Lecture 2: Data Analytics of Narrative

Data Analytics of Narrative: Pattern Recognition in Text, and Text Synthesis, Supported by the Correspondence Analysis Platform. This Lecture is presented in three parts, as follows.

Part 1 Data analytics of narrative.

Part 2 Analysis of narrative: tracking emotion in the film, Casablanca. Synthesis of narrative: collective, collaborative authoring of a novel.

Part 3 Ultrametric embedding.

Lecture 2: Data Analytics of Narrative

Data Analytics of Narrative: Pattern Recognition in Text, and Text Synthesis, Supported by the Correspondence Analysis Platform.

1. A short review of the theory and practical implications of Correspondence Analysis.
2. Analysis of narrative: tracking emotion in the film, Casablanca.
3. Synthesis of narrative: collective, collaborative authoring of a novel.
4. Towards semantic rating.

- “We call distribution of a word the set of all its possible environments” (Z.S. Harris)

- “We call distribution of a word the set of all its possible environments” (Z.S. Harris)
- Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.

- “We call distribution of a word the set of all its possible environments” (Z.S. Harris)
- Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.
- Developed in Rennes, Laboratoire de calcul de la Faculté des Sciences de Rennes, by Jean-Paul Benzécri. Subsequently in Paris, Université P. & M. Curie, Paris 6.

- “We call distribution of a word the set of all its possible environments” (Z.S. Harris)
- Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.
- Developed in Rennes, Laboratoire de calcul de la Faculté des Sciences de Rennes, by Jean-Paul Benzécri. Subsequently in Paris, Université P. & M. Curie, Paris 6.
- “The model should follow the data, not the reverse!” (In J.P. Benzécri, “Statistical analysis as a tool to make patterns emerge from data”, in *Methodologies of Pattern Recognition*, Ed. Watanable, NY: Academic, 1969.)

- “We call distribution of a word the set of all its possible environments” (Z.S. Harris)
- Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.
- Developed in Rennes, Laboratoire de calcul de la Faculté des Sciences de Rennes, by Jean-Paul Benzécri. Subsequently in Paris, Université P. & M. Curie, Paris 6.
- “The model should follow the data, not the reverse!” (In J.P. Benzécri, “Statistical analysis as a tool to make patterns emerge from data”, in *Methodologies of Pattern Recognition*, Ed. Watanable, NY: Academic, 1969.)
- So: Description first – priority. Inductive philosophy.

Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.

Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.
- By endowing the cross-tabulation matrix with the χ^2 metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.

Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.
- By endowing the cross-tabulation matrix with the χ^2 metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.
- A hierarchical clustering is induced on the Euclidean space, the factor space.

Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.
- By endowing the cross-tabulation matrix with the χ^2 metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.
- A hierarchical clustering is induced on the Euclidean space, the factor space.
- Interpretation is through projections of observations, attributes or clusters onto factors. The factors are ordered by decreasing importance.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- I is the set of observation indexes, and J is the set of attribute indexes.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- I is the set of observation indexes, and J is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- I is the set of observation indexes, and J is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- I is the set of observation indexes, and J is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously.
- What we have described here is taking numbers of occurrences into relative frequencies.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- I is the set of observation indexes, and J is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously.
- What we have described here is taking numbers of occurrences into relative frequencies.
- The conditional distribution of f_J knowing $i \in I$, also termed the jth profile with coordinates indexed by the elements of I, is:

  $$f_J^i = \{f_j^i = f_{ij}/f_i = (k_{ij}/k)/(k_i/k); f_i > 0; j \in J\}$$

  and likewise for f_I^j.

Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f_J^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.

Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f_J^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.
- Included in this expression is the fact that the cloud of observations, $N_J(I)$, is a subset of the real space of dimensionality $|J|$ where $|.|$ denotes cardinality of the attribute set, J.

Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f_J^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.
- Included in this expression is the fact that the cloud of observations, $N_J(I)$, is a subset of the real space of dimensionality $|J|$ where $|.|$ denotes cardinality of the attribute set, J.
- The overall inertia is as follows:

$$M^2(N_J(I)) = M^2(N_I(J)) = \| f_{IJ} - f_I f_J \|^2_{f_I f_J}$$

$$= \sum_{i \in I, j \in J} (f_{ij} - f_i f_j)^2 / f_i f_j \qquad (1)$$

Input 2/2

- The term $\|f_{IJ} - f_I f_J\|^2_{f_I f_J}$ is the χ^2 metric between the probability distribution f_{IJ} and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$.

Input 2/2

- The term $\|f_{IJ} - f_I f_J\|^2_{f_I f_J}$ is the χ^2 metric between the probability distribution f_{IJ} and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$.
- Decomposing the moment of inertia of the cloud $N_J(I)$ – or of $N_I(J)$ since both analyses are inherently related – furnishes the principal axes of inertia, defined from a singular value decomposition.

Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

- The χ^2 distance with center f_J between observations i and i' is written as follows in two different notations:

$$d(i,i')^2 = \|f_J^i - f_J^{i'}\|^2_{f_J} = \sum_j \frac{1}{f_j}\left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}}\right)^2 \qquad (2)$$

Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

- The χ^2 distance with center f_J between observations i and i' is written as follows in two different notations:

$$d(i,i')^2 = \| f_J^i - f_J^{i'} \|^2_{f_J} = \sum_j \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2 \tag{2}$$

- In the factor space this pairwise distance is identical. The coordinate system and the metric change. For factors indexed by α and for total dimensionality N ($N = \min \{|I|-1, |J|-1\}$; the subtraction of 1 is since the χ^2 distance is centered and hence there is a linear dependency which reduces the inherent dimensionality by 1) we have the projection of observation i on the αth factor, F_α, given by $F_\alpha(i)$:

$$d(i,i')^2 = \sum_{\alpha=1..N} \left(F_\alpha(i) - F_\alpha(i') \right)^2 \tag{3}$$

Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

- The χ^2 distance with center f_J between observations i and i' is written as follows in two different notations:

$$d(i,i')^2 = \| f_J^i - f_J^{i'} \|_{f_J}^2 = \sum_j \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2 \qquad (2)$$

- In the factor space this pairwise distance is identical. The coordinate system and the metric change. For factors indexed by α and for total dimensionality N ($N = \min \{|I| - 1, |J| - 1\}$; the subtraction of 1 is since the χ^2 distance is centered and hence there is a linear dependency which reduces the inherent dimensionality by 1) we have the projection of observation i on the αth factor, F_α, given by $F_\alpha(i)$:

$$d(i,i')^2 = \sum_{\alpha=1..N} \left(F_\alpha(i) - F_\alpha(i') \right)^2 \qquad (3)$$

- Invariance of distance in equations 2 and 3: *Parseval relation*.

Output 2/2

- In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the dual spaces.

Output 2/2

- In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the dual spaces.
- The eigenvalues associated with the factors, identically in the space of observations indexed by set I, and in the space of attributes indexed by set J, are given by the eigenvalues associated with the decomposition of the inertia.

Output 2/2

- In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the dual spaces.
- The eigenvalues associated with the factors, identically in the space of observations indexed by set I, and in the space of attributes indexed by set J, are given by the eigenvalues associated with the decomposition of the inertia.
- The decomposition of the inertia is a principal axis decomposition, which is arrived at through a singular value decomposition.

Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.
- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)

Important Consequences

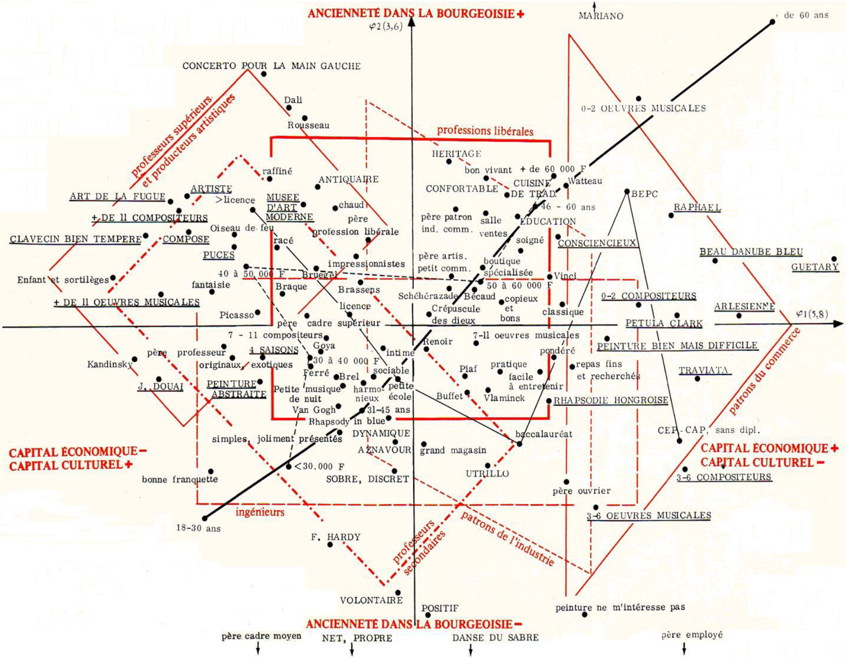
- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.
- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)
- Supplementary elements are observations or attributes retrospectively projected into the factor space.

Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.
- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)
- Supplementary elements are observations or attributes retrospectively projected into the factor space.
- Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.

Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.
- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)
- Supplementary elements are observations or attributes retrospectively projected into the factor space.
- Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.
- Following slide: from Pierre Bourdieu's *La Distinction*, 1979. *A Social Critique of the Judgment of Taste*.



ANCIENNETÉ DANS LA BOURGEOISIE +
φ2 (3,6)
MARIANO
+ de 60 ans
CONCERTO POUR LA MAIN GAUCHE
Dali
Rousseau
professeurs supérieurs et producteurs artistiques
professions libérales
0-2 OEUVRES MUSICALES
HERITAGE
bon vivant
+ de 60 000 F
raffiné
ANTIQUAIRE
CUISINE DE TRAD.
Watteau
ARTISTE
ART DE LA FUGUE
>licence
MUSEE D'ART MODERNE
chaud
CONFORTABLE
BEPC
46 - 60 ans
RAPHAEL
+ DE 11 COMPOSITEURS
père profession libérale
père patron ind. comm.
salle ventes
EDUCATION
CLAVECIN BIEN TEMPERE
COMPOSE
Oiseau de feu
racé
soigné
CONSCIENCIEUX
PUCES
impressionnistes
père artis. petit comm.
boutique spécialisée
BEAU DANUBE BLEU
GUETARY
Enfant et sortilèges
40 à 50 000 F
Bruegel
Vinci
fantaisie
Braque
Brassens
50 à 60 000 F
+ DE 11 OEUVRES MUSICALES
licence
Schéhérazade
Bécaud
copieux et bons
0-2 COMPOSITEURS
ARLESIENNE
Picasso
Crépuscule des dieux
classique
père cadre supérieur
PETULA CLARK
φ1 (5,8)
7 - 11 compositeurs
Renoir
7-11 oeuvres musicales
Goya
PEINTURE BIEN MAIS DIFFICILE
père professeur
4 SAISONS
intime
pondéré
Kandinsky
originaux, exotiques
30 à 40 000 F
sociable
pratique
repas fins et recherchés
TRAVIATA
Ferré
Brel
Piaf
facile à entretenir
J. DOUAI
PEINTURE ABSTRAITE
Petite musique de nuit
harmonieux
petite école
Buffet
Vlaminck
RHAPSODIE HONGROISE
Van Gogh
31-45 ans
Rhapsody in blue
patrons du commerce
DYNAMIQUE
baccalauréat
CEP - CAP, sans dipl.
simples, joliment présentés
AZNAVOUR
grand magasin
CAPITAL ÉCONOMIQUE −
CAPITAL CULTUREL +
CAPITAL ÉCONOMIQUE +
CAPITAL CULTUREL −
<30.000 F
3-6 COMPOSITEURS
bonne franquette
SOBRE, DISCRET
UTRILLO
père ouvrier
ingénieurs
3-6 OEUVRES MUSICALES
18-30 ans
patrons de l'industrie
F. HARDY
professeurs secondaires
VOLONTAIRE
POSITIF
peinture ne m'intéresse pas
ANCIENNETÉ DANS LA BOURGEOISIE −
père cadre moyen
NET, PROPRE
DANSE DU SABRE
père employé

Contributions, Correlations

- Contributions

Contributions, Correlations

- Contributions
- Contribution of i to moment α: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$

Contributions, Correlations

- Contributions
- Contribution of i to moment α: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$
- Correlations

Contributions, Correlations

- Contributions
- Contribution of i to moment α: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$
- Correlations
- Cosine squared of angle between i and factor α.

Contributions, Correlations

- Contributions
- Contribution of i to moment α: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$
- Correlations
- Cosine squared of angle between i and factor α.
- $\cos^2 a = F_\alpha(i)^2/\rho(i)^2$ where $\rho(i)^2 = \|f_J^i - f_J\|^2_{f_J} = \sum_{j\in J}(f_j^i - f_j)^2/f_j$

Contributions, Correlations

- Contributions
- Contribution of i to moment α: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$
- Correlations
- Cosine squared of angle between i and factor α.
- $\cos^2 a = F_\alpha(i)^2/\rho(i)^2$ where $\rho(i)^2 = \|f_J^i - f_J\|^2_{f_J} = \sum_{j \in J}(f_j^i - f_j)^2/f_j$
- Contributions *determine* the factor space, correlations *illustrate* it.

Hierarchical Clustering

- Consider the projection of observation i onto the set of all factors indexed by α, $\{F_\alpha(i)\}$ for all α, which defines the observation i in the new coordinate frame.

Hierarchical Clustering

- Consider the projection of observation i onto the set of all factors indexed by α, $\{F_\alpha(i)\}$ for all α, which defines the observation i in the new coordinate frame.
- This new factor space is endowed with the (unweighted) Euclidean distance, d.

Hierarchical Clustering

- Consider the projection of observation i onto the set of all factors indexed by α, $\{F_\alpha(i)\}$ for all α, which defines the observation i in the new coordinate frame.
- This new factor space is endowed with the (unweighted) Euclidean distance, d.
- We seek a hierarchical clustering that takes into account the observation sequence, i.e. observation i precedes observation i' for all $i, i' \in I$. We use the linear order on the observations.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
- Repeat this merge step until only one cluster remains.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
- Repeat this merge step until only one cluster remains.
- Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set I.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
- Repeat this merge step until only one cluster remains.
- Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set I.
- It can be shown that the closeness value, given by d, at each agglomerative step is strictly non-decreasing.

Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
- Repeat this merge step until only one cluster remains.
- Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set I.
- It can be shown that the closeness value, given by d, at each agglomerative step is strictly non-decreasing.
- That is, if cluster c_3 is formed earlier in the series of agglomerations compared to cluster c_4, then the corresponding distances will satisfy $d_{c3} \leq d_{c4}$. (d here is as determined in the merge step of the algorithm above.)

Example of Hierarchy Without and With Inversion

- Inversions in the sequence of agglomerations.
- That is, i and j merge, and the distance of the this new cluster to another cluster is smaller than the dening distance of the $i; j$ merger.
- Hence, there is non-monotonic change in the level index, given by the distance dening the merger agglomeration.

Hierarchy (not sequence-constrained, 30 terms)

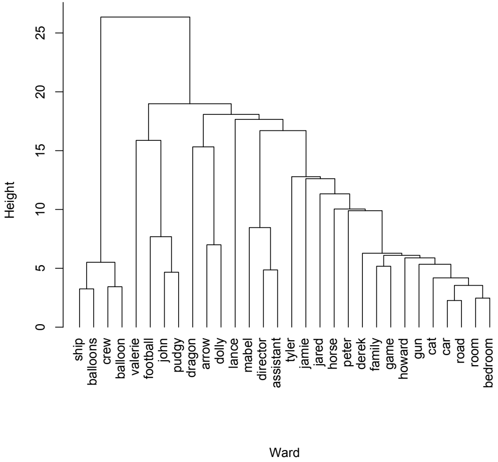


Figure : Hierarchical clustering using the Ward minimum variance agglomerative criterion.

Hierarchy (not sequence-constrained, 30 terms)

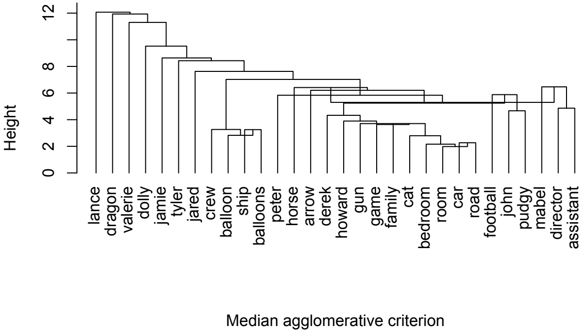


Figure : Median agglomerative criterion. (For each agglomeration, minimize the median of the pairwise dissimilarities.)