Text Mining

Emanuele Guidotti

2019/2020

Document summarization

Basically, we can regard the "summarization" as the "function" its input is document and output is summary.

There are mainly two ways to make the summary. Extractive and Abstractive.

- **Extractive**: It involves the selection of phrases and sentences from the source document to make up the new summary.
- **Abstractive**: It involves generating entirely new phrases and sentences to capture the meaning of the source document.

Extractive & Abstractive is not conflicting ways. You can use both to generate the summary. Here we cover extractive methods and, in particular, the TextRank algorithm.

TextRank

TextRank

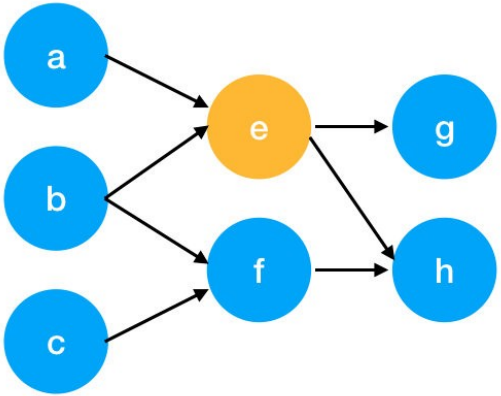
TextRank is an algorithm based on **PageRank**, which is often used in keyword extraction and text summarization.

PageRank is the first and best known algorithm used by Google to rank web pages in their search engine results. PageRank was named after Larry Page. PageRank is a way of measuring the importance of website pages. According to Google:

"*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.*"

PageRank

PageRank 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 e, it can be represented as a directed edge from a to e.



We can use a matrix to represent the inbound and outbound links among `a`, `b`, `e`, `f` in the graph.

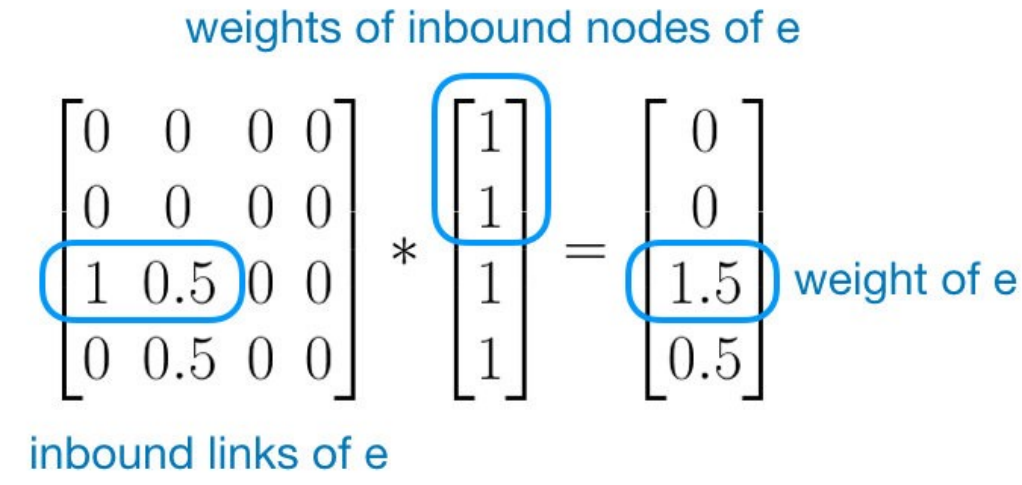
	a	**b**	**e**	**f**
a	0	0	0	0
b	0	0	0	0
e	1	1	0	0
f	0	1	0	0

Each node in a row means the inbound links from other nodes. For example, for the `e` row, node `a` and `b` have outbound links to node `e`.

To account for pages with many outbound links, we should normalize each column.

	a	**b**	**e**	**f**
a	0	0	0	0
b	0	0	0	0
e	1	0.5	0	0
f	0	0.5	0	0

We use this matrix to multiply with the weight of all nodes. In the initialization, the importance of each node is 1. We need to run this iteration many times to get the final weights.



Back to TextRank

What is the difference between TextRank and PageTank?

The simple answer is PageRank is for webpage ranking, and TextRank is for text ranking. The webpage in PageRank is the text in TextRank, so the basic idea is the same.

- Define nodes (words, sentences)
- Create a graph (link nodes)
- Calculate the weight for each node (PageRank)
- Extract the most important nodes (summary)

Keyword Extraction

- Each word is a node in PageRank
- Any two-word pairs in an *n*-gram are considered have a bidirectional link
- Based on this graph, calculate the weight for each word
- The most important words can be used as keywords

Sentence Extraction

- Tokenize words in each sentence
- Build a Similarity matrix (e.g. cosine similarity)
- Based on this graph, calculate the weight for each sentence
- The most important words can be used as summary

Take Home Concepts

Take Home Concepts

- Document summarization is the process of shortening a set of data computationally
- There are mainly two ways to make the summary. Extractive and Abstractive.
- TextRank is based on PageRank
- Keyword extraction with TextRank
- Sentence extraction with TextRank

References

References

- https://medium.com/@umerfarooq_26378/text-summarization-in-python-76c0a41f0dc4
  - (7 mins)
- https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0
  - read (7 mins)
- https://medium.com/analytics-vidhya/sentence-extraction-using-textrank-algorithm-7f5c8fd568cd
  - read (5 mins)