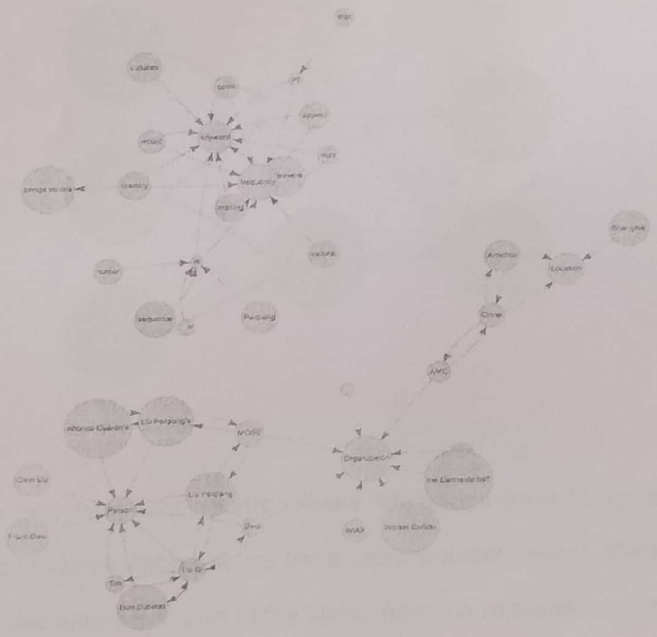
Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Artificial Intelligence (01CT0616)	Aim: TextRank for Document Summarization	
Experiment No: 10	Date:	Enrolment No:

Aim: TextRank for Document Summarization

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.



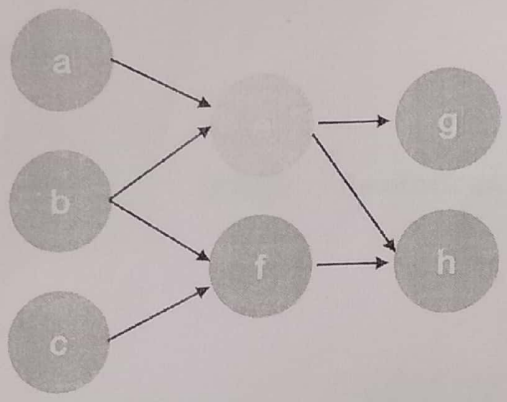
Understand PageRank

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	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Artificial Intelligence (01CT0616)	Aim: TextRank for Document Summarization	
Experiment No: 10	Date:	Enrolment No:

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

- 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.

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

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

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

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

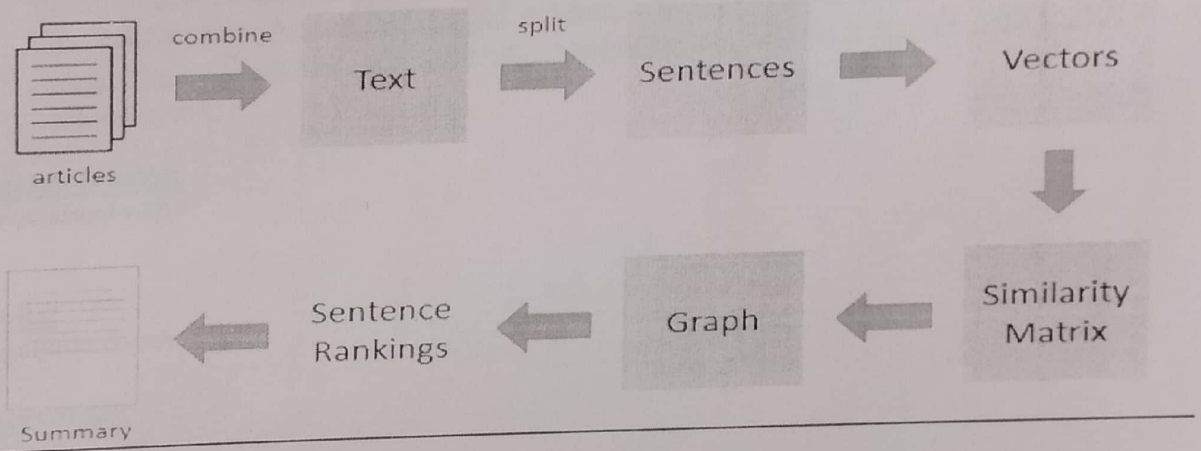
We can get the weight of webpage e by the following function.

Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Artificial Intelligence (01CT0616)	Aim: TextRank for Document Summarization	
Experiment No: 10	Date:	Enrolment No:

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

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.

TextRank for document Summarization



TextRank works in the following steps:
1. Tokenize documents into sentences.
2. Preprocess each sentence in the document.
3. Count key phrases and normalize them or produce TFIDF Matrix, you can also use any kind of vectorization such as spacy vectors.
4. Calculate the Jaccard Similarity between sentences and key phrases.
5. Rank the sentences with higher significance.

Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Artificial Intelligence (01CT0616)	Aim: TextRank for Document Summarization	
Experiment No: 10	Date:	Enrolment No:

Pre Lab Exercise:

1. How keyword extraction is similar to document summarization process?

Keyword extraction is similar to document summarization because both aim to identify the most important information from a text in a condensed

2. How keyword extraction is different to document summarization process?

Keyword extraction differs from document summarization because it selects only important words or phraphrases, which while summarization generates

3. Limitation of TextRank as a summarizer

A limitation of TextRank as a summarizer is that it may produce disjointed summaries lacking logical flow or deep understanding of the text context.

Program (Code):

To be attached with

Results:

To be attached with

Observation:

Marwadi University	Marwadi University Faculty of Technology Department of Information and Communication Technology	
Subject: Artificial Intelligence (01CT0616)	Aim: TextRank for Document Summarization	
Experiment No: 10	Date:	Enrolment No:

Post Lab Exercise:

1. Write about "your ownself" (in not more than 500 words→ You know better about you, rather than ChatGPT!!). Generate the summary of your portfolio in
   a. 50% of the size of your portfolio
   b. 25% of the size of your portfolio