Summarization of News Articles Using TextRank

—

Amy Hemmeter
Artificial Intelligence Engineer, Interactions

GitHub Repo

Problem:

- You want to follow a keyword and get news updates about that keyword
- You don’t want to read the 20 articles that pop up, but you want to get a sense for what’s happening in the articles pertaining to your keyword

Summarization

- The solution is summarization!

Summarization

- The solution is summarization!
- The task: to get 2-3 sentences from each article that sum up the content of the article

Summarization

- The solution is summarization!
- The task: to get 2-3 sentences from each article that sum up the content of the article
- How do we determine which of the sentences are “important”?

Summarization

- The solution is summarization!
- The task: to get 2-3 sentences from each article that sum up the content of the article
- How do we determine which of the sentences are “important”?
- We use a tool called TextRank (Mihalcea and Tarau 2004)

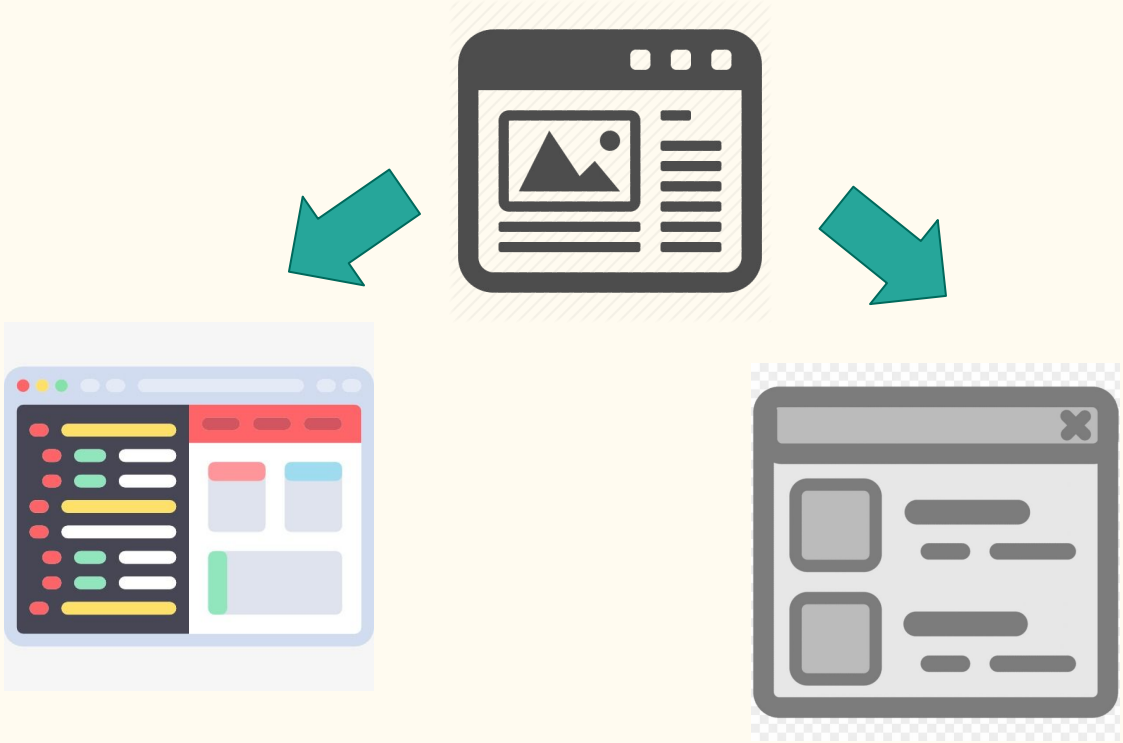
What is TextRank?

- Determines the top n most relevant sentences via a ranking system
- Draws its inspiration from Google's PageRank Algorithm

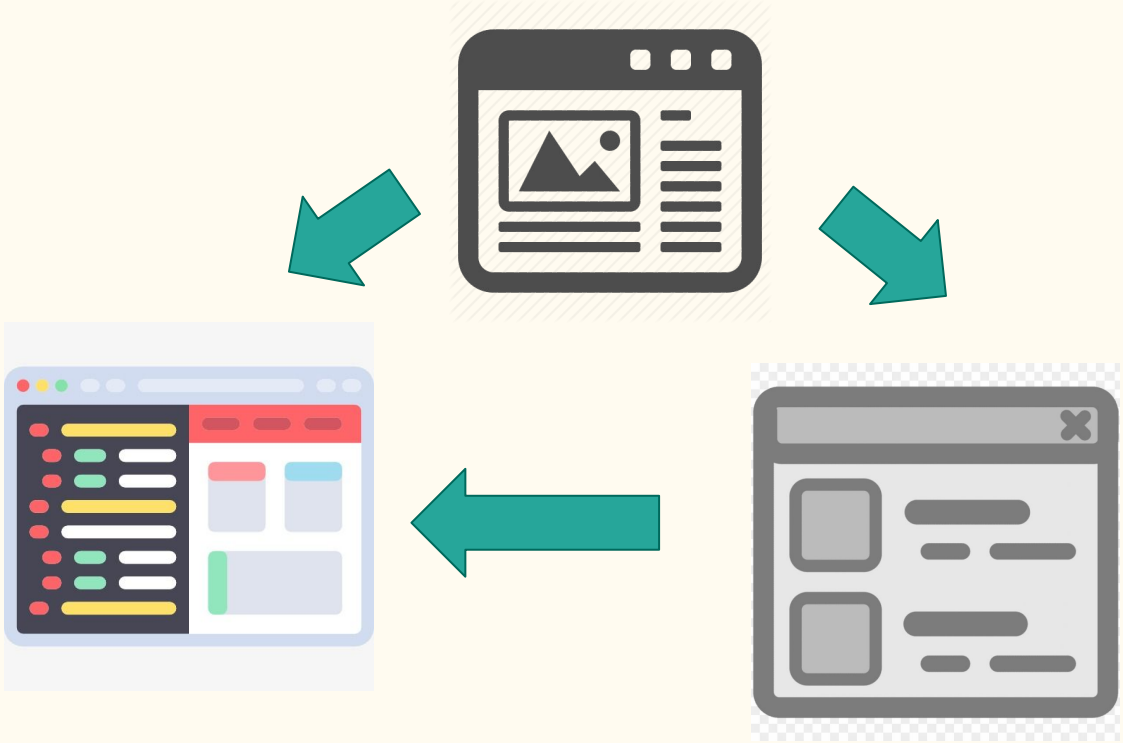
What is PageRank?

What is PageRank?

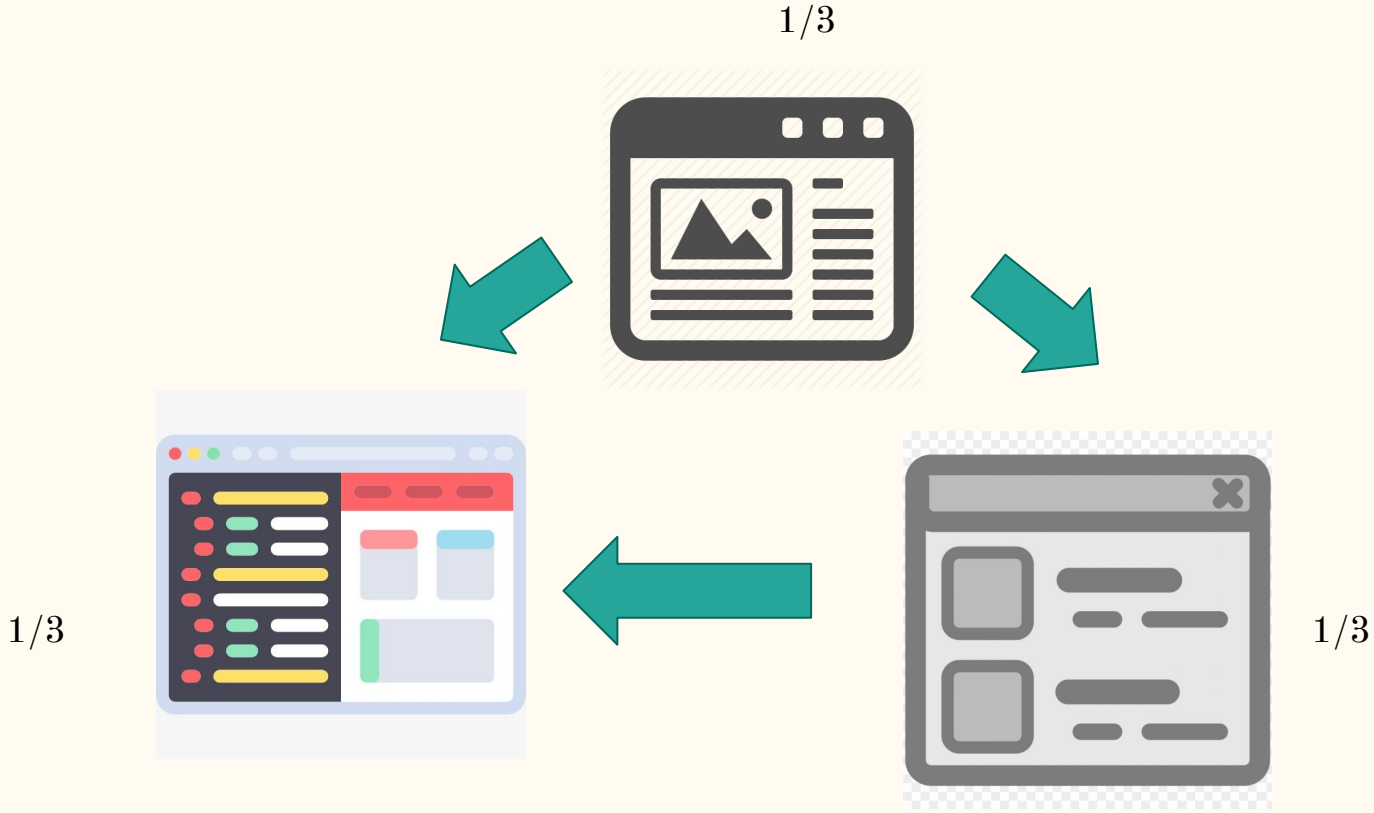
What is PageRank?



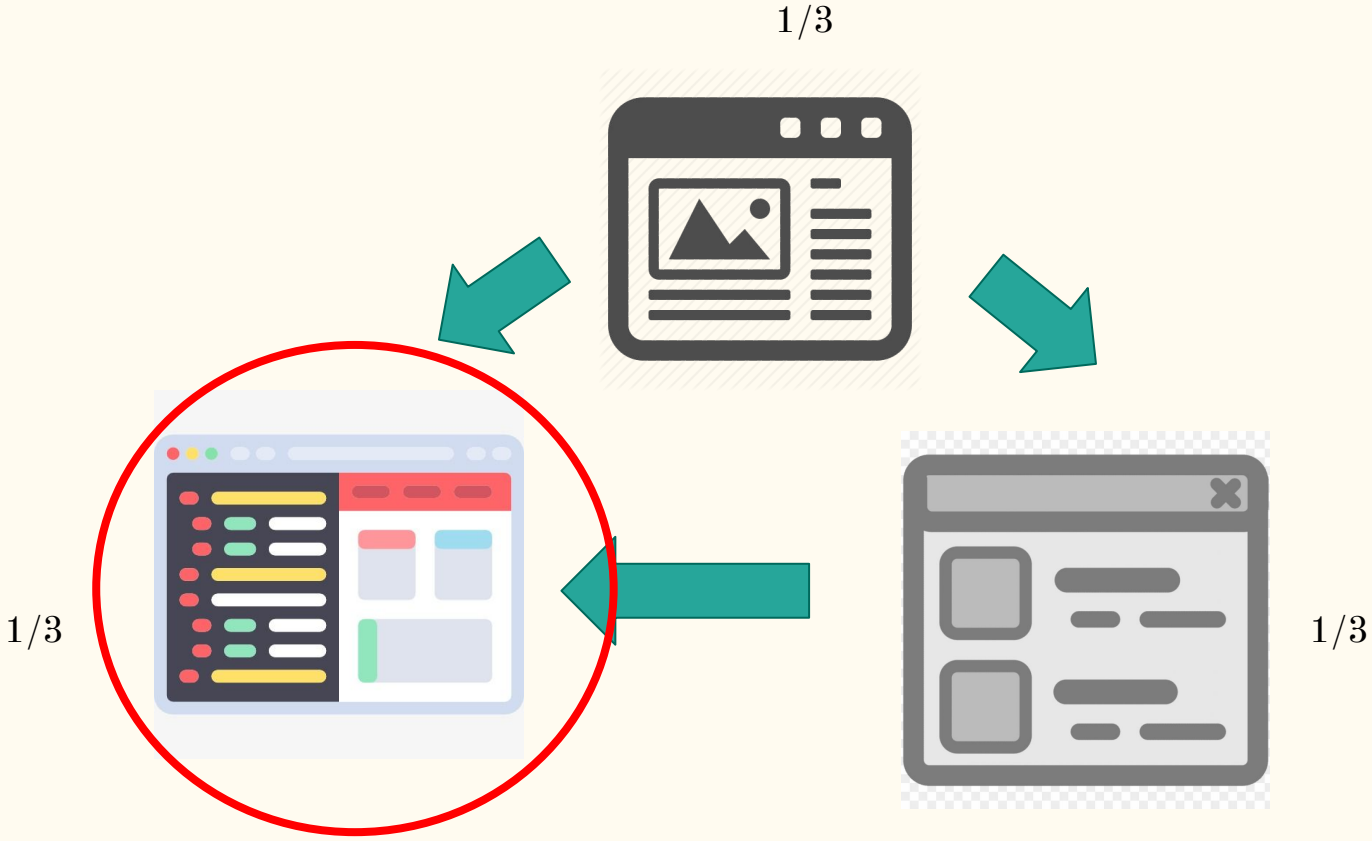
What is PageRank?



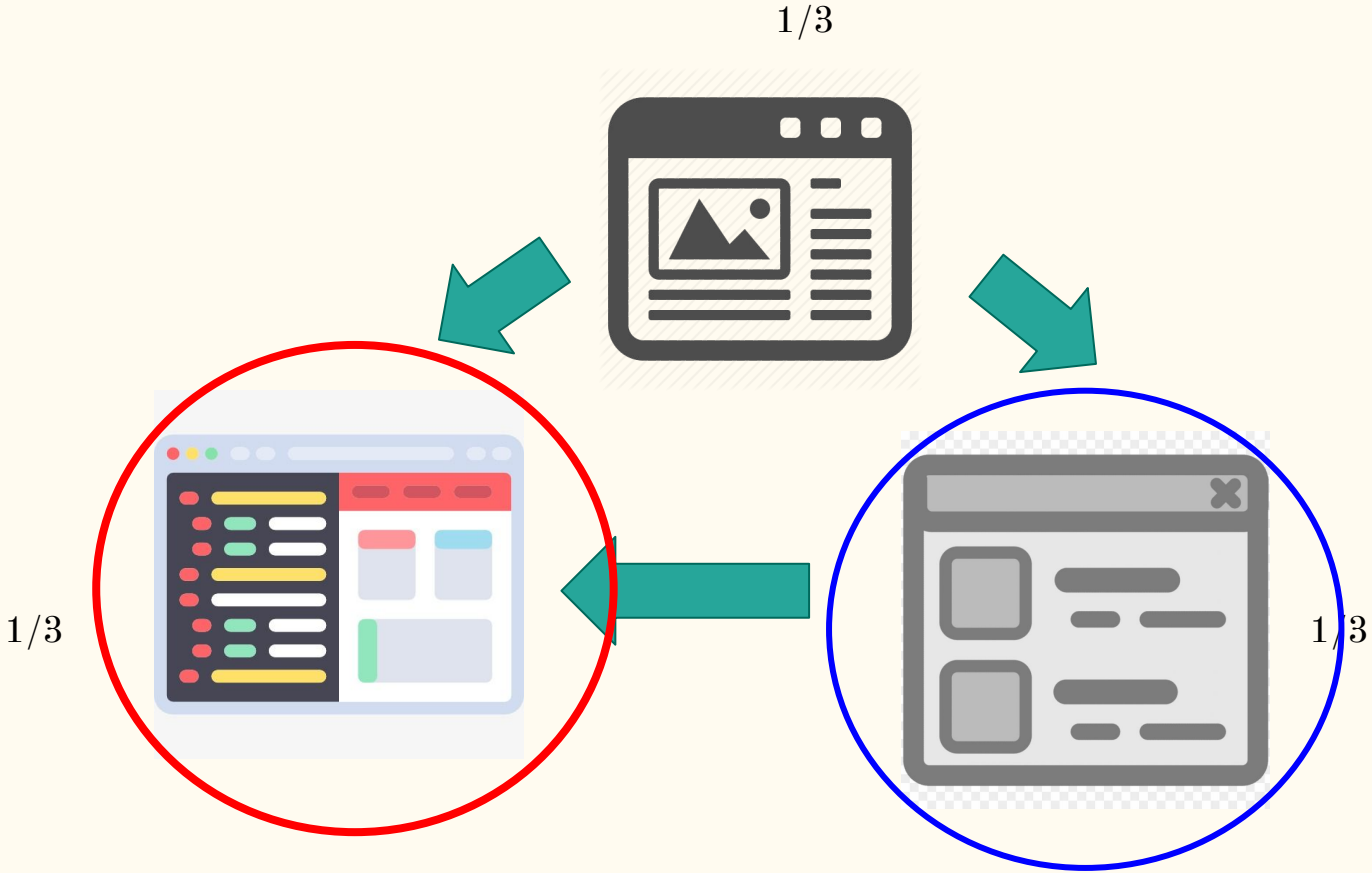
What is PageRank?



What is PageRank?



What is PageRank?



PageRank Update Formula

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

PageRank Update Formula

V_i is our target node

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

PageRank Update Formula

V_i is our target node

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

$In(V_i)$ is the set of vertices that point to V_i

PageRank Update Formula

V_i is our target node

V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

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

$In(V_i)$ is the set of vertices that point to V_i

PageRank Update Formula

V_i is our target node

V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

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

$In(V_i)$ is the set of vertices that point to V_i

This is the number of edges that are going out of Vj

PageRank Update Formula

V_i is our target node

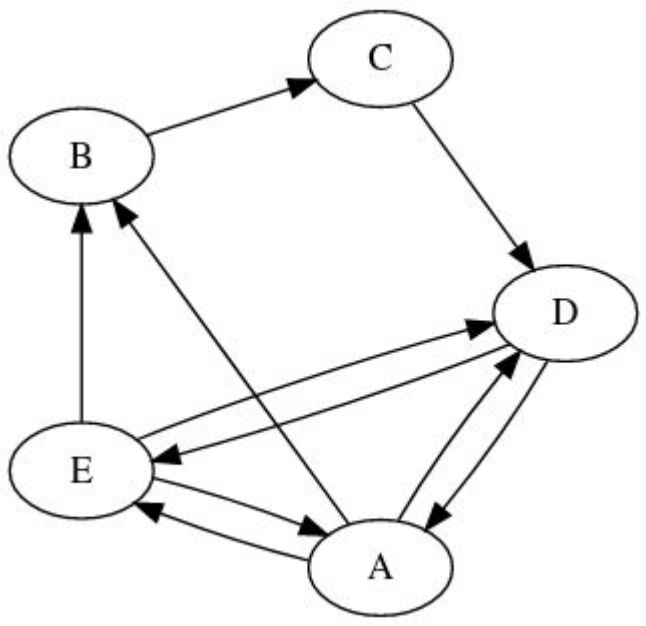
V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

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

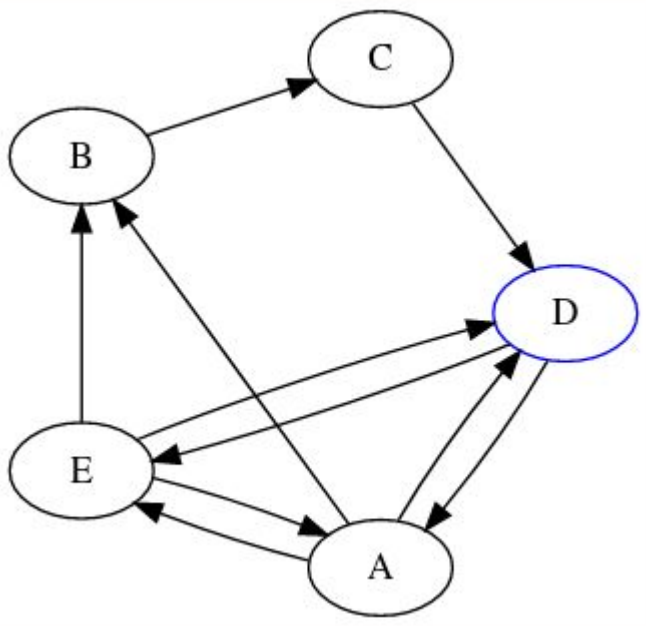
d is a damping factor (normally d=0.85)

$In(V_i)$ is the set of vertices that point to V_i

This is the number of edges that are going out of Vj

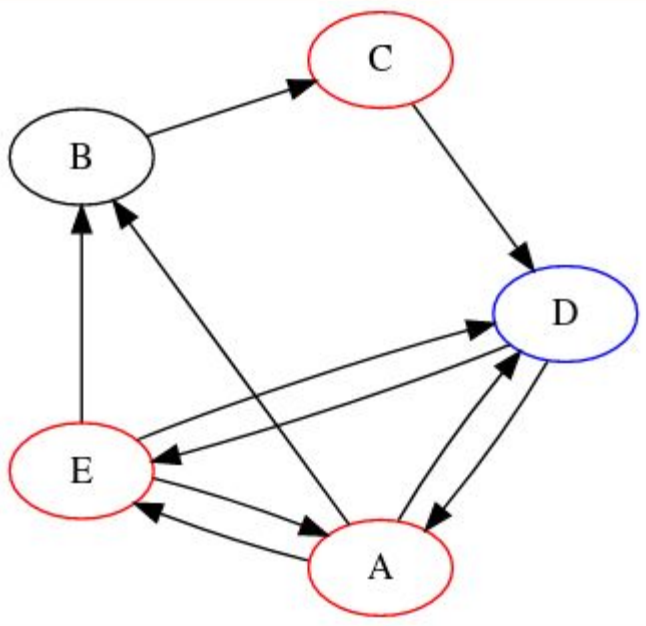


$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$



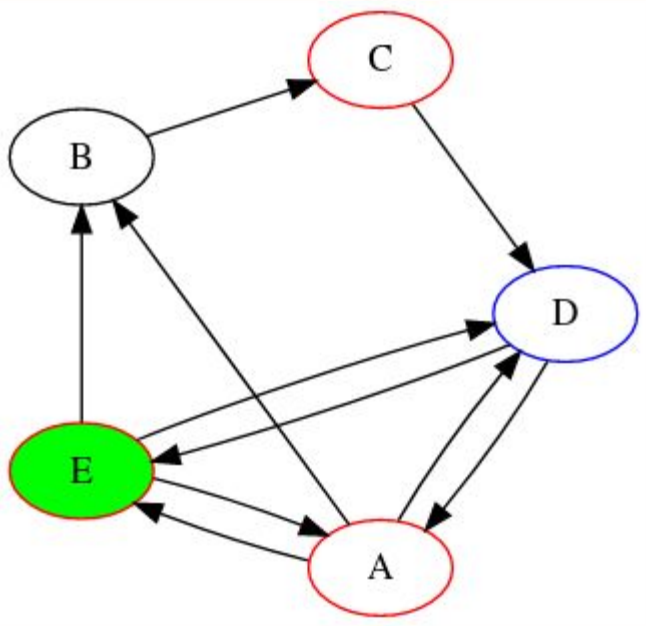
V_i is our target node

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



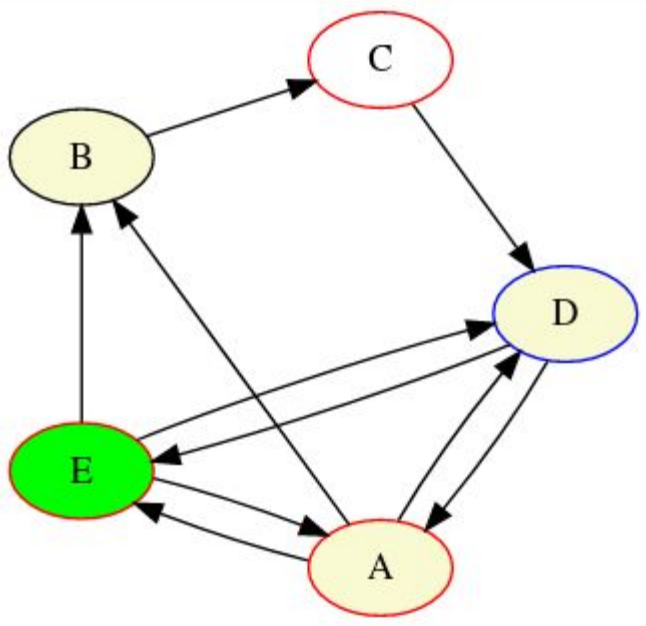
$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

$In(V_i)$ is the set of vertices that point to V_i



V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

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



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

This is the number of edges that are going out of Vj

Back to TextRank

- Also uses a graph representation where the edges “vote” for the various vertices

Back to TextRank

- Also uses a graph representation where the edges “vote” for the various vertices
- But there are no links in sentences -- what do we use for the sentences to recommend other sentences?

Back to TextRank

- Also uses a graph representation where the edges “vote” for the various vertices
- But there are no links in sentences -- what do we use for the sentences to recommend other sentences?
- We use similarity between sentences!

Back to TextRank

- Also uses a graph representation where the edges “vote” for the various vertices
- But there are no links in sentences -- what do we use for the sentences to recommend other sentences?
- We use similarity between sentences!
- Instead of a directed unweighted graph we use a weighted undirected graph

Similarity Metric

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Cras ut est a ipsum posuere auctor ac eget ex.

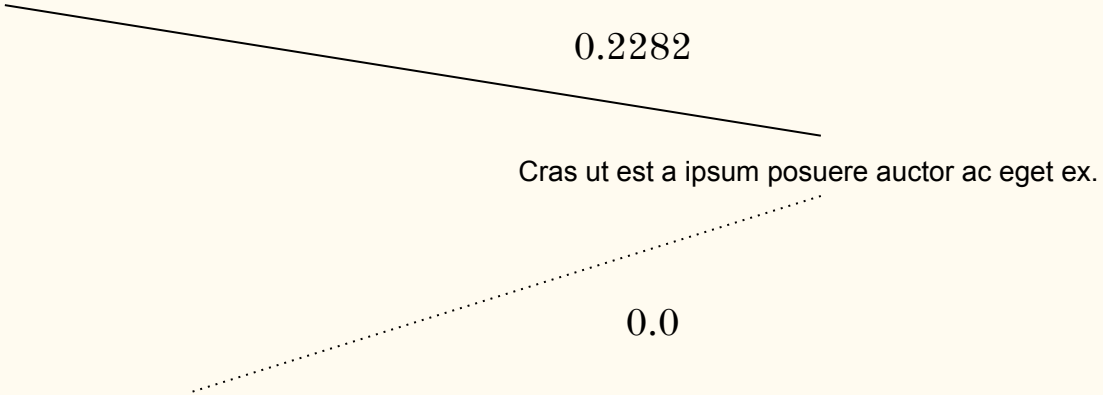
Morbi lacinia justo sit amet consectetur commodo.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

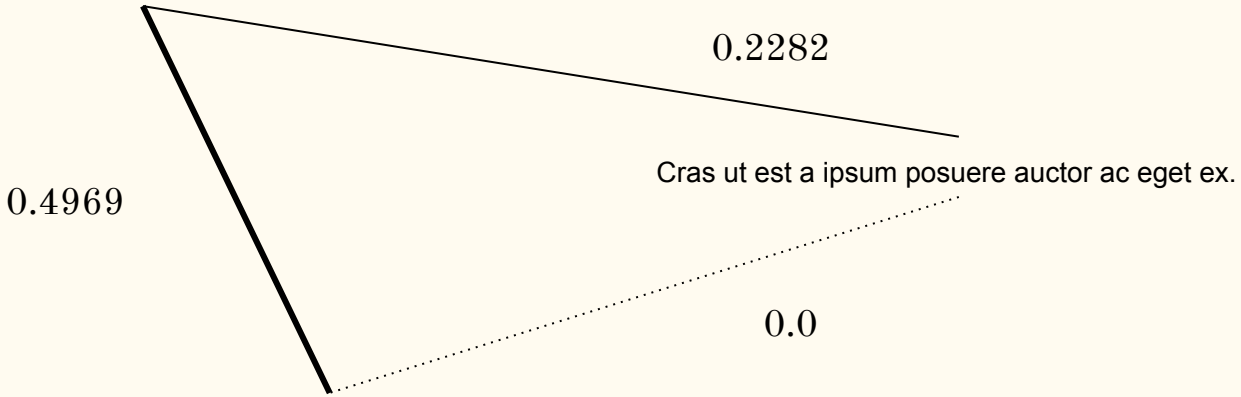
Cras ut est a ipsum posuere auctor ac eget ex.

Morbi lacinia justo sit amet consectetur commodo.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.



Lorem ipsum dolor sit amet, consectetur adipiscing elit.



Morbi lacinia justo sit amet consectetur commodo.

What is convergence in this context?

- When the error rate for any vertex falls below a certain threshold
- Because this is an unsupervised task, that is calculated as the difference between the scores in successive iterations:

$$S^{k+1}(V_i) - S^{k}(V_i)$$

TextRank Update Formula

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

TextRank Update Formula

V_i is our target node

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

TextRank Update Formula

V_i is our target node

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

$In(V_i)$ is the set of vertices that point to V_i

TextRank Update Formula

V_i is our target node

V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

$In(V_i)$ is the set of vertices that point to V_i

TextRank Update Formula

V_i is our target node

V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

$In(V_i)$ is the set of vertices that point to V_i

This is the sum of weights for all edges that are going out of Vj

TextRank Update Formula

V_i is our target node

V_j is one of the nodes in $In(V_i)$, our source node for this part of the update

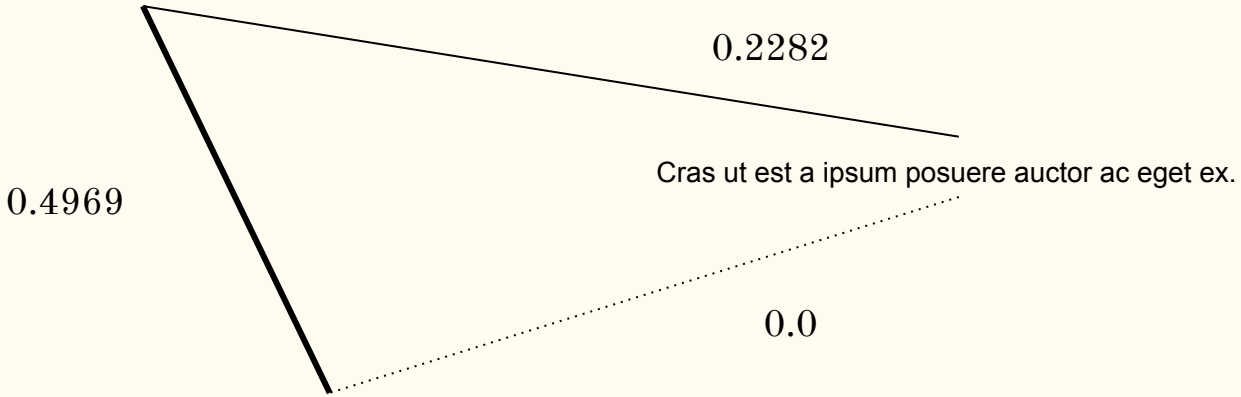
$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

d is a damping factor (normally d=0.85)

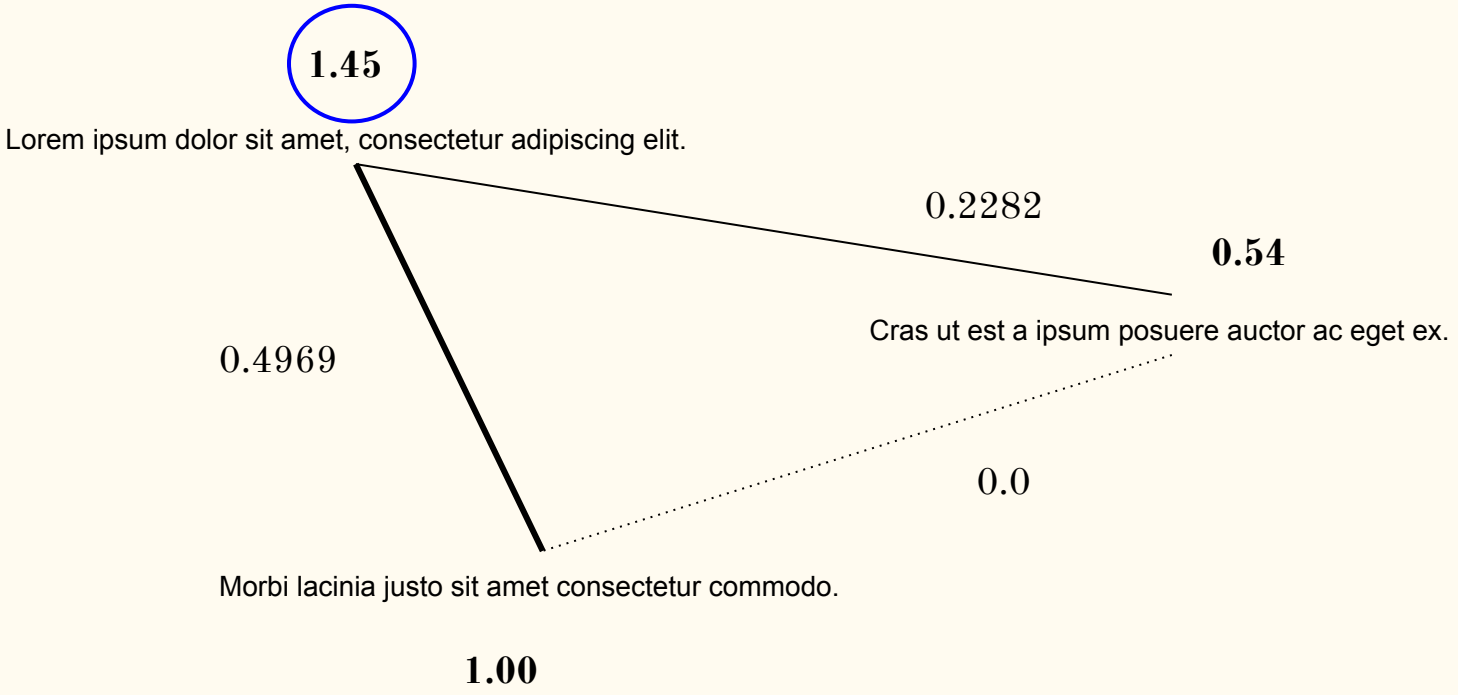
$In(V_i)$ is the set of vertices that point to V_i

This is the sum of weights for all edges that are going out of Vj

Lorem ipsum dolor sit amet, consectetur adipiscing elit.



Morbi lacinia justo sit amet consectetur commodo.


1.45
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
0.2282
0.54
Cras ut est a ipsum posuere auctor ac eget ex.
0.4969
0.0
Morbi lacinia justo sit amet consectetur commodo.
1.00

Now that we have some of the basic concepts, let's look at some code!*

* Adapted from Brian Lester's implementation of TextRank: https://github.com/blester125/text-rank

Summary #1

Volkswagen is partnering with the University of Tennessee and the Oak Ridge National Laboratory to create the company's first innovation hub for developing new technology in North America, officials said Friday in a statement. "Working with the University of Tennessee and Oak Ridge National Laboratory is a great opportunity to continue growing Volkswagen's engineering footprint in the North American region," said Wolfgang Demmelbauer-Ebner, VW's executive vice president and chief engineering officer for the region.

Summary #2

“4 Min Read MANILA (Reuters) - Schools and businesses shut across the Philippine capital on Monday as a volcano belched clouds of ash across the city and seismologists warned an eruption could happen at any time, potentially triggering a tsunami. Thousands of people were forced to evacuate their homes around Taal, one of the world’s smallest active volcanoes, which spewed ash for a second day from its crater in the middle of a lake about 70 km (45 miles) south of central Manila.”

Summary #3

“"In every country where we have a business presence, we're committed to complying with the applicable laws and regulations, and Canada is no different," Howes said. “Regardless of the decision, our position is we will continue to serve our customers in Canada with our products and networks,” Howes said.

Comparing the same story

4 Min Read MANILA (Reuters) - Schools and businesses shut across the Philippine capital on Monday as a volcano belched clouds of ash across the city and seismologists warned an eruption could happen at any time, potentially triggering a tsunami. Thousands of people were forced to evacuate their homes around Taal, one of the world's smallest active volcanoes, which spewed ash for a second day from its crater in the middle of a lake about 70 km (45 miles) south of central Manila.

"We prayed that we can rise up, put a stop to this calamity to allow us to return back to our homes," said 44-year-old evacuee Annie Villanueva. "A lot of families like us want to be together in our own homes and stand up." More than 70,000 people have been evacuated since the Taal, one of the Philippines' most active volcanoes, began spewing clouds of ash, steam and gas on Jan. 12.

Other details

- 33% of summaries contain first sentence

Other details

- 33% of summaries contain first sentence
- 17% of summaries contained the first sentence as the highest-ranked sentence

Other details

- 33% of summaries contain first sentence
- 17% of summaries contained the first sentence as the highest-ranked sentence
- Only 2 of 58 summaries had summary sentences shorter than the average sentence length for the data

Conclusion

Overall, summaries are pretty good, and surprisingly coherent for a fairly simple algorithm

Questions?