Marwadi University	Marwadi University
	Faculty of Technology
	Department of Information and Communication Technology
Subject: Artificial Intelligence (01CT0616)	Aim: Representation of a document with their keywords using TextRank
Experiment No: 09	Date: Enrolment No: 92200133030

Aim: Representation of a document with their keywords using TextRank

IDE: Google Colab

Theory:

TextRank is an algorithm based on PageRank, which often used in keyword extraction and text summarization. In this article, I will help you understand how TextRank works with a keyword extraction example and show the implementation by Python.

Understand PageRank

There are tons of articles talking about PageRank, so I just give a brief introduction to PageRank. This will help us understand TextRank later because it is based on PageRank. PageRank (PR) is an algorithm used to calculate the weight for web pages. We can take all web pages as a big directed graph. In this graph, a node is a webpage. If webpage A has the link to web page B, it can be represented as a directed edge from A to B. After we construct the whole graph, we can assign weights for web pages by the following formula.

	Marwadi University Faculty of Technology Department of Information and Communication Technology		
Subject: Artificial Intelligence (01CT0616)	Aim: Representation of a document with their keywords using TextRank		
Experiment No: 09	Date:		Enrolment No: 92200133030

$$S(V_i) = (1 - d) + d * \sum_{j \in In(v_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

- S(Vi) - the weight of webpage i
- d - damping factor, in case of no outgoing links
- In(Vi) - inbound links of i, which is a set
- Out(Vj) - outgoing links of j, which is a set
- |Out(Vj)| - the number of outbound links
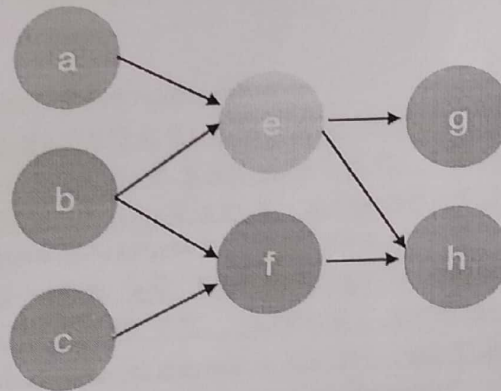
Here is an example to better understand the notation above. We have a graph to represent how web pages link to each other. Each node represents a webpage, and the arrows represent edges. We want to get the weight of webpage e. We can rewrite the summation part in the above function to a simpler version.

We can get the weight of webpage e by the following function.

$$In(v_e) = \{a, b\}, j \in \{a, b\}$$

$$\sum_{j \in \{a,b\}} \frac{1}{|Out(V_j)|} S(V_j) = \frac{1}{|Out(V_a)|} S(V_a) + \frac{1}{|Out(V_b)|} S(V_b)$$

$$= \frac{1}{|\{e\}|} S(V_a) + \frac{1}{|\{e, f\}|} S(V_b)$$

$$= S(V_a) + \frac{1}{2} S(V_b)$$

Marwadi University	Marwadi University
(logo)	Faculty of Technology
	Department of Information and Communication Technology
Subject: Artificial Intelligence (01CT0616)	Aim: Representation of a document with their keywords using TextRank
Experiment No: 09	Date: Enrolment No: 92200133030

$$S(V_e) = (1 - d) + d * \left(S(V_a) + \frac{1}{2} S(V_b) \right)$$

We can see the weight of the webpage e is dependent on the weights of its inbound pages. We need to run this iteration much time to get the final weight. In the initialization, the importance of each webpage is 1.

Pre Lab Exercise:

1. What are the key components of TextRank?

   The key components of TextRank are text prepro-cessing, graph construction, edge weighting, and applying the PageRank algorithm to rank the words.

2. What are some limitations of TextRank for keyword extraction?

   Some limitations of TextRank for keyword extraction are that it ignores word semantics, depends heavily on co-occurrence windows.

3. What are the advantages of TextRank for keyword extraction?

   The advantages of TextRank for keyword extract-ion are that it is unsupervised, does not requ-ire labeled data.

Program (Code):
To be attached with

Results:
To be attached with

Observation:

![Marwadi University logo] Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology
Subject: Artificial Intelligence (01CT0616)	Aim: Representation of a document with their keywords using TextRank
Experiment No: 09	Date:	Enrolment No: 92200133030

Post Lab Exercise:

1. Write about "your ownself" (in not more than 500 words→ You know better about you, rather than ChatGPT!!). Extract top-25 keywords for your portfolio and analyze it in 3 categories: keyword can not give the real idea, keyword gives the real idea, keyword is irrelevant. Paste the code, your portfolio, output and your analysis.