TextRank : Bringing Order into Texts

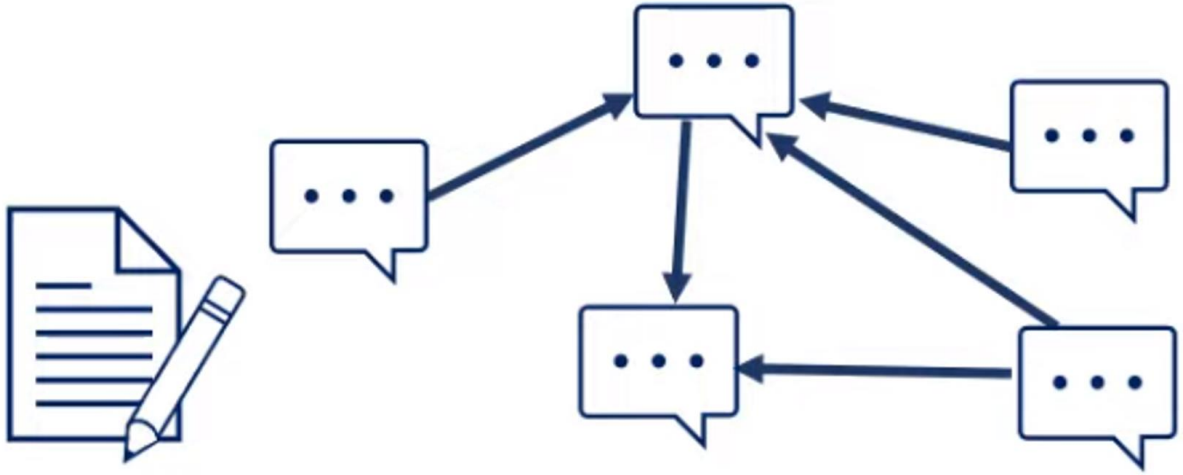
Rada Mihalcea and Paul Tarau (EMNLP, 2004)

모바일시스템공학과 32192530 양윤성

Contents

1-(1). Introduction

What is TextRank?

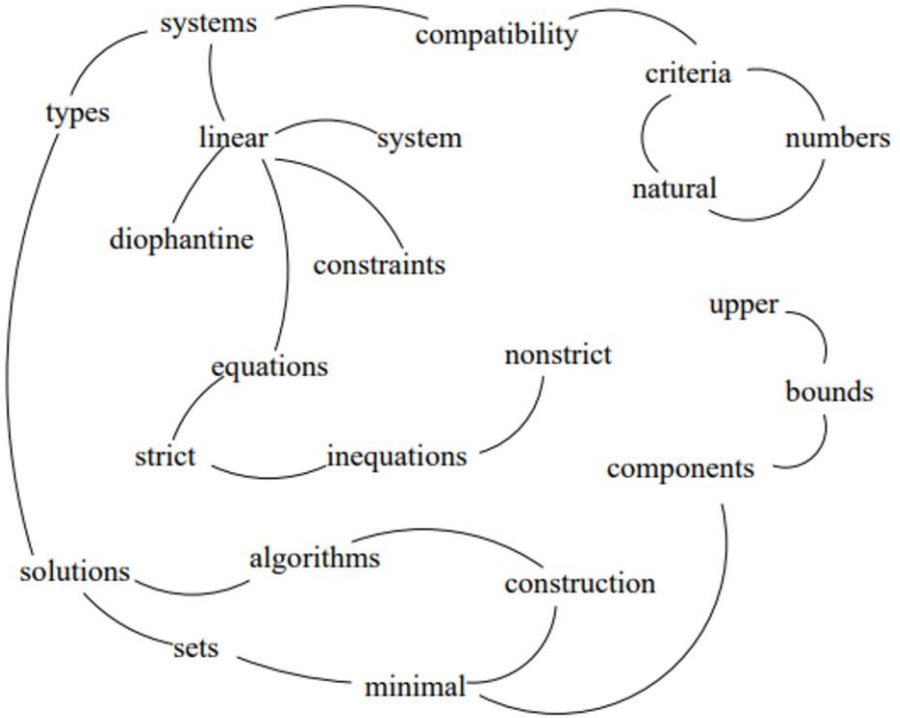


- Text-oriented graph-based ranking methods
- Find text-units relationship and represent it as graph
- Unsupervised Learning → No learning process, Depend only on the given text information

1-(2). TextRank Tasks

1. Keyword Extraction

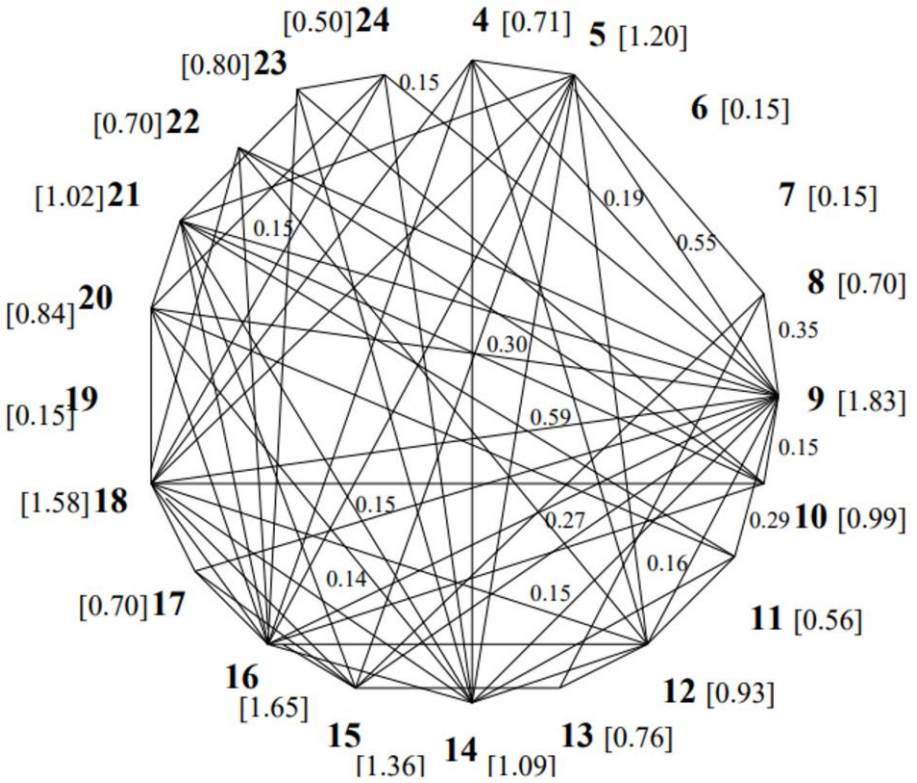
- Graph node : Each word in sentence
- Create edges using the connection of words
- Use to extract keywords in a document



1-(2). TextRank Tasks

2. Sentence Extraction

- Graph node : Sentences
- Create edges using the connection of sentences
- Use to summarize document



Contents

2-(1). Step Information

Graph Construction

- STEP 1) Basic Pre-Processing
- STEP 2) Syntactic Filtering (Create Node)
- STEP 3) Create Edge

Ranking

- STEP 4) Recursive Importance Score Calculation
- STEP 5) Selection of the Top N Words

2-(2). Basic Pre-Processing

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

Tokenizing

- Separating sentence into word(token) units
- Except prepositions and articles
- Can use no-stopwords → Remove certain words

compatibility, systems, linear, constraints, natural, numbers, criteria, system, Diophantine

2-(2). Basic Pre-Processing

Part of speech tagging

- After tokenization
- Checking each token's part of speech and tagging

compatibility, systems, linear, constraints, natural, numbers, criteria, system, Diophantine

compatibility, systems, constraints, numbers, criteria, system : NN(명사)

Linear, natural, Diophantine : JJ(형용사)

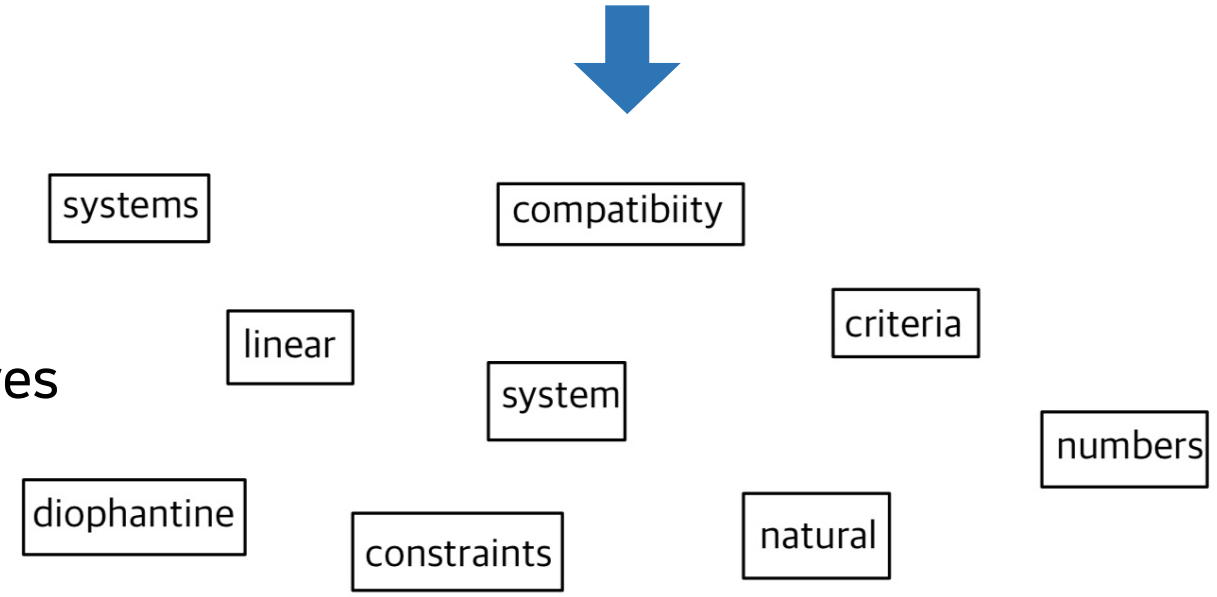
2-(3). Syntactic Filtering

Make nodes

- Save only specific part of speech and remove the rest
- Prevent excessive graph complexity
- Best result : Save only nouns and adjectives

compatibility, systems, constraints, numbers, criteria, system : NN(명사)

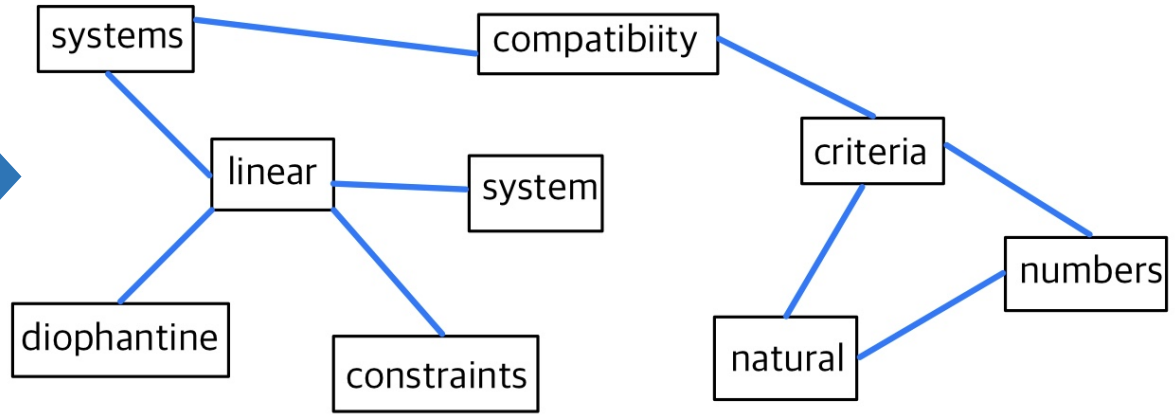
Linear, natural, Diophantine : JJ(형용사)



2-(4). Create Edge

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for

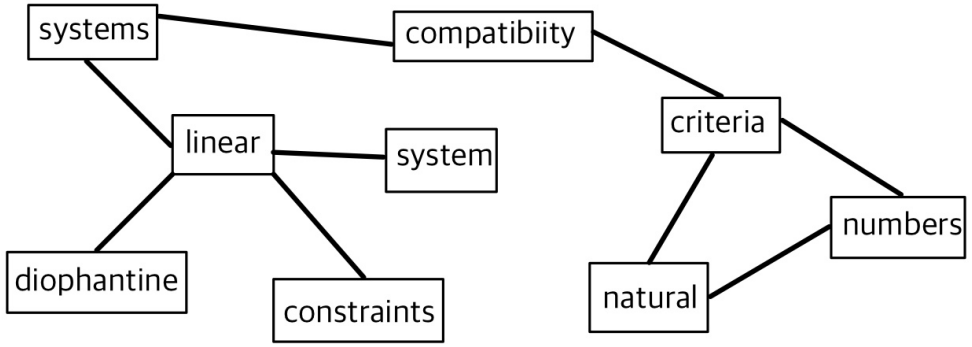
compatibility, systems, linear, constraints, natural, numbers, criteria, system, diophantine



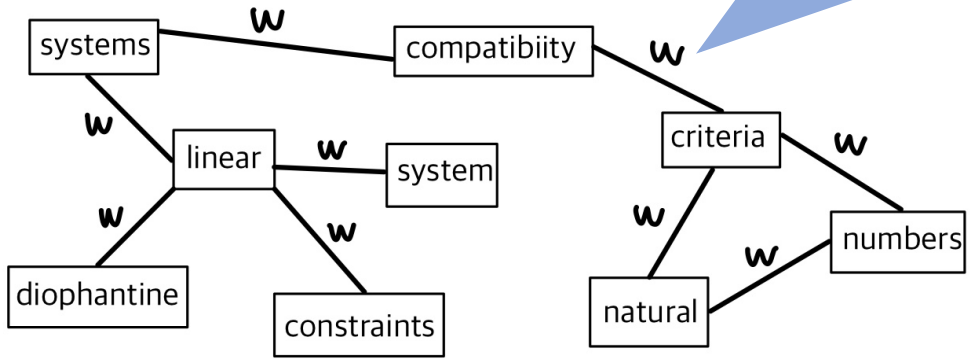
- Connecting nodes with co-occurrence relationships
- Co-occurrence : Words that appear within window size N (In paper, N : 2~10)

2-(4). Create Edge

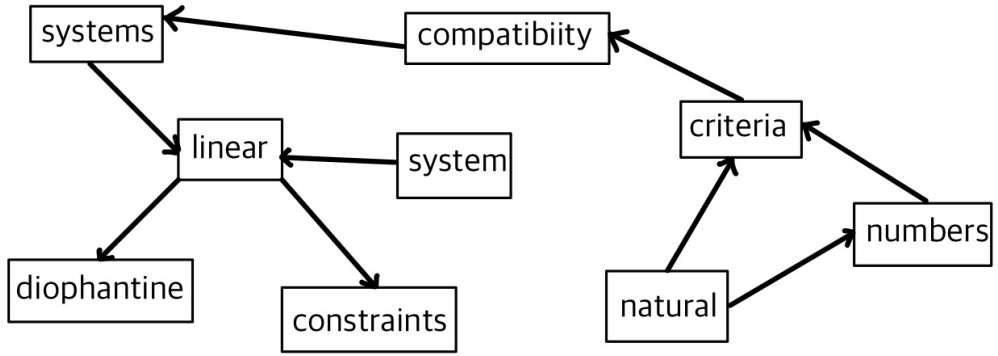
In keyword extraction, weight is random value.



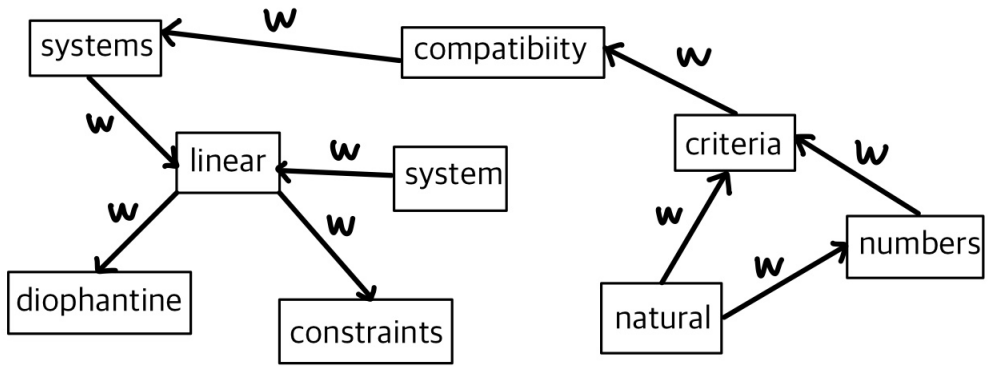
Undirected, Unweighted



Undirected, Weighted



Directed, Unweighted



Directed, Weighted

2-(5). Recursive Importance Score Calculation

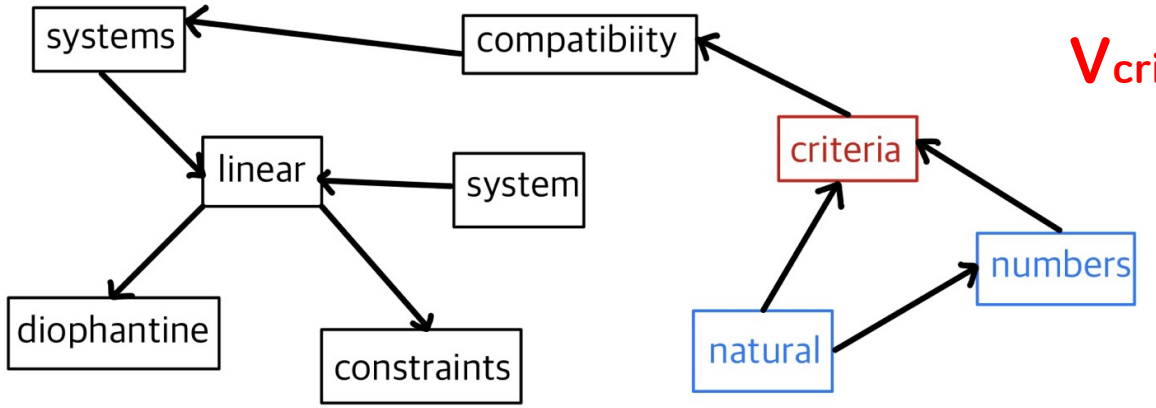
Unweight Case

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

- Recursion termination condition : $S^{k+1}(V_i) - S^k(V_i) < threshold$ (Usually 20 to 30 times)
- S(V_i) : node i 's importance score → Initial value is 1 or random
- d : Damping factor (Same as d in page rank) → Setting 0.85
- In(V_i) : Set of nodes entering node i
- Out(V_j) : Set of nodes exiting node j

2-(5). Recursive Importance Score Calculation

Unweight Case Example



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

$$= (1-d) + d * \frac{S(V_{natural})}{|Out(V_{natural})|} + \frac{S(V_{numbers})}{|Out(V_{numbers})|}$$

- In($V_{criteria}$) = {$V_{natural}$, $V_{numbers}$}
- Out($V_{natural}$) = {$V_{criteria}$, $V_{numbers}$} → |Out($V_{natural}$)| = 2
- Out($V_{numbers}$) = {$V_{criteria}$} → |Out($V_{natural}$)| = 1

2-(5). Recursive Importance Score Calculation

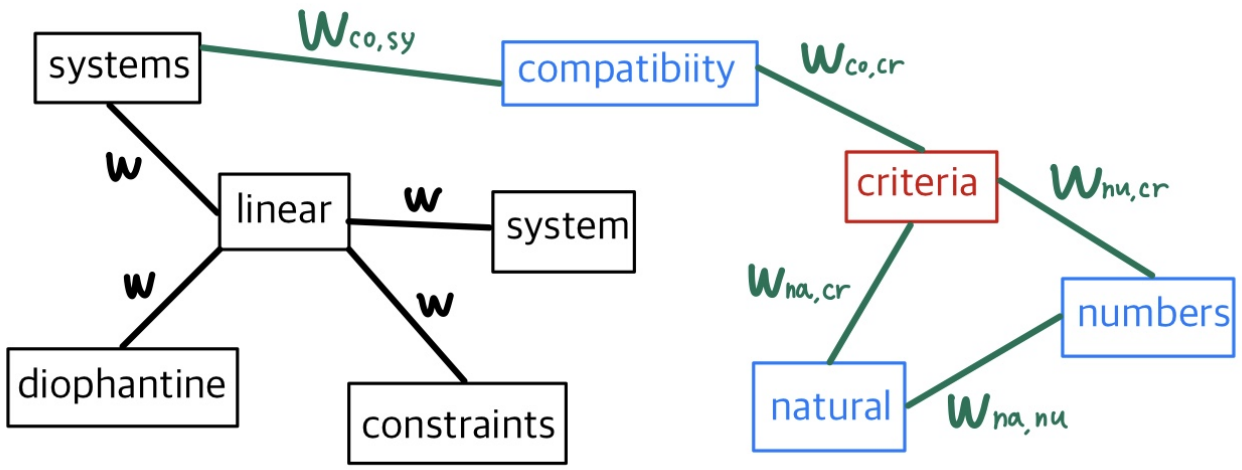
Weight Case

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

- Similar to the case of unweight, but
  - ✓ Numerator : Multiply node j 's importance score by the weight from node j to i
  - ✓ Denominator : For all nodes k belonging to Out(Vj), sum all weights from node k to j

2-(5). Recursive Importance Score Calculation

Weight Case Example



- In(Vcriteria) = {Vnatural, Vnumbers, Vcompatibility}
- Out(Vnatural) = {Vcriteria, Vnumbers} → Denominator : Wna,cr + Wna,cu
- Out(Vnumbers) = {Vcriteria, Vnatural} → Denominator : Wnu,cr + Wnu,na
- Out(Vcompatibility) = {Vcriteria, Vsystems} → Denominator : Wco,cr + Wco,sy

$$V_{criteria} = (1-d) + d * \left(\frac{w_{na,cr}}{w_{na,cr}+w_{na,nu}} WS(V_{\text{natural}}) + \frac{w_{nu,cr}}{w_{nu,cr}+w_{nu,na}} WS(V_{\text{numbers}}) + \frac{w_{co,cr}}{w_{co,cr}+w_{co,sy}} WS(V_{\text{compatibility}}) \right)$$

2-(6). Selection of the Top N Words

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

단어	점수
numbers	1.46
inequations	1.25
linear	1.29
:	:

Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; non−strict inequations; set of natural numbers; strict inequations; upper bounds

- Select N words in order of highest importance score
- If adjacent words are selected together in Top N,
  combine them into one to create keyword

Contents

3-(1). Step Information

STEP 1) Graph Construction

STEP 2) Recursive Importance Score Calculation

STEP 3) Make summary to use Top N sentences

3-(2). Graph Construction

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

	1	2	3	⋯	23	24
1	1	0.43	2	⋯	0	0.1
2		1	0.2	⋯	0	0
3						⋮
⋮				⋱		⋮
23						
24						1

- No need pre-processing in sentence extraction → All sentence are nodes.
- Edge connection : Measure how many words overlap in two sentences
  - ❖ Log in denominator : Normalization factor
- In similarity matrix, create edge between nodes above min_sim(criteria)
- Weight : Set in proportion to similarity between nodes
- Direction : Set in order of sentences in the document

3-(3). Recursive Importance Score Calculation

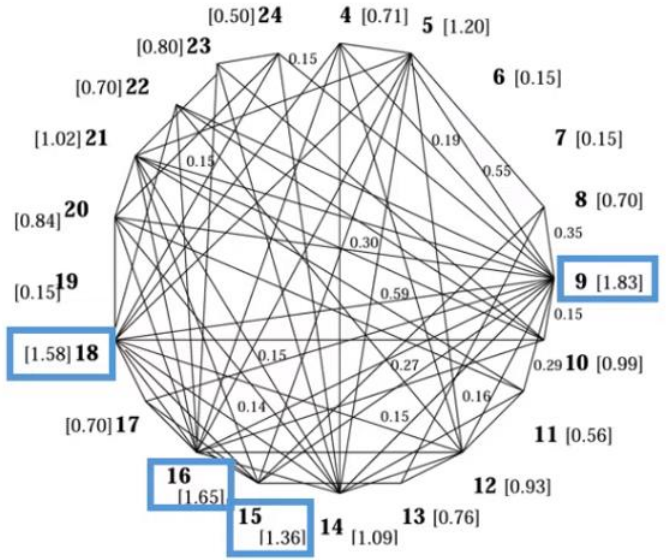
$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

- In sentence extraction, only weight importance score is used.
- Other is same as keyword extraction.

3-(4). Make summary to use Top N sentences

Document

3: BC-HurricaineGilbert, 09-11 339
4: BC-Hurricane Gilbert, 0348
5: Hurricane Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricane Gilbert Swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14: Tropical storm Gilbert formed in the eastern Carribean and strenghtened into a hurricane Saturday night.
15: The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17: The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.



문장 번호	점수
9	1.83
16	1.65
18	1.58
15	1.36
:	:

Summarization by TextRank:

(9) Hurricaine Gilbert Swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas. (15) The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. (16) The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather“ rotating around the center of the storm. (18) Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico’s south coast.

- Select N sentences in order of highest importance score and make summary

Contents

4-(1). Keyword Extraction

- **Evaluation Metric**
  - Precision : Percentage of keyword extracted through TextRank that matches the human summary (Human summary = Assigned)
  - Recall : Percentage of keyword in the human summary that matches keyword extracted through TextRank
  - F-measuere : Coordinated average of Precision and Recall ($2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$)

4-(1). Keyword Extraction

Method	Assigned		Correct		Precision	Recall	F-measure
	Total	Mean	Total	Mean			
TextRank							
Undirected, Co-occ.window=2	6,784	13.7	2,116	4.2	**31.2**	43.1	**36.2**
Undirected, Co-occ.window=3	6,715	13.4	1,897	3.8	28.2	38.6	32.6
Undirected, Co-occ.window=5	6,558	13.1	1,851	3.7	28.2	37.7	32.2
Undirected, Co-occ.window=10	6,570	13.1	1,846	3.7	28.1	37.6	32.2
Directed, forward, Co-occ.window=2	6,662	13.3	2,081	4.1	31.2	42.3	35.9
Directed, backward, Co-occ.window=2	6,636	13.3	2,082	4.1	31.2	42.3	35.9
Hulth (2003)							
Ngram with tag	7,815	15.6	1,973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4,788	9.6	1,421	2.8	29.7	37.2	33.0
Pattern with tag	7,012	14.0	1,523	3.1	21.7	39.9	28.1

Table 1: Results for automatic keyword extraction using TextRank or supervised learning (Hulth, 2003)

- Undirected & window size 2 method is the best.
- Increase window size → Performance decrease
  - Words located far away are less relevant to each other.
- Directed graph < Undirected graph
  - More accurate by comparing similarity in both directions.

4-(2). Sentence Extraction

System	ROUGE score – Ngram(1,1)		
	basic (a)	stemmed (b)	stemmed no-stopwords (c)
S27	0.4814	0.5011	0.4405
S31	0.4715	0.4914	0.4160
TextRank	**0.4708**	**0.4904**	**0.4229**
S28	0.4703	0.4890	0.4346
S21	0.4683	0.4869	0.4222
Baseline	*0.4599*	*0.4779*	*0.4162*
S29	0.4502	0.4681	0.4019

Table 2: Results for single document summarization: TextRank, top 5 (out of 15) DUC 2002 systems, and baseline. Evaluation takes into account (a) all words; (b) stemmed words; (c) stemmed words, and no stop-words.

- Rogue Score : Calculate n-gram by comparing model summary with human summary
  - Example) Model : I am really hungry / Human : I am very hungry
  - 1-gram(uni-gram) : 3/4, 2-gram(bi-gram) : 1/3
- TextRank is 3rd place in performance

Thank You!

TextRank : Bringing Order into Texts