12.0 Latent Semantic Analysis for Linguistic Processing

References: 1. “ Exploiting Latent Semantic Information in Statistical Language Modeling”, Proceedings of the IEEE, Aug 2000

2. “ Special Issue on Language Modeling and Dialogue Systems”, IEEE Trans. on Speech & Audio Processing, Jan 2000

3. “ A Multi-span Language Modeling Framework for Large Vocabulary Speech Recognition”, IEEE Trans. on Speech & Audio Processing, Sept 1998

4. Golub & Van Loan, “ Matrix Computations ”, 1989

Word-Document Matrix Representation

- **Vocabulary V of size M and Corpus T of size N**
  - $V=\{w_1,w_2,\ldots w_i,..w_M\}$, w_i: the i-th word ,e.g. $M=2\times10^4$
    $T=\{d_1,d_2,\ldots d_j,..d_N\}$, d_j: the j-th document ,e.g. $N=10^5$
  - c_{ij}: number of times w_i occurs in d_j
    n_j: total number of words present in d_j
    $t_i= \sum_j c_{ij}$: total number of times w_i occurs in T

$$\Rightarrow \varepsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\left(\frac{c_{ij}}{t_i}\right)\log\left(\frac{c_{ij}}{t_i}\right),\quad \text{normalized entropy (indexing power) of } w_i \text{ in T}$$

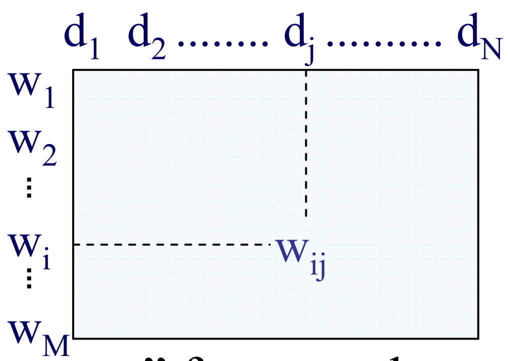
$$0\le\varepsilon_i\le1\ ,\qquad \varepsilon_i=0 \quad \text{if } c_{ij}=t_i \text{ for some j and } c_{ij}=0 \text{ for other j}$$

$$\varepsilon_i=1 \quad \text{if } c_{ij}=t_i/\text{N for all j}$$

  - $w_{ij}=(1-\varepsilon_i)\dfrac{c_{ij}}{n_j}$, word frequencies in doucments, but normalized with document length and word entropy

- **Word-Document Matrix W**

  $W = [w_{ij}]$



  - each row of W is a N-dim "feature vector" for a word w_i with respect to all documents d_j
    each column of W is a M-dim "feature vector" for a document d_j with respect to all words w_i

Dimensionality Reduction

- $WW^T = \overline{U}\,\overline{S}_1^2\,\overline{U}^T$
  - $\overline{U} = [e_1, e_2, \ldots e_M]$, $\overline{S}_1^2 = [s_i^2]_{M\times M}$, s_i^2 : eigenvalues of WW^T, $s_i^2 \geq s_{i+1}^2$

    (i, j) element of WW^T : inner product of i-th and j-th rows of W

    "similarity" between w_i and w_j

    $WW^T = \sum_i s_i^2 e_i e_i^T$, e_i : orthonormal eigenvectors , $\overline{U}^T\overline{U} = I_M$

    s_i^2 : weights (significance of the "component matrices" $e_i\, e_i^T$)
  - dimensionality reduction: selection of R largest eigenvalues (R=800 for example)

    $W_{M\times N} W_{N\times M}^T \approx U_{M\times R} S_{R\times R}^2 U_{R\times M}^T$, $U_{M\times R} = [e_1, e_2, \ldots. e_R]$

    R "concepts" or "latent semantic concepts"
- $W^T W = \overline{V}\,\overline{S}_2^2\,\overline{V}^T$
  - $\overline{V} = [e'_1, e'_2, \ldots. e'_N]$, $\overline{S}_2^2 = [s_i^2]_{N\times N}$, s_i^2 : eigenvalues of $W^T W$, $s_i^2 \geq s_{i+1}^2$, $s_i^2 = 0$ for $i > \min(M, N)$

    (i, j) element of $W^T W$: inner product of i-th and j-th columns of W

    "similarity" between d_i and d_j

    $W^T W = \sum_i s_i^2 e'_i e_i'^T$, e'_i : orthonormal eigenvectors, $\overline{V}^T\overline{V} = I_N$

    s_i^2 : weights (significance of the "component matrices" $e'_i\, e_i'^T$)
  - dimensionality reduction: selection of R largest eigenvalues

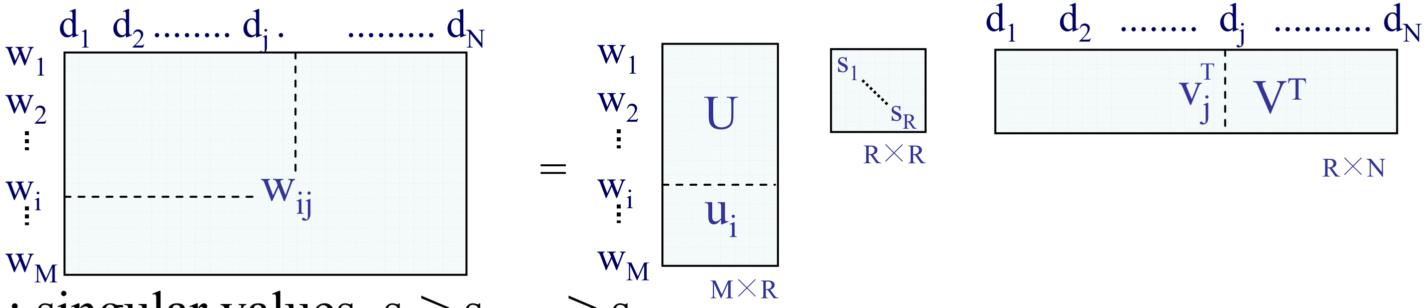
    $W_{N\times M}^T W_{M\times N} \approx V_{N\times R} S_{R\times R}^2 V_{R\times N}^T$, $V_{N\times R} = [e'_1, e'_2 \ldots e'_R]$

    R "concepts" or "latent semantic concepts"

Singular Value Decomposition (SVD)

- **Singular Value Decomposition (SVD)**

$$W_{M\times N} \approx \hat{W}_{M\times N} = U_{M\times R}\, S_{R\times R}\, V^{T}_{R\times N}$$



  – s_i: singular values, $s_1 \geq s_2 .... \geq s_R$
  U: left singular matrix, V: right singular matrix

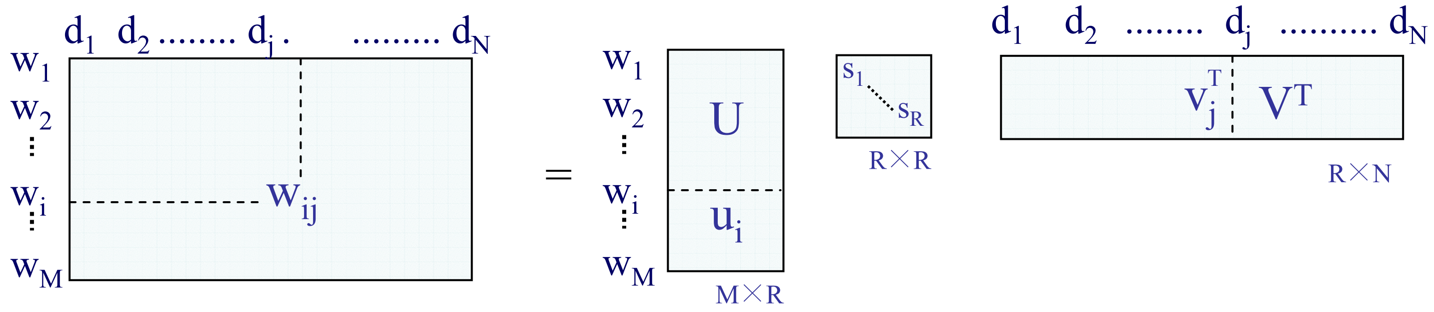
- **Vectors for word w_i: $u_i S = \underline{u}_i$ (a row)**
  – a vector with dimentionality N reduced to a vector $u_i S = \underline{u}_i$ with dimentionality R
  – "discrete" dimentionality defined by N documents reduced to "continuous" dimentionality defined by R "concepts"
  – the R row vectors of V^T, or column vectors of V, or eigenvectors $\{e'_1,..e'_R\}$, are the R orthouormal basis for the "latent semantic space" with dimentionality R, with which $u_i S = \underline{u}_i$ is represented

- **The Association Structure between words w_i and documents d_j is preserved with noisy information deleted, while the dimensionality is reduced to a common set of R "concepts"**

Singular Value Decomposition (SVD)

- **Singular Value Decomposition (SVD)**

$$W_{M\times N} \approx \hat{W}_{M\times N} = U_{M\times R} S_{R\times R} V^{T}_{R\times N}$$



  - s_i: singular values, $s_1 \geq s_2 .... \geq s_R$
    U: left singular matrix, V: right singular matrix

- **Vectors for document d_j: $v_j S = \underline{v}_j$ (a row, or $\underline{v}_j^T = S\, v_j^T$ for a column)**
  - a vector with dimentionality M reduced to a vector $v_j S = \underline{v}_j$ with dimentionality R
  - "discrete" dimentionality defined by M words reduced to "continuous" dimentionality defined by R "concepts"
  - the R columns of U, or eigenvectors $\{e_1, ... e_R\}$, are the R orthonormal basis for the "latent semantic space" with dimensionality R, with which $v_j S = \underline{v}_j$ is represented

- **The Association Structure between words w_i and documents d_j is preserved with noisy information deleted, while the dimensionality is reduced to a common set of R "concepts"**

Example Applications in Linguistic Processing

- **Word Clustering**
  - example applications: class-based language modeling, information retrieval ,etc.
  - words with similar "semantic concepts" have "closer" location in the "latent semantic space"
    - they tend to appear in similar "types" of documents, although not necessarily in exactly the same documents
  - each component in the reduced word vector $u_j S = \underline{u}_j$ is the "association" of the word with the corresponding "concept"
  - example similarity measure between two words:

$$sim(w_i, w_j) = \frac{\underline{u}_i \cdot \underline{u}_j}{\left|\underline{u}_i\right| \cdot \left|\underline{u}_j\right|} = \frac{u_i S^2 u_j^T}{\left|u_i S\right| \cdot \left|u_j S\right|}$$

- **Document Clustering**
  - example applications: clustered language modeling, language model adaptation, information retrieval, etc.
  - documents with similar "semantic concepts" have "closer" location in the "latent semantic space"
    - they tend to include similar "types" of words, although not necessarily exactly the same words
  - each component on the reduced document vector $v_j S = \underline{v}_j$ is the "association" of the document with the corresponding "concept"
  - example "similarity" measure between two documents:

$$sim(d_i, d_j) = \frac{\underline{v}_i \cdot \underline{v}_j}{\left|\underline{v}_i\right| \cdot \left|\underline{v}_j\right|} = \frac{v_i S^2 v_j}{\left|v_i S\right| \cdot \left|v_j S\right|}$$

Example Applications in Linguistic Processing

- **Information Retrieval**
  - "concept matching" vs "lexical matching" : relevant documents are associated with similar "concepts", but may not include exactly the same words
  - example approach: treating the query as a new document (by "folding-in"), and evaluating its "similarity" with all possible documents
- **Fold-in**
  - consider a new document outside of the training corpus T, but with similar language patterns or "concepts"
  - construct a new column d_p ,p>N, with respect to the M words
  - assuming U and S remain unchanged

    $d_p = USv_p^T$ (just as a column in $W = USV^T$)

    $\underline{v}_p = v_p S = d_p^T U$ as an R-dim representation of the new document

    (i.e. obtaining the projection of d_p on the basis e_i of U by inner product)

Integration with N-gram Language Models

- **Language Modeling for Speech Recognition**
  - $\text{Prob}(w_q|d_{q-1})$
    w_q: the q-th word in the current document to be recognized (q: sequence index)
    d_{q-1}: the recognized history in the current document
    $\underline{v}_{q-1}=d_{q-1}^T U$: representation of d_q by v_q (folded-in)
  - $\text{Prob}(w_q|d_{q-1})$ can be estimated by $\underline{u}_q$ and $\underline{v}_{q-1}$ in the R-dim space
  - integration with N-gram
    $\text{Prob}(w_q|H_{q-1}) = \text{Prob}(w_q|h_{q-1}^{(n)}, d_{q-1})$
    H_{q-1}: history up to w_{q-1}
    $h_{q-1}^{(n)}$: $< w_{q-n+1}, w_{q-n+2}, \ldots w_{q-1} >$
  - N-gram gives local relationships, while d_{q-1} gives semantic concepts
  - d_{q-1} emphasizes more the key content words, while N-gram counts all words similarly including function words
- **$\underline{v}_{q-1}$ for d_{q-1} can be estimated iteratively**
  - assuming the q-th word in the current document is w_i

$$d_q = \left(\frac{q-1}{q}\right)d_{q-1} + \left(\frac{1-\varepsilon_i}{q}\right)[00\ldots 0\underset{\uparrow}{1}00\ldots.0]^T$$

i-th dimentionality out of M

$$\underline{v}_q = d_q^T U = \left(\frac{q-1}{q}\right)\underline{v}_{q-1} + \left(\frac{1-\varepsilon_i}{q}\right)u_i \quad \text{,updated word - by - word}$$

$\underline{v}_q$ moves in the R-dim space initially, eventually settle down somewhere