Unsupervised Keyword Extraction from Single Document

Swagata Duari
Aditya Gupta
Vasudha Bhatnagar

Presentation Outline

- Introduction and Motivation
- Statistical Methods for Automatic Keyword Extraction
- Graph-based Methods for Automatic Keyword Extraction
- Hands-on/Demo Session
- Conclusion

NATIONAL POLICY ON EDUCATION

(WITH MODIFICATIONS UNDERTAKEN IN 1992)

PART-I

INTRODUCTORY

1.1 Education has continued to evolve, diversify and extend its re
dawn of human history. Every country develops its system of educ
its unique socio-cultural identity and also to meet the challeng
moments in history when a new direction has to be given to an ag
today.

1.2 The country has reached a stage in its economic and tec
effort must be made to derive the maximum benefit from
ensure that the fruits of change reach all sections. Ed

South African Women Unite For A Non-Sexist Post-Apartheid South Africa!

For the first time women from different organisations all over South Africa and South African women from the African National Congress (ANC) outside of the country, met together for a conference. The conference was called 'Malibongwe,' and the theme was 'Women united for a unitary, non-racial

women. South African women visited Dutch women's organisations and organisations which contributed funds for Malibongwe from all over Holland. On these visits Dutch and South African women shared experiences. South African women were moved to feel the strong support Dutch people have for the struggle against apartheid. South African women called on Dutch people to increase international pressure to end apartheid and injustice in South Africa.

Women's Lives

During the the first week of Malibongwe there were many discussions. One of the discussions was about women and the media. Women spoke about the way the media only talks about rich and successful women. The media never says anything about the lives and problems of

@TheSageBrush | nevadasagebrush.com TUESDAY, SEPTEMBER 11, 2012

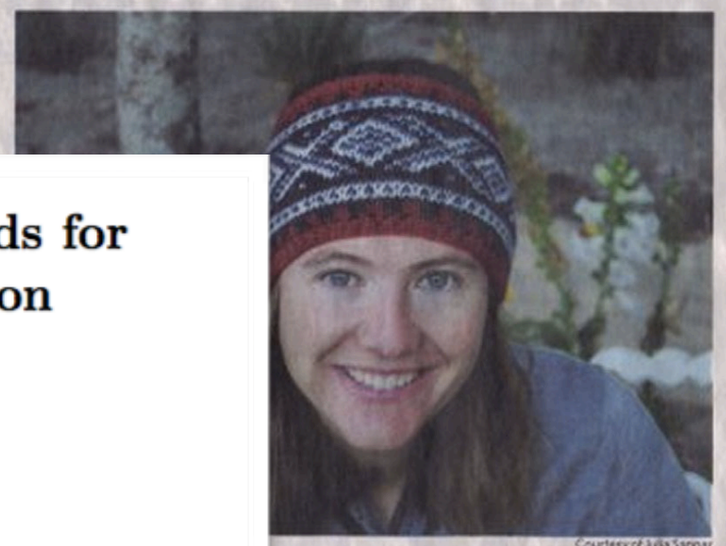
udent wins knitting scholarship

olly Moser

e for a student to ssory while simul- nning a scholarship or classes, and that's t 31-year-old Cas- en did.

"I wanted to put modern colors to it instead of just traditional reds, blues and whites," Hansen said.

When she graduated from high school, Hansen lived in Norway for two years at a private Norwegian school to major in



Courtesy of Julia Sannar

e winners of the Bears For Brains Scholarship for knitting a headband shown above is one of the many creations Hansen knitted.

Main Core Retention on Graph-of-words for Single-Document Keyword Extraction

François Rousseau and Michalis Vazirgiannis

LIX, École Polytechnique, France

Abstract. In this paper, we apply the concept of *k-core* on the *graph-of-words* representation of text for single-document keyword extraction, retaining only the nodes from the main core as representative terms. This approach takes better into account proximity between keywords and variability in the number of extracted keywords through the selection of more *cohesive* subsets of nodes than with existing graph-based approaches solely based on *centrality*. Experiments on two standard datasets show statistically significant improvements in F1-score and AUC of precision/recall curve compared to baseline results, in particular when weighting the edges of the graph with the number of co-occurrences. To the best of our knowledge, this is the first application of graph degeneracy to natural language processing and information retrieval.

Keywords: single-document keyword extraction; graph representation of text; weighted graph-of-words; k-core decomposition; degeneracy

www.cambridge-news.co.uk/news

al on Iraq

lisastrous

course that includes weapons of mass destruction, which demonstrates how stupid Saddam Hussein would be to use any he might have. Let us remember, also, that the Americans supported Saddam during his war against Iran and the Taliban government of Afghanistan when it was invaded by Russia.

It may well be argued that Iraq is in breach of United Nations resolutions. However, the same applies to Israel. But similar action has not been taken.

It is time we stopped letting our politicians pick our enemies for us then frightening us to kill people we have no knowledge of. Let them instead accept the resources of the world are there for all to be shared and the richest nations are in a position to do something about it.

Sterne Close

Blair hiding?

From Gus Jeevar

IF BLAIR knows that people's opinion on the war will shift, and that Saddam Hussein has weapons of mass destruction, then he has knowledge that he won't share and knows he can use it to manipulate people.

He says that disclosing his knowledge would reveal his source. I expect it would reveal that Saddam Hussein got much of his weaponry from rogue countries in the West, and rogue companies in England.

He might even fear it would reveal the West's links with Bin Laden which in the 1980s helped anti-Communist rebels in Afghanistan pave the way for

Children are being killed

From Dr Beatrice Boctor

WE DEFINITELY do not agree with Blair about war with Iraq.

While Israel flouts all UN resolutions, breaks every code in international law, and threatens its neighbours with nuclear annihilation, Blair and Bush find it quite acceptable.

But when 8,000 children die weekly for 12 years because of the weapons of mass destruction the U.S./UK used in Gulf War 1, the suffering Iraqi population is treated to an even faster genocide.

In the name of what god can Blair claim these things, or is the god he worships insatiably thirsty for oil?

Girton Road
Cambridge

No support

From E Bruce

I DO NOT support war in Iraq.

Lode Avenue
Waterbeach

Let's keep out of U.S. fight

The Courier-Mail

Wednesday September 12, 2001

•TRADE TOWERS COLLAPSE •PENTAGON ON FIRE •WHITE HOUSE EVAC

3AM WORLD SPECIAL

US ATTACKE

Suicide terrorists hit cities

Experience Vietnam and Angkor Wat

ple left in limbo

Rescue bid ruled out

3

Image sources: Google images

Introduction

- Keywords represent a small set of important and relevant terms
- Sufficiently describes a document
- Keywords are an everyday part of looking up topics and specific content.
- Huge number of digital documents do not have keywords assigned to them.

Problem: **Automatic extraction of keywords from text documents with minor or no human intervention.**

Automatic Keyword Extraction:Evolution

- Motivation back then …. (mid 50's – early 90's)
  - Construction of Book index
  - Indexing technical documents for efficient retrieval

- Motivation Post 90's…. (After the internet boom!..)
  - Indexing digital documents for search engines
  - Demand for higher levels of searching efficiency
  - Organizing massive repositories of all types
  - Classification of documents in news industry

Automatic Keyword Extraction:Evolution

- Motivation back
  - Construction of
  - Indexing techni
- Motivation Post
  - Indexing digital
  - Demand for hig
  - Organizing mas
  - Classification o

H. P. Luhn

A Statistical Approach to Mechanized Encoding and Searching of Literary Information*

Abstract: Written communication of ideas is carried out on the basis of statistical probability in that a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning. Since this process varies among individuals and since similar ideas are therefore relayed at different levels of specificity and by means of different words, the problem of literature searching by machines still presents major difficulties. A statistical approach to this problem will be outlined and the various steps of a system based on this approach will be described. Steps include the statistical analysis of a collection of documents in a field of interest, the establishment of a set of "notions" and the vocabulary by which they are expressed, the compilation of a thesaurus-type dictionary and index, the automatic encoding of documents by machine with the aid of such a dictionary, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

1. Introduction

The essential purpose of literature searching is to find those documents within a collection which have a bearing on a given topic. Many of the systems and devices, such as classifications and subject-heading lists, that have been developed in the past to solve the problems encountered be found in automation, there is a real danger that the demand for professional talent will become too great to fill. In view of the foreseeable strain, the most efficient use of talent will have to be made even by automatic systems. The operating requirements of these systems

Automatic Keyword Extraction: Challenges

- Subjectivity due to variation in
  - Perception of subject
  - Ideas on subject
  - Language capabilities
  - Bias due to current interests, experiences, and points of view.

- Massive volume of text to be analyzed
  - Time requirements are pressing

Automatic Keyword Extraction: Flavours

- Supervised vs. **Unsupervised**
- **Single** vs. Multi-Document
- **Extractive** vs. Abstractive

We are focusing on Unsupervised, Single-Document, Extractive Keyword Extraction

Automatic Keyphrase Extraction

- Variant of Automatic Keyword Extraction
  - Phrases are retrieved instead of just unigrams
- Can be performed as a post-processing step
  - Collapse co-occurring candidate keywords together to form phrases
  - Rank phrases based on a function of their candidate scores

Statistical Methods for Automatic Keyword Extraction

Statistical Methods

- Performs statistical analysis of a single/collection of documents
- Strengths
  - Language and domain independent
- TF (Term Frequency)
  - relevance of a term is proportional to the frequency of its occurrence

---

[1] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309-317.

σ-index

- Takes into account word spatial distribution
- Relevant words follow in-homogeneous distribution and tend to form clusters
- Calculates standard deviation, **σ**, of distribution of distances of successive occurrences
  - **σ** > 1: a word attracts itself
  - **σ** = 1: Poisson Distribution
  - **σ** < 1: a word repels itself

---

[2] Ortuno, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *EPL (Europhysics Letters)*, 57(5), 759.

σ-index contd.

- Algorithm Pseudocode
  - Obtain the sequence of occurrence t_i ($i = 1, 2, \ldots, m$) of a word w_i and its inter-token distances
  - Calculate the average distance between two successive word tokens

$$\mu = \frac{(t_2 - t_1) + \cdots + (t_m - t_{m-1})}{m - 1} = \frac{t_m - t_1}{m - 1}$$

  - Calculate normalized Standard deviation,

$$\hat{\sigma} = \frac{\sigma}{\mu} \quad \text{, where} \quad \sigma = \sqrt{\frac{1}{m-2}\sum_{i=2}^{m}((t_i - t_{i-1}) - \mu)^2}$$

– large values correspond to terms relevant to the text considered

Table 1: Words with larger σ in the Bible

Word	Frequency	$\hat{\sigma}$	Word	Frequency	$\hat{\sigma}$
jesus	983	24.18	david	1064	8.86
christ	571	18.42	king	2542	8.15
paul	162	11.56	pharisees	87	8.06
disciples	244	10.88	jeremiah	148	8.00
peter	164	10.17	gospel	104	7.91
joab	145	10.03	solomon	305	7.67
faith	247	9.34	mordecai	60	7.45
saul	420	9.17	esther	57	7.43
absalom	108	9.12	joshua	217	7.42
john	137	9.03	elisha	58	7.39

Source:

[2] Ortuno, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *EPL (Europhysics Letters)*, 57(5), 759.

Γ-index

- Consider the spacings, $w_i = t_i - t_{i-1}$, with $i = 1, \ldots, m+1$
- Calculate average separation around t_i

$$d(t_i) \equiv \frac{w_{i+1} + w_i}{2} = \frac{t_{i+1} - t_{i-1}}{2}, \quad i = 1, \ldots, m \, .$$

- Local cluster index at position t_i

$$\gamma(t_i) = \begin{cases} \dfrac{\hat{\mu} - d(t_i)}{\hat{\mu}} & \text{if } t_i \text{ is a cluster point} \, , \\ 0 & \text{else} \, . \end{cases}$$

- Average of all cluster indices

$$\Gamma(W) = \frac{1}{m} \sum_{i=1}^{m} \gamma(t_i)$$

---

[3] Zhou, H., & Slater, G. W. (2003). A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1), 309-327.

Table 2: Five words of the Holy Bible for which the σ and Γ metrics strongly disagree

word:*m*	$\hat{\sigma}$:Rank	Γ:Rank
sirach:52	0.25:9543	0.93:9
deuteronomy:34	0.37:9412	0.91:24
jude:11	0.74:8737	0.90:32
leviticus:27	0.42:9346	0.90:33
maccabees:31	0.28:9513	0.89:46

Source:

[3] Zhou, H., & Slater, G. W. (2003). A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1), 309-327.

Graph-based Methods for Automatic Keyword Extraction

Graph-based Methods

- Graph representation of text
- Strengths
  - Totally unsupervised – no need of training corpus
  - Scale to any document length

Graph-based Methods

- KeyGraph [4]- earliest method
  - based on segmentation of graph into clusters
  - clusters represent co-occurrence between terms in a document
  - Each cluster corresponds to a concept
  - Extracts top ranked terms per cluster based on a statistic
    - Uses term frequency and term location, but no corpus
  - Content-sensitive and domain independent device of indexing

- State-of-the-art
  - KeyWorld [6], TextRank [7], DegExt [9], PositionRank [10], *k*-core retention [11]

---

[4] Ohsawa, Y., Benson, N. E., & Yachida, M. (1998, April). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. *In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings*. IEEE International Forum on (pp. 12-18). IEEE.

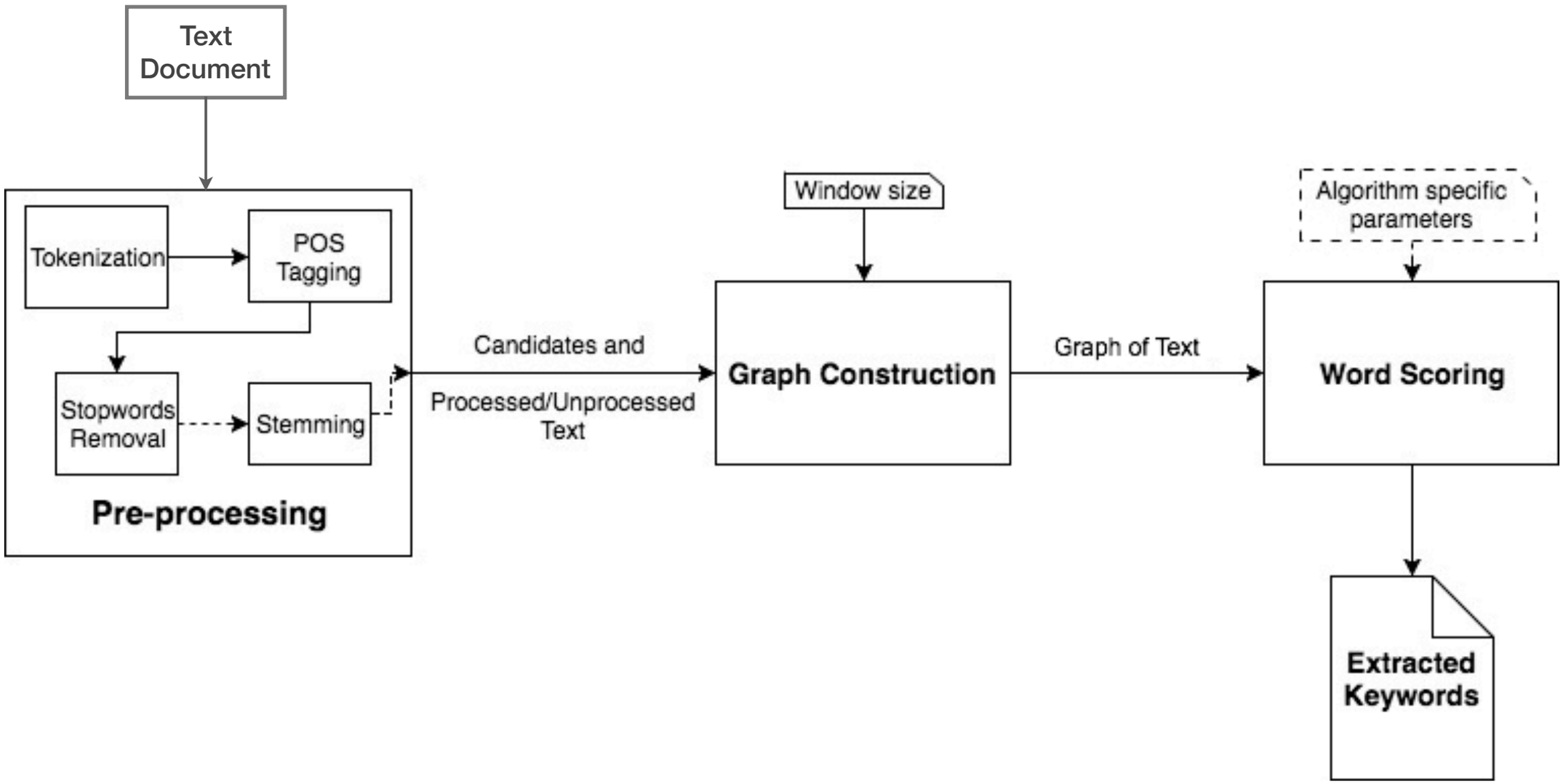


Fig 1. Sequence of sub-tasks in graph-based keyword extraction methods

Cleaning of Text Data

- Documents should be in text format
- Remove non-text elements - like tables, figures, etc.
  - Remove references
  - Remove document title and all section headings
  - Remove all captions
  - Consider only parts of document - title and abstract, full text, etc.

Preprocessing

- Tokenization
- Part-of-Speech (POS) tagging
  - Stanford POS tagger[1], Apache OpenNLP POS tagger[2]
- Stopwords removal[3]
- Stemming
  - Porter Stemmer[4]

---

1. https://nlp.stanford.edu/software/tagger.shtml
2. https://opennlp.apache.org/docs/1.8.0/manual/opennlp.html#tools.postagger
3. http://www.lextek.com/manuals/onix/stopwords1.html
4. https://tartarus.org/martin/PorterStemmer/

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.

Input Text

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.

Input Text

1. Tokenization

[1] “compatibility" "of" "systems" "of" "linear"

[6] "constraints" "over" "the" "set" "of"

[11] "natural" "numbers" ……..

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.

Input Text

2. POS Tagging

[1] “compatibility/NN" “of/IN" “systems/NNS” “of/IN"

[5] “linear/JJ” “constraints/NNS” “over/IN” “the/DT"

[9] “set/NN” “of/IN” “natural/JJ” "numbers/NNS"

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.

Input Text

3. Selection

[1] “compatibility/NN" “of/IN" “systems/NNS” “of/IN"

[5] “linear/JJ” “constraints/NNS” “over/IN” “the/DT"

[9] “set/NN” “of/IN” “natural/JJ” "numbers/NNS"

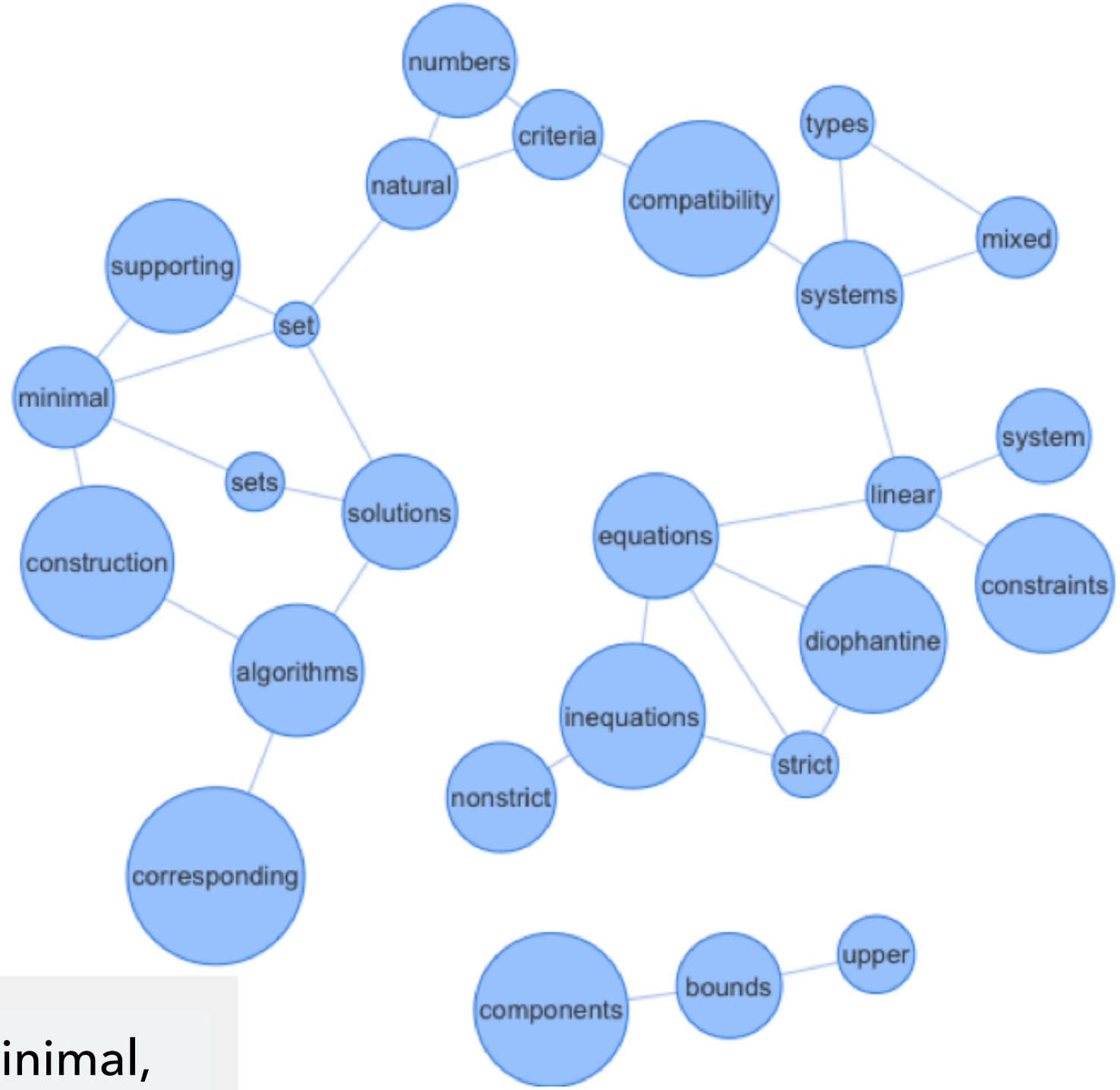
Graph Construction

- Nodes: can be words, sentences, etc.
- Edges: Define relationship between nodes
  - Relationship: co-occurrence, syntactical, semantic, etc. [5]
- Edge weight
  - e.g. frequency of co-occurrence
- Edge direction
  - e.g. directed edge from node *a* to node *b* if *a* is followed by *b* in the original text.

Word Scoring and Ranking

- Small-world property
- Centrality measures
  - Degree centrality,
  - eigenvector centrality, etc.
- Hierarchical decomposition of text graphs
  - *k-core*
  - *k*-truss

Graph Construction



Keywords

[1] linear, [2] systems, [3] minimal,

[4] equations, [5] sets, [6] criteria,

[7] inequations

Keyword Extraction Using Small World Property

KeyWorld

- Small-world structure exists in text graphs
- Cleaning of text document
- Document title, section headings, and all captions of figures and tables are considered as single sentences
- Exclude figures, tables, and references

---

[6] Matsuo, Y., Ohsawa, Y., & Ishizuka, M. (2001, November). Keyworld: extracting keywords from a document as a small world. *In Discovery Science* (Vol. 2226, pp. 271-281)

KeyWorld contd.

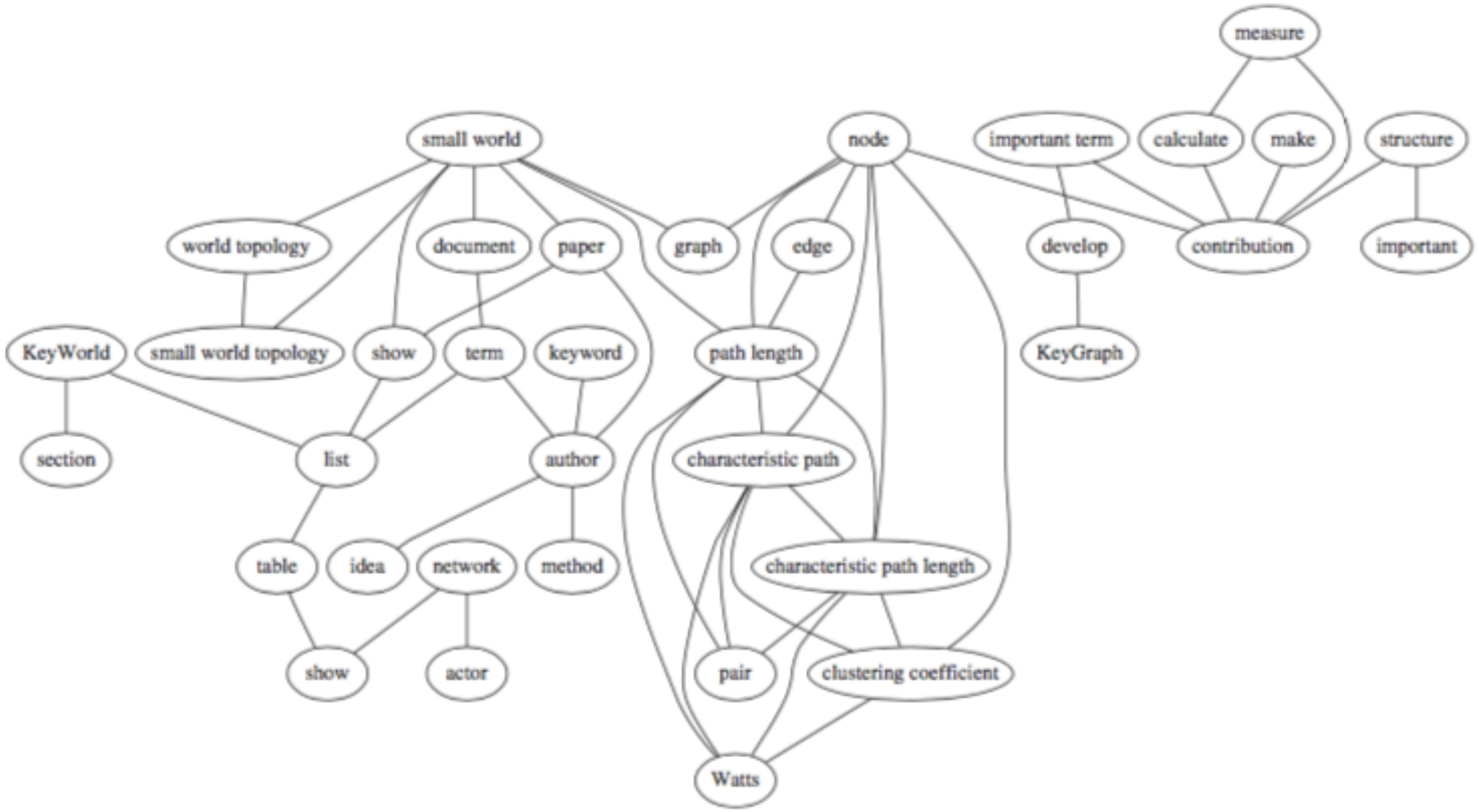
- Text Preprocessing
  - *Candidates identification:* Cleaning, stopwords removal, stemming, counting frequency
  - Candidates are terms with frequency of occurrence above a user-given threshold
- Graph Construction
  - Undirected, unweighted, simple co-occurrence graph - sentence-based co-occurrence
  - Candidates are nodes
  - Co-occurrence relations between nodes are represented as edges with following considerations
    - Co-occurrence is observed in a sentence-based span
    - An edge is added only if the Jaccard coefficient exceeds a certain threshold

KeyWorld contd.

- Word Scoring
  - Calculates the contribution CB_v of a node v to make the graph small-world as follows:

$$CB_v = L'_{G_v} - L'_v$$

  - Difference of average path length of the graph without node v and average path length of the graph with v



[6] Matsuo, Y., Ohsawa, Y., & Ishizuka, M. (2001, November). Keyworld: extracting keywords from a document as a small world. *In Discovery Science* (Vol. 2226, pp. 271-281)

Keyword Extraction using Centrality Measures

TextRank

- Text Preprocessing
  - *Candidates identification:* Tokenization, POS tagging (and retaining only nouns and adjectives), stopwords removal
- Graph Construction
  - Undirected co-occurrence graph with window size 2-10 sliding over the original text
  - Candidates are nodes and co-occurrence relations between nodes are represented as edges
  - Edge weight: frequency of co-occurrence
- Word Scoring
  - PageRank [8] (with $p = 0.85$ and convergence threshold = 0.0001)

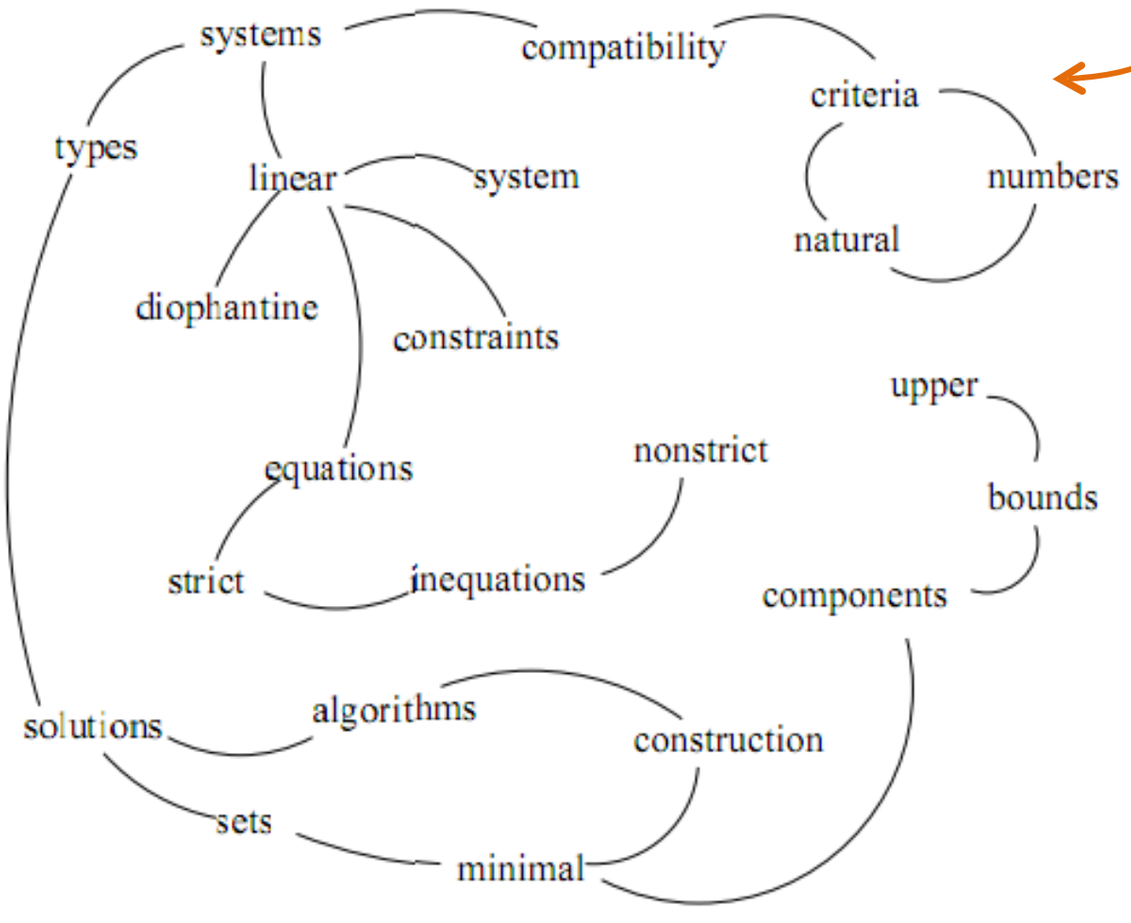
---

[7] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *In Proceedings of the 2004 conference on empirical methods in natural language 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.

The original text

The Constructed graph



[7] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *In Proceedings of the 2004 conference on empirical methods in natural language processing*.

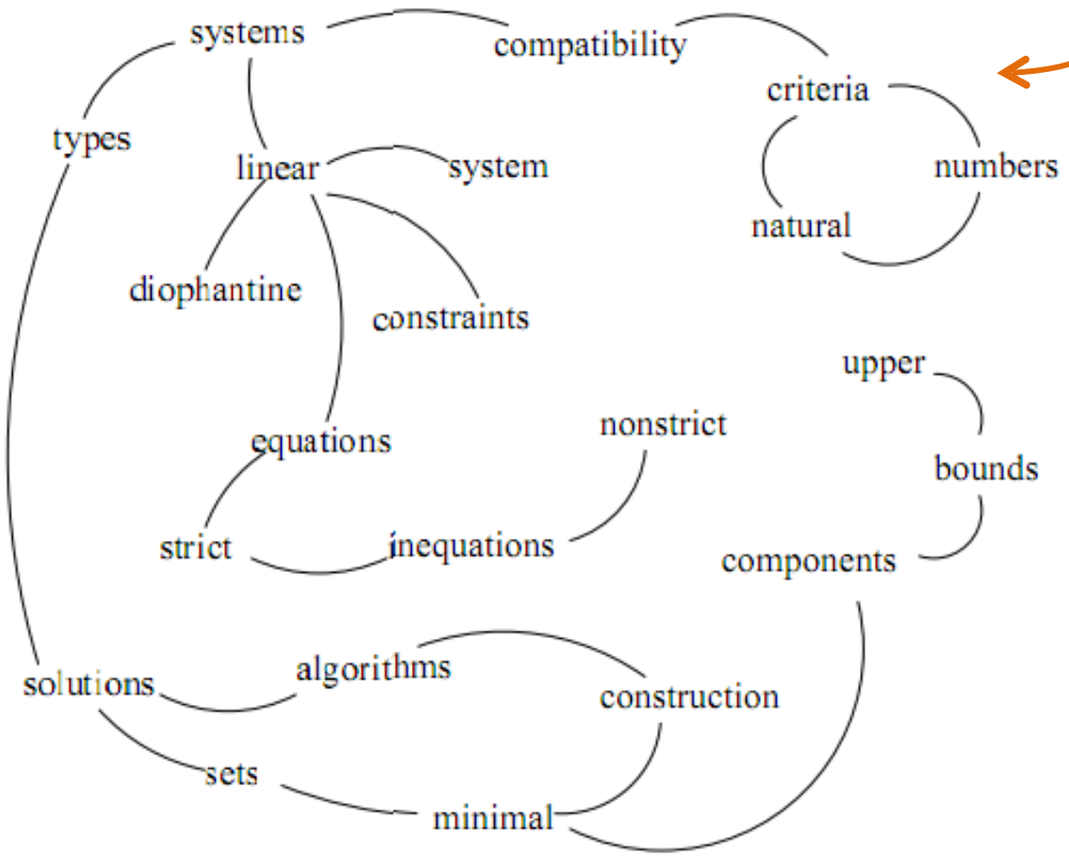
TextRank contd.

- Calculation of score of term t_i

$$score(t_i) = TR(t_i) = (1-p) + p * \sum_{e_{t_i,t_j} \in \mathcal{E}} \frac{w_{t_i,t_j}}{\sum_{e_{t_j,t_k} \in \mathcal{E}} w_{t_j,t_k}} TR(t_j) \ .$$

- Post-processing
  - Mark the potential keywords in the text
  - Sequences of adjacent keywords are collapsed into a multi-word keyword

The Constructed graph



The Extracted keywords

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

[7] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *In Proceedings of the 2004 conference on empirical methods in natural language processing*.

Degree-based Extractor (DegExt) [9]

- Text Preprocessing
  - *Candidates identification:* Tokenization, Stopwords removal
- Graph Construction
  - Directed co-occurrence graph with window size 2 sliding over the processed text without over-spanning (do not connect words separated by punctuation marks)
  - Candidates are nodes and co-occurrence relations between nodes are represented as edges
  - Edge weight: frequency of co-occurrence/sentence id
- Word Scoring
  - Degree Centrality

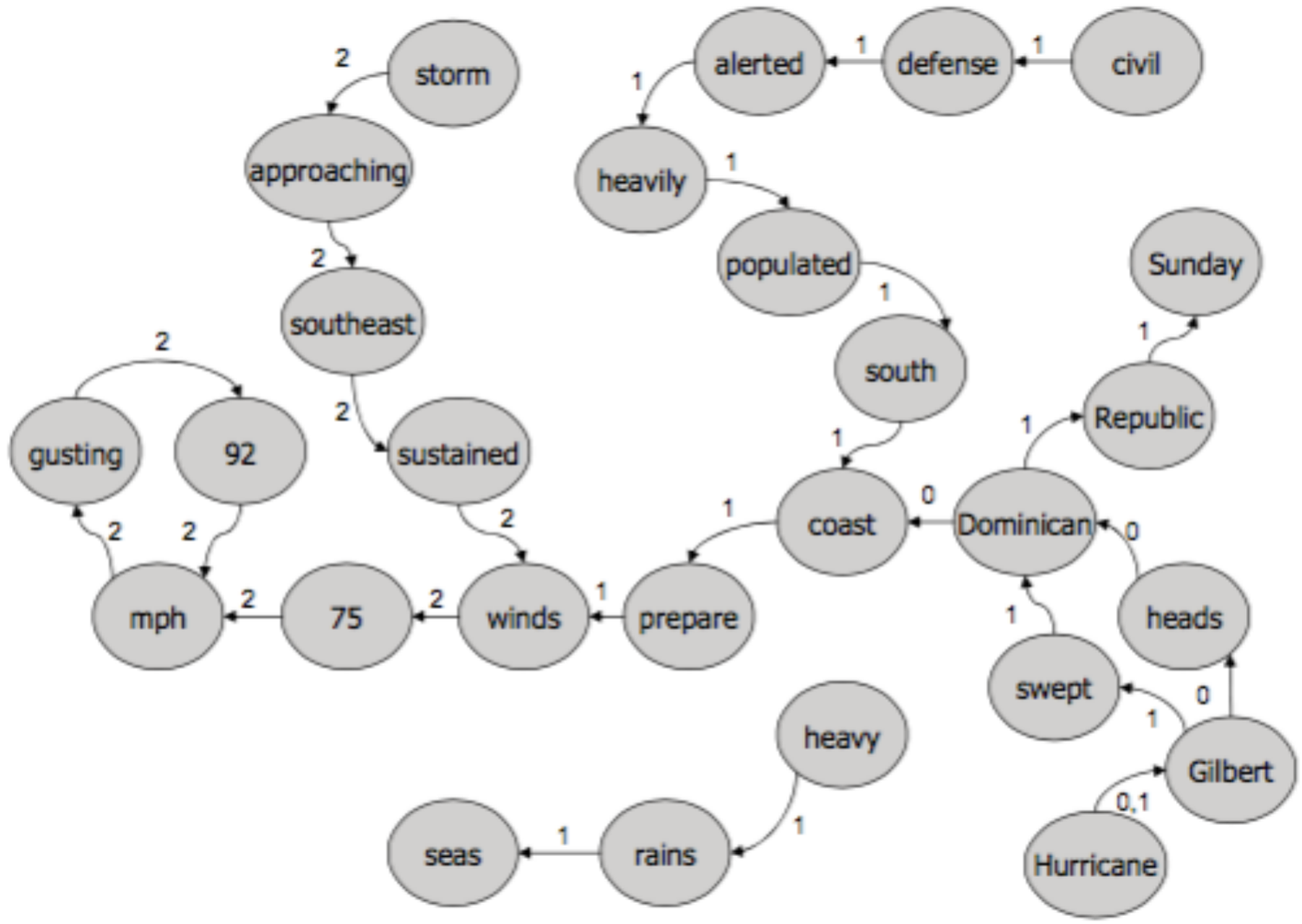
---

[9] Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). DegExt—A language-independent graph-based keyphrase extractor. In *Advances in Intelligent Web Mastering–3* (pp. 121-130). Springer, Berlin, Heidelberg.

0	Hurricane Gilbert Heads Toward Dominican Coast.
1	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.
2	The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.

The Text Document

The Constructed graph



[9] Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). DegExt—A language-independent graph-based keyphrase extractor. In *Advances in Intelligent Web Mastering–3* (pp. 121-130). Springer, Berlin, Heidelberg.

PositionRank [10]

- Text Preprocessing and Graph construction is same as TextRank [Mihalcea, 2004]
- Word Scoring
  - Position-biased PageRank (with α = 0.15 and convergence threshold = 0.001)
  - Assumption: keywords tend to occur at the beginning of the text
    - Assign positional weight to each candidate
      - For a word *w* occurring at *n* positions p_i (*i* = 1, 2, 3, … n), its positional weight $\tilde{p}_i$ is calculated as follows:

$$\tilde{p}_i = \sum_{i=1}^{n} 1/p_i$$

---

[10] Florescu, C., & Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1105-1115).

Duari et al.

PositionRank contd.

- Computation of PositionRank score of a word w_i (node v_i)

$$S(v_i) = (1-\alpha)\cdot \widetilde{p}_i + \alpha \cdot \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{O(v_j)} S(v_j)$$

where $O(v_j) = \sum_{v_k \in Adj(v_j)} w_{jk}$ and $\widetilde{p}_i$ is the weight vector for v_i.

Geographically$^{0.274}$ Focused$^{0.134}$ **Collaborative**$^{0.142}$ **Crawling**$^{0.165}$
by Weizheng Gao, Hyun Chul Lee and Yingbo Miao A **collaborative**$^{0.142}$ **crawler**$^{0.165}$ is a group$^{0.025}$ of crawling$^{0.165}$ nodes$^{0.033}$, in which each crawling$^{0.165}$ node$^{0.033}$ is responsible$^{0.012}$ for a specific$^{0.010}$ portion$^{0.010}$ of the web$^{0.015}$. We study the problem$^{0.007}$ of collecting$^{0.011}$ geographically$^{0.274}$ aware$^{0.006}$ pages$^{0.018}$ using **collaborative**$^{0.142}$ **crawling**$^{0.165}$ strategies$^{0.017}$. We first propose several **collaborative**$^{0.142}$ **crawling**$^{0.165}$ strategies$^{0.017}$ for the **geographically**$^{0.274}$ **focused**$^{0.134}$ **crawling**$^{0.165}$, whose goal$^{0.004}$ is to collect web$^{0.015}$ pages$^{0.018}$ about specified$^{0.010}$ geographic$^{0.274}$ locations$^{0.003}$ by considering features$^{0.005}$ like URL$^{0.006}$ address$^{0.005}$ of page$^{0.018}$ [...] More precisely, features$^{0.005}$ like URL$^{0.006}$ address$^{0.005}$ of page$^{0.018}$ and extended$^{0.004}$ anchor$^{0.004}$ text$^{0.004}$ of link$^{0.004}$ are shown to yield the best overall performance$^{0.003}$ for the **geographically**$^{0.274}$ **focused**$^{0.134}$ **crawling**$^{0.165}$.

Author-input keyphrases: *collaborative crawling, geographically focused crawling, geographic entities*

[10] Florescu, C., & Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1105-1115).

Keyword Extraction using Hierarchical Decomposition of Text Graph

Keyword Extraction using Hierarchical Decomposition of Text Graph

- Hypothesis 1: Keywords occur in constrained environment
- Hypothesis 2: Keywords co-occur with each other more often than chance
- Hypothesis 3: No. of retrieved keywords depends on the document

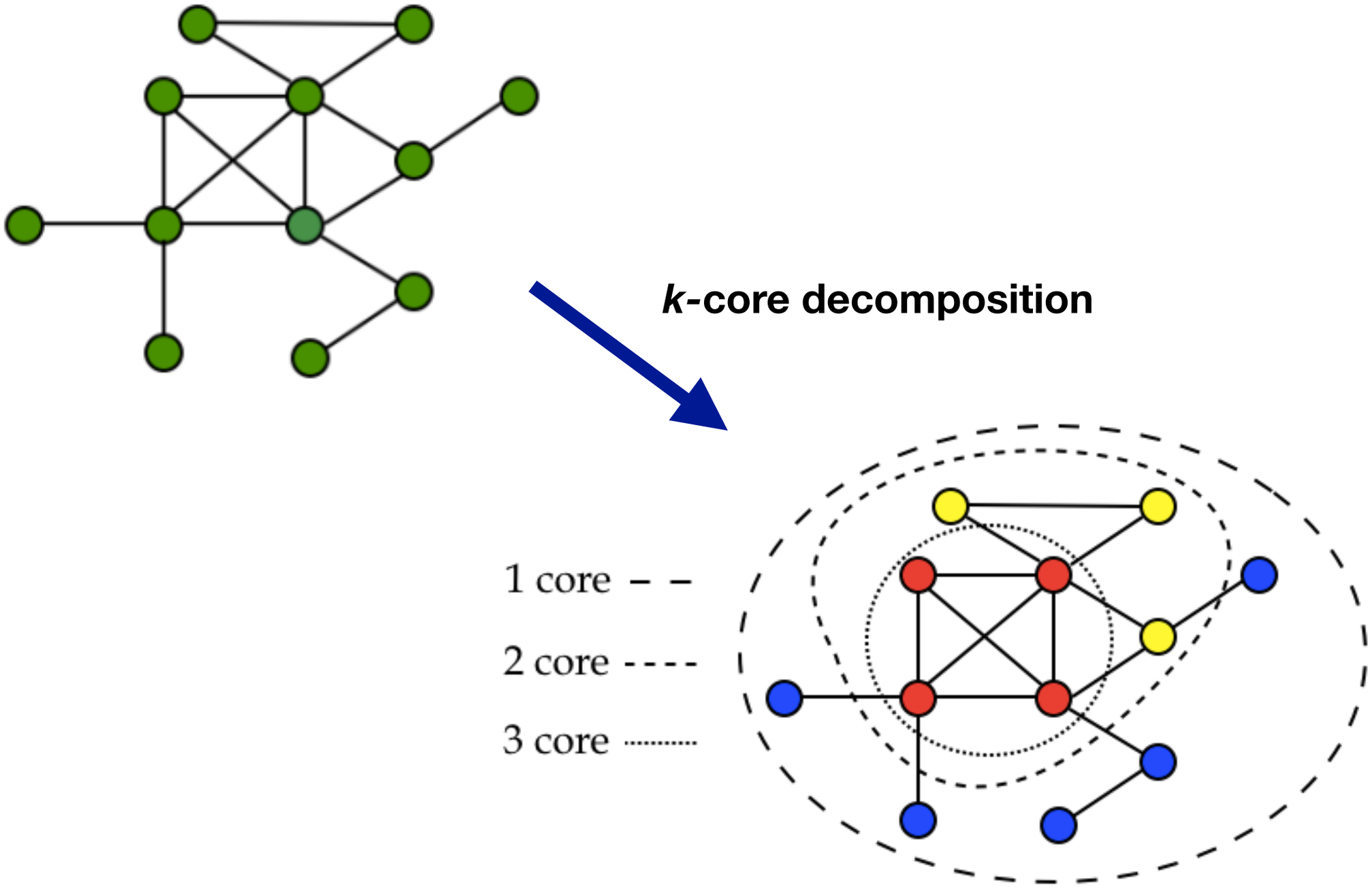
- Challenges the *term independence assumption* of *bag-of-words* model
- Takes into account *term dependence* and *term order*

k-core retention [11]

- Text Preprocessing
  - *Candidates identification:* Tokenization, POS tagging (and retaining only nouns and adjectives), Stopwords removal, Stemming
- Graph Construction
  - Undirected co-occurrence graph with window size 4 sliding over the processed text
  - Candidates are nodes and co-occurrence relations between nodes are represented as edges
  - Edge weight: frequency of co-occurrence
- Word Scoring
  - *k-core* decomposition: Retains the main core after performing graph decomposition

---

[11] Rousseau, F., & Vazirgiannis, M. (2015, March). Main core retention on graph-of-words for single-document keyword extraction. *In European Conference on Information Retrieval (pp. 382-393). Springer International Publishing*.



Source: https://chaoslikehome.wordpress.com/tag/k-core/

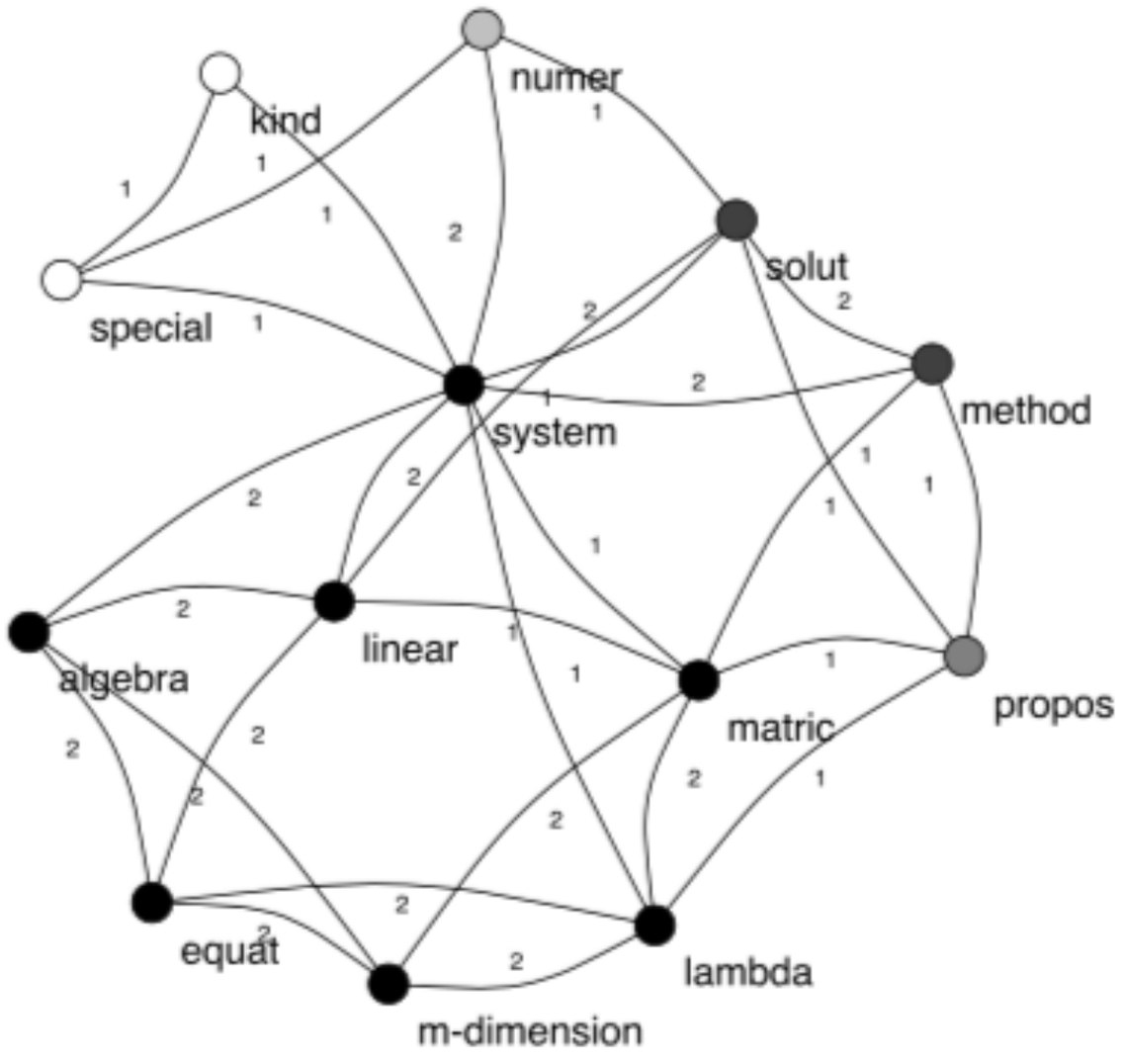
k-core retention contd…

_k_-core decomposition of graph G(V,E) [11]:

A subgraph $Hk = (V', E')$, induced by the subset of vertices $V' \subseteq V$ is called a *k*-core or a core of order *k* if and only if $\forall v \in V'$, $deg_{H_k}(v) \geq k$ and H_k is the maximal subgraph with this property.

- Keywords are more densely connected with each other in the text graph
- Keywords are at the top of hierarchy, if text graph is decomposed on the basis of connectivity, i.e., they lie in the most cohesive connected component(s) of the graph
- Parameter free method – no. of extracted keywords adapts to the structure of each graph

```
A method for solution of systems of linear algebraic equations
with m-dimensional lambda -matrices
A system of linear algebraic equations with m-dimensional lambda
-matrices is considered. The proposed method of searching for
the solution of this system lies in reducing it to a numerical
system of a special kind
```



WK-core		PageRank	
system	6	**system**	1.93
matric	6	**matric**	1.27
lambda	6	solut	1.10
linear	6	**lambda**	1.08
equat	6	**linear**	1.08
algebra	6	**equat**	0.90
m-dim...	6	**algebra**	0.90
method	5	**m-dim...**	0.90
solut	5	propos	0.89
propos	4	method	0.88
numer	3	special	0.78
specia	2	**numer**	0.74
kind	2	kind	0.55

[11] Rousseau, F., & Vazirgiannis, M. (2015, March). Main core retention on graph-of-words for single-document keyword extraction. *In European Conference on Information Retrieval (pp. 382-393). Springer International Publishing*.

References

1. Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309-317.
2. Ortuno, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., & Somoza, A. M. (2002). Keyword detection in natural languages and DNA. *EPL (Europhysics Letters)*, 57(5), 759.
3. Zhou, H., & Slater, G. W. (2003). A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1), 309-327.
4. Ohsawa, Y., Benson, N. E., & Yachida, M. (1998, April). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. *In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings*. IEEE International Forum on (pp. 12-18). IEEE.
5. Blanco, R., & Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information retrieval*, 15(1), 54-92.
6. Matsuo, Y., Ohsawa, Y., & Ishizuka, M. (2001, November). Keyworld: extracting keywords from a document as a small world. *In Discovery Science* (Vol. 2226, pp. 271-281).
7. Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. *In Proceedings of the 2004 conference on empirical methods in natural language processing*.
8. Brin, S., & Page, L. (1998). The anatomy of a large-scale hyper- textual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7)
9. Litvak, M., Last, M., Aizenman, H., Gobits, I., & Kandel, A. (2011). DegExt—A language-independent graph-based keyphrase extractor. In *Advances in Intelligent Web Mastering–3* (pp. 121-130). Springer, Berlin, Heidelberg.
10. Florescu, C., & Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1105-1115).
11. Rousseau, F., & Vazirgiannis, M. (2015, March). Main core retention on graph-of-words for single-document keyword extraction. *In European Conference on Information Retrieval (pp. 382-393). Springer International Publishing*.

TELL ME AND I FORGET. TEACH ME AND I REMEMBER. INVOLVE ME AND I LEARN.

Benjamin Franklin

For further queries, please contact:

vbhatnagar@cs.du.ac.in

sduari@cs.du.ac.in

aditya.mcs.du.2014@gmail.com