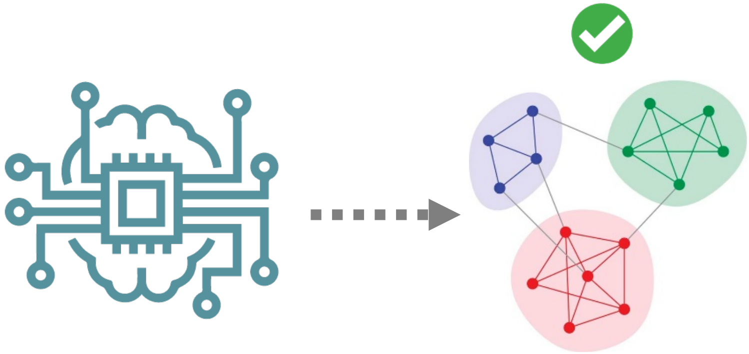
Graph and Centroid-based Word Clustering

Santipong Thaiprayoon

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

4th International Conference on Natural Language Processing and Information Retrieval

Motivations



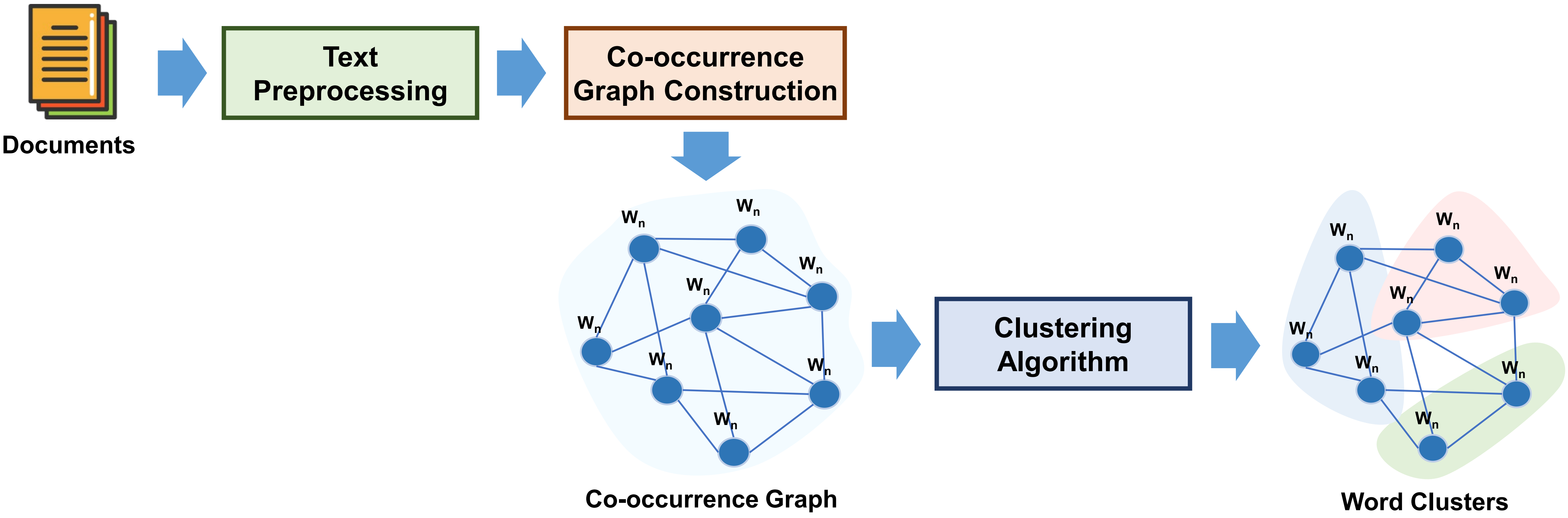
- Word clustering is the process of grouping similar words into distinct clusters, essential approach for NLP tasks.
  - ✓ Reducing the dimensionality for words in documents
  - ✓ Enhancing feature selection for words

Limitations of Word Clustering

- There are some drawbacks of word clustering.
  - ⚠ Require to specify the number of clusters
  - ⚠ Unable to handle noises and outliers

The Process Overview

- The method automatically builds a collection of word clusters containing the most similar words based on their word similarity.



Documents

Text Preprocessing

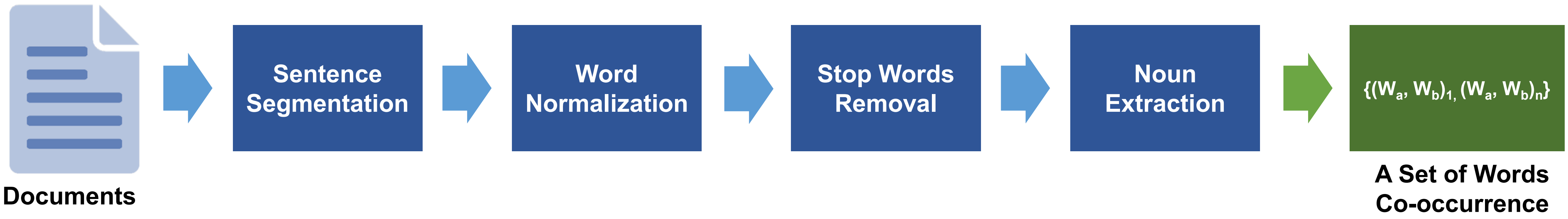
Co-occurrence Graph Construction

Co-occurrence Graph

Clustering Algorithm

Word Clusters

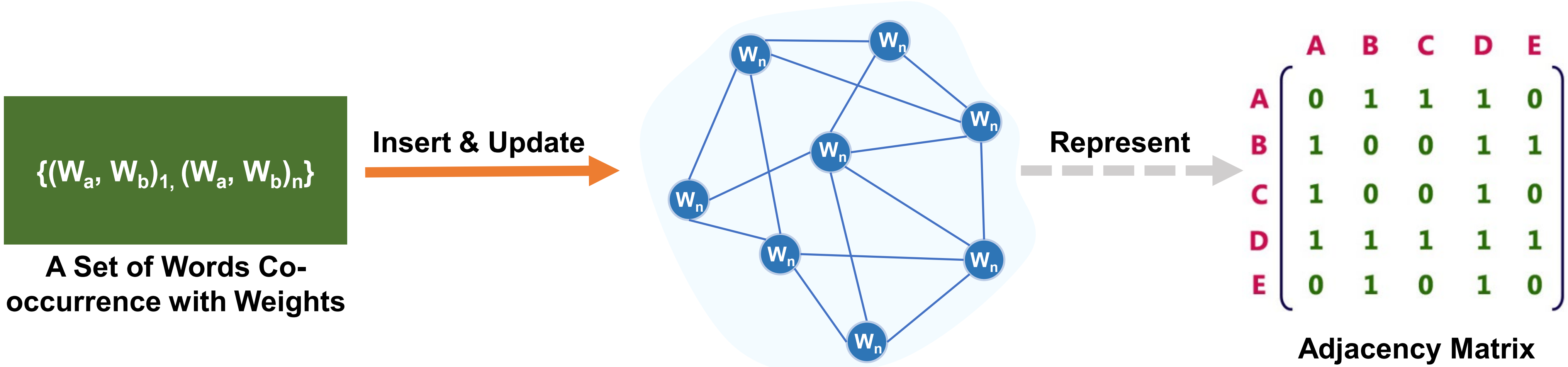
Text Preprocessing



- ✓ Sentence segmentation: The text document is split into a set of sentences.
- ✓ Word normalization: Stemming and lemmatization are applied to process individual words in a sentence by reducing different forms of words to their base form.
- ✓ Stop words removal: A list of stop words including punctuation marks, spaces, and word terminators is used to filter out stop words in the sentence.
- ✓ Noun extraction: A part-of-speech tagging (POS) is used to identify and select the only noun in the sentence.

Building a Co-occurrence Graph

- ✓ A set of co-occurrence of words and weights are inserted and updated into an undirected weighted co-occurrence graph *G = (V, E, W)*.
  - The nodes *V* is a set of nodes representing a term that occurs in documents.
  - The edges *E* is a set of edges representing relationships between every pair of nodes.
  - The weighted W is a set of weights assigning to the edges of the graph *G*.
- ✓ The edges are weighted by a distance score that represents the strength of the relationship of the connected nodes.
- ✓ The distance score is the frequency of occurrences of a pair of words in a sentence.



A Set of Words Co-occurrence with Weights

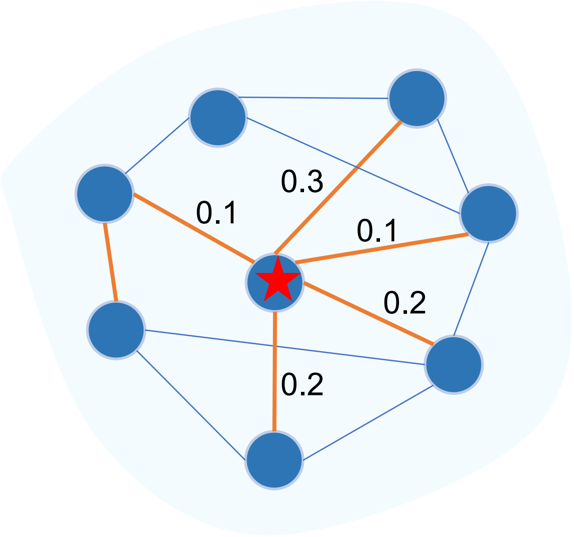
A Co-occurrence Graph

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 1 & 1 & 1 & 0 \\ B & 1 & 0 & 0 & 1 & 1 \\ C & 1 & 0 & 0 & 1 & 0 \\ D & 1 & 1 & 1 & 1 & 1 \\ E & 0 & 1 & 0 & 1 & 0 \end{array}$$

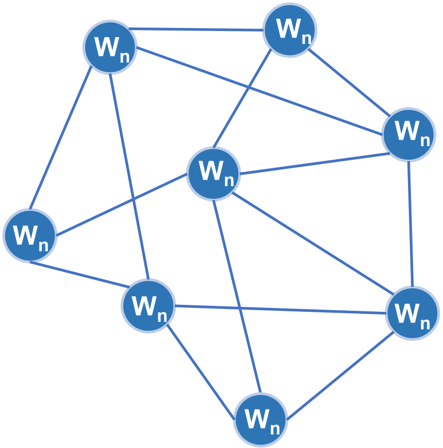
Adjacency Matrix
(Co-occurrence Matrix)

Finding a Cluster Center (Centroid)

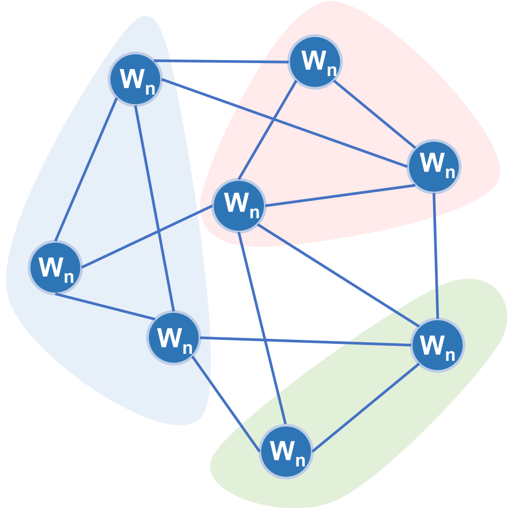
A **centroid** of a document is the term with the minimum average distance to all words in the co-occurrence graph.



Building Word Clusters



Before
Applying Word Clustering

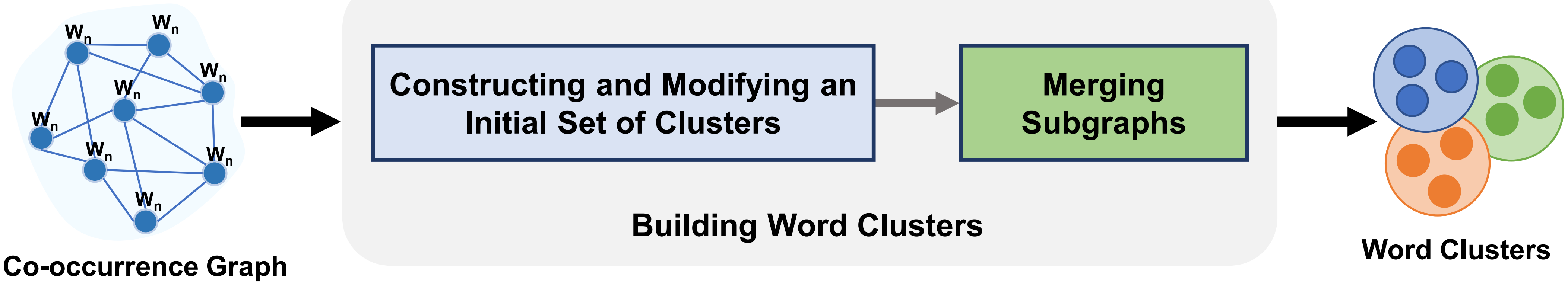


After
Applying Word Clustering

Building Word Clusters

✓ The proposed method performs two-step clustering to create word clusters.

- **Step 1:** Constructing and Modifying an initial set of clusters using hybrid hierarchical k-means clustering for optimizing clustering outputs.
- **Step 2:** Grouping small clusters and isolated words into clusters which aim at reducing the number of clusters.



Co-occurrence Graph

Constructing and Modifying an Initial Set of Clusters

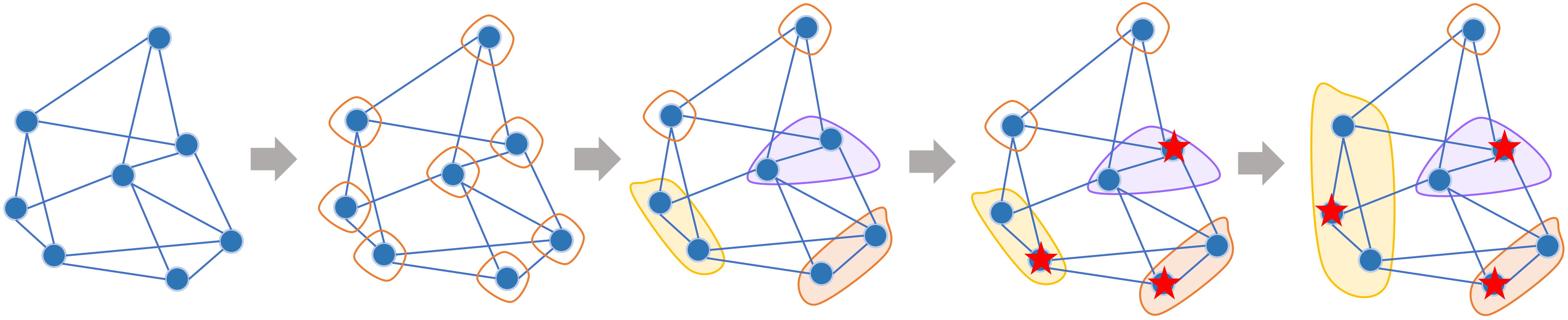
Merging Subgraphs

Building Word Clusters

Word Clusters

Building Word Clusters

STEP: 1



(1) Original Graph

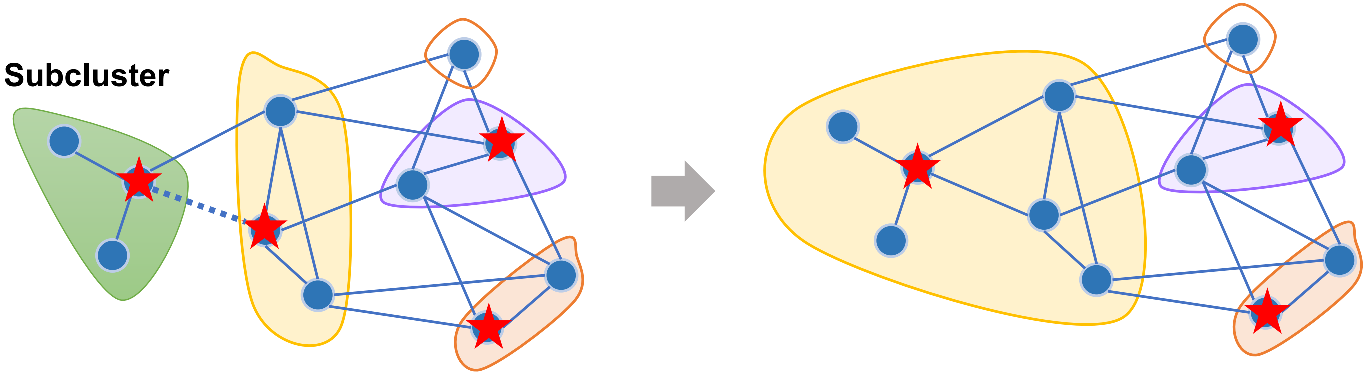
(2) Each word is considered as an individual cluster

(3) Combining the two closest clusters (distance < threshold)

(4) Assigning the cluster to the closest centroid (distance < threshold)

(5) Recomputing the centroids of newly formed clusters

STEP: 2



(1) Finding the two most similar clusters and merge them into a new bigger cluster based on the distance between centroid of clusters

(2) Recomputing the centroids of the newly formed cluster

◉ **Isolated Subcluster**

★ **Cluster Center (Centroid)**

Experimental Design

- ✓ The method is benchmarked against four classical clustering algorithms.
- ✓ Each of the algorithms generates word clusters with default parameters.

K-means

The task of dividing the data points into k clusters which each data point belongs to the cluster with the cluster centers.

K-means++

K-Means++ is designed to improve the centroid initialization. The basic assumption is that initial centroids should be distant from each other.

DBSCAN

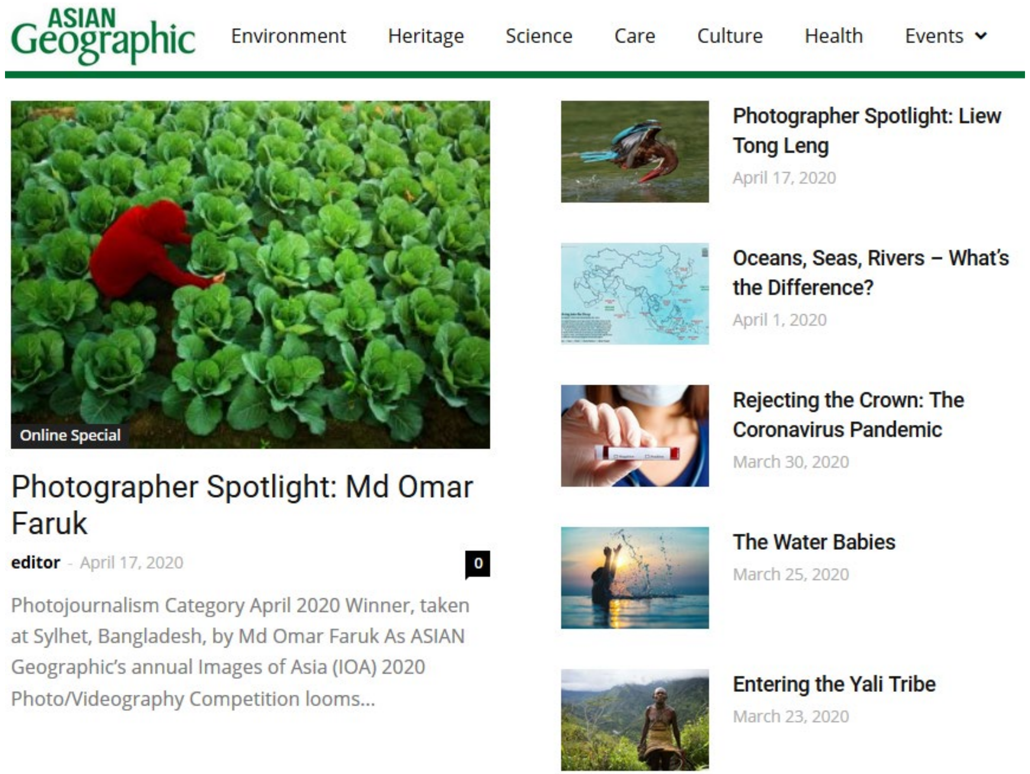
The task of grouping together data points that are close to each other based on a distance measurement.

LSI

The task of identifying relationships between a set of concepts related documents and the terms based on singular value decomposition.

Dataset

- The dataset contains 100 articles from Asian Geographic magazine in 2019.
- The articles include the topic of art, car, computer, leisure, and sport.
- Each of the articles is converted from a PDF into a plain text format.



ASIAN Geographic

Environment Heritage Science Care Culture Health Events

Online Special

Photographer Spotlight: Md Omar Faruk

editor - April 17, 2020 0

Photojournalism Category April 2020 Winner, taken at Sylhet, Bangladesh, by Md Omar Faruk As ASIAN Geographic's annual Images of Asia (IOA) 2020 Photo/Videography Competition looms...

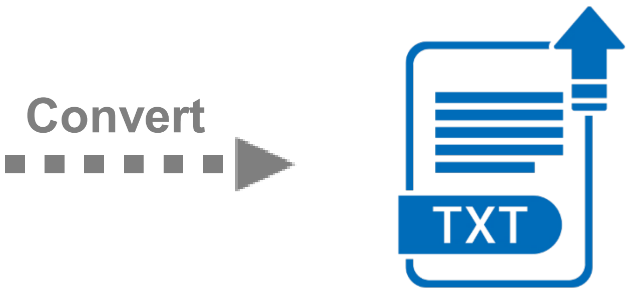
Photographer Spotlight: Liew Tong Leng
April 17, 2020

Oceans, Seas, Rivers – What's the Difference?
April 1, 2020

Rejecting the Crown: The Coronavirus Pandemic
March 30, 2020

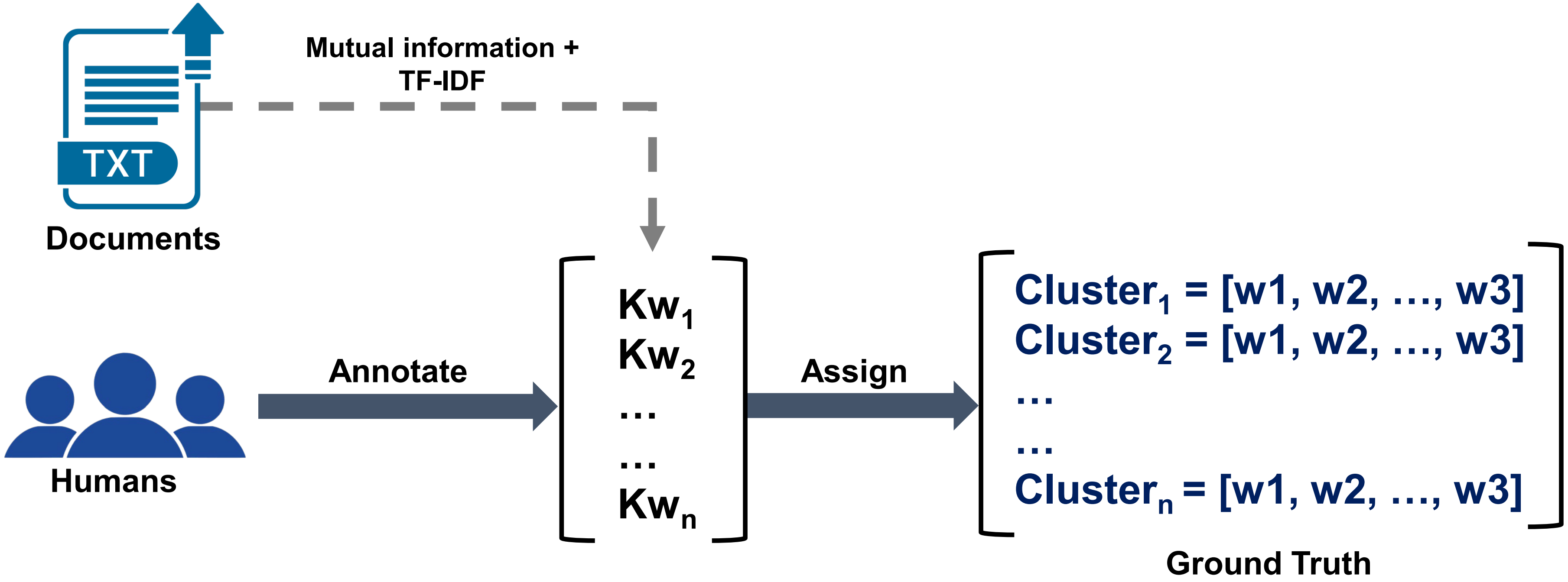
The Water Babies
March 25, 2020

Entering the Yali Tribe
March 23, 2020



Building the Ground Truth Dataset

- ✓ Each article is automatically generated the top 30 words using mutual information with TF-IDF as feature selection.
- ✓ Human annotators manually assign relevant words into predefined clusters as ground truth using a vote system.


TXT
Documents
Mutual information + TF-IDF
Humans
Annotate
Kw_1
Kw_2
…
…
Kw_n
Assign
$Cluster_1$ = [w1, w2, …, w3]
$Cluster_2$ = [w1, w2, …, w3]
…
…
$Cluster_n$ = [w1, w2, …, w3]
Ground Truth

Experimental Results

✓ The purity measure is employed to evaluate the accuracy of the word clustering algorithms.

Algorithms	Purity
Proposed method	**0.383**
K-means	0.104
K-means++	0.125
DBSCAN	0.115
LSI	0.117

An Example of Word Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
airport	*pig*	*dessert*	*leg*	*seafood*
provider	buffalo	roti	jaw	ingredient
speedboat	deer	appetizer	chin	cocktail
vessel	macaque	spread	ear	curry
drive	bird	vegetable	neck	menu
vehicle	turtle	snack	stroke	salad

Discussions

The method can automatically group similar words into the same cluster and more robust noisy text and outliers due to merging small clusters and using an actual node in the cluster to represent the mean point as the center of a cluster.

Cannot handle semantic relationships with other words

PROS

- Easily implement
- Merge small clusters and isolated words
- No need to specific initial parameters
- Not require the certain number of clusters
- Handle noisy data or outliers

CONS

- Cannot capture semantic relationships
- Time complexity for a larger graph
- Computational time is very expensive

Future Works

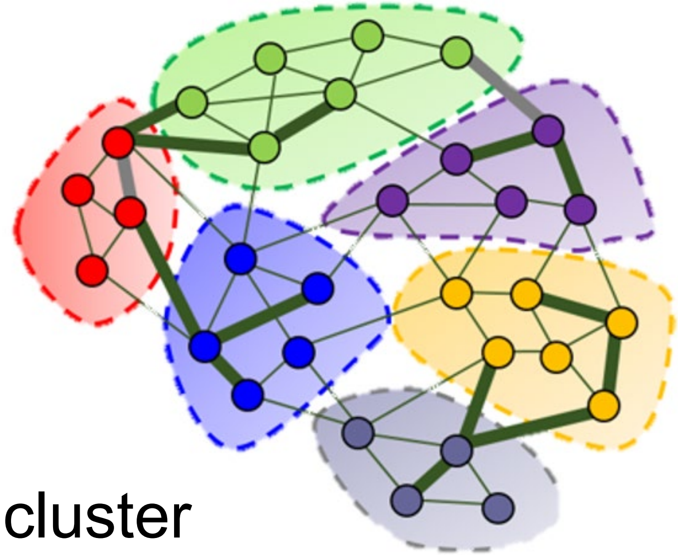
Name Entity Recognition (NER) to identify named entities such as person names, organizations, and locations

Word embeddings with TF-IDF to measure the similarity distance between words

PageRank algorithm to calculate the importance of centroid terms in texts

Graph embeddings to reduce the computational time of the shortest path algorithm

Conclusions



- ✓ A new method of the graph and centroid-based word clustering, which address the problem of specifying initial parameters and outliers
- ✓ Finding similar words according to the criterion of distance to the cluster center (centroid)
- ✓ Small clusters and isolated words merged into another cluster
- ✓ The method outperforms traditional clustering algorithms
- ✓ The method can be easily integrated with NLP applications to further improve the performance of downstream tasks

Thank you

Santipong Thaiprayoon

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

✉ **Email:** santipong.thaiprayoon@fernuni-hagen.de