# Dynamic Clustering for Segregation of Co-Occcurrence Graphs

Supaporn Simcharoen

Lehrgebiet Kommunikationsnetze,
Fakultät für Mathematik und Informatik,
FernUniversität in Hagen

## Contents

✓ Introduction

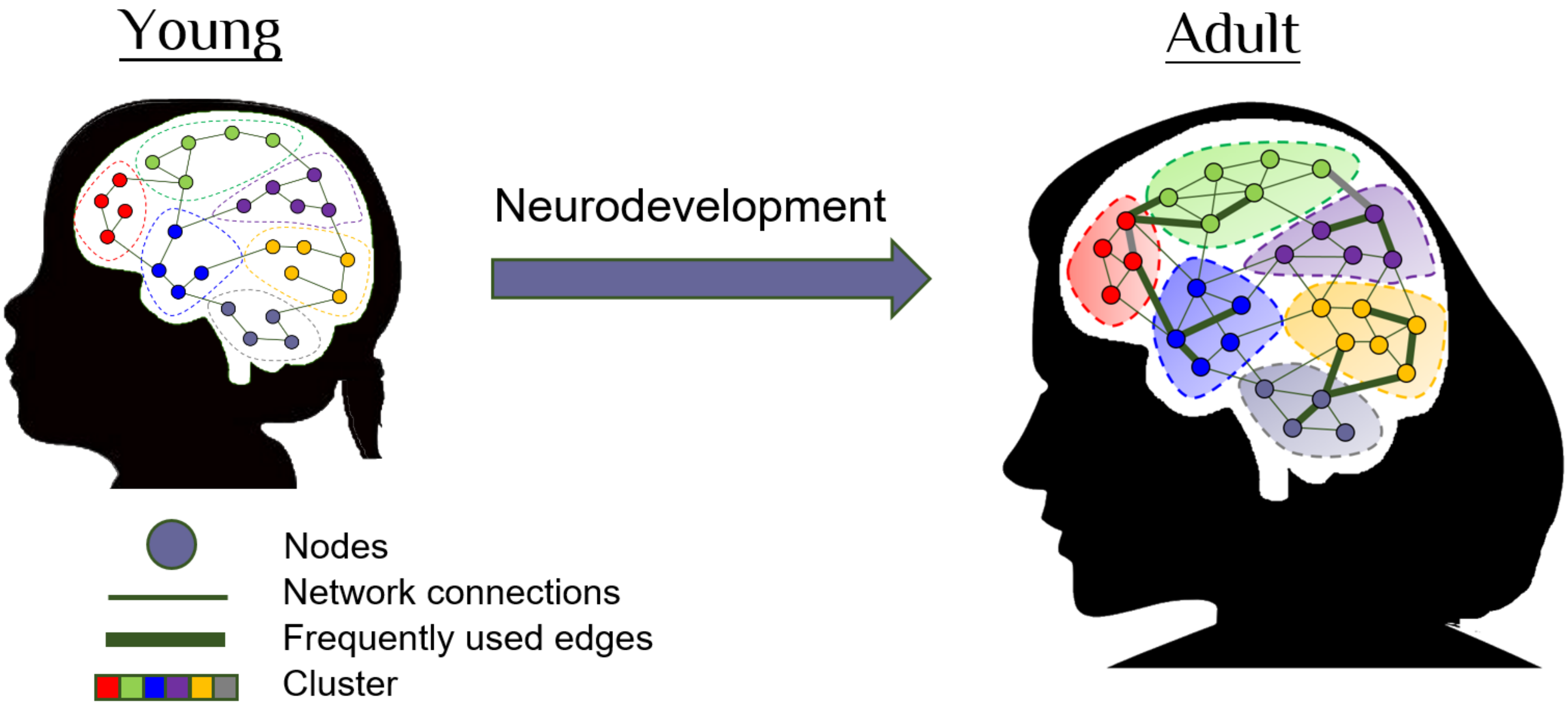
✓ Idea

✓ The Cluster Building

✓ Simulation Results

✓ Conclusion

# Introduction



Young

Neurodevelopment

Adult

● Nodes

— Network connections

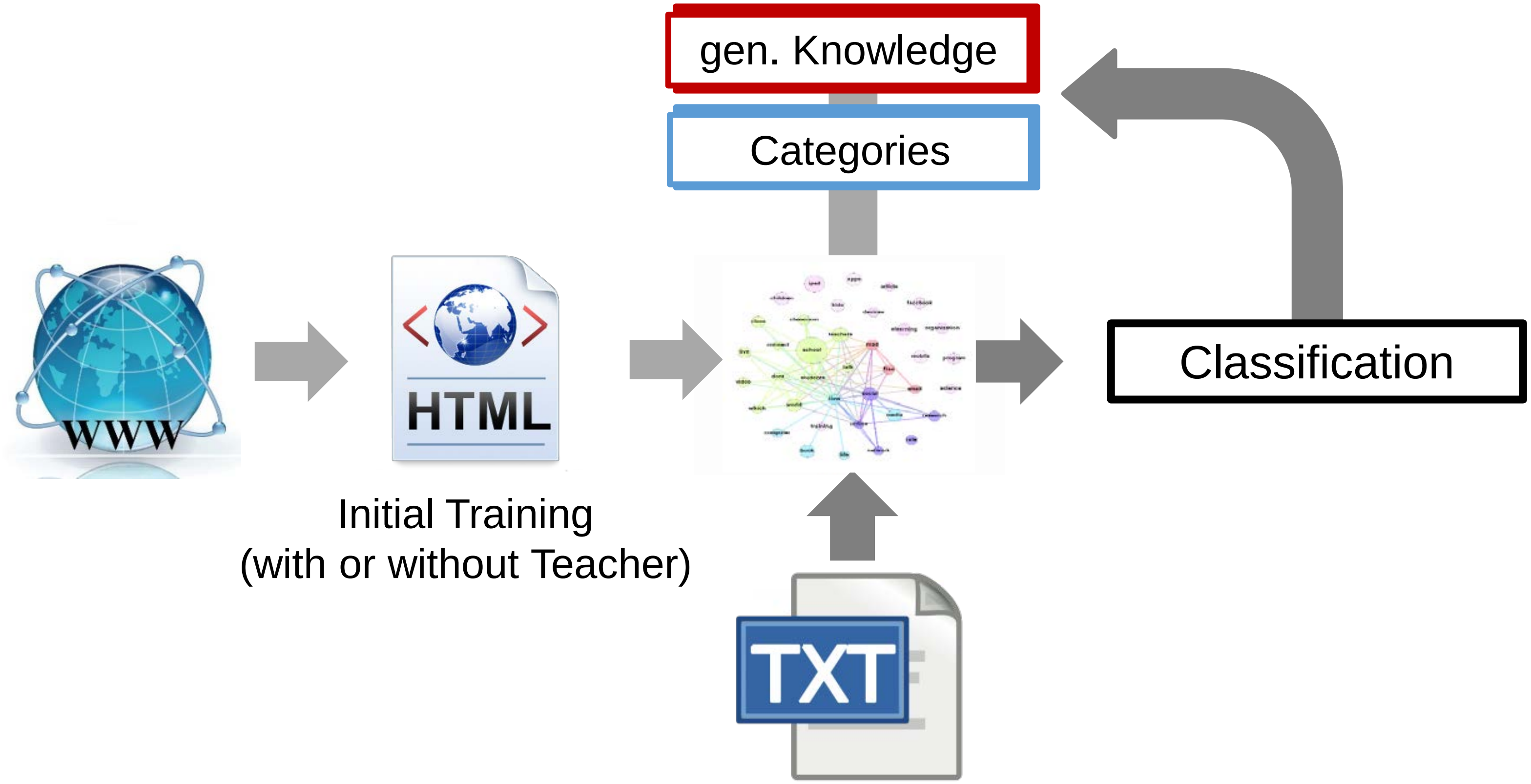
▬ Frequently used edges

🟥🟩🟦🟪🟨⬜ Cluster

“The development of structural brain networks”

# Introduction

To formalize this process in a model,



gen. Knowledge

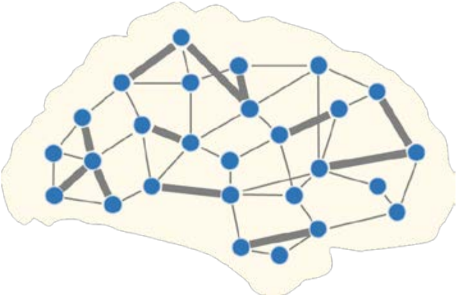
Categories

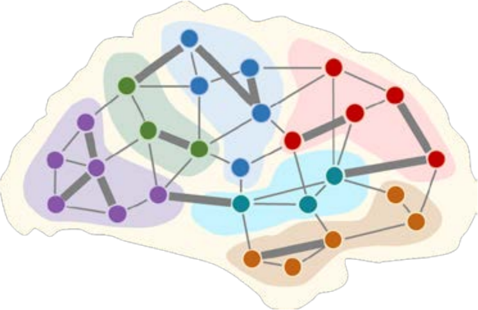
Classification

Initial Training
(with or without Teacher)

“Processes in the brain **while reading**”

# Introduction





## Four significant **processes** appear in the brain

1. New **words** are learnt.

   **Nodes** in co-occ. graph

2. **Relations** between words are added.
   A complex Network of connections appears.

   **Edges** in co-occ. graph

3. **Clusters** emerge in this networks.

   **Clusters** in co-occ. graph

4. Depending on the recent state of the network,
   new texts can be **categorised**.

   By looking, in which cluster
   **a centroid term** of a text is a node.

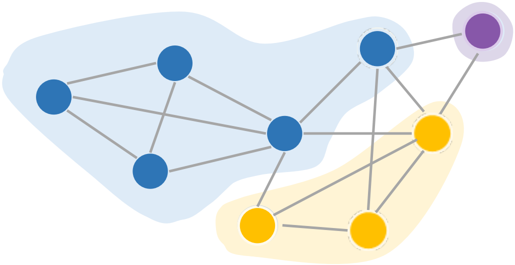
Let's do all of this **processes** in a model…

# Idea

- ✓ In the beginning, one document after the other is read (sentence by sentence). New words of each sentence and relations (edges) between words are added, **a co-occurrence graph** is built.

- ✓ New words must **find** for clusters to assign. The **distance** to **the cluster center** (Centroid) ensures that the word is a member of that cluster.

  - ➢ Some new words may add to existing clusters.
  - ➢ If any word far away from the cluster center (Centroid), a new cluster must be **created.**

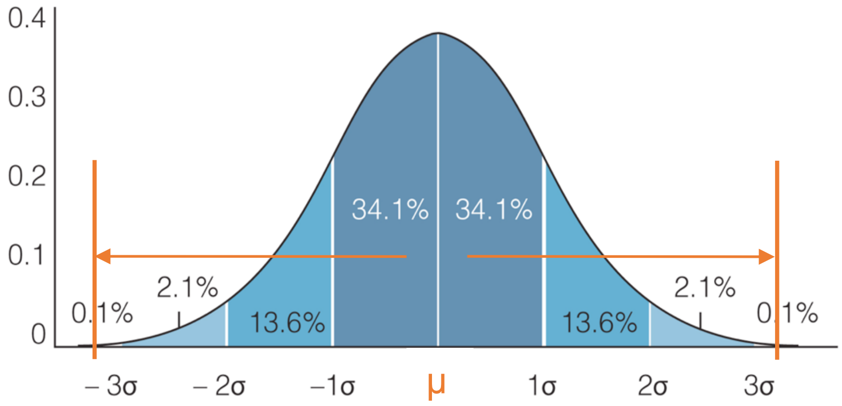


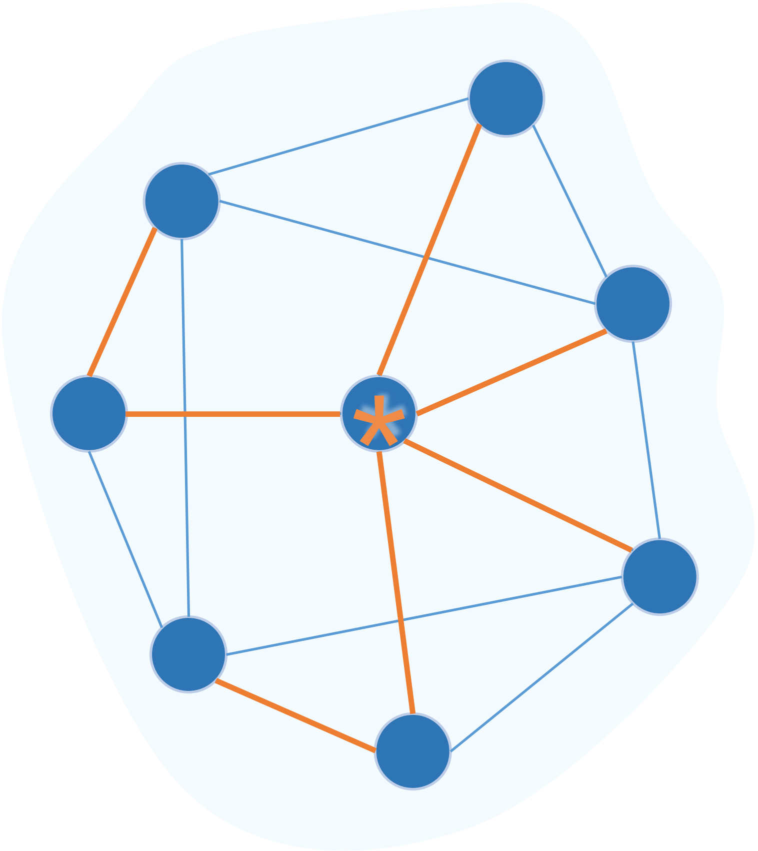
- ✓ Each **addition** of a new word or **change** of an edge weight, change the evaluation of the clustering in the co-occurrence graph.

- ✓ Clusters can be joined, divided, and restructured (add or remove nodes)

# The Cluster Building - Idea

- ✓ The Cluster Center is a Node of the cluster with the shortest average distance to every other node in the graph.



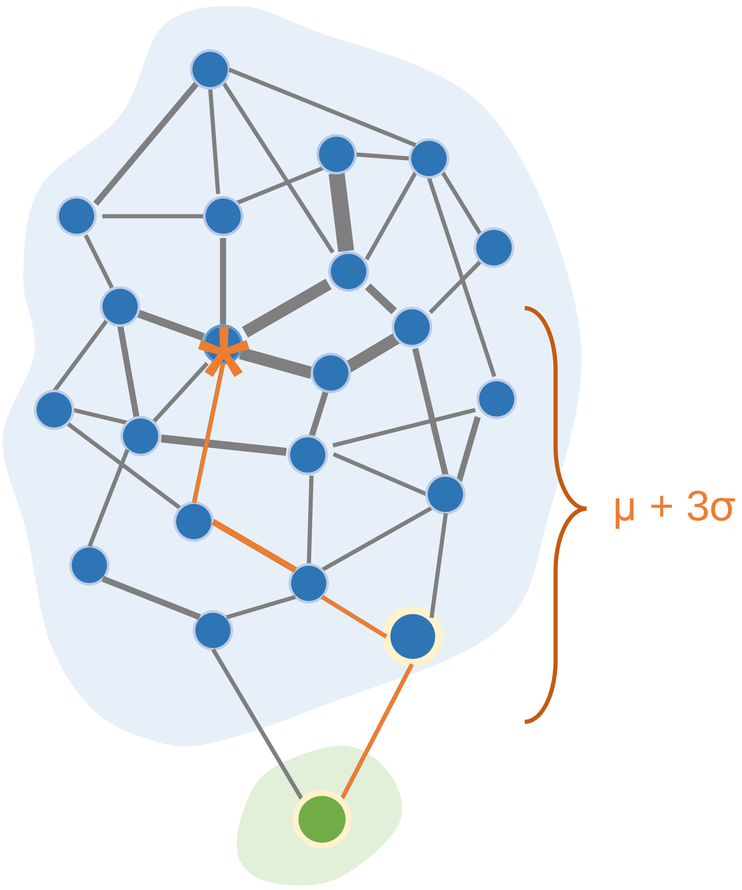


- ✓ The average distance (μ) of all nodes from the cluster center can be calculated.
- ✓ Furthermore, a standard deviation (σ) of distances can be obtained in each cluster.
- ✓ Only nodes within the

  distance range (μ + 3σ)

  from the centroid shall be a member of the respective cluster.

# The Cluster Building - Growth



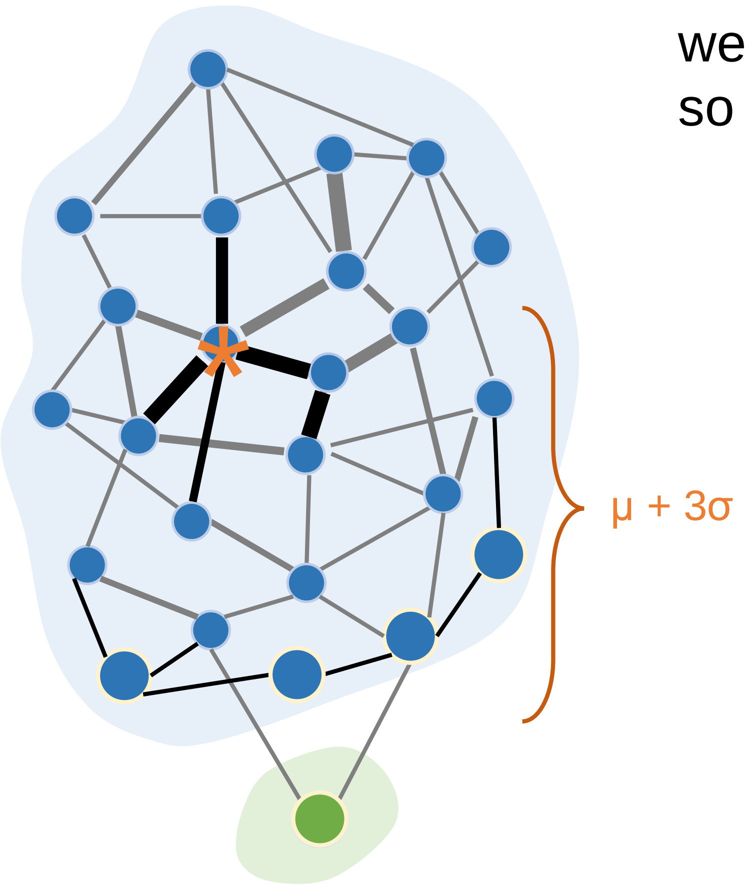
* is a position of the cluster center (Centroid)

➢ While reading documents, new nodes (words) and edges are added in the co-occurrence graph. Every new word **finds** the cluster where shall be assign to.

✓ If the shortest distance less than the **range** value (μ+3σ), then this word is **added** to that cluster.

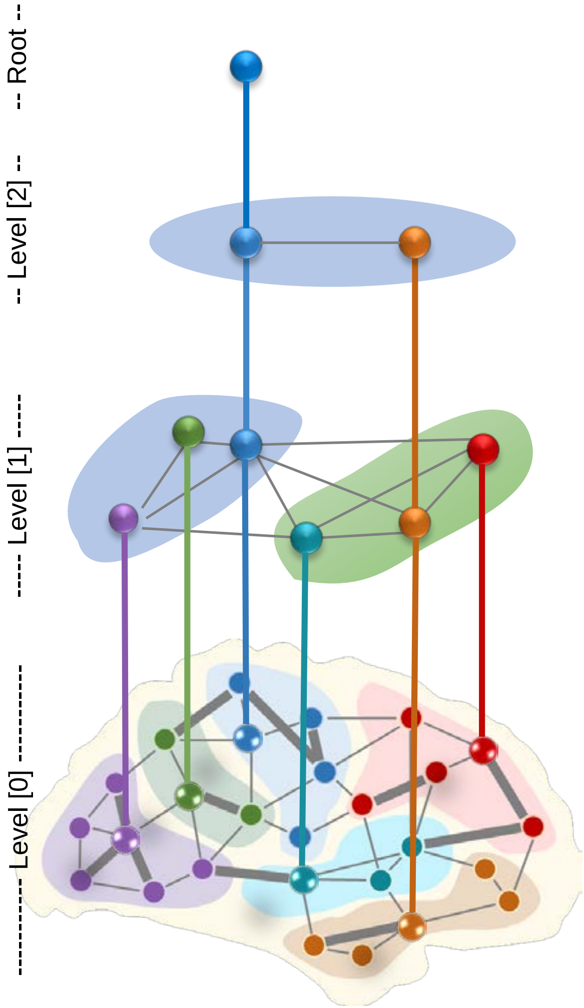
✓ If the shortest distance more than the **range** value (μ+3σ), then this word is **moved** to another cluster or a new cluster *(if no cluster with ≤ μ+3σ* ).

# The Cluster Building - Update



- Newly added nodes (words) and changes of the weight of edges change the situations within the so far built clusters of the co-occurrence graph.
  - ✓ It is possible, that the cluster center is **moved** to another node.
    - ✓ The average distance (μ) and the standard deviation (σ) need to be re-calculated.
    - ✓ Nodes with a bigger distance than (μ + 3σ) shall be removed from the cluster.
- Re-calculation must be done for all affected clusters in a repeated manner until a stable state is reached.

# The Cluster Building - Hierarchies



Reading the centroid term of each cluster (node by node) for building the next hierachy level,

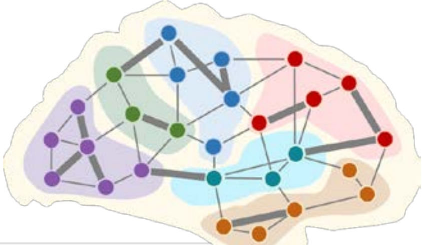
- ✓ Add the first node to a new **inter-cluster** then to set an average distance (μ), a standard deviation (σ), and distance range (μ + 3σ).
- ✓ Every next node **finds** the inter-cluster to add by calculate the shortest distance with the cluster center of related existing inter-clusters,
  - ✓ If less than (μ + 3σ), then this node is **added**
  - ✓ If more than (μ + 3σ), then to **move** to another inter-cluster or a new inter-cluster.
- ✓ Recalculator must be done for all affected inter-clusters in a repeated manner until a stable state is reached.

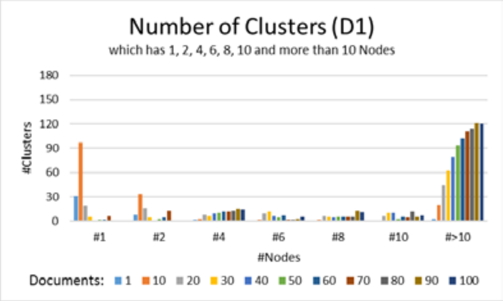
To **repeat** until only one node remains.

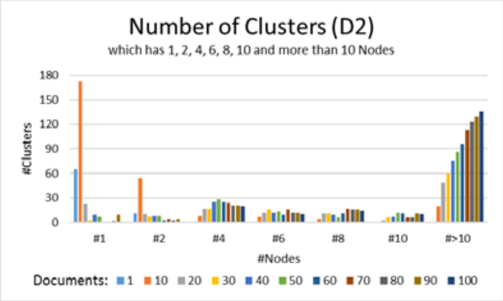
# Simulation Results

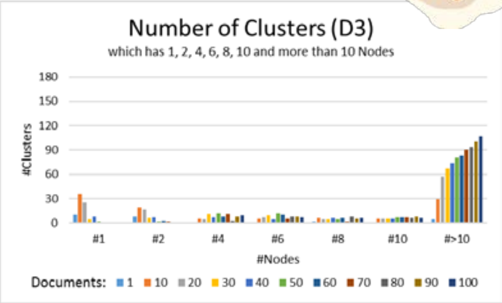
## Experimental Results:: *Dataset(D1-D10)*

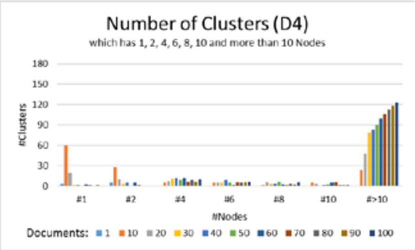
*(Each Dataset* consists of 100 articles covering topic 'art' (20), 'car' (20), 'computer' (20), 'leisure' (20), 'sport' (20))

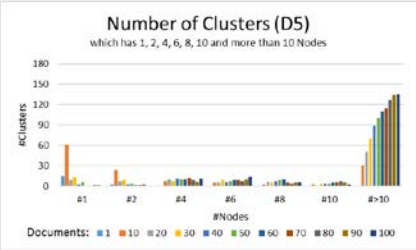


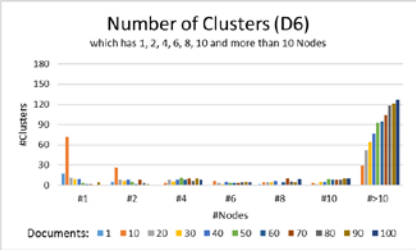

Number of Clusters (D1)
which has 1, 2, 4, 6, 8, 10 and more than 10 Nodes
#Clusters
180
150
120
90
60
30
0
#1
#2
#4
#6
#8
#10
#>10
#Nodes
Documents: 1 10 20 30 40 50 60 70 80 90 100

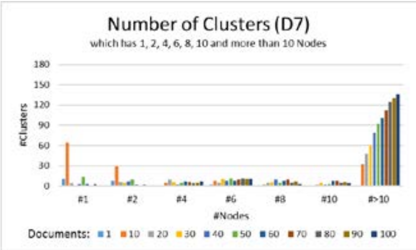


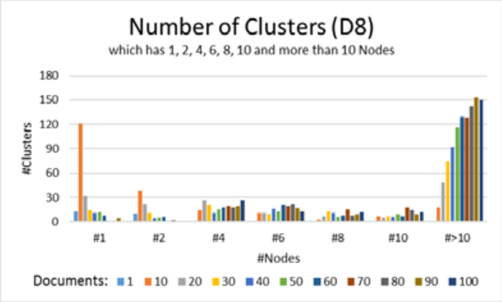


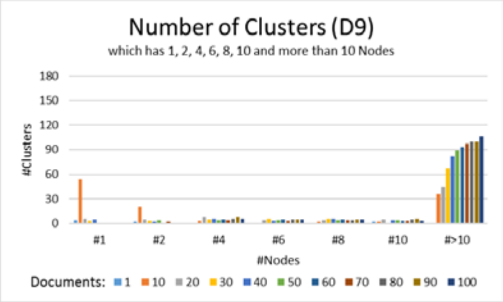


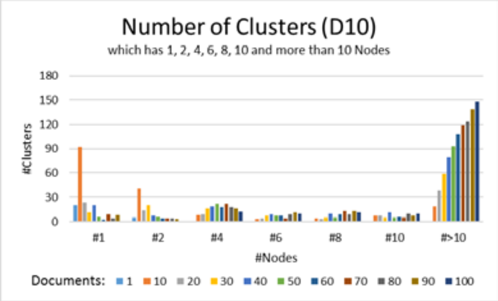








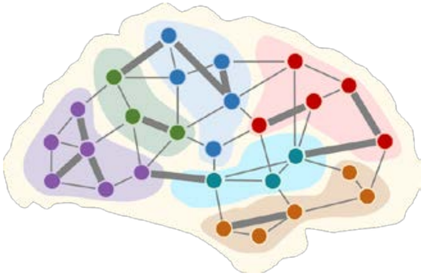




# Simulation Results

## Experimental Results:: *Dataset(D1-D10)*

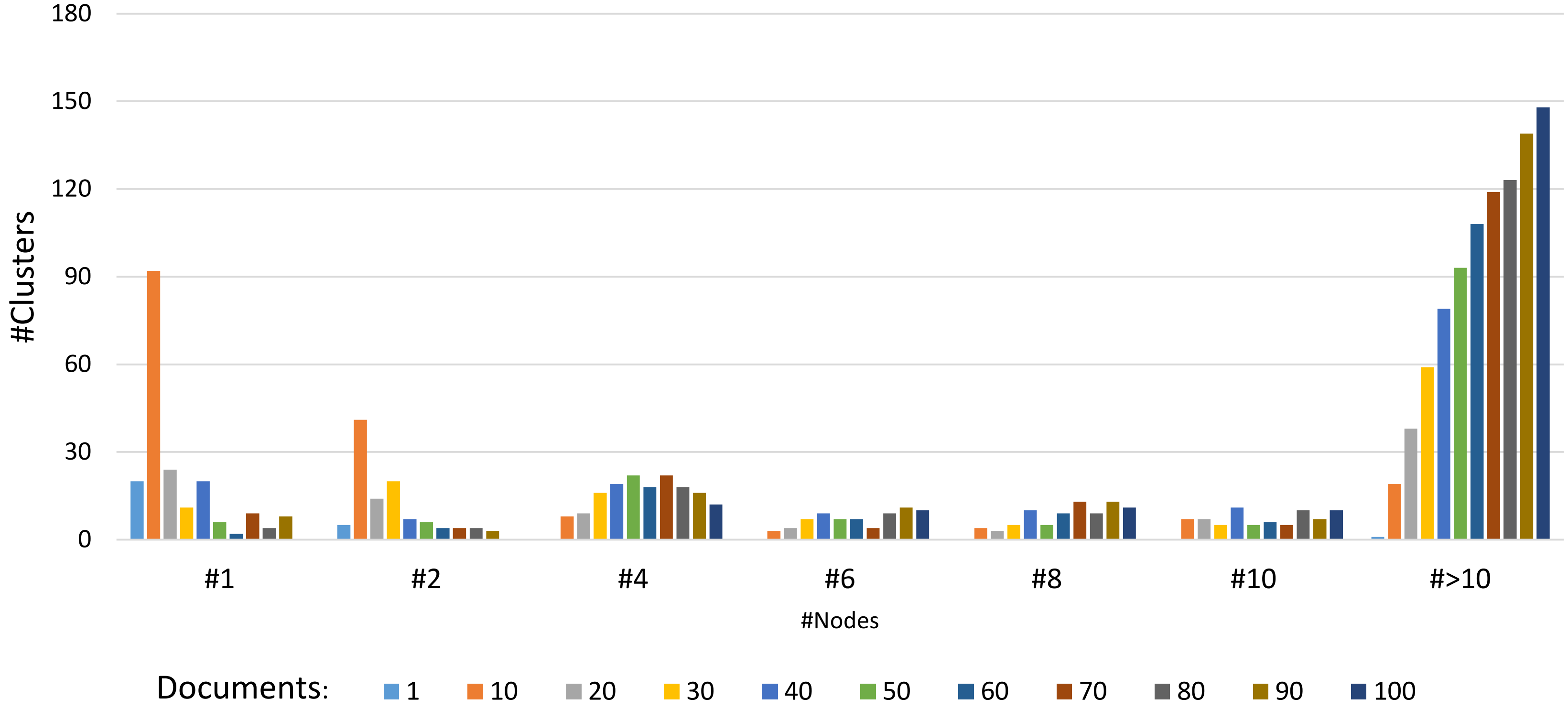
*(Each Dataset* consists of 100 articles covering topic 'art' (20), 'car' (20), 'computer' (20), 'leisure' (20), 'sport' (20))



### Number of Clusters (D10)

which has 1, 2, 4, 6, 8, 10 and more than 10 Nodes

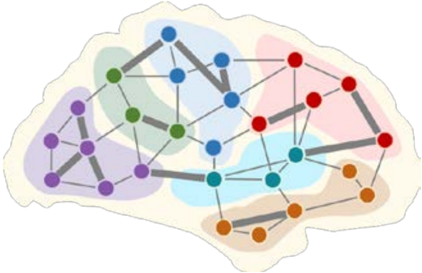
(4683 Words, 233 Clusters)



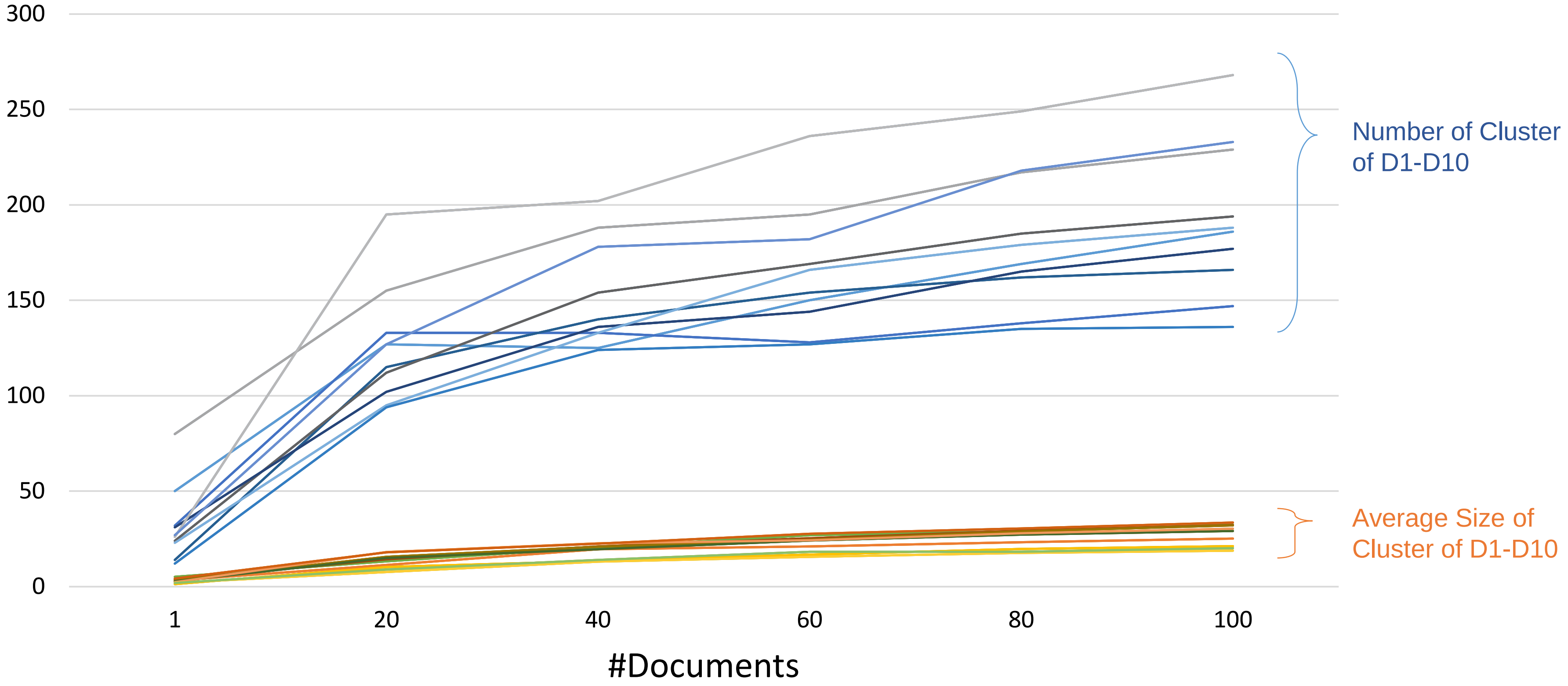
# Simulation Results

## Experimental Results:: *Dataset(D1-D10)*

*(Each Dataset* consists of 100 articles covering topic 'art' (20), 'car' (20), 'computer' (20), 'leisure' (20), 'sport' (20))



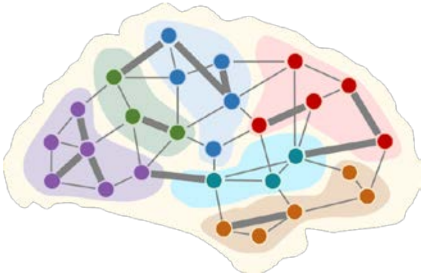
Number of Clusters and Average Size of Clusters



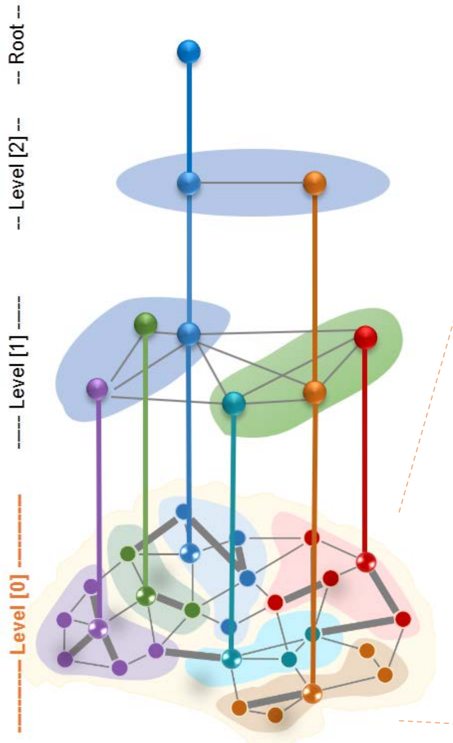
# Simulation Results

## Experimental Results:: *Dataset(D1-D10)*

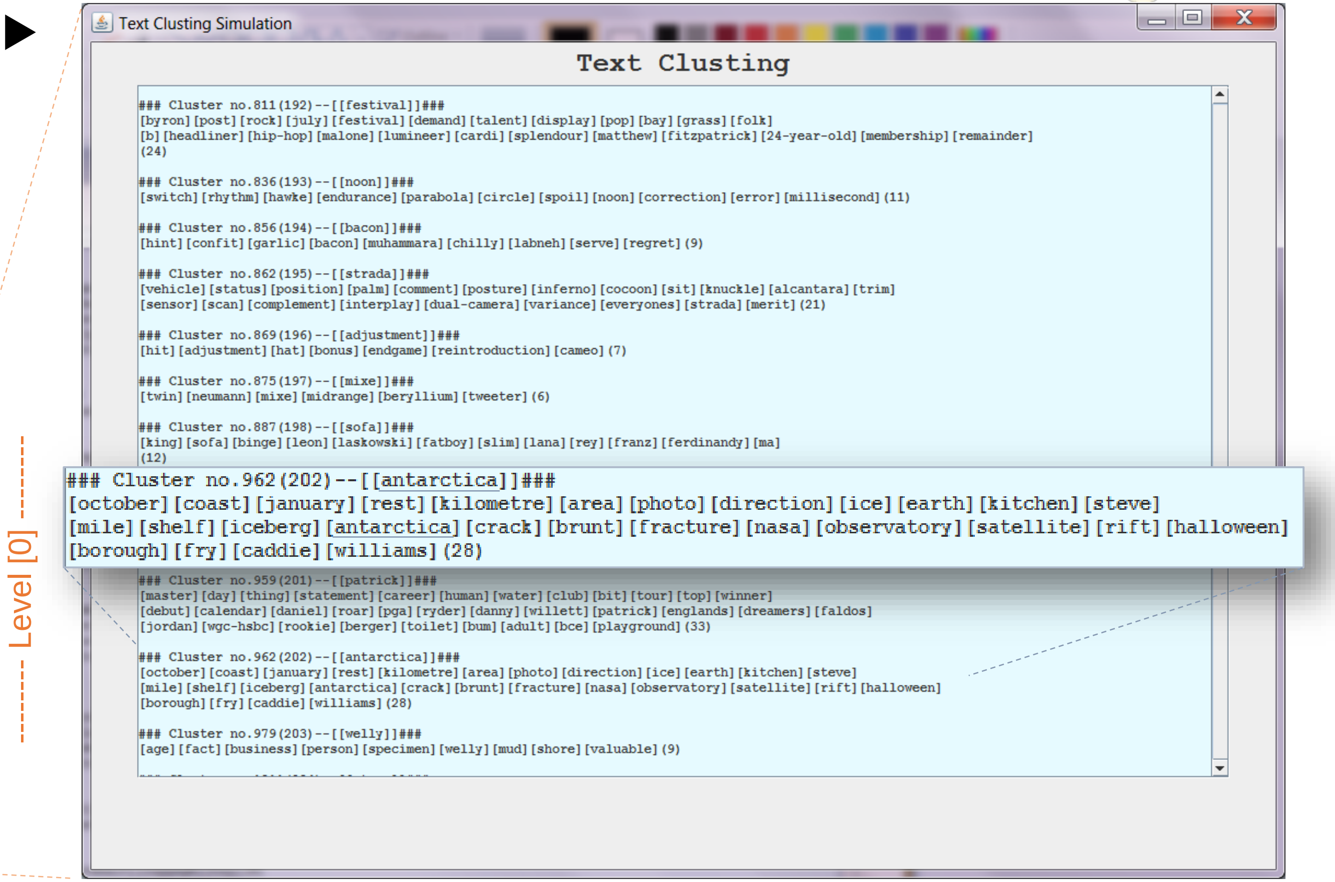
(*Each Dataset* consists of 100 articles covering topic 'art' (20), 'car' (20), 'computer' (20), 'leisure' (20), 'sport' (20))



Member of Clusters ▶
of Dataset10 (D10)



„The hierarchy creation“



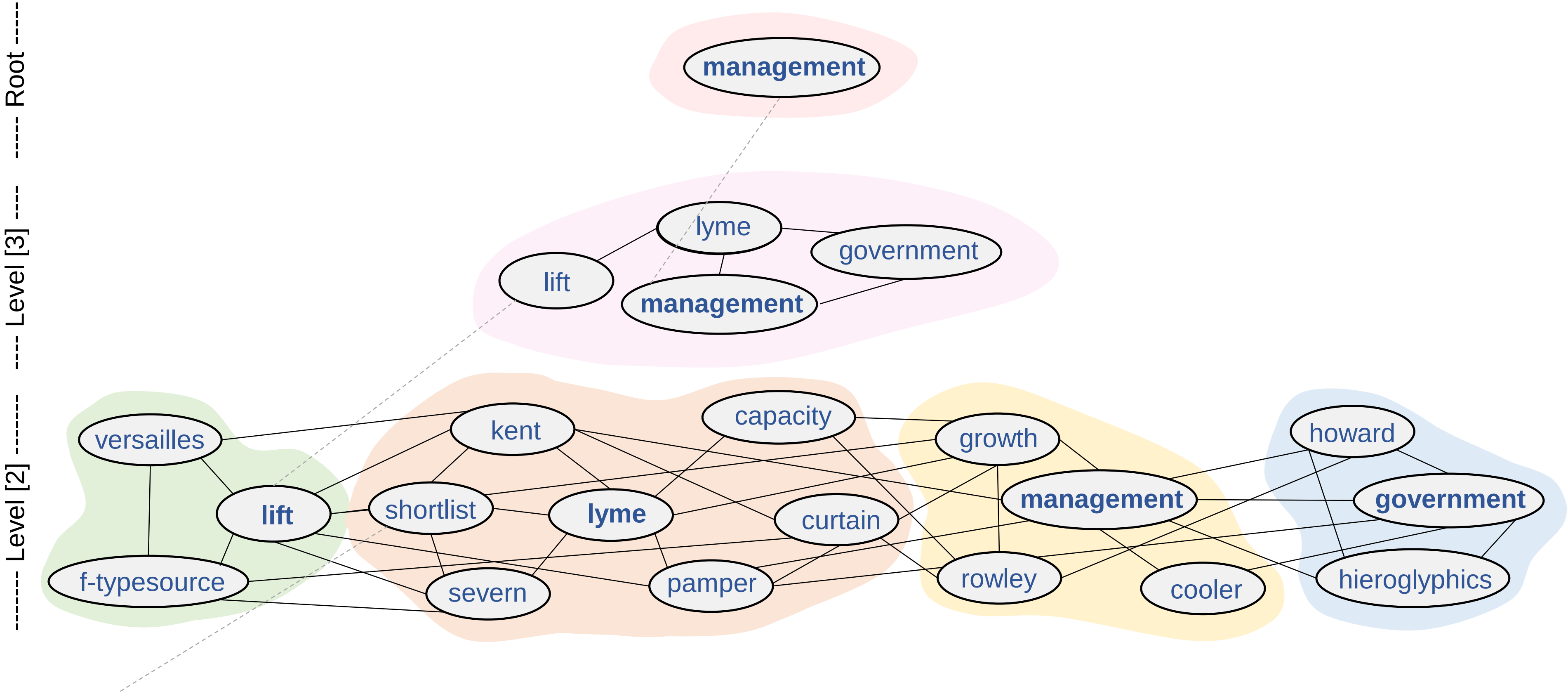
Text Clusting

```
### Cluster no.811(192)--[[festival]]###
[byron][post][rock][july][festival][demand][talent][display][pop][bay][grass][folk]
[b][headliner][hip-hop][malone][lumineer][cardi][splendour][matthew][fitzpatrick][24-year-old][membership][remainder]
(24)

### Cluster no.836(193)--[[noon]]###
[switch][rhythm][hawke][endurance][parabola][circle][spoil][noon][correction][error][millisecond](11)

### Cluster no.856(194)--[[bacon]]###
[hint][confit][garlic][bacon][muhammara][chilly][labneh][serve][regret](9)

### Cluster no.862(195)--[[strada]]###
[vehicle][status][position][palm][comment][posture][inferno][cocoon][sit][knuckle][alcantara][trim]
[sensor][scan][complement][interplay][dual-camera][variance][everyones][strada][merit](21)

### Cluster no.869(196)--[[adjustment]]###
[hit][adjustment][hat][bonus][endgame][reintroduction][cameo](7)

### Cluster no.875(197)--[[mixe]]###
[twin][neumann][mixe][midrange][beryllium][tweeter](6)

### Cluster no.887(198)--[[sofa]]###
[king][sofa][binge][leon][laskowski][fatboy][slim][lana][rey][franz][ferdinandy][ma]
(12)

### Cluster no.959(201)--[[patrick]]###
[master][day][thing][statement][career][human][water][club][bit][tour][top][winner]
[debut][calendar][daniel][roar][pga][ryder][danny][willett][patrick][englands][dreamers][faldos]
[jordan][wgc-hsbc][rookie][berger][toilet][bum][adult][bce][playground](33)

### Cluster no.962(202)--[[antarctica]]###
[october][coast][january][rest][kilometre][area][photo][direction][ice][earth][kitchen][steve]
[mile][shelf][iceberg][antarctica][crack][brunt][fracture][nasa][observatory][satellite][rift][halloween]
[borough][fry][caddie][williams](28)

### Cluster no.979(203)--[[welly]]###
[age][fact][business][person][specimen][welly][mud][shore][valuable](9)
```

```
### Cluster no.962(202)--[[antarctica]]###
[october][coast][january][rest][kilometre][area][photo][direction][ice][earth][kitchen][steve]
[mile][shelf][iceberg][antarctica][crack][brunt][fracture][nasa][observatory][satellite][rift][halloween]
[borough][fry][caddie][williams](28)
```

# Hierarchies - Results

Create a hierarchy of 100 documents (On Level 0 : 4688 words, 186 clusters) ▼

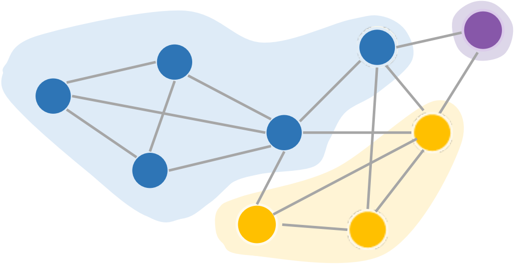


------ Level [1] ------

[ **shortlist**, ffp, 4200mm, objective, ir, swazis, association, authority ] [8]
[ **lyme**, darcy, district, pemberley, connection, worcestershire, villa, fashion ] [8]
[ **kent**, sussex, gainsborough, footpath, cleanliness, miracle ] [6]
[ **lift**, vauxhalls, fault ] [3]
[ **growth**, fiesta, today, hasnt, ct, harald, felipe, mg2, hamlet, kai-shek, volunteer ] [11]
[ **management**, incarnation, affair, conduct, colonisation, staple, photoshop ] [7]
[ **pamper**, wallow, cupful ] [3]
[ **government**, delivery, textbook ] [3]
[ **curtain**, filigree, stead ] [3]
[ **severn**, worshipper, cloister, triad, beefeater, aswan ] [6]

[ **f-typesource**, ollington, shift, oil-pan, wait, touche, tempest ] [7]
[ **versailles**, diana, secrecy ] [3]
[ **howard**, marquess, documentation, maker, steam, stephenson, sorbonne, backer, greenwich, cream, cape, horseshoe-shap, coffin, butter, navigate, medway, oasis ] [17]
[ **hieroglyphics**, manuscript, sub-brand ] [3]
[ **rowley**, sf, relation, plexiglass ] [4]
[ **capacity**, osborne, benchrest, raise, americas, woking ] [6]
[ **cooler**, block, wilderness ] [3] [ **lift**, vauxhalls, fault ] [3]

# Conclusion



- ✓ Words and Relations between words were **inserted** into a co-occurrence graph.
- ✓ The Clusters start to **build** after a new word was added.
- ✓ Words can **find** the cluster where shall be **assigned** to.
  - ✓ Words were added or moved to another cluster.
  - ✓ A new cluster was **created**, if it don’t have cluster to assign.
  - ✓ When the cluster center (Centroid) **move**, the range of that cluster was **recalculated**.
- ✓ Hierarchical clustering was **build** after a stable state of the lower clusters was reached (repeat until only one node remains).

# Thank you

for your attention

Contact:

Supaporn Simcharoen

Lehrgebiet Kommunikationsnetze,
Fakultät für Mathematik und Informatik,
FernUniversität in Hagen

supaporn.simcharoen@gmail.com