Master of Technology in Knowledge Engineering

Text Mining

Clustering

Institute of Systems Science
National University of Singapore

Clustering in General

Motivation Example

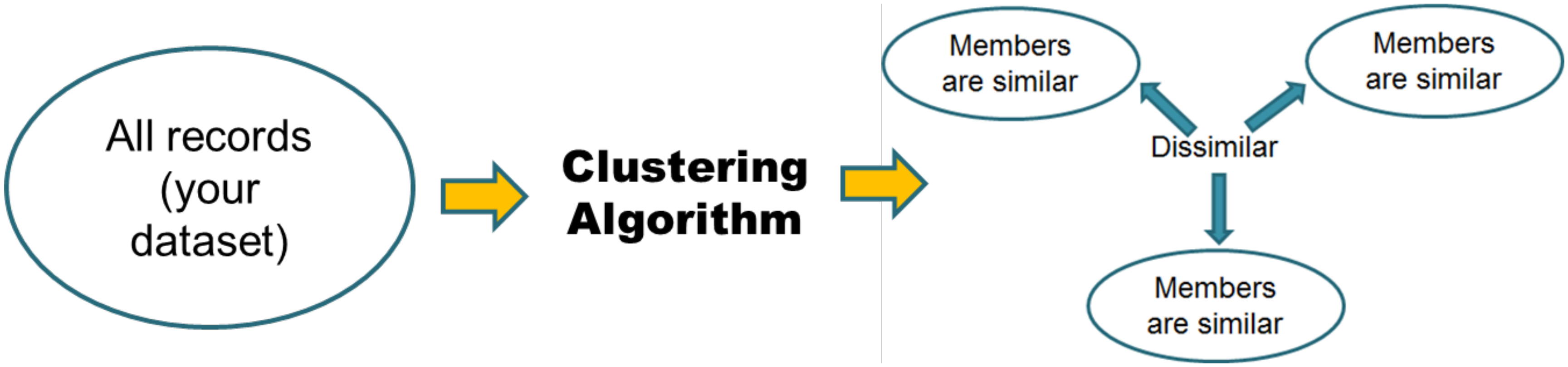
- **<u>Who</u>** are in the queue
  - Age? Gender? Race? Political-alignment? Employed? HDB? Own-car?
- **<u>What</u>** do they bet on:
  - 4D? Toto? Sweep? Sports?
- **<u>How</u>** often and How much do they bet
- You **intuitively know** that there are different types of betters:
  - We often talk about “addicts”
  - What about other stereotypes?

Source: asiaone

How do we go about grouping people?

Clustering

- *Clustering*, or cluster analysis, is the process of automatically identifying similar items to group them together into clusters.
  - *Unsupervised learning* –no labeled training examples need to be supplied; no prior knowledge of the number of groups,
  - Originated in the fields of statistics and data mining, used on numerical data



Applications of Cluster Analysis

Cluster Analysis is versatile and can be used in many business problems across many domains:

- Sales & Marketing: help marketers discover groups in their customer databases, and then use this insight to develop more targeted marketing campaigns
- Fraud Detection: Identify groups of customers whose transaction behavior is uncharacteristic
- Balanced portfolios: Selecting securities from different clusters can help create a balanced portfolio (for a better risk management)

Major Clustering Algorithms

- **Hierarchical Clustering**
  - Iteratively groups documents into cascading sets of clusters.
  - Top-down (divisive) approach – Items are split iteratively based on their similarity measures.
  - Bottom-up (agglomerative) – Items are joined together iteratively.
- **Partitioning Clustering**
  - Constructs various partitions and then evaluate them by some criterion
  - Most popular type – k-means and its variants (k-medoids and k-medians)

Hierarchical Clustering Example

Agglomerative Clustering:

There are *N* records in the dataset.

Step 1: Assign each record as its own cluster (i.e. N clusters)

Step 2: Calculate the **distances*** between each cluster

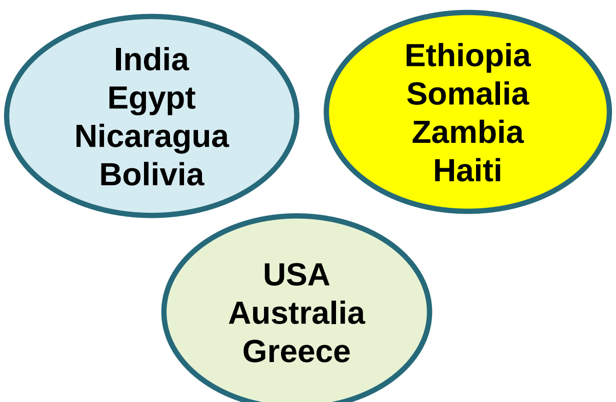
Step 3: Find the closest pair of clusters and merge them into a single (larger) cluster

Step 4: Repeat Step 2-3, until all records are clustered into a single cluster of size *N*.

***e.g. Single link method**- The distance between two clusters is equal to the distance between the two closest records in them, aka *nearest neighbor method*.

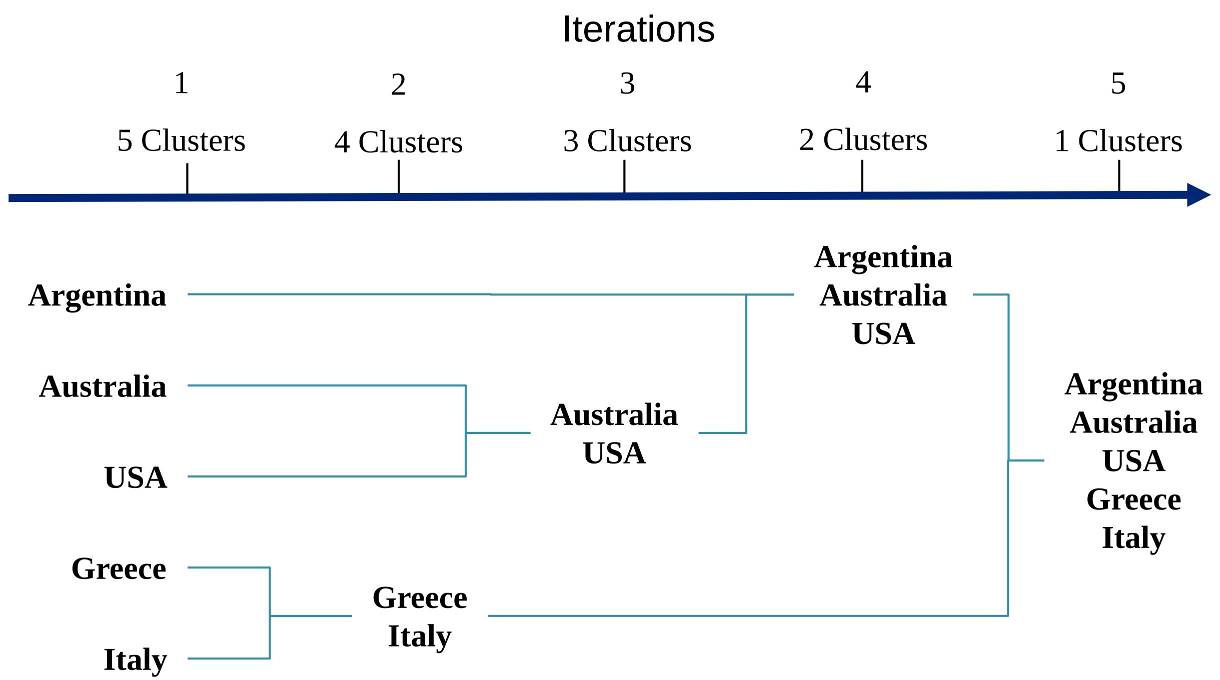
Example Data

- Objective is to find groupings of countries that share similar characteristics in terms of:
  - Literacy
  - Baby Mortality
  - Births
  - Deaths



Country	Literacy	Baby Mort	Birth Rate	Death Rate
Argentina	95	25.6	20	9
Australia	100	7.3	15	8
Bolivia	78	75	34	9
Cameroon	54	77	41	12
Chile	93	14.6	23	6
China	78	52	21	7
Costa Rica	93	11	26	4
Egypt	48	76.4	29	9
Ethiopia	24	110	45	14
Greece	93	8.2	10	10
Haiti	53	109	40	19
India	52	79	29	10
Indonesia	77	68	24	9
Italy	97	7.6	11	10
Kenya	69	74	42	11
Kuwait	73	12.5	28	2
Mexico	87	35	28	5
Nicaragua	57	52.5	35	7
Nigeria	51	75	44	12
Phillippine:	90	51	27	7
Somalia	24	126	46	13
Thailand	93	37	19	6
USA	97	8.1	15	9
Vietnam	88	46	27	8
Zambia	73	85	46	18

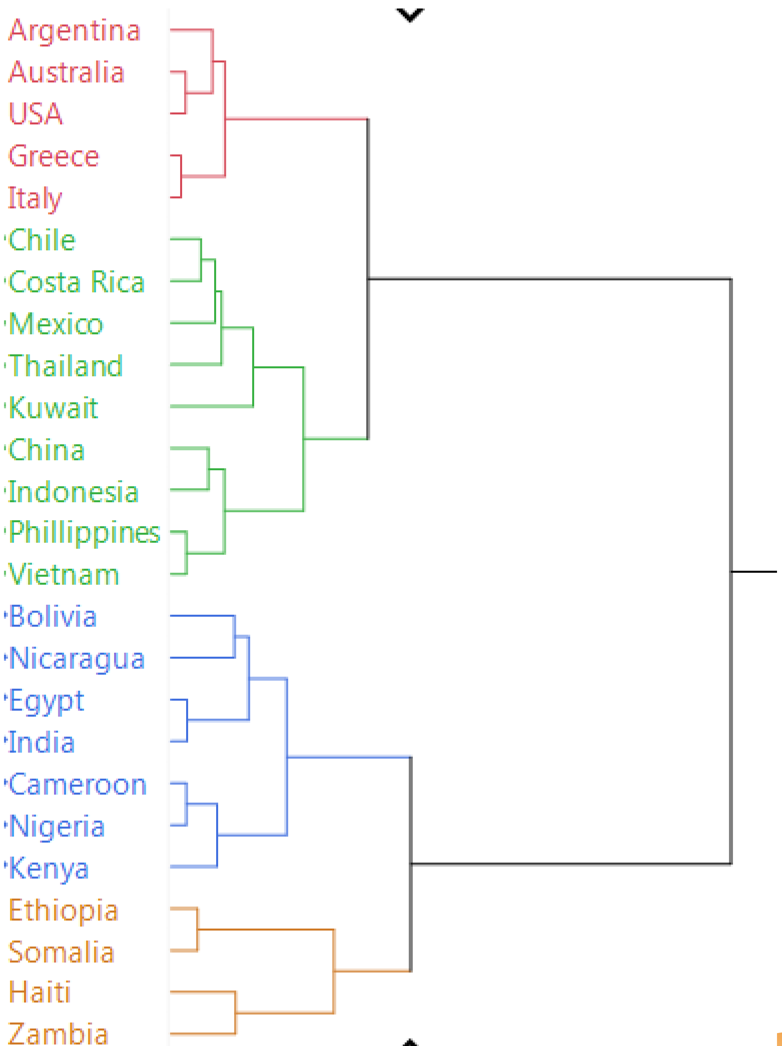
Agglomerative Example



In the first iteration, Greece & Italy are the closest

Output of Hierarchical Clustering

- Notice the pairing sequence from left to right.
  1. Greece + Italy
  2. Australia + USA
  3. Philippines + Vietnam
  4. Cameroon + Nigeria …
- This color display is for 4 clusters
  - C1: Argentina … Italy
  - C2: Chile … Vietnam
  - C3: Bolivia … Kenya
  - C4: Ethiopia … Zambia



K-means (Partition) Clustering

Specify *K* number of clusters that you want.

Step 1: Arbitrarily designate *k* records as seed points

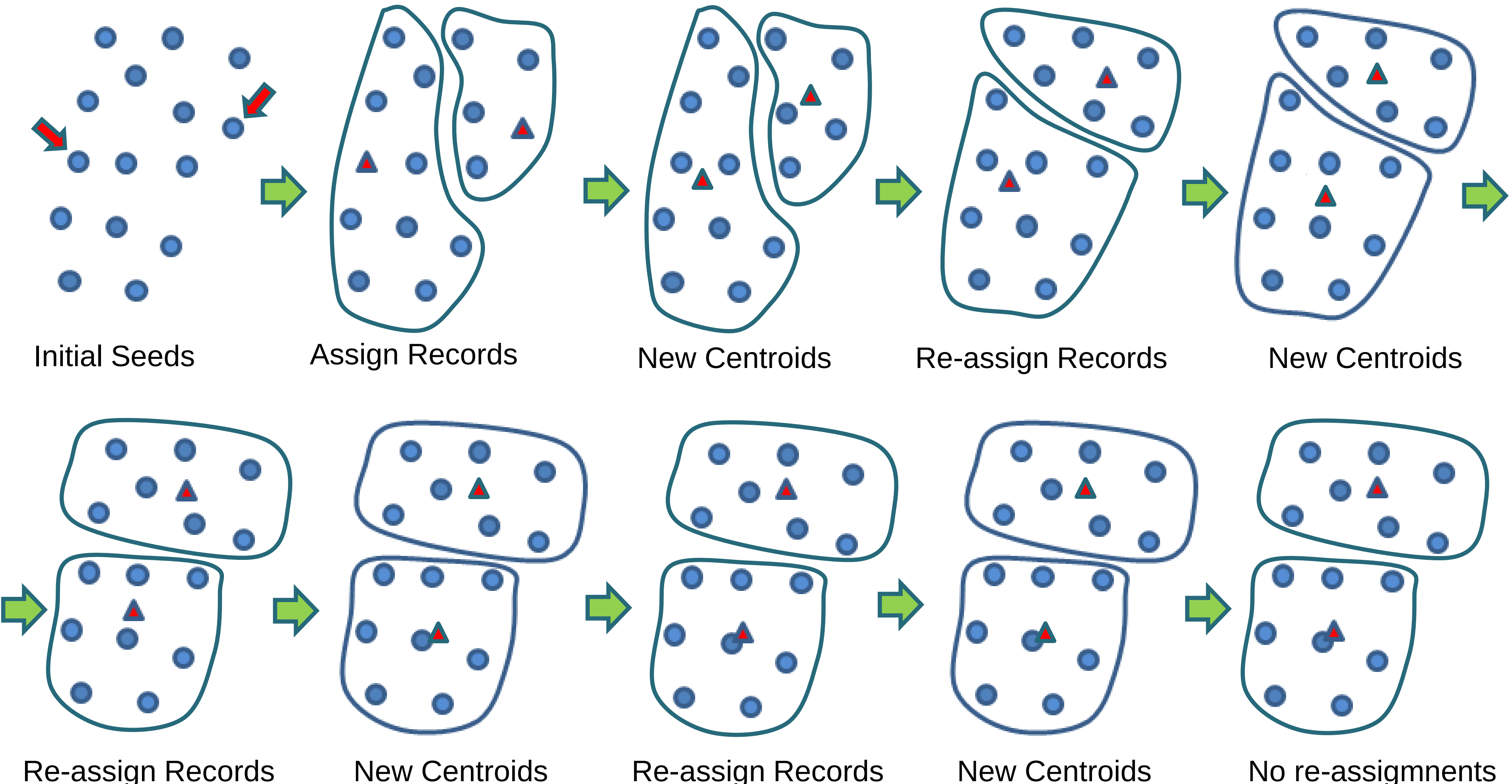
Each record is then assigned to the nearest seed and clusters are created

Step 2: Calculate the cluster centroids and the distances between each record and the centroids

Step 3: Re-assign all records to the nearest cluster (some records may remain in the same cluster)

Step 4: Repeat Step 2-3, until no more re-assignment is possible

K-Means Example



Partition vs Hierarchical

- Hierarchical clustering algorithm generates small clusters of homogenous records that are nested within large clusters of less homogenous records.
  - Advantage – ability to analyze sub-clusters in the hierarchy
  - Disadvantage – more computing resources, slower
  - Useful when there is no intuition on the # of groups
- Partition clustering algorithm generates partitions of non-overlapping records with no hierarchical relationships between them.
  - Advantage – fast clustering
  - Disadvantage – no flexibility to analyze sub-clusters
  - Useful when you have an intuition on # of groups

Interpreting Cluster Analysis Output

- Can the clusters be explained in practical terms by experience, expectations or **domain knowledge**
  - Provide **descriptive labels** for each cluster
- Look at **distinguishing** characteristics of each cluster
  - There should be substantial differences between clusters
- Look at the **cluster quality** or goodness-of-fit
  - This is a measure of similarity and dissimilarity (**Silhouette** refers to a method of interpretation and validation of consistency within clusters of data) Poor: -1 to 0.2; Fair: 0.2 to 0.5; Good: 0.5 to 1

Clustering in Text Mining

-Clustering similar documents
- Clustering similar words

Document Clustering

Document Clustering vs. Text Classification

- Document/Text Classification (*supervised* learning)
  - Looks at stored examples with correct answers and projects answers for new examples.
  - The answers, or predetermined class labels, must be available.

- Document Clustering (*unsupervised* learning)
  - Groups together documents with similar content into the same cluster.
  - The number of the clusters and their labels are not know before clustering.
  - Ideally each document is very similar to the other documents in its cluster and much less similar to documents in other clusters

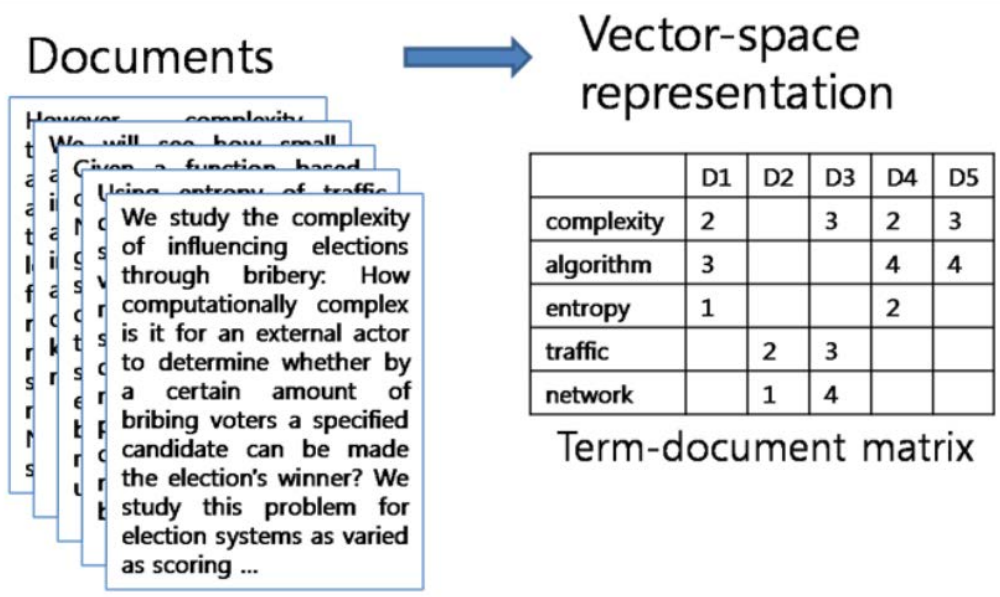
Document Clustering

Summarize the reviews;

Summarize students profiles;

Fake claim, fake review;

Document Clustering

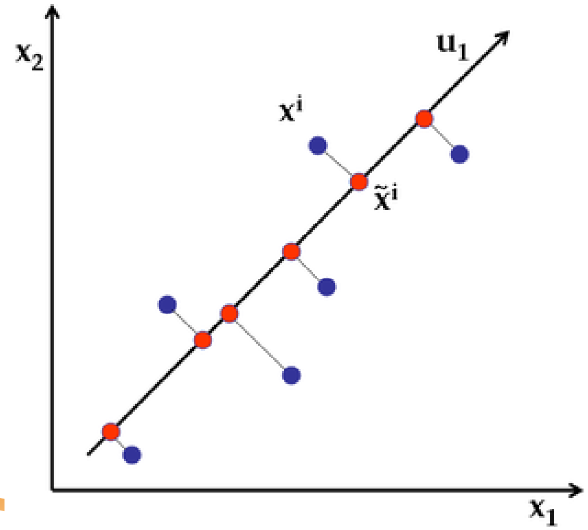


	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

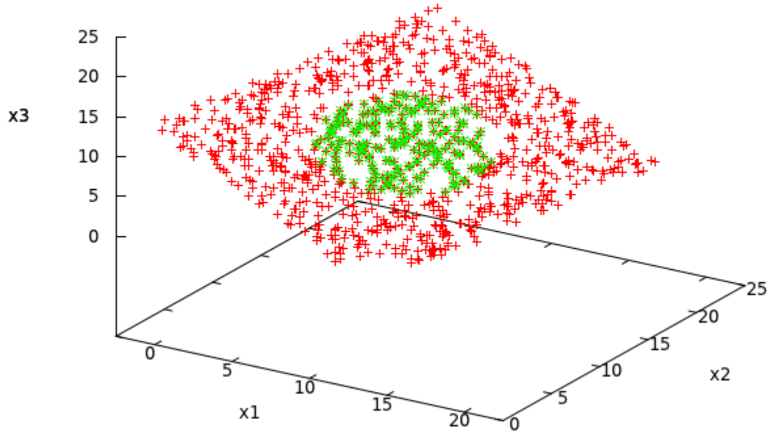
Dimensional Reduction

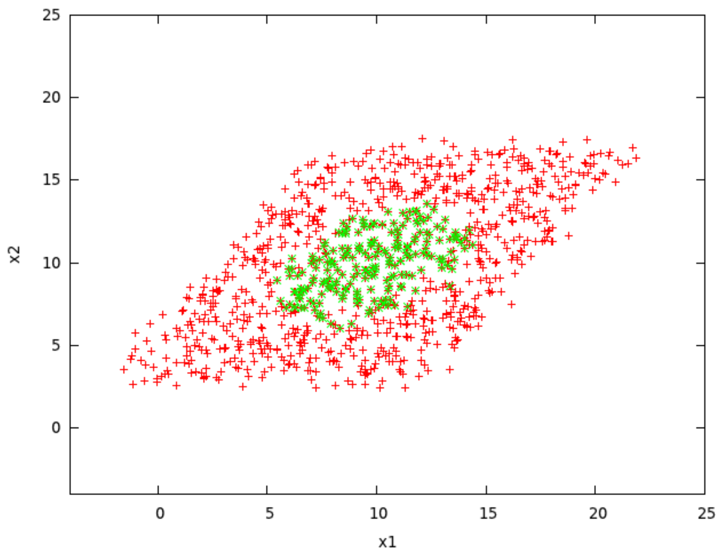
- With big document collections, the dimension of the vector space may easily range into tens of thousands.
- Approach – dimension reduction
  - By mapping a high-dimensional feature space to a much lower dimensional subspace
  - *Singular Value Decomposition*.



Dimensional Reduction

- Approach – dimension reduction





Singular Value Decomposition

- The singular value decomposition of a matrix A is the factorization of A into the product of three matrices $A = UDV^T$ where the columns of U and V are orthonormal and the matrix D is diagonal with positive real entries.

- A: Input data matrix. E.g., m documents, n terms

- It is always possible to decompose a real matrix A into $A = UDV^T$

Singular Value Decomposition

- U, D, V unique
- U, V:

  Columns are orthogonal and unit vectors

  U data record similarity

  V variable similarity

- D: diagonal

  Entries (singular values) are positive and sorted in decreasing order

Singular Value Decomposition

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Singular Value Decomposition

- UD Gives the coordinates of the points in the projection axis

Project of the data records on the new axis:

$$\begin{bmatrix} 1.61 & 0.19 & -0.01 \\ 5.08 & 0.66 & -0.03 \\ 6.82 & 0.85 & -0.05 \\ 8.43 & 1.04 & -0.06 \\ 1.86 & -5.6 & 0.84 \\ 0.86 & -6.93 & -0.87 \\ 0.86 & -2.75 & 0.41 \end{bmatrix}$$

Singular Value Decomposition

How to further do dimension reduction?

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Singular Value Decomposition

How to further do dimension reduction?

$$
? = \begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}
$$

Singular Value Decomposition

How to further do dimension reduction?

$$\begin{bmatrix} 0.92 & 0.95 & 0.92 & 0.01 & 0.01 \\ 2.91 & 3.01 & 2.91 & -0.01 & -0.01 \\ 3.90 & 4.04 & 3.90 & 0.01 & 0.01 \\ 4.82 & 5.00 & 4.82 & 0.03 & 0.03 \\ 0.7 & 0.53 & 0.7 & 4.11 & 4.11 \\ -0.69 & 1.34 & -0.69 & 4.78 & 4.78 \\ 0.32 & 0.23 & 0.32 & 2.01 & 2.01 \end{bmatrix} \approx \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}$$

Singular Value Decomposition

- D_new= [5,0,0,0,0]
- Then project into new space

- D_new_space=D_new*V=[2.8, 0.6]

- D_new2=[0,4,5,0,0]
- D_new_space=[5.2, 0.4]

Usefulness of SVD

- Generally appropriate for data reduction in text mining
- Not useful if the purpose of the analytical project is to identify the specific phrases or terms that are important and related to key performance indicators (e.g., which phrases in physicians' notes are predictive of subsequent health care costs)
- Computationally expensive

Labeling the Clusters

- A cluster can be labeled with a small number of carefully selected words distinguishing the cluster from others.
  - Documents are composed of words and the distribution of words is the basis of document clustering
  - We can select:
    - Most frequent words in a cluster
- One or more exemplar documents may also be selected as "typical" documents to represent the cluster
  - E.g. the document that is most similar to the cluster mean vector

Topic Modeling

What is topic modeling?

Topic modelling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents.

Latent Dirichlet Allocation(LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

What is topic modeling?

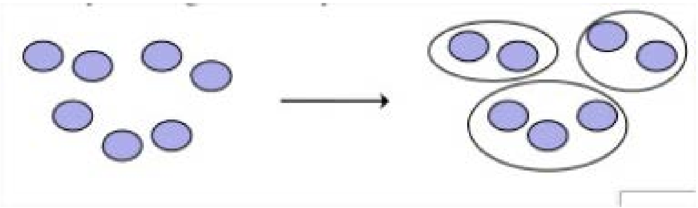
LDA:

The number of topics is difficult to decide;

Bag of words (the sentence structure is not modeled);

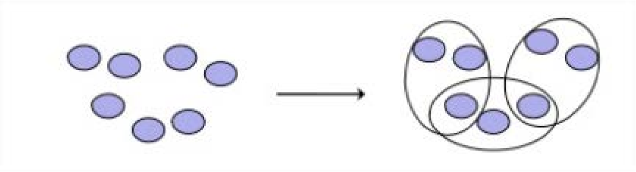
Hard Clustering:

Every object belong to one cluster



Soft Clustering:

Objects may belong to several clusters



What is topic modeling?

LDA:

Document 1: I had a peanut butter sandwich for breakfast.
Document 2: I like to eat almonds, peanuts and walnuts.
Document 3: My neighbor got a little dog yesterday.
Document 4: Cats and dogs are mortal enemies.
Document 5: You mustn't feed peanuts to your dog.

Topic 1: 30% peanuts, 15% almonds, 10% breakfast… (you can interpret that this topic deals with food)
Topic 2: 20% dogs, 10% cats, 5% peanuts… (you can interpret that this topic deals with pets or animals)

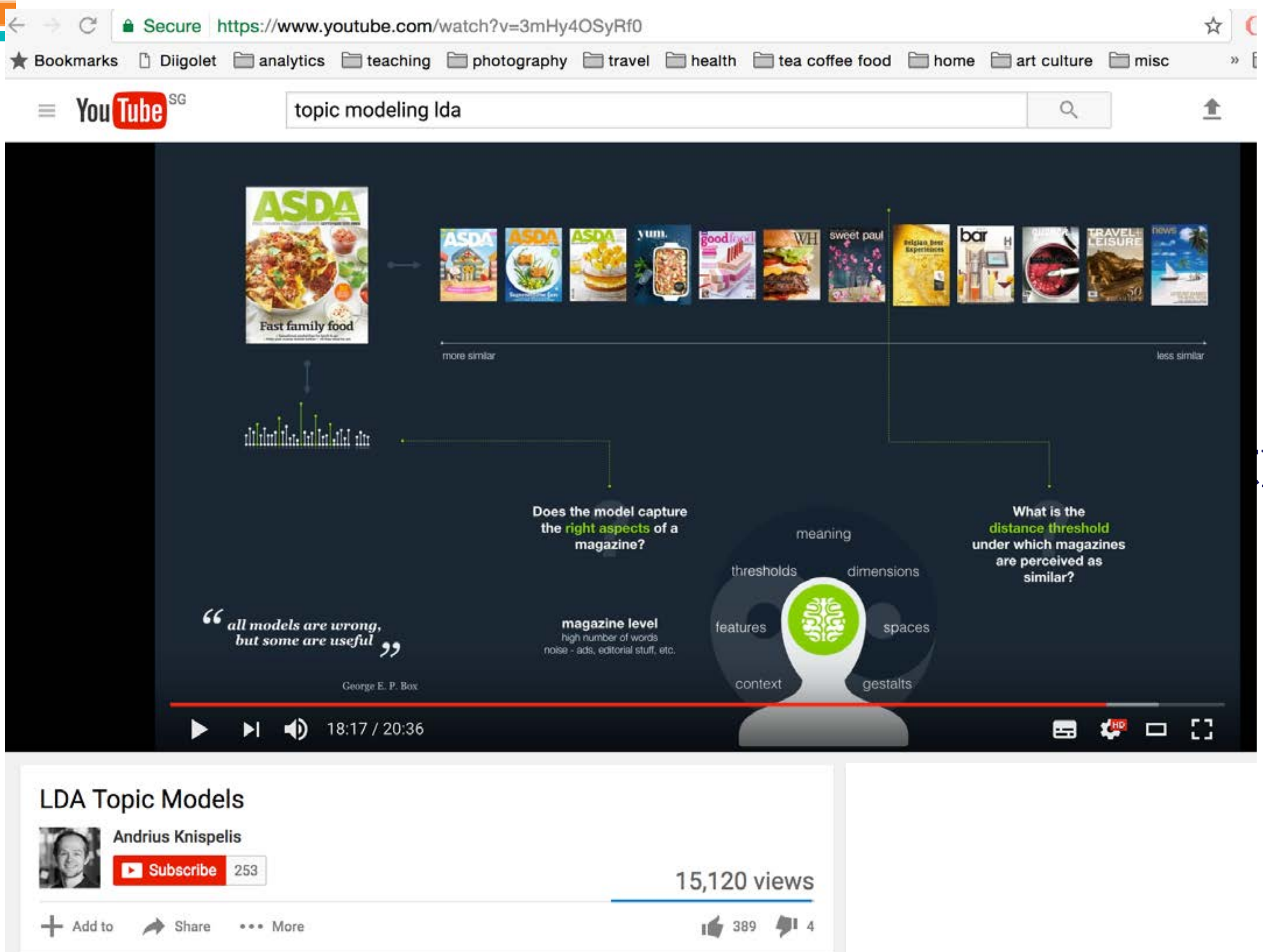
What is topic modeling?

LDA:

Document 1: I had a peanut butter sandwich for breakfast.
Document 2: I like to eat almonds, peanuts and walnuts.
Document 3: My neighbor got a little dog yesterday.
Document 4: Cats and dogs are mortal enemies.
Document 5: You mustn't feed peanuts to your dog.

Documents 1 and 2: 100% Topic 1
Documents 3 and 4: 100% Topic 2
Document 5: 70% Topic 1, 30% Topic 2

...ion



From: https://www.youtube.com/watch?v=3mHy4OSyRf0

What is topic modeling?

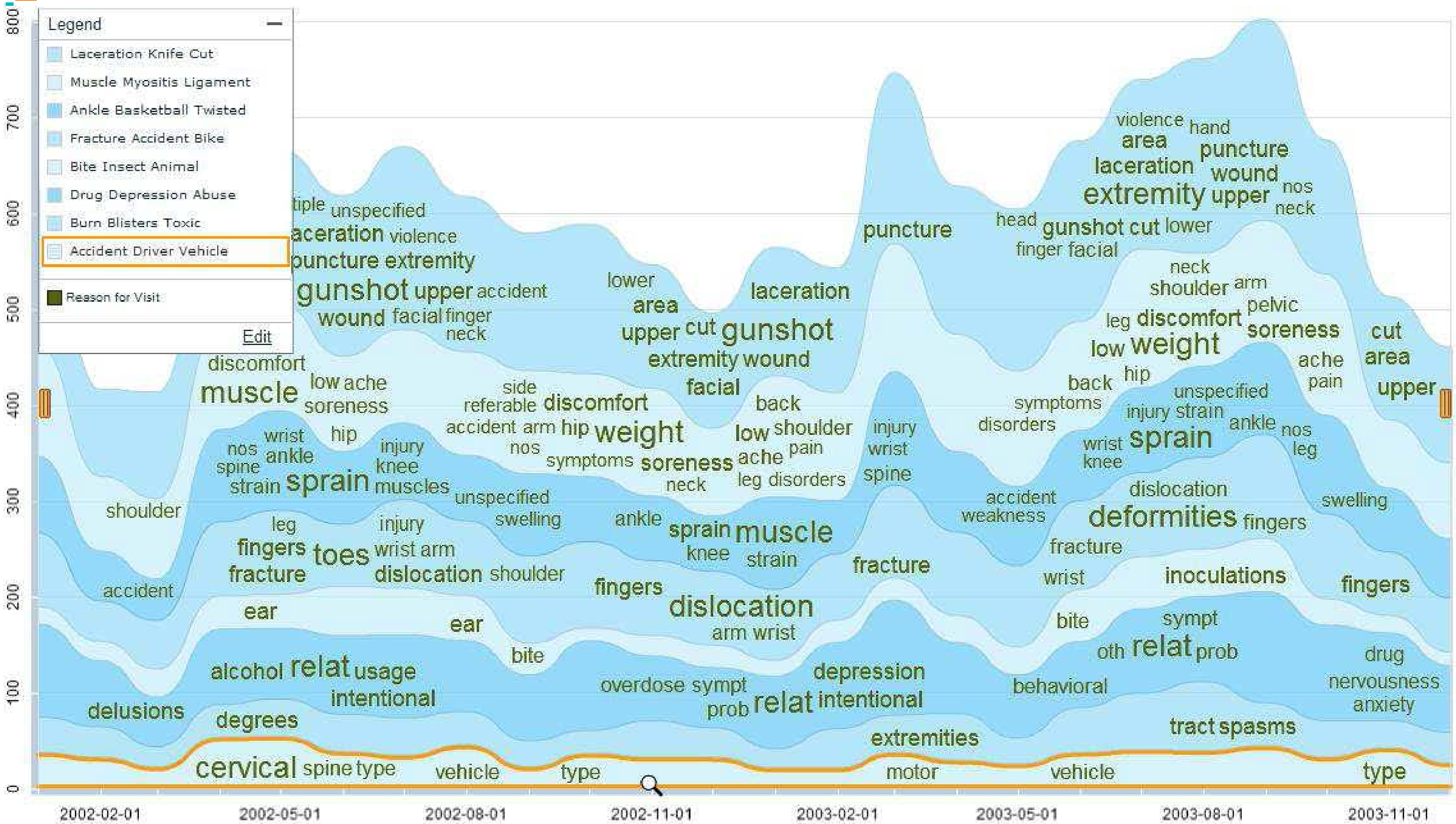
LDA:

Alpha:

Turn it down and the documents will likely have less of a mixture of topics. Turn it up and the documents will likely have more of a mixture of topics.

.

TIARA: Summary of "Reason for Visit"

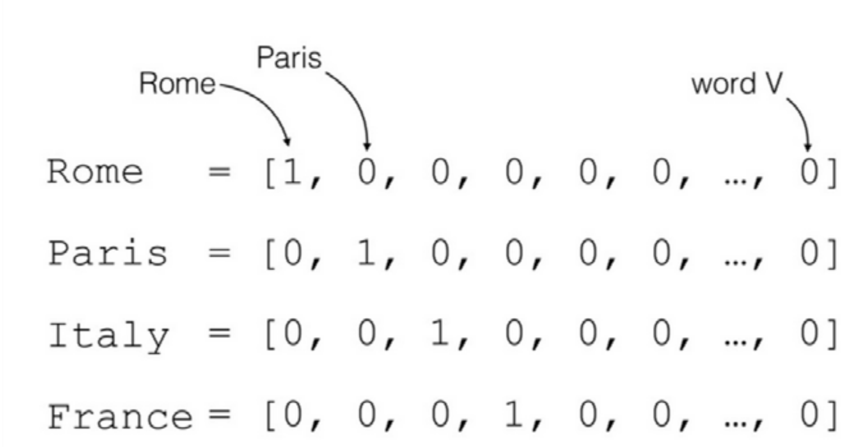


Word Clustering

Word Embedding

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where **words or phrases from the vocabulary are mapped to vectors of real numbers.**

One-Hot encoding:
A one hot encoding is a representation of categorical variables as binary vectors. Each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.



Word Embedding

Count Vector:

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

TF-IDF Vector:

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

tf-idf

	angeles	los	new	post	times	york
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	0.584
d3	1.584	1.584	0	0	0.584	0

Word Embedding

Co-occurrence Vectors:

I love Programming. I love Math. I tolerate Biology

Define window size = 1

This means that each word will be defined by its neighboring word to the left as well as the one to the right.

	I	love	Programming	Math	tolerate	Biology	.
I	0	2	0	0	1	0	2
love	2	0	1	1	0	0	0
Programming	0	1	0	0	0	0	1
Math	0	1	0	0	0	0	1
tolerate	1	0	0	0	0	1	0
Biology	0	0	0	0	1	0	1
.	1	0	1	1	0	1	0

Word Embedding

Co-occurrence Vectors:

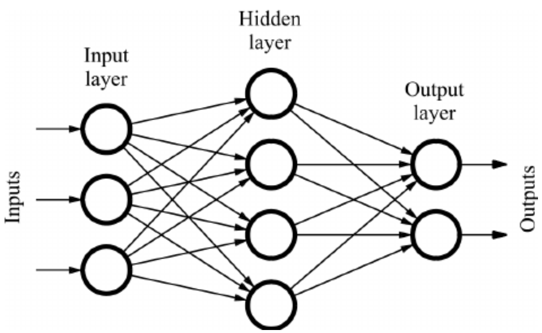
Programming’ and ‘Math’ share the same co-occurrence values, they would be placed in the same place; meaning that in this context they mean the same thing

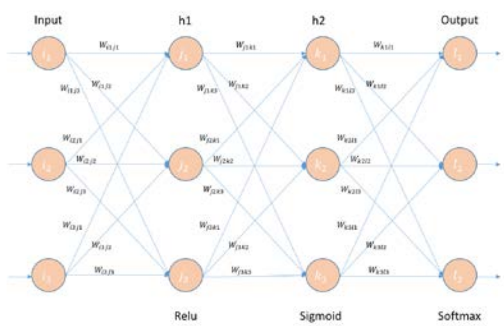
It preserves the semantic relationship between words.

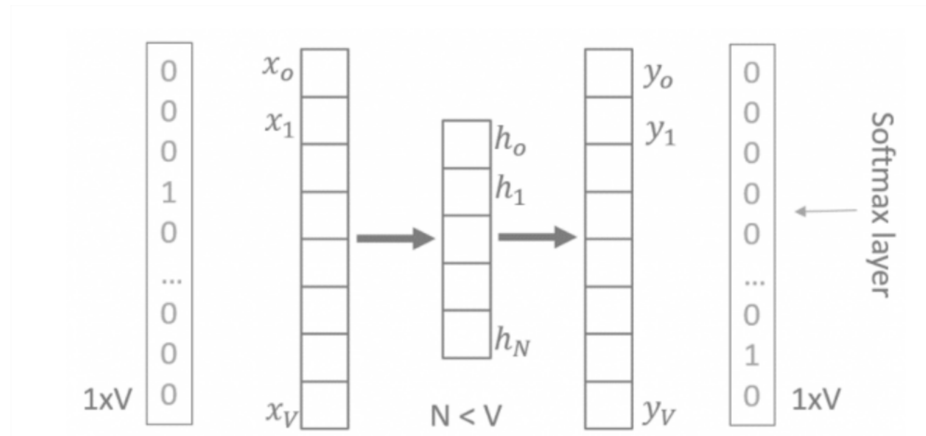
Computationally expensive since we are talking about a very high-dimensional space.

Word Embedding

Word2Vec Embedding

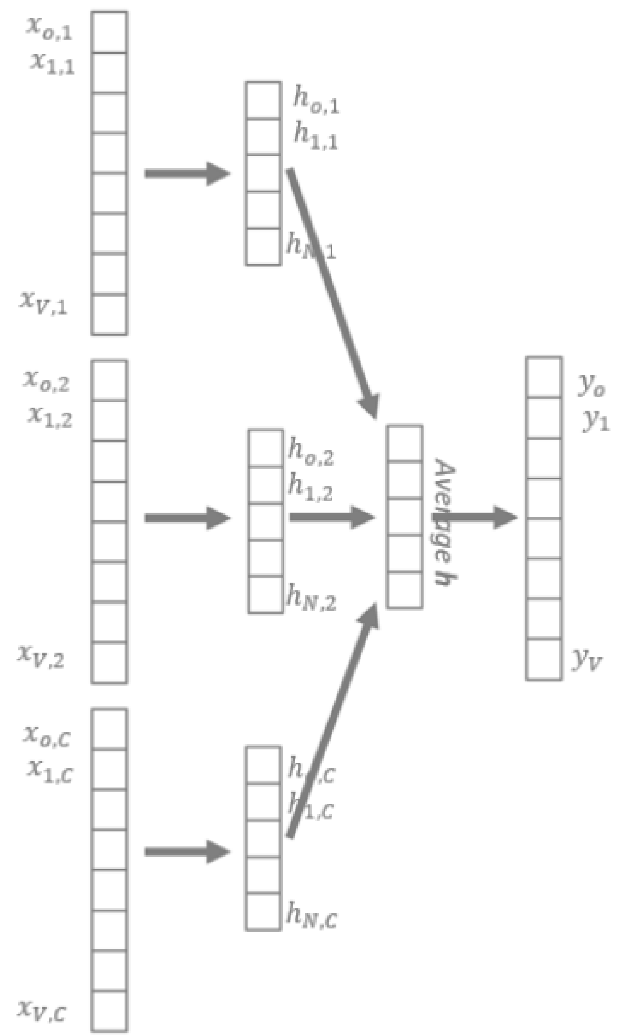






Word Embedding

Word2Vec Embedding



Word Clustering

- Words can be clustered in two ways.

1. By meaning
   - Grouping together semantically similar words into a cluster (or concept)

2. By co-occurrence
   - Grouping words that commonly appear together

Clustering semantically similar words

- It's also referred to as *Concept Extraction* in some literature.
- Useful in grouping and typing domain concepts.
- The context-dependent nature of word meaning

  "You shall know a word by the company it keeps."

  – J. R. Firth (1957)

- Words with similar meaning appear in similar context

  E.g. "dogs", "cats", "fish", "birds", "hamsters"…

  – Referring to household pets

  – Used in the same context

How to cluster semantically similar words?

- Clustering on the similarity between the contexts in which the words appear
- Check the context by using the Co-occurrence matrix
- Apply clustering algorithms (e.g. *k-means*) with an appropriate distance measure (e.g. *cosine distance*)

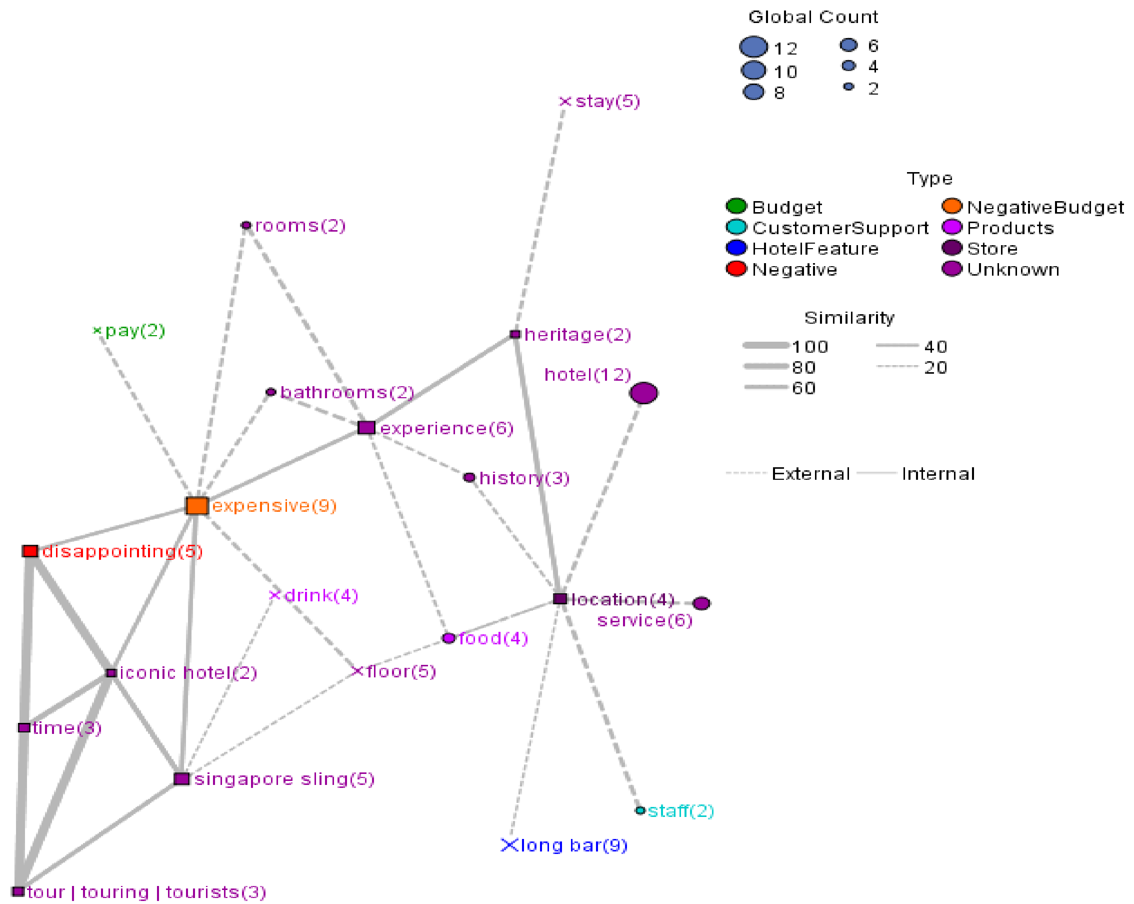
How to cluster based co co-occurrence words?

- Words appearing together in the same document

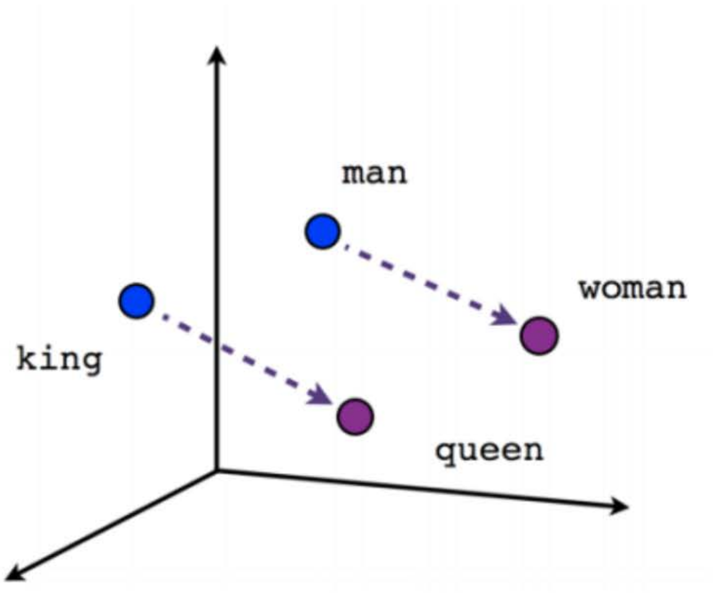
- Apply clustering algorithms (e.g. *k-means*))

Word Cluster Visualization – Co-occurrence Based

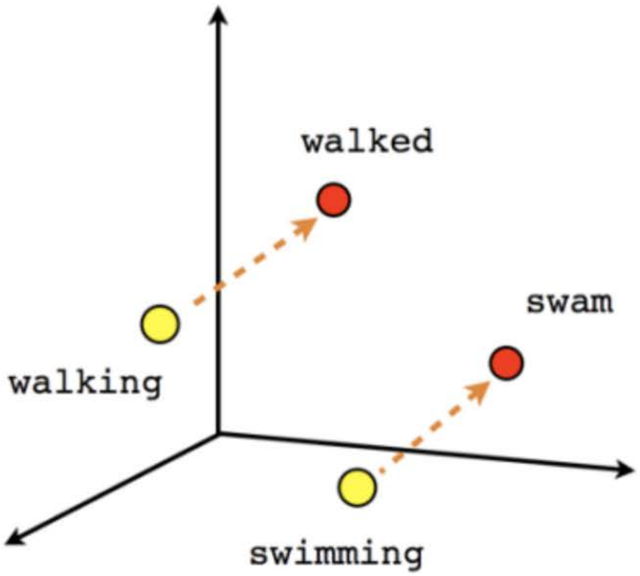
- From SPSS Modeler TA



Word Cluster Visualization Word2Vec Based



Male-Female



Verb tense

Summary

- Clustering is an important technique for data exploration and understanding.
- Cluster requires functions to measure the similarity between data objects, and algorithms to efficiently compare and cluster them.
- Text clustering is used to group together documents or words based on similarity. Document clustering is useful for exploring and understanding how documents are related, whereas word clustering can discover words sharing topical or semantic meaning and words that co-occur frequently.

References

- P. Arabie, L.J. Hubert, G. De Soete. Singapore; River Edge (Ed). *Clustering and classification* , NJ : World Scientific, 1996.
- Gary Miner, John Elder IV et. al. Chapter 13 Clustering Words and Documents, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, 2012
- Weiss, Indurkhya, & Zhang. Chapter 5 Finding Structure in a Document Collection, *Fundamentals of Predictive Text Mining*, Springer, 2010.
- R. Mack, M Hehenberger. Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discovery Today*, 7 (11), 2002
- Albright, Russell. Taming Text with the SVD. SAS, *January* 7, 2004.
- Manning, Chris, and Hinrich Schütze. Collocations. *Foundations of statistical natural language processing* (1999): 141-77.