Low rank approximation of the term-document matrix

- We saw in the TF-IDF demo that documents can be characterized by its most heavily TF-IDF weighted words

- Idea: what if we were to project a document onto a set of basis' comprised of linear combinations of the constituent words such that the basis themselves capture semantic relationships between words?

- How would we learn such a transformation?
  - PCA?
  - SVD?

- Why is PCA not suitable for this problem?

- How many dimensions would we need for our sub manifold to capture most of the information in our sparse term-document matrix?

Latent semantic analysis (LSA)

- Also known as Latent semantic indexing (LSI)
- Used extensively in search engines
- Factors the document-term matrix, $\mathbf{X} \in \mathbb{R}^{M \times N}$, by computing its SVD

Descending order —>

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \begin{bmatrix} u_1^{(1)} & \dots & u_1^{(M)} \\ \vdots & \vdots & \vdots \\ u_M^{(1)} & \dots & u_M^{(M)} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & 0 \\ 0 & \dots & \sigma_M & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1^{(1)} & \dots & v_N^{(1)} \\ \vdots & \vdots & \vdots \\ v_1^{(N)} & \dots & v_N^{(N)} \end{bmatrix}$$

where $\mathbf{X} \in \mathbb{R}^{M \times N}$ term-document matrix

$\mathbf{U} \in \mathbb{R}^{M \times M}$ left singular vectors

$\boldsymbol{\Sigma} \in \mathbb{R}^{M \times N}$ diagonal matrix of singular values

$\mathbf{V} \in \mathbb{R}^{N \times N}$ right singular vectors

$N =$ number of words

$M =$ number of documents