Singular Value Decomposition (SVD)

* Same as PCA when the mean of each attribute is zero
* SVD does not subtract the mean
  + Appropriate if values close to zero should not be influential
  + PCA puts them at in the extreme negative side

SVD often used for text analysis

* Values close to zero are frequent and should not affect the analysis

A picture containing text

Description automatically generated

Text

Description automatically generated

Text Processing

Latent Semantic Analysis

* Create an occurrence matrix (term-document matrix)
  + Words (term t) are the rows
  + Paragraphs (documents d) are the columns
  + Uses the term frequency-inverse document frequency (tf-idf) metric
  + Tf(t,d) = simplest form is frequency of t in d = f(t, d)
  + Idf is a measure of term rareness, it’s 0 when term occurs in all of D
  + Important terms get a higher tf-idf
* Use SVF to reduce the number of rows
  + Preserves similarity of colums
  + Diagram

    Description automatically generated

Visualizing the Concept Space

* Concept to use when approximating the matrix
  + If too few, important patterns are left out
  + If too many, noise caused by random word choices will creep in
  + Can use elbow method in the scree plot
* Throw out the 1st dimension in U and V
  + In U it is correlated with document length
  + In V it is correlated with the number of times a term was mentioned
* Project the k-D concept space into 2D and visualize as a map
  + Can cluster the map
  + The cluster of documents are then labeled by the terms

LSA Disadvantages

* LSA assumes a Gaussian distribution and Fresenius norm
  + This may not fit all problems
* LSA cannot handle polysemy effectively
  + Need LDA (Latent Dirichlet Allocation) for this
* LSA depends heavily on SVD
  + Computationally intensive
  + Hard to update as new document appear
  + But faster algorithms have emerged recently

Chi-square Test (Nominal Data)

* Used to investigate relationships
* Relationships between categorical, or nominal-scale, variables representing attributes of people, interaction technique, system, etc
* Data organized in a contingency table – cross tabulation containing counts (frequency data) for number of observation in each category
* Compares the observed values against expected values
* Expected values assume “no difference”

Correspondence Analysis (CA)

* 1
  + Compute distance matrix of the rows CC^T
  + Compute Eigenvector matrix U and the eigenvalue matrix D
  + Sort eigenvectors by values, pick two major vectors, create 2D plot
* 2
  + Compute distance matrix of the column C^T C
  + Compute eigenvector matrix V (gives the eigenvalues matrix D
  + Sort eigenvectors by value
  + Pick two major vectors, create 2D plot
* 3
  + Combines the plots of U and V
  + If chi-square test statistics was significant we should see some dependencies

Multiple Correspondence Analysis

* Extension where there are more than 2 categorical variables
  + Matrix X
* Compute X’X to get the Burt Table
* Compute Eigenvector and Eigenvalues
  + Keep top two Eigenvectors/values
  + Visualize the attribute loadings of these two Eigenvectors into the Burt Table plot (the loadings are the coordinates

Gartner Magic Quadrant

* a culmination of research in a specific market, providing a wide-angle view of the relative positions of the market's competitors