief a document or group of sentences. Nowadays it is used by lot of companies to improve customer satisfaction. Summarization can be implemented in Extractive or Abstractive methods. Extractive methods use the available sentences from the document in providing summary. In this project we implemented text summarization in two different ways using extractive techniques; we developed Latent Semantic Analysis and Text Rank algorithms to achieve summarization. To get little idea on working of algorithms we used ROUGE score to evaluate the model.
Index Terms – Text Summarization, Latent Semantic Analysis, Text Rank, Rouge, News Articles dataset

I. INTRODUCTION

Text Summarization can be used to output a meaningful summary of large amounts of texts. There is a huge amount of data getting generated using software applications, so it is difficult and time-consuming for humans to write a meaningful summary to large amounts of text data, and also generating summary automatically will help companies to take necessary actions to improve the experience to users1. For example, It can be used by product owners to get an overall review of the product given by multiple users, It can be used by the financial sector to get a brief report of analysis instead of a detailed one. It can also be used to search engines to search for a document from a summary of the page.

Text Summarization can be implemented in two ways, they are called extractive and abstractive methods. Extractive methods summarize texts or documents by selecting more weighted sentences from given data where Abstractive methods try to capture the meaning of sentences and output summary with new set of sentences like the way human reviewers extract from data, As abstractive methods are more complex to implement we can see the popularity extractive summarizers [1].

We implemented this project in two different ways, one using the Text Rank algorithm and the other is using Latent Semantic Analysis. Algorithms are developed without using any libraries. News articles dataset is used to generate summary for documents and results are compared against pre-defined summaries to get the performance of the model.

II. LITERATURE SURVEY

Single document summarization can be used to get an overview of a topic. Early work on summaries started with calculating the frequency of words in a document and the occurrence of words with frequency in each sentence after removing stop words. In addition to the above, one more feature was introduced to detect the positioning of the sentence. For Example, It is proved the majority of paragraphs have a topic sentence as first and few have a topic sentence at the last. Two more additional features were introduced later to this, one is the presence of cue words and the other is structure of the document. Weights were given to sentences based on these four features to derive a summary [1, 2].

III. PROPOSED METHODOLOGY

3.1 Understanding the Dataset

This project can be used with smaller to larger datasets. We can find a wide range of datasets differing in categories like new articles, amazon product reviews, email summarization etc. The limitations we have with respect to infrastructure, having only basic versions of aws clusters and limitations from jupyter notebook we are using medium sized dataset News Articles. The dataset taken from UCI ML consists of about 28000 news articles content along with pre-defined summary. We used pre-defined summaries to know the model implementation even though accuracy for text summaries cannot be done

3.2 Text Rank

While supervised algorithms have properties to produce interpretable rules which characterize a 'key phrase' they need more training data with lots of key phrases to understand the content to prepare the summary. Instead, Text Rank algorithm learns key phrases from the data itself and we don't need any references for key phrases. Thus, it can be implemented easily on all the documents. Text Rank is a graph-based summarization algorithm or ranking algorithm. It builds a graph using the word list or text units from a sentence as the nodes. Edges between word list or text units are based some measure of the relationship between or simply similarity between the wordlist vectors [5].

We have to create a graph that takes wordlist(nodes) and all other sentences in a file (all other

nodes) as input and creates an adjacency list of nodes based on the similarity between the selected node and all other nodes. There exists an edge between the nodes only if they both share some common words and the weight of an edge is based on the relation/similarity between them. There won't be any edge between the wordlists that do not share any common words between them. The similarity coefficient between the wordlist is calculated using the following formula:

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

Equation.1. Image displaying formula for similarity calculation

In our implementation above, we have added a smoothing factor of 1 in the denominator to avoid any cases of divide by zero. Unlike page rank here in text rank all the edges are mostly undirected and have a weight equal to the similarity between them. Once we build the graph using all the nodes, we now apply text rank algorithm on it.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Equation.2. Image displaying Text Rank formula implementation

3.3 Implementation of Text Rank

To start with we pick the content from the data file (to which the summary has to be generated) and send it for processing. The first step includes separating the individual sentences from the content and labelling them. The summary includes some of these sentences which describe or summarize the whole content. The Next step is to preprocess these sentences, which includes removing stop words, converting to lower case tokenize, removing unwanted characters, and lemmatization. Now we are left with pre-processed sentences from the content. We now convert these sentences into wordlist6.

We now apply text rank algorithm to these wordlists of sentences considering each sentence as a node. Every neighbour is assigned a value of 0.15 and itself a value of 1. We run the text rank algorithm iteratively on these sentences to yield the sentences with the highest rank. Next, sort the rank rdd according to the rank value and will look out for the top n sentences with the highest rank. By using the labels of these sentences, we will look out for the original sentences that we have from our content. At last, we append all such sentences to the output file. This output file contains the summary in 'n' sentences. We can change the value of n if we have larger content and the value of n does not sufficiently fulfil the summary.

[Note: When we consider running the text rank on all sentences without limiting the length of sentences it did not give better results. So, we pick sentences that are long enough for a sentence in a summary. Typically sentences with a minimum of 10 words (including stop words.).]

International Journal of Core Engineering & Management
Volume-6, Issue-09, 2020 ISSN No: 2348-9510

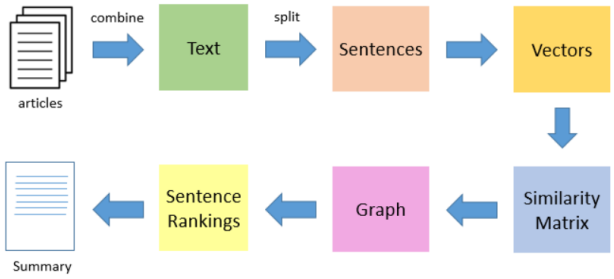

Fig.1. Image displaying Text Rank model implementation [9]

3.4 Latent Semantic Analysis

Latent Semantic Analysis algorithm is used to deduce conclusions from the text. It is a statistical unsupervised algorithm. LSA doesn't require any prior knowledge to extract hidden summary from the articles. LSA does the extraction using the words that are used jointly and with the words which are identified in various sentences. If there is a high correlation among words in different sentences, then those can be put together to form a summary. The decision is made based on the usage of words. For example, a word might have two different meanings based on the context. Only when the context is same, they are considered3.

LSA uses Singular Value Decomposition internally to find the relation between words and sentences. The popularity for LSA is due to the utilization of SVD algorithm. It can reduce noise, which leads to high accuracy [3, 4].

3.5 Data Preparation for LSA

LSA needs the data to be in format of a matrix. Initially the data is pre-processed. Entire article is split into sentences using tokenizer. Then the sentences are broken into words. Using word lemmatization, if any word is meaningless then it is removed, and any other stop words are also removed. Sentences of certain length suppose of length 5 or 10 can be removed. Because they might not contribute to a summary of less information. But we have not followed this approach because, there are articles having sentences of length very small. In order to implement our algorithm for general purpose, we have not followed any strict pre-processing techniques.

The input text given to the algorithm is converted into a matrix. Initially created a term frequency vector which tells about the frequency of phrase in a sentence. And then document frequency vector is created which shows the total frequency of particular phrase in entire article. Then we can calculate the inverse document frequency vector and the tf-idf matrix is formed. The columns represent the sentence numbers and the rows represent the phrases. The value corresponding to a particular row and column defines the importance of word in particular sentence. There are various ways in filling out the matrix. Sparse matrix is created in general.

SVD performance highly relies on the size of the matrix. The complexity increases with increase in size of matrix. In order to decrease to size, data is pre-processed before proceeding with creation of matrix4.

Different methods in creating the matrix

- Phrase Frequency: matrix cells are filed with frequency of phrases in particular sentence.

- Representing in Binary: (0,1) values are inserted based on the occurrence of phrase in sentence.
- Tf-idf: The cells are filled with tf-idf values. Higher the value, it's occurrence is more in particular sentence than in others.
- Log Entropy: Cells are filled calculating log entropy. It provides us the importance of word in that sentence.
- Root Type: If the phrase is a noun then its frequency is inserted. Otherwise '0' is inserted into particular cell.

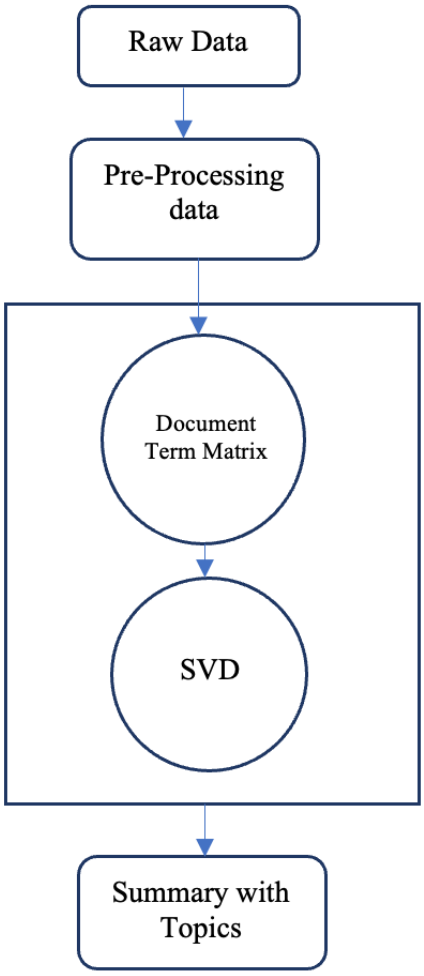

Fig.2. Flow diagram of LSA Algorithm for Text Summarization

3.6 Single Value Decomposition (SVD)

SVD is mathematical model that helps us to establish relationship between phrases and sentences. The document term matrix is represented as vectors (points) in Euclidean space. These vectors denote the sentences.

International Journal of Core Engineering & Management
Volume-6, Issue-09, 2020 ISSN No: 2348-9510

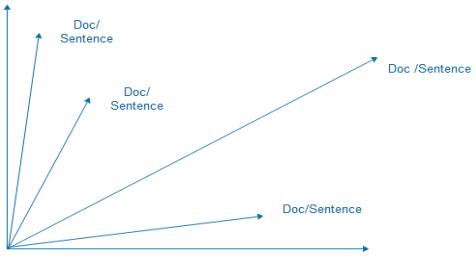

Fig.3. Euclidean Plane

The tf-idf matrix given to SVD algorithm is decomposed into three new matrices:

$A = U_k \sum_k V^T$, where

- A is the tf-idf input matrix of size m x n
- U is phrases * derived topics of size m x n
- ∑_k is a diagonal matrix consisting of scaling values of size n x n
- V^T is a matrix having sentences * derived topics of size n x m

-	Sen 1	Sen 2	Sen 3	Sen 4
Term 1	1	0	1	1
Term 2	0	0	1	1
Term 3	1	1	1	0
Term 4	0	0	0	1

Table.1. Input Term document (sentence) matrix m x m (A).

	Topic 1	Topic 2
Term 1	0.2	0.1
Term 2	0.4	0.5
Term 3	0.2	0.1
Term 4	0.1	0.3

Table.2. Phrases association with topics m x n (U_k)

	Topic 1		Topic 2
Topic 1	0.8		0
Topic 2	0		0.2

Table.3. importance of topics n x n diagonal matrix ($\sum_k$).

International Journal of Core Engineering & Management
Volume-6, Issue-09, 2020 ISSN No: 2348-9510

	Sen 1	Sen 2	Sen 3	Sen 4
Topic 1	0.4	0.7	0.3	0.7
Topic 2	0.2	0.1	0.4	0.2

Table.4. Distribution of topics among documents n x m (V^T).

3.7 LSA Topics Selection

The algorithm provides us with many topics based on the column count of matrix. We have selected only one topic per summary and it has five sentences. This is followed to maintain uniformity across all articles. We get many topics based on correlation, they might be sub-topics of all already selected topic for the summary. There is no repetition of content in our summary because we have selected only one topic.

The topic x topic matrix is created by finding the topics which have common sentences in them. After this score for each topic is calculated. Based on the highest value we pick the topic. It indicates higher correlation with the others. After the topic is selected, we go to matrix and choose the first five sentences having the highest value for the chosen topic. By combing these sentences our desired summary is extracted from the article.

IV. IMPLEMENTATION RESULTS

4.1 Observations

a. Text Rank Algorithm

```
In [62]: hypothesis[4]

Out[62]: 'Risk Warning: The price and value of investments and their income fluctuates; you may get back less than the amount you invest
edThe value of international investments may be affected by currency fluctuations which might reduce their value in sterling Pl
ease note, the tax treatment of these products depends on the individual circumstances of each customer and may be subject to c
hange in future If you are uncertain about the tax treatment of the products you should contact HMRC or seek independent tax ad
vice If you are unsure about the suitability of a particular investment or think that you need a personal recommendation, you s
hould speak to a suitably qualified financial adviser'

In [63]: reference[4]

Out[63]: 'The value of international investments may be affected by currency fluctuations which might reduce their value in sterling.\nF
oreign markets will involve different risks from the UK markets.\nRisk Warning: The price and value of investments and their in
come fluctuates; you may get back less than the amount you invested.\nPlease note, the tax treatment of these products depends
on the individual circumstances of each customer and may be subject to change in future.\nIf you are uncertain about the tax tr
eatment of the products you should contact HMRC or seek independent tax advice.'
```

Fig.4. Above displays result of summary from Text Rank Algorithm.

International Journal of Core Engineering & Management
Volume-6, Issue-09, 2020 ISSN No: 2348-9510

b. LSA Algorithm

```
In [20]: print("Displaying first 5 values of generated summary by LSA algorithm:")
         summary[1:5]

Displaying first 5 values of generated summary by LSA algorithm:

Out[20]: [["All rights reserved.© 2019 Thomson/Reuters.A government report on Friday showed U.S. employers added 175,000 jobs last mo
         nth after creating 129,000 new positions in January.Aside from the reference to the jobs report, Plosser's speech was largel
         y the same as a speech he gave in London on March 6.In a speech at the Bank of France, Philadelphia Fed President Charles Pl
         osser pointed to U.S. payroll gains in December, January and February."],
          ['More must-reads on MarketWatch:\n\nWSJ's Hilsenrath: Fed is shifting its exit strategy\n\nWarning sign for stocks: High m
         argin levels\n\nJeff Reeves: Bull market's biggest winners and losers"Knock on wood it would all go very smoothly, but you n
         ever know," he said in a question-and-answer session after the speech.As the economic outlook improves, the Fed announced in
         January its second cut to its monthly purchase program to $65 billion.Real output showed growth of 3.3% from 1.8% in the fir
         st half.He also raised concerns about how the Fed might ultimately reduce the size of its balance sheet without disrupting i
         nflation or the economy.'],
          ["He also suggested that the US economy would expand by around 3% this year and that upside risks to inflation existed in t
         he longer term.Plosser, a critic of the Fed's quantitative easing program, also expressed hope that the central bank's monet
         ary policy would return to "normalcy" as soon as possible.The voting FOMC member suggested that the last three Nonfarm Payro
         lls monthly readings were weaker due to the "effect of the unusually severe winter weather," but that he expected the US lab
         or market to improve from now on.FXStreet (Łódź) - Philadelphia Fed President Charles Plosser said today in a speech deliver
         ed at the Bank of France alongside Governor Christian Noyer that he expected US unemployment to drop even lower than 6.2% by
         the end of the year."],
          ['Please note, the tax treatment of these products depends on the individual circumstances of each customer and may be subj
         ect to change in future.If you are uncertain about the tax treatment of the products you should contact HMRC or seek indepen
         dent tax advice.In some cases the risks will be greater.Foreign markets will involve different risks from the UK markets.If
         you are unsure about the suitability of a particular investment or think that you need a personal recommendation, you should
         speak to a suitably qualified financial adviser.']]
```

Fig.5. Above displays result of only summary from LSA Algorithm

```
In [24]: final_summary_df.show(20)

+--------------------+--------+--------------------+--------------------+--------------------+
|                 URL|CATEGORY|             CONTENT|             SUMMARY|   generated_summary|
+--------------------+--------+--------------------+--------------------+--------------------+
|http://www.livemi...|business|Paris/London/Atla...|Paris/London/Atla...|It's OK to contin...|
|http://www.moneyn...|business|Severe winter wea...|Severe winter wea...|All rights reserv...|
|http://www.market...|business|PARISn – The Fede...|"We must back awa...|More must-reads o...|
|http://www.fxstre...|business|FXStreet (Łódź) -...|FXStreet (Łódź) -...|He also suggested...|
|http://www.iii.co...|business|The value of inte...|The value of inte...|Please note, the ...|
|http://in.reuters...|business|BANGALORE (Reuter...|The euro sign lan...|In order to keep ...|
|http://blogs.reut...|business|The European Unio...|The European Unio...|The EU is doing a...|
|http://in.reuters...|business|* Countries grapp...|Policymakers agre...|TUG OF WAR

Offic...|
|http://in.reuters...|business|FRANKFURT, March ...|FRANKFURT, March ...|That seemed to ha...|
|http://www.nasdaq...|business|Shutterstock phot...|Shutterstock phot...|The views and opi...|
|http://blogs.wsj....|business|The European Unio...|The European Unio...|The money won't r...|
|http://www.fxstre...|business|Outlook

Attentio...|As everyone notes...|Noyer says everyt...|
|http://www.busine...|business|Good morning. Her...|Here's what you n...|China's consumer ...|
|http://www.binary...|business|The euro retreate...|"When the euro te...|Ultimately, with ...|
|http://in.reuters...|business|By Laura Noonan

...|By Laura NoonanDU...|The ECB declined ...|
|http://in.reuters...|business|* Impairments and...|As well as the in...|The ECB declined ...|
|http://www.rte.ie...|business|The European Cent...|This is according...|The ECB declined ...|
|http://www.reuter...|business|Bank of France Go...|Bank of France Go...|"Monetary policy ...|
|http://in.reuters...|business|* Noyer sees "per...|Noyer, the govern...|(Reporting by Lei...|
|http://in.reuters...|business|PARIS, March 10 (...|PARIS, March 10 (...|(Reporting by Lei...|
+--------------------+--------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Fig.6. Above displays result of output from LSA Algorithm

4.2 Algorithm Evaluations

a. Text Rank Algorithm Evaluation

Metrics are required to compare a system-generated summary or translation against a single or a set of human-produced references or translation.

ROUGE7 determines the quality of an automatic summary by comparing common units such as n-

grams, word sequences, and word pairs with human-produced summaries.
There are many different metrics in rouge alone we use 3 of them which are:

- ROUGE-N denotes the overlap of N-grams (n-words) between hypothesis and reference summaries.
- ROUGE-1 denotes the overlap of unigram (each word) between hypothesis (generated summary) and reference (original summary)
- ROUGE-1 denotes the overlap of unigram (each word) between hypothesis (generated summary) and reference (original summary)

ROUGE does not try to assess how consistent is the summary, it only tries to assess the similarity or common words just by simply counting how many n-grams in the generated summary matches the n-grams in the original summary.
Illustration:

- Hypothesis: he got first rank in the college
- Reference: he got first in the college

If we consider the number of common individual words in hypothesis and reference as our metric, is not good way of evaluation and it does not work as metric. We get a quantitative value by computing the Precision and Recall using overlap in the context of ROUGE. With multiple references, scores of ROUGE-1 is averaged. Rouge can determine if same general concepts are discussed between the generated and original summary but, it cannot detect weather the generated result is logical or not as it is based on the common words between the summaries. Also, I it found that higher order n-gram measures or try to measure the fluency to some extent [7].

ROUGE	Results
Rouge-1	f-score: 0.570 precision: 0.632 recall:0.516
Rouge-2	f-score: 0.0.523 precision: 0.626 recall:0.449
Rouge-L	f-score: 0.636 precision: 0.691 recall:0.568

Table.5. Rouge evaluation

Note that ROUGE is like that of BLEU measure for machine translation, but BLEU is precision based. To overcome these difficulties, we have chosen another way of evaluating the generated caption know as BLEU score
Another metric that can be used for evaluation is BLUE8 score. BLUE score is calculated by generating the individual segments from hypothesis and comparing them with the reference segments. These scores are averaged over the whole summary to reach an estimate of the translation's overall quality. BLUE score does not consider grammatical correctness. The BLEU score is a number between 0 and 1. This value indicates how similar the hypothesis is to the reference, with values closer to 1 indicates more similarity between them [8].

BLEU score: 0.480

b. LSA Algorithm Evaluation

LSA provides us with so many topics, in that list we are taking only one topic among them, and it is not feasible to make evaluations for LSA

V. CONCLUSION

- Text Rank: With increase in number of iterations we can get good results. The limitation here is, generated summary contains only 5 sentences, but in real world there might be more sentences in existing summary. This leads to decrease in accuracy. Also, the rouge and bleu score are calculated for the combined results, as we cannot display the results for all individual 28000 articles. We can improve the processing time by using the graph frames.
- LSA: There are few advantages of LSA, and they are, Dimensionality reduction of the articles, Helps us to get insights from each topic of the article, Easy to train and adjust by tweaking pre-processing techniques, Easy to obtain relation between phrases. Limitations of LSA are, It just takes the knowledge from the article, doesn't consider real world knowledge. With large amounts of data, there is decrease in performance of LSA because of complex SVD is being used, It doesn't know any relation between words such as ordering, syntactic association, Unable to interpret different meanings of a phrase.

REFERENCES

1. Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef, Saeid Safaei, Text Summarization Techniques: A Brief Survey. https://arxiv.org/pdf/1707.02268.pdf
2. Gupta, V., Lehal, G.S.: A survey of Text Summarization Extractive Techniques.
3. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.348.3840&rep=rep1&type=pdf #page=104
4. Y. Gong, X. Liu: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of the 24 th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States 2001.
5. Thomas K Landauer, Peter W. Foltz, An Introduction to Latent Semantic Analysis
6. http://lsa.colorado.edu/papers/dp1.LSAintro.pdf
7. Text Summarization https://medium.com/@umerfarooq_26378/text-summarization-in-python-76c0a41f0dc4
8. Rada Mihalcea, Paul Tarau, TextRank: Bringing Order into Texts, [ONLINE] Available.
9. https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
10. https://en.wikipedia.org/wiki/ROUGE_(metric)
11. https://en.wikipedia.org/wiki/BLEU
12. https://fptsoftware.com/resource-center/blogs/text-summarization-in-machine-learning