# Automatic Labeling of Word Clusters

Group 11
Rajeshreddy Datla(CS17RESCH11007)
Dewalkar Swapnil Ashok(CS17MTECH11004)

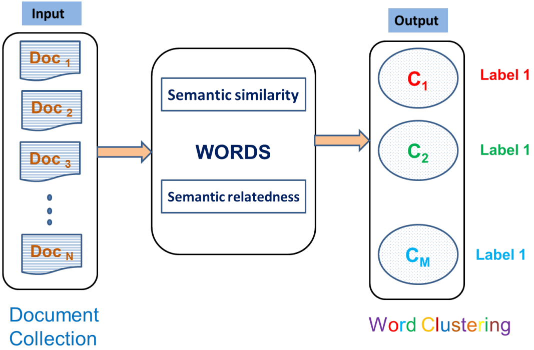
21 November 2017

*Information Retrieval project.*

# Problem statement

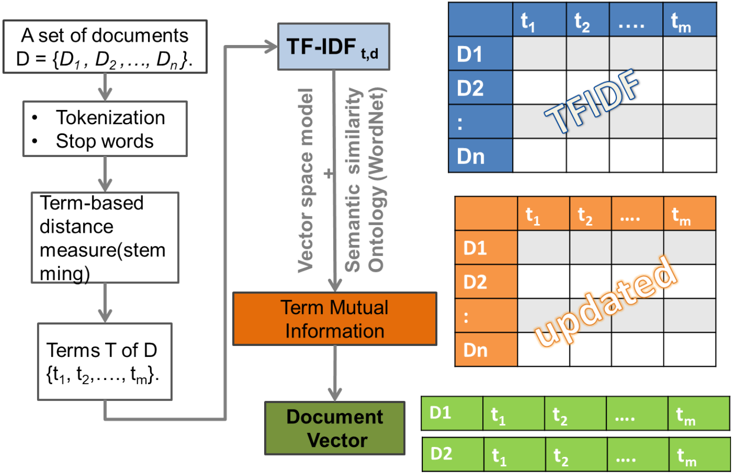
- Given a Document collection, Partition words into clusters such that words that belong to the same cluster are semantically similar and related.
- Assign labels automatically to the clusters.

# Introduction



- Word clustering is a technique for partitioning sets of words into subsets of semantically similar words.

# Data processing


A set of documents D = {$D_1$, $D_2$,…, $D_n$}.
• Tokenization
• Stop words
Term-based distance measure(stemming)
Terms T of D {$t_1$, $t_2$,…., $t_m$}.
TF-IDF $_{t,d}$
Vector space model
+
Semantic similarity Ontology (WordNet)
Term Mutual Information
Document Vector
TFIDF
updated

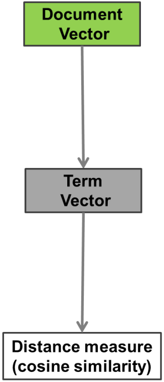
|  | $t_1$ | $t_2$ | …. | $t_m$ |
|---|---|---|---|---|
| D1 |  |  |  |  |
| D2 |  |  |  |  |
| : |  |  |  |  |
| Dn |  |  |  |  |

|  | $t_1$ | $t_2$ | …. | $t_m$ |
|---|---|---|---|---|
| D1 |  |  |  |  |
| D2 |  |  |  |  |
| : |  |  |  |  |
| Dn |  |  |  |  |

|  |  |  |  |  |
|---|---|---|---|---|
| D1 | $t_1$ | $t_2$ | …. | $t_m$ |
| D2 | $t_1$ | $t_2$ | …. | $t_m$ |

L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In Workshop on Text Mining, SIAM International Conference on Data Mining, Bethesda, MD, 2006. SIAM

# Data processing Contd..



| | | | | |
|---|---|---|---|---|
| D1 | $t_1$ | $t_2$ | .... | $t_m$ |
| D2 | $t_1$ | $t_2$ | .... | $t_m$ |
| Dn | $t_1$ | $t_2$ | .... | $t_m$ |

| | | | | |
|---|---|---|---|---|
| $t_1$ | D1 | D2 | .... | Dn |
| $t_2$ | D1 | D2 | .... | Dn |
| $t_m$ | D1 | D2 | .... | Dn |

| | $t_1$ | $t_2$ | ... | $t_m$ |
|---|---|---|---|---|
| $t_1$ | | | | |
| $t_2$ | | | | |
| : | | | | |
| $t_m$ | | | | |

L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In Workshop on Text Mining, SIAM International Conference on Data Mining, Bethesda, MD, 2006. SIAM

# Clustering methods

- Agglomerative Clustering
- Markov Clustering
- Affinity Propagation Clustering

# Agglomerative Clustering

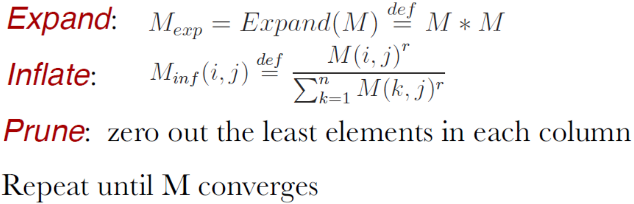
**Basic algorithm:**

- Compute the distance matrix between the input data points
- Let each data point be a cluster
- Repeat
  - Merge the two closest clusters
  - Update the distance matrix
- Until only a single cluster remains
- Key operation is the computation of the distance between two clusters
- Different definitions of the distance between clusters lead to different algorithms

- R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" (PDF). The Computer Journal. British Computer Society. 16 (1): 3034. doi:10.1093/comjnl/16.1.30.
- D. Defays (1977). "An efficient algorithm for a complete-link method". The Computer Journal. British Computer Society. 20 (4): 364366. doi:10.1093/comjnl/20.4.364.

# Markov Clustering

## Basic Algorithm



*Expand*: $M_{exp} = Expand(M) \overset{def}{=} M * M$

*Inflate*: $M_{inf}(i,j) \overset{def}{=} \frac{M(i,j)^r}{\sum_{k=1}^{n} M(k,j)^r}$

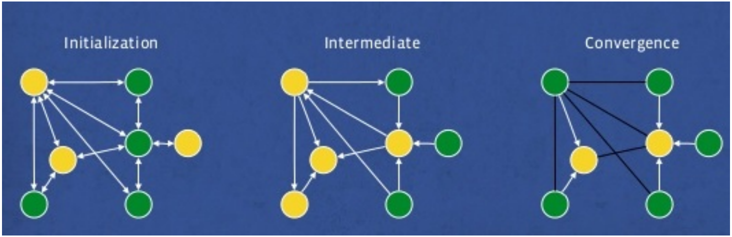
*Prune*: zero out the least elements in each column

Repeat until M converges

Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.

# Affinity propagation clustering

- It is a clustering algorithm based on the concept of ”message passing” between data points.
- Discover exemplars based on similarity.



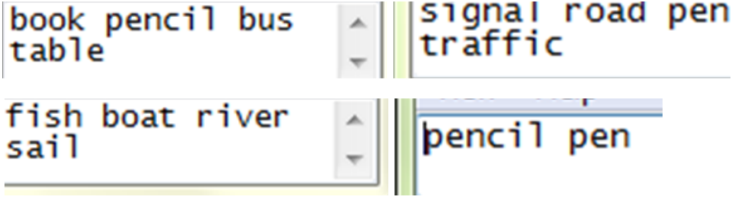
Frey, B. J. and Dueck, D, Clustering by passing messages between data points. 315, 972–976, Science 2007.

# Labeling

*Labeling using Heuristics based on keywords*

- Input: List of keywords from Document collection.
- Select top scored words from each cluster.
- Similarity measure between these top scored words in a clusetr and keyword.
- Select the keyword with best similarity score.
- Obtain hypernym of the selected keyword and consider that as label for that cluster

# Experiments

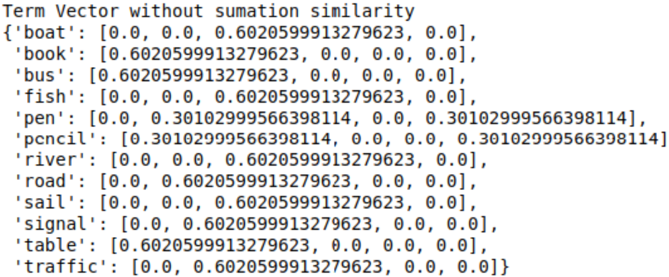


book pencil bus table

signal road pen traffic

fish boat river sail

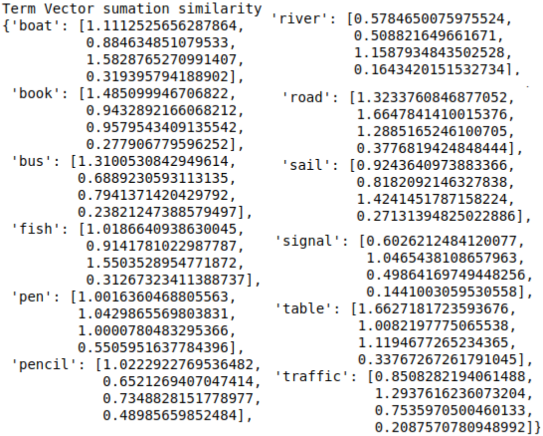
pencil pen



```
Term Vector without sumation similarity
{'boat': [0.0, 0.0, 0.6020599913279623, 0.0],
 'book': [0.6020599913279623, 0.0, 0.0, 0.0],
 'bus': [0.6020599913279623, 0.0, 0.0, 0.0],
 'fish': [0.0, 0.0, 0.6020599913279623, 0.0],
 'pen': [0.0, 0.30102999566398114, 0.0, 0.30102999566398114],
 'pencil': [0.30102999566398114, 0.0, 0.0, 0.30102999566398114]
 'river': [0.0, 0.0, 0.6020599913279623, 0.0],
 'road': [0.0, 0.6020599913279623, 0.0, 0.0],
 'sail': [0.0, 0.0, 0.6020599913279623, 0.0],
 'signal': [0.0, 0.6020599913279623, 0.0, 0.0],
 'table': [0.6020599913279623, 0.0, 0.0, 0.0],
 'traffic': [0.0, 0.6020599913279623, 0.0, 0.0]}
```

# Results

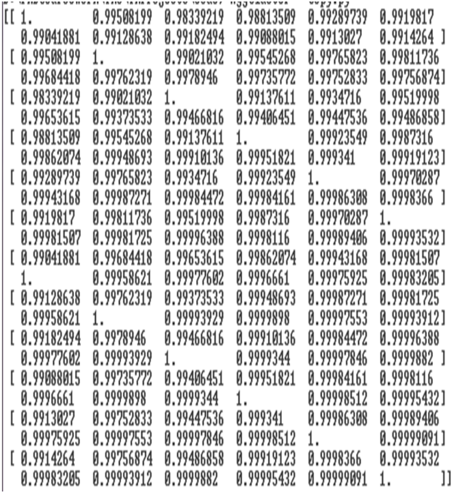
## Mutual Information Matrix



```
Term Vector sumation similarity
{'boat': [1.112525656287864,
          0.884634851079533,
          1.5828765270991407,
          0.319395794188902],
 'book': [1.485099946706822,
          0.9432892166068212,
          0.9579543409135542,
          0.277906779596252],
 'bus': [1.3100530842949614,
         0.6889230593113135,
         0.7941371420429792,
         0.23821247388579497],
 'fish': [1.0186640938630045,
          0.9141781022987787,
          1.5503528954771872,
          0.31267323411388737],
 'pen': [1.0016360468805563,
         1.0429865569803831,
         1.0000780483295366,
         0.550595163778439 6],
 'pencil': [1.0222922769536482,
            0.6521269407047414,
            0.7348828151778977,
            0.48985659852484],
 'river': [0.5784650075975524,
           0.508821649661671,
           1.1587934843502528,
           0.1643420151532734],
 'road': [1.3233760846877052,
          1.6647841410015376,
          1.2885165246100705,
          0.3776819424848444],
 'sail': [0.9243640973883366,
          0.8182092146327838,
          1.4241451787158224,
          0.27131394825022886],
 'signal': [0.6026212484120077,
            1.0465438108657963,
            0.49864169749448256,
            0.1441003059530558],
 'table': [1.6627181723593676,
           1.0082197775065538,
           1.1194677265234365,
           0.33767267261791045],
 'traffic': [0.8508282194061488,
             1.2937616236073204,
             0.7535970500460133,
             0.2087570780948992]}
```

# Results

## Similarity Matrix

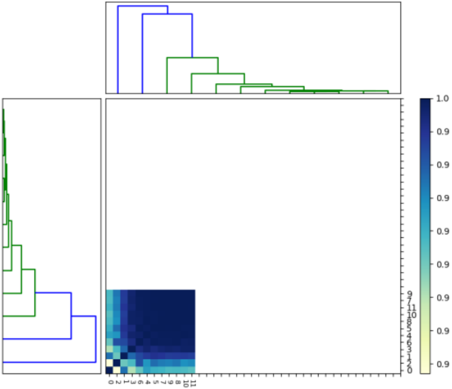


```
[[ 1.          0.99508199  0.98339219  0.98813509  0.99289739  0.9919817
   0.99041881  0.99128638  0.99182494  0.99088015  0.9913027   0.9914264 ]
 [ 0.99508199  1.          0.99021032  0.99545268  0.99765823  0.99811736
   0.99684418  0.99762319  0.9978946   0.99735772  0.99752833  0.99756874]
 [ 0.98339219  0.99021032  1.          0.99137611  0.9934716   0.99519998
   0.99653615  0.99373533  0.99466816  0.99406451  0.99447536  0.99486858]
 [ 0.98813509  0.99545268  0.99137611  1.          0.99923549  0.9987316
   0.99862074  0.99948693  0.99910136  0.99951821  0.999341    0.99919123]
 [ 0.99289739  0.99765823  0.9934716   0.99923549  1.          0.99970287
   0.99943168  0.99987271  0.99984472  0.99984161  0.99986308  0.9998366 ]
 [ 0.9919817   0.99811736  0.99519998  0.9987316   0.99970287  1.
   0.99981507  0.99981725  0.99996388  0.9998116   0.99989406  0.99993532]
 [ 0.99041881  0.99684418  0.99653615  0.99862074  0.99943168  0.99981507
   1.          0.99958621  0.99977602  0.9996661   0.99975925  0.99983205]
 [ 0.99128638  0.99762319  0.99373533  0.99948693  0.99987271  0.99981725
   0.99958621  1.          0.99993929  0.9999898   0.99997553  0.99993912]
 [ 0.99182494  0.9978946   0.99466816  0.99910136  0.99984472  0.99996388
   0.99977602  0.99993929  1.          0.9999344   0.99997846  0.9999882 ]
 [ 0.99088015  0.99735772  0.99406451  0.99951821  0.99984161  0.9998116
   0.9996661   0.9999898   0.9999344   1.          0.99998512  0.99995432]
 [ 0.9913027   0.99752833  0.99447536  0.999341    0.99986308  0.99989406
   0.99975925  0.99997553  0.99997846  0.99998512  1.          0.99999091]
 [ 0.9914264   0.99756874  0.99486858  0.99919123  0.9998366   0.99993532
   0.99983205  0.99993912  0.9999882   0.99995432  0.99999091  1.        ]]
```

# Results

Agglomerative Clustering

- Single link

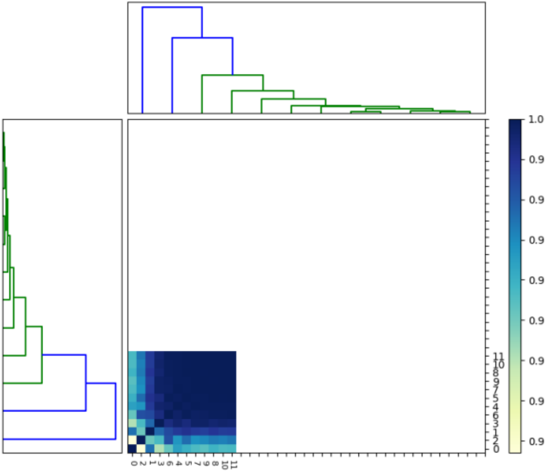


**0: 'pencil', 1: 'bus', 2: 'signal', 3: 'sail', 4: 'pen', 5: 'book', 6: 'traffic', 7: 'boat', 8: 'table', 9: 'river', 10: 'fish', 11: 'road'**

# Results

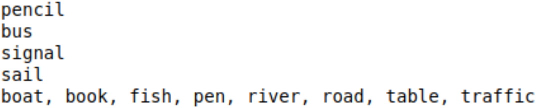
## Agglomerative Clustering

- complete link



# Results

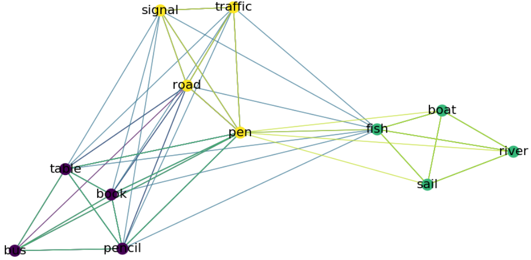
Affinity propagation Clustering Number of clusters: 5



```
pencil
bus
signal
sail
boat, book, fish, pen, river, road, table, traffic
```

# Results

Markov Clustering

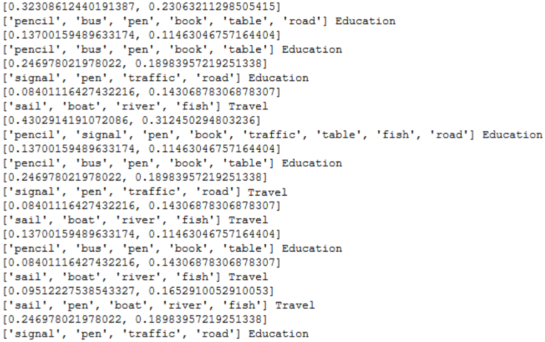


# Cluster validity

**Silhoutte score:** The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample.
The Silhouette Coefficient for a sample is $(b - a)/max(a, b)$. Silhoutte score: 0.59327577 (Affinity propagation)

# Labeling

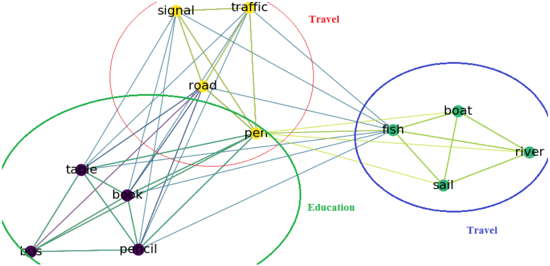
Based on markov clustering



```
[0.32308612440191387, 0.23063211298505415]
['pencil', 'bus', 'pen', 'book', 'table', 'road'] Education
[0.13700159489633174, 0.11463046757164404]
['pencil', 'bus', 'pen', 'book', 'table'] Education
[0.246978021978022, 0.18983957219251338]
['signal', 'pen', 'traffic', 'road'] Education
[0.08401116427432216, 0.14306878306878307]
['sail', 'boat', 'river', 'fish'] Travel
[0.4302914191072086, 0.312450294803236]
['pencil', 'signal', 'pen', 'book', 'traffic', 'table', 'fish', 'road'] Education
[0.13700159489633174, 0.11463046757164404]
['pencil', 'bus', 'pen', 'book', 'table'] Education
[0.246978021978022, 0.18983957219251338]
['signal', 'pen', 'traffic', 'road'] Travel
[0.08401116427432216, 0.14306878306878307]
['sail', 'boat', 'river', 'fish'] Travel
[0.13700159489633174, 0.11463046757164404]
['pencil', 'bus', 'pen', 'book', 'table'] Education
[0.08401116427432216, 0.14306878306878307]
['sail', 'boat', 'river', 'fish'] Travel
[0.09512227538543327, 0.1652910052910053]
['sail', 'pen', 'boat', 'river', 'fish'] Travel
[0.246978021978022, 0.18983957219251338]
['signal', 'pen', 'traffic', 'road'] Education
```

# Labeling

Based on markov clustering



# References

- Lichi Yuan,Word Clustering Algorithms Based on Word Similarity, 7th IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015.
- L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang. Ontology-based distance measure for text clustering. In Workshop on Text Mining, SIAM International Conference on Data Mining, Bethesda, MD, 2006. SIAM
- R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" (PDF). The Computer Journal. British Computer Society. 16 (1): 3034. doi:10.1093/comjnl/16.1.30.
- D. Defays (1977). "An efficient algorithm for a complete-link method". The Computer Journal. British Computer Society. 20 (4): 364366. doi:10.1093/comjnl/20.4.364.

# THANK YOU