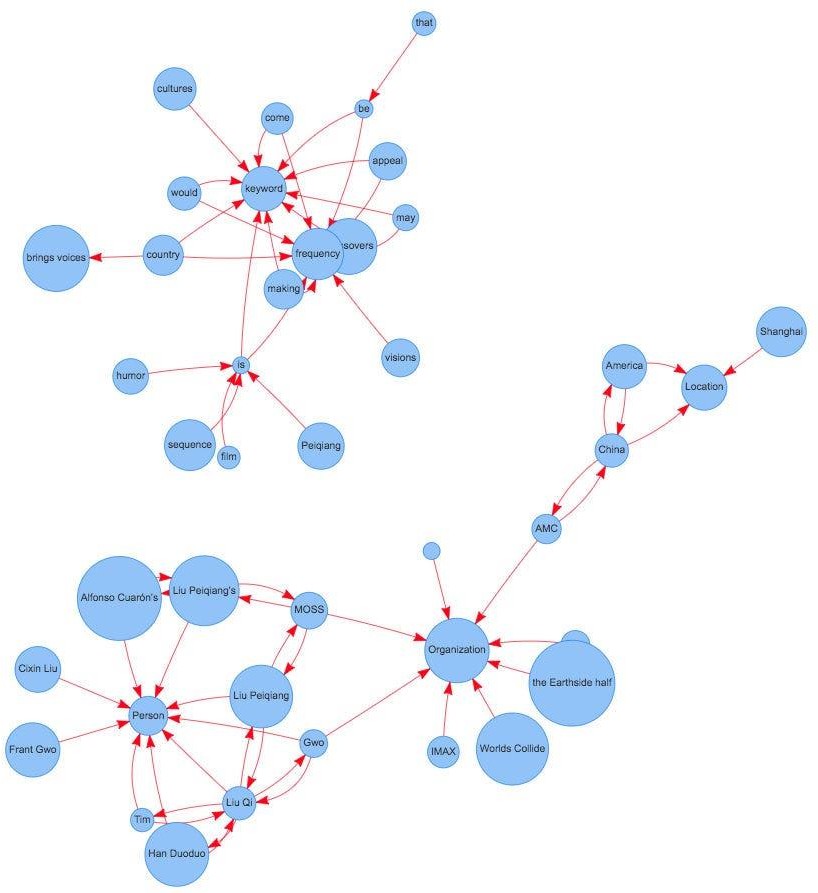
**Aim:** TextRank for Document Summarization

**IDE:** Google Colab

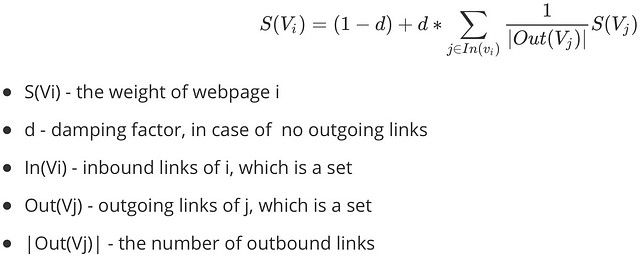
# Theory:

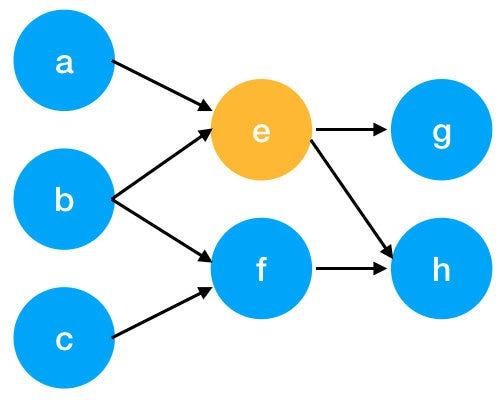
TextRank is an algorithm based on PageRank, which often used in keyword extraction and text summarization. In this article, I will help you understand how TextRank works with a keyword extraction example and show the implementation by Python.

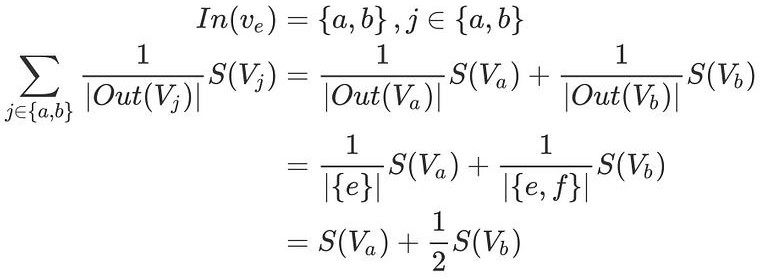


## Understand PageRank

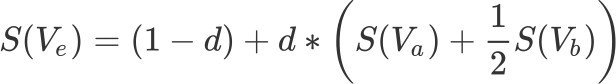
There are tons of articles talking about PageRank, so I just give a brief introduction to PageRank. This will help us understand TextRank later because it is based on PageRank. PageRank (PR) is an algorithm used to calculate the weight for web pages. We can take all web pages as a big directed graph. 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.





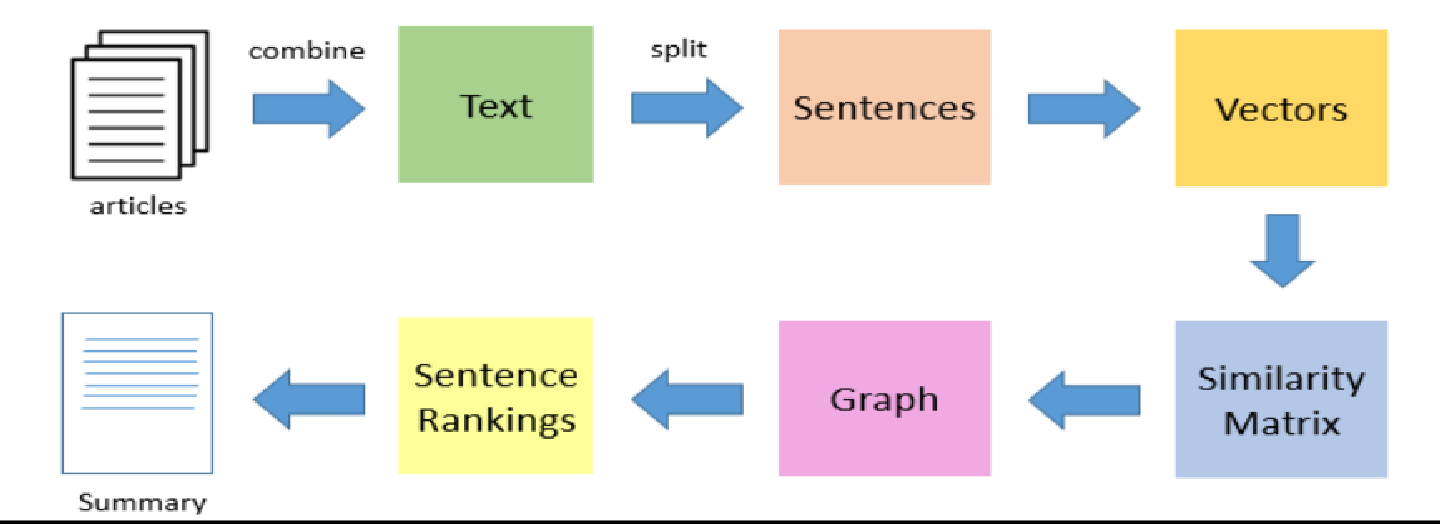
Here is an example to better understand the notation above. We have a graph to represent how web pages link to each other. Each node represents a webpage, and the arrows represent edges. We want to get the weight of webpage e. We can rewrite the summation part in the above function to a simpler version.

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



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.

# Pre Lab Exercise:

* 1. How keyword extraction is similar to document summarization process?
  2. How keyword extraction is different to document summarization process?
  3. Limitation of TextRank as a summarizer

# Program (Code):

To be attached with

# Results:

To be attached with

# Observation:

**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
   1. 50% of the size of your portfolio
   2. 25% of the size of your portfolio

Comment over the summary obtained. Which sentence you think should be there in your summary, but was not spotted by textrank? Rate the overall summary in the scale of 1-5 (with 1 as least and 5 as highest rating). Paste the code, your portfolio, output and your analysis.