COLLEGE OF ENGINEERING GUINDY
ANNA UNIVERSITY

CS6030
NATURAL LANGUAGE PROCESSING

MINI PROJECT

TOPIC WISE MULTI-DOCUMENT SUMMARIZATION

Q BATCH
DATE: 20- 01- 2021

INTRODUCTION

Text summarization is widely used as a technique for generating concise and important information present within source text. In recent times, due to the advancement in hardware and Natural Language Processing techniques, computers are able to identify and pick out noteworthy information present within text documents, generating shorter versions of text documents in an efficient and faster way compared to humans.

Expanding the scope of typical summarization techniques, topic-wise summarization is a very useful and productive technique which involves identifying dominant topics in the source text and generating a summary of information discussed under each topic. This eliminates the need for readers who look for information on a specific topic in the source text to read the entire document or the overall summary of the document. General overall summary of the document tends to miss out important details on the topic of interest of the reader in the favor of brevity. Topic wise text summarization aims to capture and generate all the varied, salient and relevant information available in the source text on each topic without losing important information.

In this project, we propose a method for performing topic wise summarization on multiple documents. This technique identifies and produces the summary of important details and information on each available topic present in different source documents. ***The idea is to summarize multiple text documents by clustering their contents based on latent topics produced using topic modeling techniques and by generating extractive summaries for each of the identified text clusters. All extractive***

sub-summaries are later combined to generate a summary for each topic in the source documents.

The dataset utilized for our approach is public comments on federal regulations. This is less used and more challenging than more commonly used CNN news dataset for text summarization. The well-known news datasets present their most important information in the first few lines of their source texts, which make their summarization a lesser challenging task when compared to summarizing the federal regulation comments dataset. Contrary to these news datasets, the documents in our dataset are written using a generalized approach, have lesser abstraction and higher compression ratio, thus proposing a greater challenge to generate summaries.

Our model produces encouraging ROUGE results and summaries when compared to the other available extractive and abstractive text summarization models.

OBJECTIVE

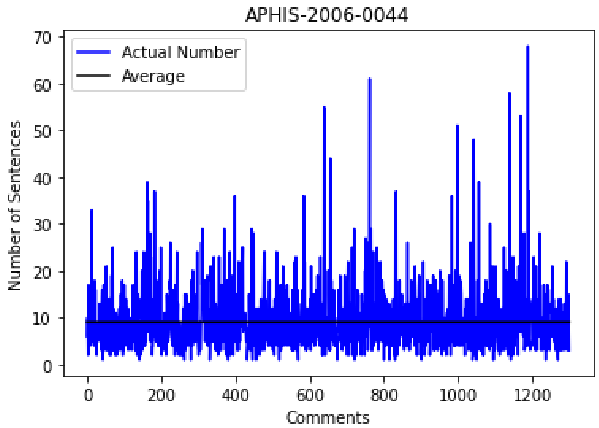
" TO IDENTIFY DOMINANT TOPICS IN THE SOURCE TEXT DOCUMENTS AND GENERATE SUMMARY OF INFORMATION PRESENT IN THE DOCUMENT REFLECTING EACH TOPIC "

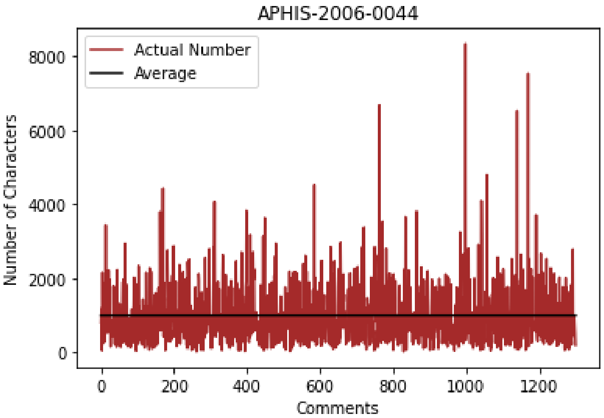
DATASET DESCRIPTION

The dataset considered here contains public comments on federal regulations downloaded from the API at https://www.regulations.gov/. It consists of four documents with each having a list of comments. Some insights into the dataset is given below:

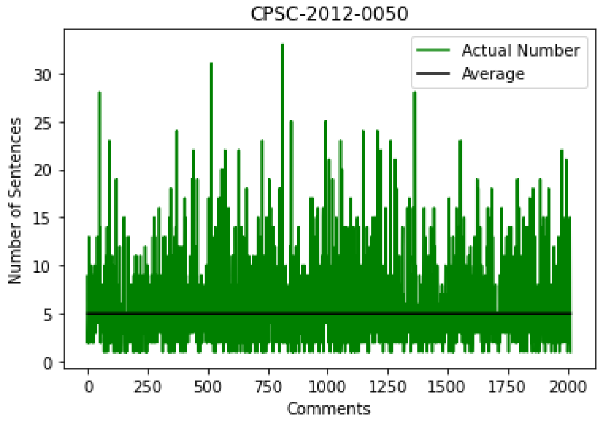
Documents	Number of comments	Average number of sentences in each comment	Average number of characters in each comment
Document 1	1300	9	1000
Document 2	2014	5	503
Document 3	1576	15	1808
Document 4	762	7	646

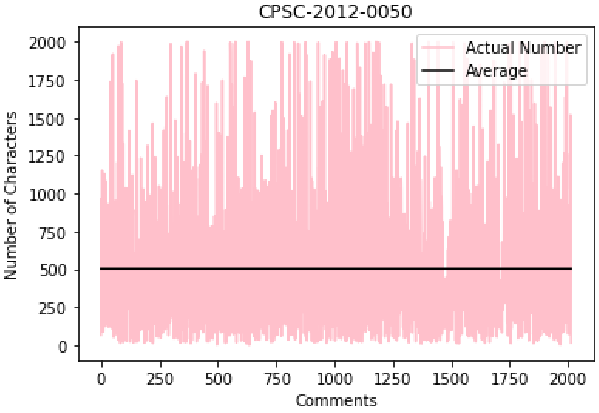
Document 1



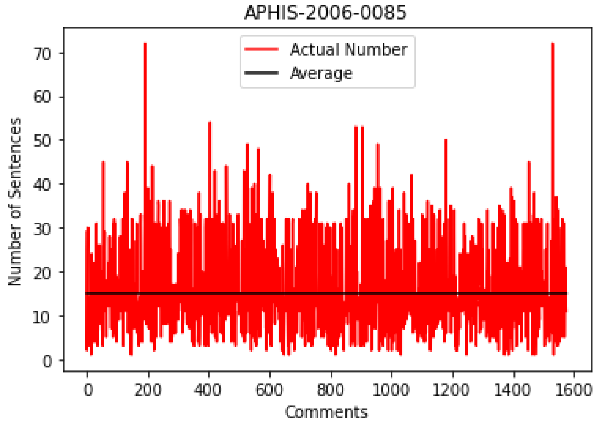


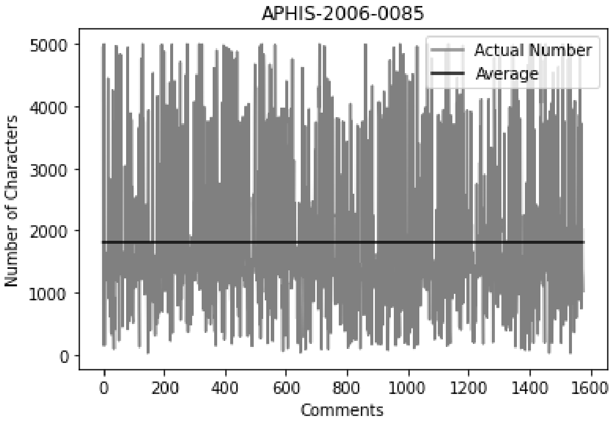
Document 2



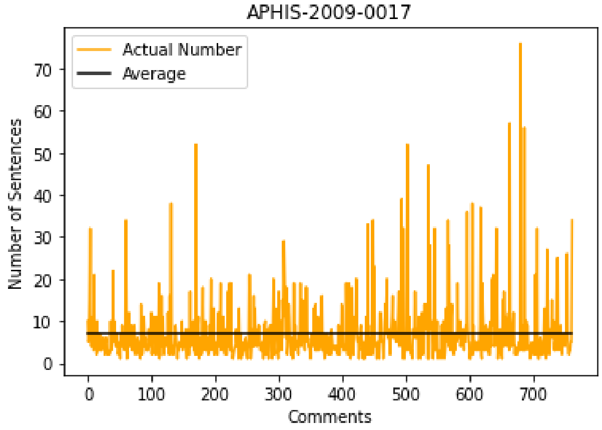


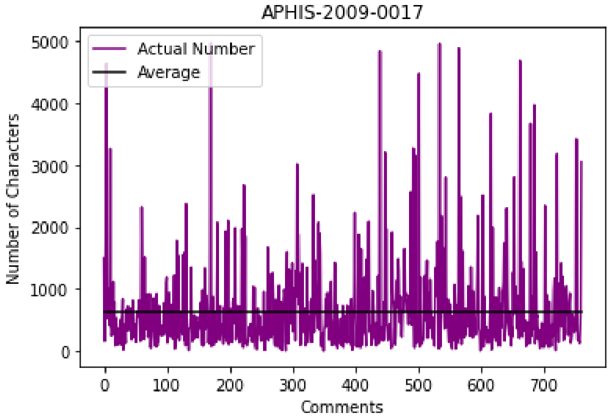
Document 3

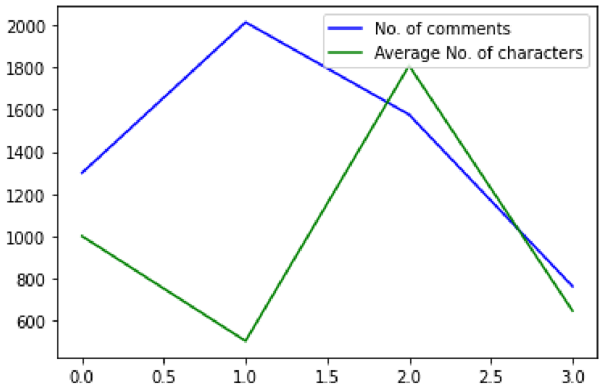


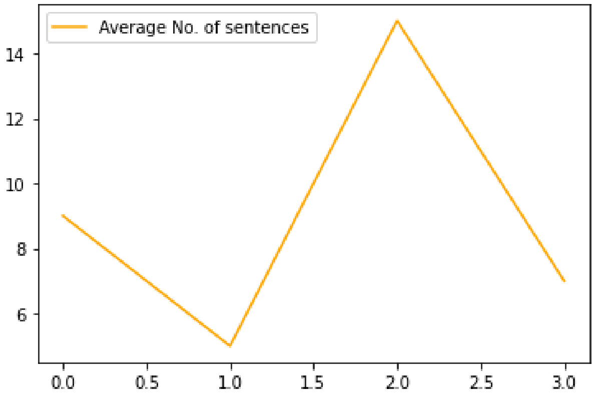


Document 4







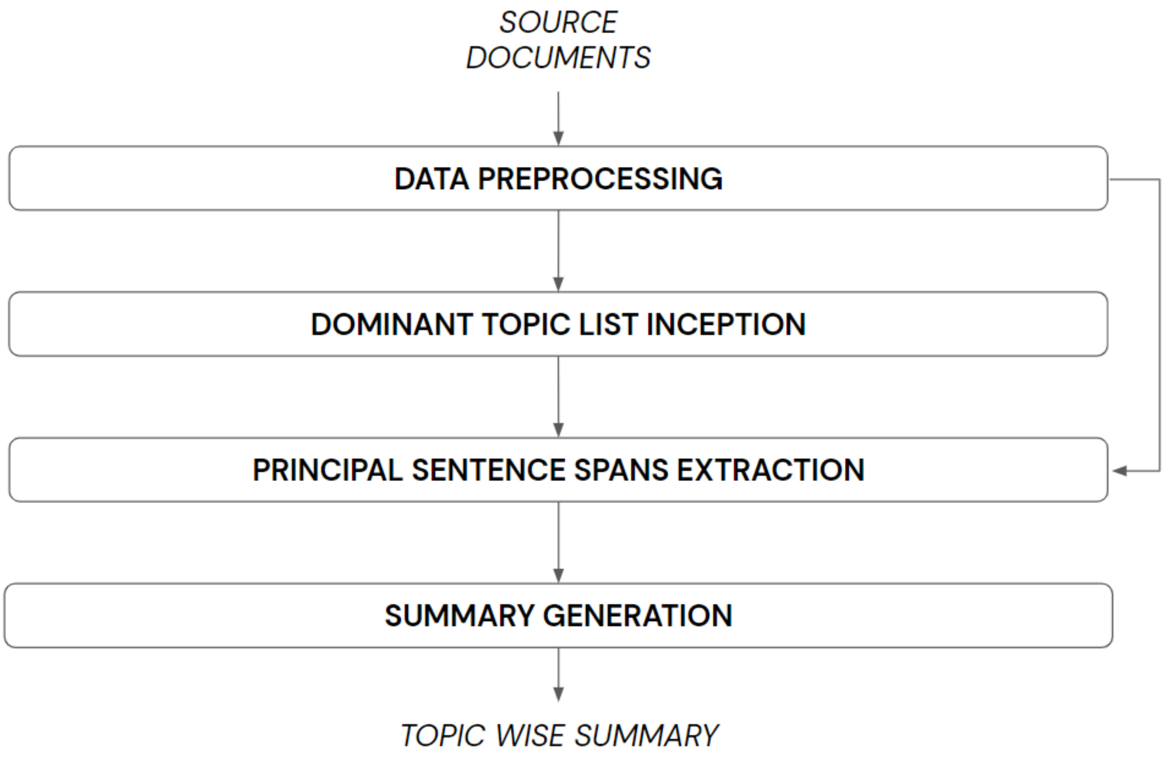


LITERATURE SURVEY

S.NO.	NAME OF PAPER	METHODOLOGY	LIMITATIONS
1.	“Topic Modeling Based Extractive Text Summarization”, Kalliath Abdul Rasheed Issam, Shivam Patel, Subalalitha C. N.	It is a single document summarization technique where: (1) Dominant topics are identified using topic modeling (2) Sentences in the source text are clustered among the different topics (3) Important sentences are selected, producing top few of them as summary	The technique performs poorly for topics with few sentences
2.	"Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization", Libin Yang, Xiaoyan Cai, Yang Zhang, and Peng	It presents the technique of multi-document summarization using topic ranking and by generating topic clusters. Their model is successful in reducing redundancy in the generated summaries and it also produces	Since it lacks topic diversity, it presents challenges to generate WikiHow summaries of high coverage.

	Shi	summaries that are of high quality.	
3.	"A topic-based sentence representation for extractive text summarization", Nikolaos Gialitsis, Nikiforos Pittaras, and Panagiotis Stamatopoulos	An approach to extractive text summarization using binary classification modeling and topic-based sentence extraction using LDA	They use supervised learning to perform binary classification on sentences, where one class holds sentences to be included in the summary and the other class includes the sentences that should not be present in the summary.
4.	"Multiple documents summarization based on evolutionary optimization algorithm"Rasim M.Alguliev, Ramiz M.Aliguliyev, and Nijat R.Isazade	An evolutionary algorithm is proposed by Rasim M. A. et al. [19] to perform summarization on multiple documents. They claim their model to have a superior correlation between sentences with a low rate of redundancy.	Their model requires a large computational overhead. Aiming to have good topic coverage, this approach cannot be directly applied to datasets like WikiHow for generating topic-based extractive summaries.

MODULE DIAGRAM



MODULE DESCRIPTION

DATA PREPROCESSING:

INPUT: Text Documents

OUTPUT: Preprocessed text

The functionality of this module is to preprocess the text documents in order to make it suitable for further processing. Specifically, the following preprocessing steps are carried out on the text:

- stripping out punctuation and non-alphabetical characters
- tokenization
- lemmatization using the NLTK WordNetLemmatizer
- bi-grammization via Gensim
- removal of stopwords using an augmented version of the NLTK English stopwords corpus
- removal of low and high frequency tokens
- removal of documents that are too short and too long (this step is specifically geared towards public feedback or product reviews)

DOMINANT TOPIC LIST INCEPTION:

INPUT: Text Documents

OUTPUT: List of Dominant topics

The functionality of this module is to identify and create a list of topics dominant in the entire text document collection. For this purpose, LDA (Latent Dirichlet Allocation) is used. It reads, preprocesses and vectorizes the list of documents, performs LDA computation and identifies dominant topics.

PRINCIPAL SENTENCE SPANS EXTRACTION

INPUT: Topic list and Source Documents

OUTPUT: Most important and relevant sentences extracted from the source documents reflecting each topic in the Topic list

The functionality of this module is to capture most relevant and critical sentences from all the source documents which most reflect each topic in the input Topic list. For this purpose, Extractive Summarization(ES) is used. ES is the technique of creating a summary of source text by identifying chief and valuable sentences and displaying them as they are. Here, the ES technique is changed in such a way, so that it produces sentences that most reflect and are most relevant to a given topic. The algorithm performs the following steps to extract sentences for each topic:

(i) Pass the individual comments in the documents to the LDA to determine the distribution of topics for each comment.

(ii) Filter out the comments whose dominant topic is not equal to the given topic. As a result, a subset of topics that reflect the given topic is left out

(iii) For each comment within this subset:

(i) Split the comment up into sentences, using the NLTK sentence tokenizer

(ii) Feed the sentences to the LDA object to determine the topic
distribution of each sentence.

(iii) Filter out the sentences whose dominant topic is not equal to the given topic, as well as sentences that are too short or sentences that are too long. As a result, a subset of sentences that reflect the given topic are left out.

SUMMARY GENERATION

INPUT: Extracted sentences under each topic

OUTPUT: Summary of information under each topic

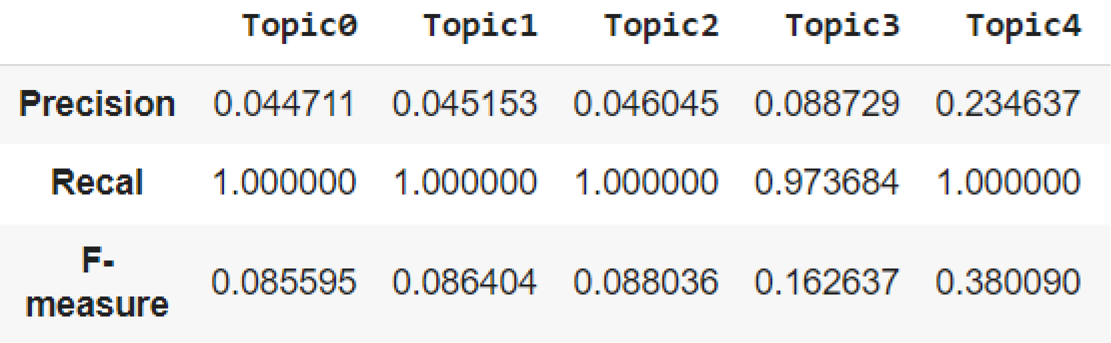
The functionality of this module is to summarize the extracted sentences under each topic. For this purpose, we use an abstractive summarization method BART. Abstractive summarization is a task inNLP that aims to generate a concise summary of a source text. Unlike extractive summarization, **abstractive summarization does not simply copy important phrases from the source text but also potentially come up with new phrases that are relevant**, which can be seen as paraphrasing. Texts summarized using this technique look more human-like and produce more condensed summaries. The **Bidirectional and Auto-Regressive Transformer** or **BART** is a Transformer that combines the Bidirectional Encoder (i.e. BERT like) with an autoregressive decoder (i.e. GPT like) into one seq2seq model. BART can be used with an arbitrary noising scheme. It is a Seq2Seq model thus learns the original text better.

PERFORMANCE METRICS

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring algorithm evaluates the similarity between a candidate document and a collection of reference documents. Use the ROUGE score to evaluate the quality of document translation and summarization models.

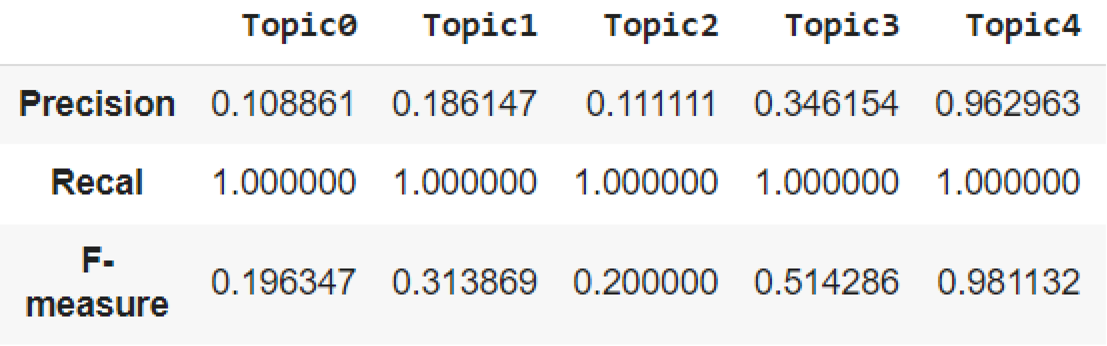
The considered dataset has been divided into 5 sets and the proposed algorithm is executed and the evaluated as follows:

Set 1



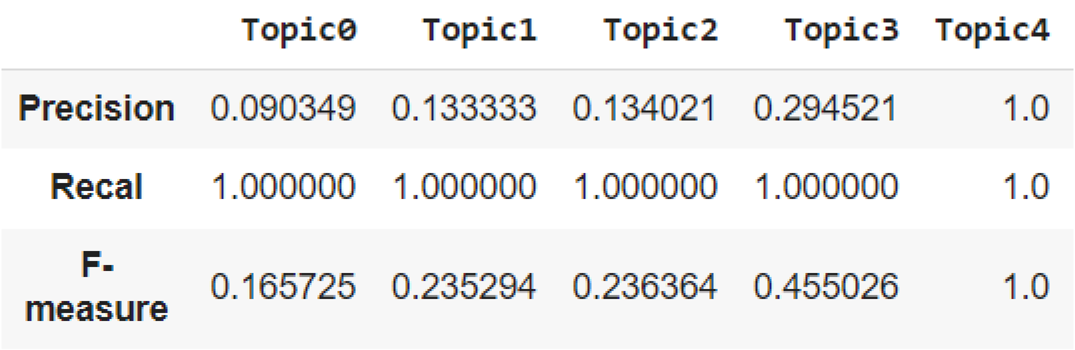
	Topic0	Topic1	Topic2	Topic3	Topic4
Precision	0.044711	0.045153	0.046045	0.088729	0.234637
Recal	1.000000	1.000000	1.000000	0.973684	1.000000
F-measure	0.085595	0.086404	0.088036	0.162637	0.380090

Set 2



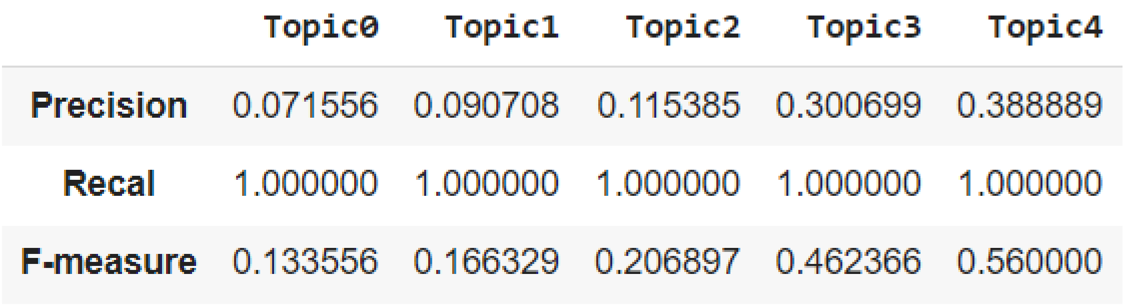
	Topic0	Topic1	Topic2	Topic3	Topic4
Precision	0.108861	0.186147	0.111111	0.346154	0.962963
Recal	1.000000	1.000000	1.000000	1.000000	1.000000
F-measure	0.196347	0.313869	0.200000	0.514286	0.981132

Set 3



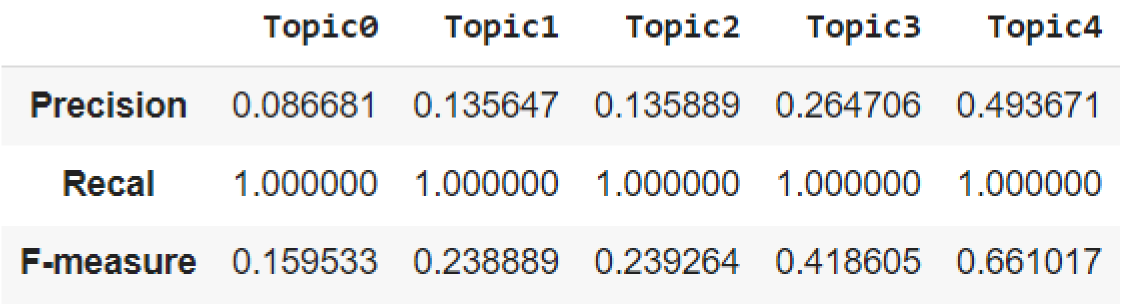
	Topic0	Topic1	Topic2	Topic3	Topic4
Precision	0.090349	0.133333	0.134021	0.294521	1.0
Recal	1.000000	1.000000	1.000000	1.000000	1.0
F-measure	0.165725	0.235294	0.236364	0.455026	1.0

Set 4

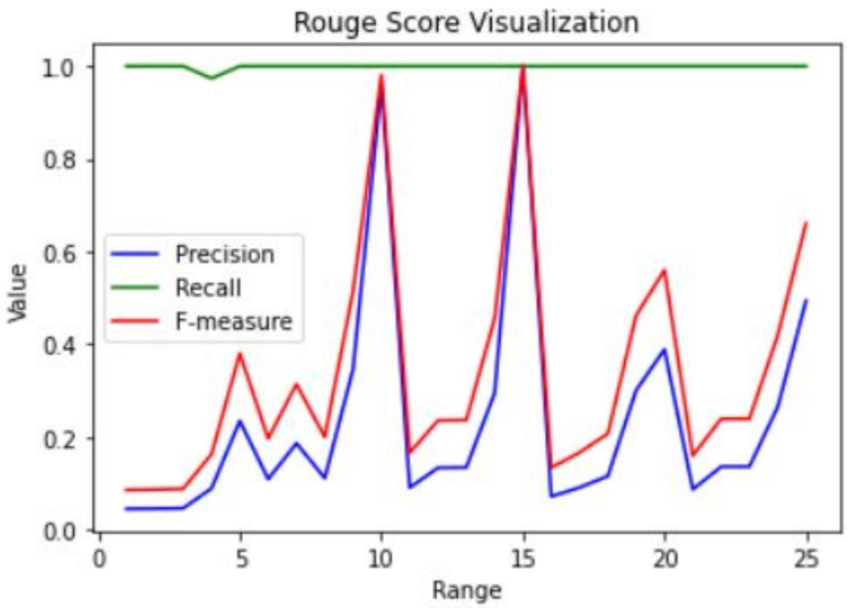


	Topic0	Topic1	Topic2	Topic3	Topic4
Precision	0.071556	0.090708	0.115385	0.300699	0.388889
Recal	1.000000	1.000000	1.000000	1.000000	1.000000
F-measure	0.133556	0.166329	0.206897	0.462366	0.560000

Set 5



	Topic0	Topic1	Topic2	Topic3	Topic4
Precision	0.086681	0.135647	0.135889	0.264706	0.493671
Recal	1.000000	1.000000	1.000000	1.000000	1.000000
F-measure	0.159533	0.238889	0.239264	0.418605	0.661017


Rouge Score Visualization
Value
Range
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
Precision
Recall
F-measure

REFERENCES

1) Kalliath Abdul Rasheed Issam, Shivam Patel, Subalalitha C, “Topic Modeling Based Extractive Text Summarization”, International Journal of Innovative Technology and Exploring Engineering, Volume-9 Issue-6, April 2020, Page No. 1710-1719

2) Libin Yang, Xiaoyan Cai, Yang Zhang, and Peng Shi, "Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization", Information Sciences, Elsevier, Volume 260, 2014, pp. 37-50

3) Nikolaos Gialitsis, Nikiforos Pittaras, and Panagiotis Stamatopoulos, "A topic-based sentence representation for extractive text summarization", Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources, INCOMA Ltd., 2019, pp. 26–34

4) Rasim M. Alguliev, Ramiz M.Aliguliyev, and Nijat R.Isazade, "Multiple documents summarization based on evolutionary optimization algorithm", Expert Systems with Applications, Elsevier, Volume 40, Issue 5, Apr 2013, pp. 1675-1689

– – – – x – – – – x – – – – x – – – – x – – – –