LSI (Latent Semantic Indexing) with SVD (Singular Value Decomposition)

Vector Space model of documents:

Given a collection of documents:

Let w be the number of unique words over the entire document collection.

Let d be the number of documents in a collection.

Define:

Matrix **A** as a $w \times d$ matrix.

- Each row of A represents a word over the entire collection of documents.
- Each column of A represents one of the documents in the collection.
- A_{ij} is proportional to the number of occurrences of word i in document j.

Various methods for scaling the document vectors making up the matrix were discussed in previous classes.

Define

A query vector q as a w dimensional vector such that q_i is nonzero when word i is a part of the query.

Then the vector $A^t q$ gives *exact keyword matching* over the document collection.

- That is, the j–th entry of the vector $A^t q$ is nonzero when document j contains one or more words in the query.

Is it possible to do better?

Use $\boldsymbol{A_k}$ instead of $\boldsymbol{A}$ → Calculate $\boldsymbol{A^t{}_k q}$ rather than $\boldsymbol{A^t q}$.

$\boldsymbol{A_k}$ is defined at the k-truncated SVD of $\boldsymbol{A}$.

Note: Both $\boldsymbol{A}$ and $\boldsymbol{A_k}$ are w x d matrices.

Derivation of A_k:

$$A_k = U_k \Sigma_k V_k^t$$

$\boldsymbol{U_k}$ = [$\boldsymbol{u_1}$ $\boldsymbol{u_2}$ . . . $\boldsymbol{u_k}$]

Σ_k =

$$\begin{matrix} \delta_1 & 0 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & \delta_k \end{matrix}$$

$$\boldsymbol{V^t{}_k} = \begin{pmatrix} \boldsymbol{v^t{}_1} \\ \boldsymbol{v^t{}_2} \\ . \\ . \\ \boldsymbol{v^t{}_k} \end{pmatrix}$$

- $\boldsymbol{U_k}$ is the matrix of left singular vectors, while $\boldsymbol{V_k}$ is the matrix of right singular vectors,
- $\boldsymbol{u}$'s and $\boldsymbol{v}$'s are orthogonal to each other,
- σ_1, σ_2, .. σ_k are real.

Note: $\boldsymbol{k}$=1 → $\boldsymbol{A_k} = \sigma_1 \boldsymbol{u_1 v^t{}_1}$ == (<u>*w X d)*</u> rank 1 matrix
General k → $\boldsymbol{A_k} = \Sigma(\sigma_i \boldsymbol{u_i v^t{}_i})$ == (k-dimensional <u>*w X d*</u> matrix)

Why are singular vectors better than a random vector?

- It can be proven that $\boldsymbol{A_k}$ is the closest/best rank-$\boldsymbol{k}$ approximation to $\boldsymbol{A}$ (i.e. Norm($\boldsymbol{A}-\boldsymbol{A_k}$) <= Norm($\boldsymbol{A}-\boldsymbol{B}$) for any rank $\boldsymbol{k}$ matrix $\boldsymbol{B}$

Why might this approximation be better?

- In normal language, there is a variability in word usage. Different people use different words for the same concepts (synonymy). Also, the same word sometimes has many different meanings (polysemy). The claim is that use of $\boldsymbol{A_k}$ removes this "noise" from the documents. $\boldsymbol{A_k}$ only includes important conceptual information.

Structure and computation of A, A_k.

While $\boldsymbol{A}$ is sparse, $\boldsymbol{A_k}$ in general is not.
$\mathbf{A_k}$ is a dense, **w X d** matrix.
$\mathbf{A_k}$ requires **wd** storage while **A** might have required (.01 * **wd**) storage.

Problem: 1) It is difficult to store such a potentially large matrix.
2) Computation of the SVD takes much time.

Efficient Computation of $\boldsymbol{A^t{}_k q}$:

Write $\mathbf{A_k^T q} = \mathbf{V_k \Sigma_k U_k^T q}$
Compute $\mathbf{x} = \mathbf{U_k^T q}$ (a **k x 1** vector)
$\mathbf{Y} = \mathbf{V_k \Sigma_k}$ (a **d x k** matrix)
$\mathbf{A_k^T q} = \mathbf{YX}$

Note: $\boldsymbol{V_k \Sigma_k}$ is not dependent on the query $\boldsymbol{q}$ →
It is only necessary to compute it once.
$\boldsymbol{V_k \Sigma_k}$ is a ***d x k*** matrix.

Geometric interpretation

$$\boldsymbol{U_k^t q} = \begin{pmatrix} u^t{}_1 \\ u^t{}_2 \\ . \\ . \\ u^t{}_k \end{pmatrix} * \boldsymbol{q} = \begin{pmatrix} u^t{}_1 q \\ u^t{}_2 q \\ \\ \\ u^t{}_k q \end{pmatrix}$$

Each of the components of this vector is the projection of $\boldsymbol{q}$ onto each singular vector $\boldsymbol{u_i}$

This is a $\boldsymbol{w}$ → $\boldsymbol{k}$-dimensional transformation ($\boldsymbol{U_k^t q}$ is $\boldsymbol{k}$-dimensional, $\boldsymbol{q}$ is $\boldsymbol{w}$-dimensional)

The singular vectors ($\boldsymbol{u_i}$'s) are not a basis of the document space, but in a sense cover the "important" part of the space.

Also, $V_k\Sigma_k$ projects document vectors onto this k-dimensional space as well.

<u>Algebraic justification</u>

Let e_i be a vector such that $e_{ij} = 1$ when i=j and 0 otherwise.

Then Ae_i = the i^{th} document vector.
$U^t{}_k Ae_i$ = projection of i^{th} document vector onto $u_1, u_2 \ldots u_k$

$$A = U_k\Sigma_k V^t{}_k = \underbrace{U_k\Sigma_k V^t{}_k}_{\text{Signal}} + \underbrace{U_{(d-k)}\Sigma_{(d-k)} V^t{}_{(d-k)}}_{\text{Noise}}$$

where:

$U_{(d-k)} = [u_{(k+1)}\ u_{(k+2)}\ \ldots\ u_d]$
all of which are orthogonal to $u_1, u_2 \ldots u_k \rightarrow U_k^T U_{(d-k)} = 0$

$V_{(d-k)} = [v_{(k+1)}\ v_{(k+2)}\ \ldots\ v_d]$ (assuming **w*(**log **d)** <= **w**)

$\Sigma_{(d-k)}$ =

$$\begin{pmatrix} \delta_{k+1} & 0 & 0 & 0 & 0 \\ 0 & \delta_{k+2} & 0 & 0 & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & \delta_d \end{pmatrix}$$

Then $U^t{}_k Ae_i = (\Sigma_k V^t{}_k + U^t{}_k U_{(d-k)}\Sigma_{(d-k)} V^t{}_{(d-k)})e_i$
$= \Sigma_k V^t{}_k e_I = Y^T e_i$

$V_k\Sigma_k$ has as it i^{th} row the projection of each document vector onto the space spanned by $u_1, u_2 \ldots u_k$.

The w-dimensional i^{th} document vector Ae_i is represented as a k-dimensional document vector $Y^T e_i$.

Thus, we have gone from a very large (w-dimensional) space to a much smaller k-dimensional space. This is known as *Dimensionality Reduction*.