

Singular Value Decomposition

With Applications to IR and Text Clustering

Jiangsheng Yu

©School of Electronics Engineering and Computer Science

Peking University, Beijing, 100871

yujs@pku.edu.cn, <http://icl.pku.edu.cn/yujs>



Topics

1. Preliminary linear algebra
2. Latent semantic indexing (LSI)
3. Singular Value Decomposition (SVD) of matrix in the view of data compression
4. Information retrieval (IR) and text clustering by SVD method
5. Clustering of terms
6. Further discussions
7. Conclusion
8. References

Preliminary Linear Algebra

The following content can be found in any textbook of Linear Algebra.

1. Vector Space
2. Eigenvector and Eigenvalue
3. Frobenius Norm and p Norm
4. Hölder Inequality
5. Singular Value Decomposition (SVD)
6. Eckart-Young Theorem
7. Least Square Method

Vector Space

Definition 1 A **vector space** over \mathbb{R}^n is a set of vectors $\mathbf{x} \in \mathbb{R}^n$, say V , satisfying that:

Abelian group

$(V; +)$ is an additive group

Associativity of Scalar Multiplication

$$r(s\mathbf{x}) = (rs)\mathbf{x}, \forall r, s \in \mathbb{R}$$

Distributivity of scalar sums

$$(r + s)\mathbf{x} = r\mathbf{x} + s\mathbf{x}, \forall r, s \in \mathbb{R}$$

Distributivity of vector sums

$$r(\mathbf{x} + \mathbf{y}) = r\mathbf{x} + r\mathbf{y}, \forall r \in \mathbb{R}$$

Scalar multiplication identity

$$1\mathbf{x} = \mathbf{x}$$

Orthogonal Matrix

Definition 2 A matrix $A = (a_{ij})_{m \times n}$ is **orthogonal**^a if

$$A^T A = \mathbf{1}_{n \times n} \quad (1)$$

where A^T is the transpose of A .

Definition 3 $x \in \mathbb{R}^n$ is an **eigenvector** of $A_{n \times n}$ if $\lambda \in \mathbb{R}$ such that

$$Ax = \lambda x \quad (2)$$

where λ is called an **eigenvalue** of A .

Homework 1 Why we have to study eigenvector?

Hint: In the view of transformation.

^aIf $A_{n \times n} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ is orthogonal, then $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$ is an orthogonal basis of \mathbb{R}^n , where $\mathbf{a}_i = (a_{1i}, a_{2i}, \dots, a_{ni})^T$.

p Norm of Vector

Definition 4 The p norm of vector $\mathbf{x} \in \mathbb{R}^n$ is

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} \quad (3)$$

For instance, the usual p norm is

- $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
- $\|\mathbf{x}\|_2 = (\mathbf{x}^T \mathbf{x})^{\frac{1}{2}}$ and
- $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Hölder Inequality

Theorem 1 (Hölder Inequality)

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q \quad (4)$$

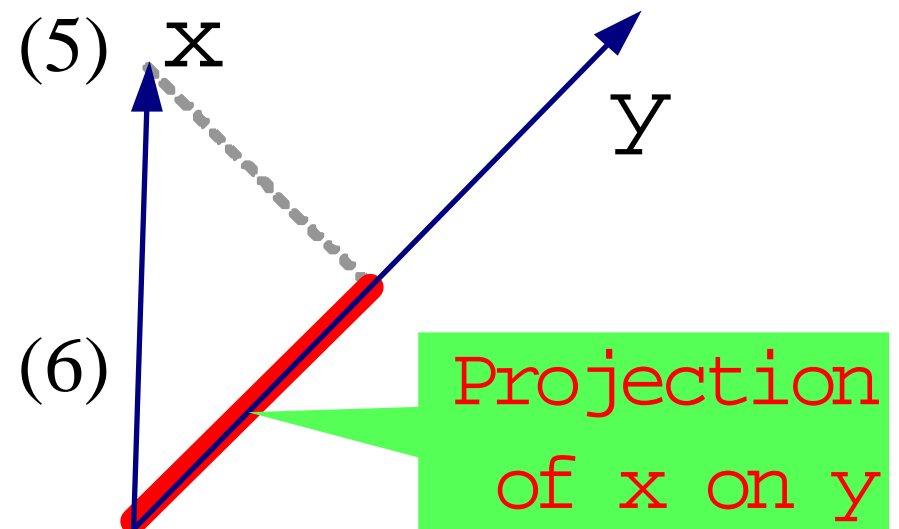
where $\frac{1}{p} + \frac{1}{q} = 1$. Especially,

Corollary 1 (Cauchy-Schwartz Inequality)

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \quad (5)$$

Note that

$$\cos \theta = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$



Frobenius Norm and p Norm

Definition 5 Given a matrix $\mathbf{A} = (a_{ij})_{m \times n}$, the **Frobenius norm** (or F norm) of \mathbf{A} is

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} \quad (7)$$

The **p norm** of \mathbf{A} is

$$\begin{aligned} \|\mathbf{A}\|_p &= \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \\ &= \max_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_p \end{aligned} \quad (8)$$

Properties of Norms

Homework 2 $\forall \mathbf{A} \in \mathbb{R}^{m \times n}$, we have

- $\max_{i,j} |a_{ij}| \leq \|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{n} \|\mathbf{A}\|_2 \leq n\sqrt{m} \max_{i,j} |a_{ij}|$
- $\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$ and $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$
- $\frac{1}{\sqrt{m}} \|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_2 \leq \sqrt{n} \|\mathbf{A}\|_1$
- $\frac{1}{\sqrt{n}} \|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq \sqrt{m} \|\mathbf{A}\|_\infty$

Homework 3 The length of projection of $\mathbf{x} \in \mathbb{R}^n$ on $\mathbf{y} \in \mathbb{R}^n$ is $|\mathbf{x}^\top \mathbf{y}| / \|\mathbf{y}\|_2$.
Hint: By (6).

Singular Value

Definition 6 Given a matrix $\mathbf{A}_{m \times n}$ whose rank is r , then the eigenvalues of $\mathbf{A}^T \mathbf{A}$ are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$$

$\sigma_i = \sqrt{\lambda_i}$ is called the **singular value** of \mathbf{A} where $i = 1, 2, \dots, n$.

We have the properties

- $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$

Proof The coefficient of λ^{n-1} in $|\mathbf{A}^T \mathbf{A} - \lambda \cdot \mathbf{1}_{n \times n}|$ is

$\|\mathbf{A}\|_F^2$, which is also $\sum_{i=1}^r \sigma_i^2$.

- $\|\mathbf{A}\|_2 = \sigma_1$

Latent Semantic Indexing

[2] says that “LSI tries to overcome the problems of **lexical matching** by using statistically derived conceptual indices instead of individual words for retrieval. LSI assumes that there is some underlying or **latent structure in word usage** that is partially obscured by variability in word choice. . . . Performance data shows that these statistically derived vectors are more robust indicators of meaning than individual terms.”

Homework 4 Read the original papers on LSI and the statistical models for detecting collocations.

Homework 5 Discuss the virtue and disadvantage of Vector Space Model (VSM) justly.

History Notes about LSI

1. S. T. Dumais, G. W. Furnas, T. K. Landauer, and S. Deerwester (1988), *Using latent semantic analysis to improve information retrieval*. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285.
2. S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990), *Indexing by latent semantic analysis*. Journal of the Society for Information Science, 41(6), 391-407.
3. P. W. Foltz (1990), *Using Latent Semantic Indexing for Information Filtering*. In R. B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA, 40-47.
4. J. S. Yu, Z. H. Jin, and Z. S. Wen (2003), *Automatic Detection of Collocation*. Report at the seminar of Statistical Machine Learning, Peking University, <http://icl.pku.edu.cn/yujs>

Why LSI?

The serious shortcomings of cosine similarity in VSM include, at least,

Synonymy: distinct names of the same object (e.g., car and automobile) have small cosine similarity, which leads to poor recall.

Polysemy: most words have more than one meaning (e.g., model, \dots . *Markov* is cosine related to *model*), which leads to poor precision.

High-dimensional: The dimension of VSM usually leads to an unsuccessful clustering.

Homework 6 How to utilize a knowledge base, e.g. WordNet, in IR or text clustering?

Singular Value Decomposition

Theorem 2 Given a matrix $A_{m \times n}$ whose rank is r and $m \geq n$, there exist two orthogonal matrixes $U_{m \times n} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$ (**term vectors**) and $V_{n \times n} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ (**document vectors**) s.t.

$$\begin{aligned} A &= U \Sigma V^T \\ &= \sum_{i=1}^r \mathbf{u}_i \cdot \sigma_i \cdot \mathbf{v}_i^T \end{aligned} \quad (9)$$

where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n) \in \mathbb{R}^{n \times n}$ and σ_i is the singular value of A . (9) is called the Singular Value Decomposition (SVD) of A .

Term-Document Matrix

Definition 7 Given a term bank with m terms and a set of n documents. The **term-document matrix** (TD matrix) is $A = (a_{ij})_{m \times n}$, where a_{ij} is the frequency (count, or Boolean value) of term i in document j .

TERM	d_1	d_2	\cdots	d_n
t_1	a_{11}	a_{12}	\cdots	a_{1n}
t_2	a_{21}	a_{22}	\cdots	a_{2n}
\vdots	\vdots			
t_m	a_{m1}	a_{m2}	\cdots	a_{mn}

Homework 7 How to extract the terms?

Hint: The collocations and distribution of terms.

Terms and Documents

Example 1 All the terms in the following two documents (Cited from the **MEE2002**) are underlined:

1. Integer, any number that is a natural number (the counting numbers 1, 2, 3, 4, \dots), a negative of a natural number ($-1, -2, -3, -4, \dots$), or zero. A large proportion of mathematics has been devoted to integers because of their immediate application to real situations.
2. Any integer greater than 1 that is divisible only by itself and 1 is called a prime number (see Number Theory). Every integer has a unique set of prime factors, that is, a list of prime numbers that when multiplied together produce the integer concerned. For example, the prime factors of the integer 42 are 2, 3 and 7.
3. All mesons must have spins equal to integers (0, 1, 2, and so on). Particles with spins equal to integers are called bosons. Bosons differ from particles with noninteger spins, called fermions, in that bosons do not obey a rule of physics called the Pauli exclusion principle.

Example of TD Matrix

Example 2 The TD matrix is

TERM	d_1	d_2	d_3
integer	2	3	2
natural number	2	0	0
mathematics	1	0	0
prime number	0	2	0
prime factor	0	2	0
Number Theory	0	1	0
meson	0	0	1
Boson	0	0	3
fermion	0	0	1
particle	0	0	2
physics	0	0	1
spin	0	0	2
Pauli exclusion principle	0	0	1

$\rightarrow \mathbf{A} =$

$$\begin{pmatrix} 2 & 3 & 2 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

Eckart-Young Theorem

Theorem 3 (Eckart-Young) Let the SVD of \mathbf{A} be given by (9) with $r = \text{rank}(\mathbf{A}) \leq p = \min\{m, n\}$ and define

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top = \sum_{i=1}^k \mathbf{u}_i \cdot \sigma_i \cdot \mathbf{v}_i^\top \quad (10)$$

then \mathbf{A}_k is the optimal approximation of \mathbf{A} in the view of

$$\begin{aligned} \min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F &= \|\mathbf{A} - \mathbf{A}_k\|_F \\ &= \sqrt{\sum_{i=k+1}^p \sigma_i^2} \\ \min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2 &= \|\mathbf{A} - \mathbf{A}_k\|_2 \\ &= \sigma_{k+1} \end{aligned} \quad (11)$$

Data Compression

The red part renews $A_{m \times n}$ with rank $k \leq r$ the best.

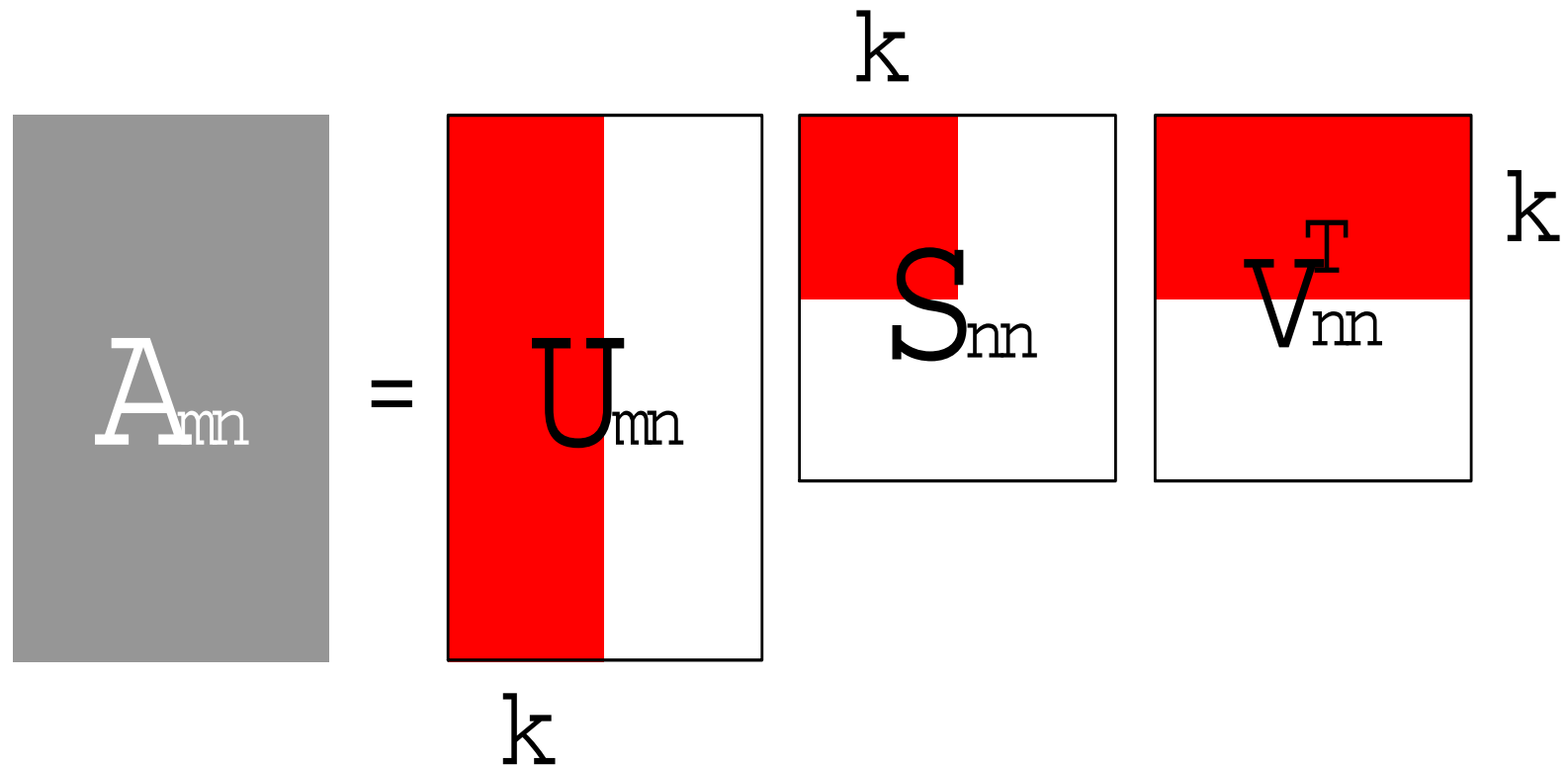


Figure 1: SVD is a way of data compression

Least Square Method

Problem 1 Given a finite training set, say $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, we intend to find a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that simulates T the best.

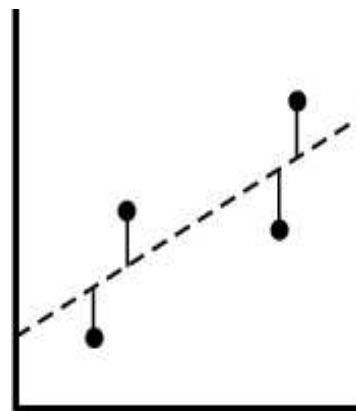
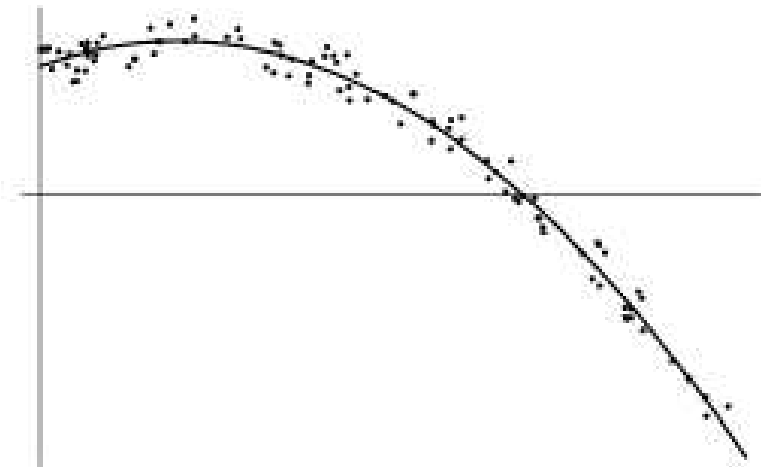
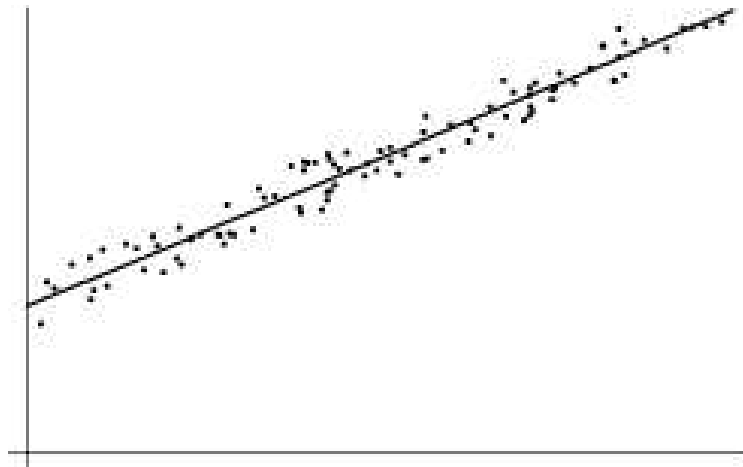
Solution The explanation of “the best” is

$$f = \operatorname{argmin}_{g \in C(\mathbb{R}^n)} \sum_{i=1}^m [g(\mathbf{x}_i) - y_i]^2 \quad (12)$$

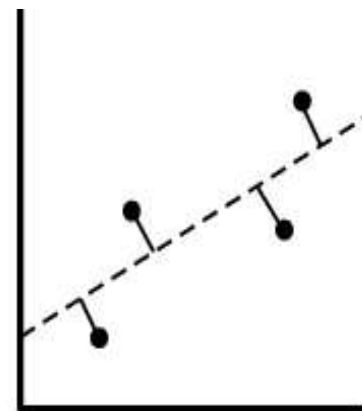
where $C(\mathbb{R}^n)$ is the set of continuous functions on \mathbb{R}^n . The optimal approximation is determined by the precise standard.^a

^aSee the excellent [3] for more details on LSM.

Linear Regression



vertical offsets



perpendicular offsets

Figure 2: Linear regression — a particular LSM

Best Approximation of SVD

Example 3 Eckart-Young Theorem guarantees that A_k , the truncated matrix of A , is the closest k -rank matrix to A by both F norm and 2 norm.

Definition 8 The **document vector** d_j is projected to a **k -document vector** on the space of $\text{span}(\mathbf{U}_k)$:

$$\hat{d}_j = \sum_{i=1}^k c_{ij} \mathbf{u}_i \quad (13)$$

where c_{ij} is the entry of $\Sigma_k \mathbf{V}_k^T$, a $k \times n$ matrix. Or, the j^{th} column of $\Sigma_k \mathbf{V}_k^T$ is the coordinates of d_j projected on $\text{span}(\mathbf{U}_k) = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$.

Explanation of $\text{span}(\mathbf{U}_k)$

\mathbf{u}_i has coordinates for each term, the i^{th} character for clustering, which is orthogonal to \mathbf{u}_j if $i \neq j$.

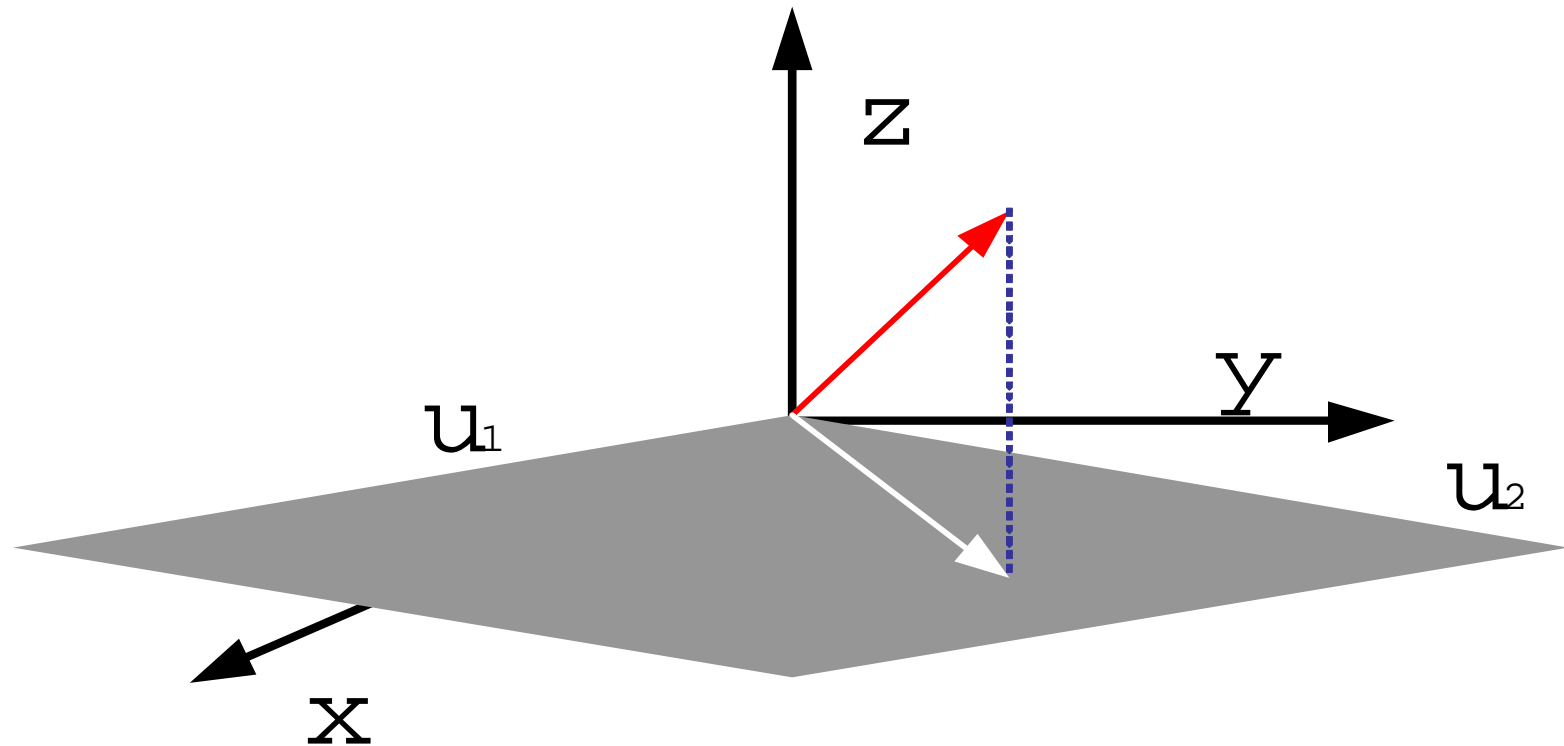


Figure 3: Meaning of $\text{span}(\mathbf{U}_k)$

Parameters of SVD Method

parameter	explanation
m	number of terms
n	number of documents
k	number of factors
r	rank of A
U	term vectors
Σ	matrix of singular values
V	document vectors
$A = U\Sigma V^T$	term-document matrix
$A_k = U_k \Sigma_k V_k^T$	the best k -rank approximation to term-document matrix A
$\Sigma_k V_k^T$	k -document vectors

Example of SVD

Example 4 The SVD of term-document matrix $A = U\Sigma V^T$ calculated by MATLAB:

$$U = \begin{pmatrix} 0.6824 & 0.4111 & -0.1760 \\ 0.1073 & 0.1551 & -0.7419 \\ 0.0536 & 0.0776 & -0.3710 \\ 0.1909 & 0.3736 & 0.3461 \\ 0.1909 & 0.3736 & 0.3461 \\ 0.0954 & 0.1868 & 0.1730 \\ 0.1444 & -0.1523 & 0.0234 \\ 0.4332 & -0.4568 & 0.0702 \\ 0.1444 & -0.1523 & 0.0234 \\ 0.2888 & -0.3045 & 0.0468 \\ 0.1444 & -0.1523 & 0.0234 \\ 0.2888 & -0.3045 & 0.0468 \\ 0.1444 & -0.1523 & 0.0234 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 5.5182 & 0 & 0 \\ 0 & 3.9498 & 0 \\ 0 & 0 & 2.4390 \end{pmatrix}$$
$$V = \begin{pmatrix} 0.2959 & 0.3063 & -0.9048 \\ 0.5266 & 0.7379 & 0.4221 \\ 0.7969 & -0.6014 & 0.0570 \end{pmatrix}$$

Document Similarity

After SVD, the dimension is reduced and

$$\text{sim}(\mathbf{d}_i, \mathbf{d}_j) \approx \text{sim}(\hat{\mathbf{d}}_i, \hat{\mathbf{d}}_j) \quad (14)$$

The granularity of similarity should be defined by the user, not the designer of IR or text clustering.

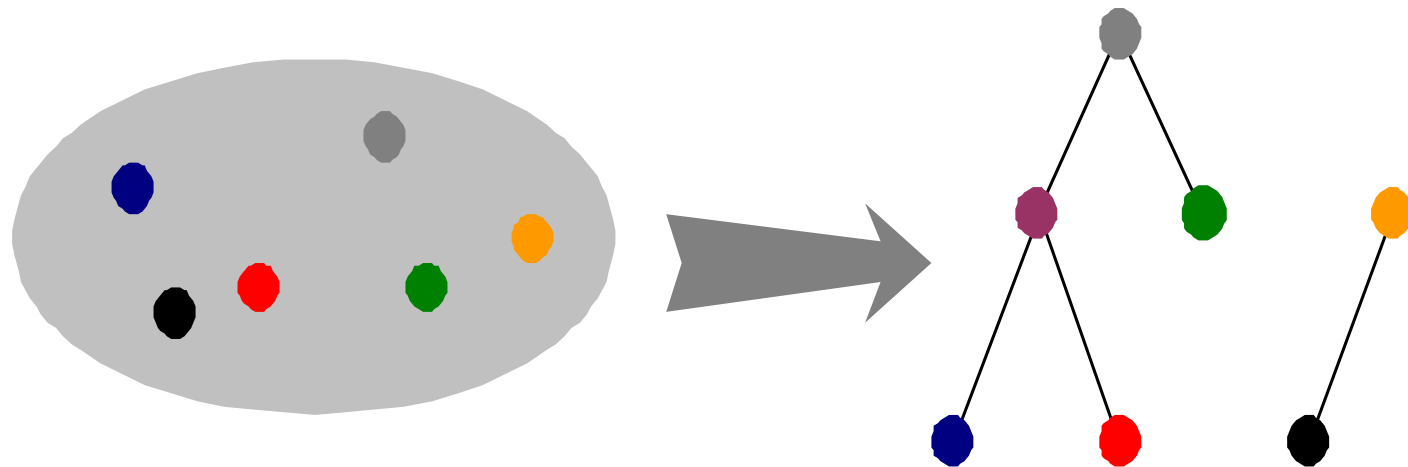


Figure 4: Hypernymy relation described by WordNet

SVD Method for IR

Definition 9 The k -query vector is the projection of query vector q on $\text{span}(\mathbf{U}_k)$, that is,

$$\hat{q} = (q^\top u_1, q^\top u_2, \dots, q^\top u_k)^\top \quad (15)$$

Regard q as a special document vector, then by (14) the relativity between query and document is done.

Homework 8 Let the user's query, in Example 1, be prime number,^a i.e., $q^\top = (0, 0, 0, 1, 0, \dots, 0)$.

Calculate the cosine similarity between 2-query vector and 2-document vectors and explain your result.

^aNotice that "prime number" does not appear in doc 1 and doc 3.

Cluster Package in MatLab

In Statistics Toolbox, we can use

1. **pdist** to compute the pairwise distance between vectors. There are various distances, such as, Euclidean distance (default), Mahalanobis distance, City Block metric, Minkowski metric, etc.
2. **dendrogram** to generate the hierarchical, binary cluster tree.
3. **cophenet** to compare two sets of values and compute their correlation, returning a value called the cophenetic correlation coefficient.

Dendrogram

The following dendrogram is generated by MatLab:

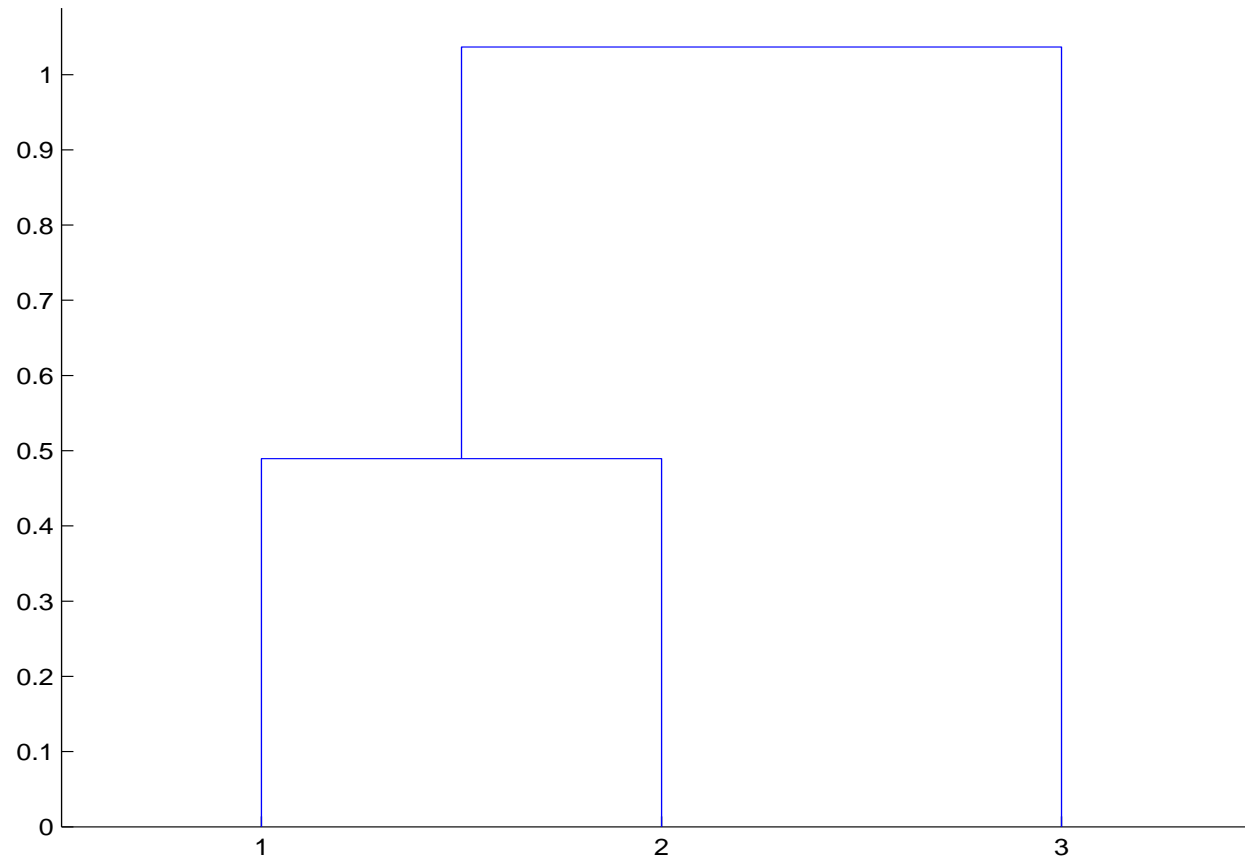


Figure 5: Dendrogram of the three documents

Experiment of Text Clustering

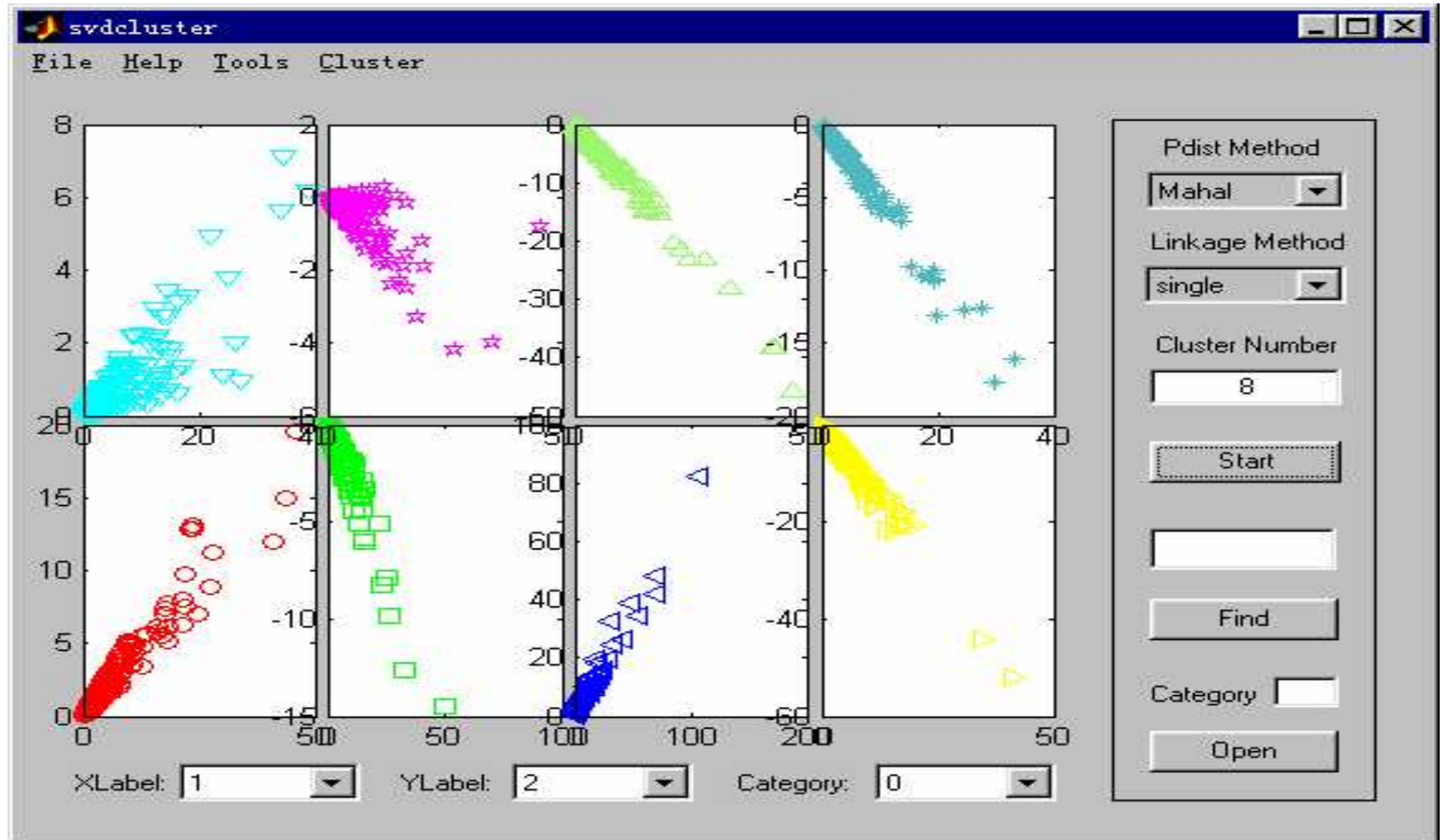


Figure 6: Text clustering of People's Daily

Clustering of Terms

Definition 10 The projection of **term vector**^a $\mathbf{t}_{n \times 1}$ on the space of $\text{span}(\mathbf{V}_k)$, denoted by $\hat{\mathbf{t}}$, is called a **k -term vector**.

Similar to the discussion of k -document vector, $\hat{\mathbf{t}}$ is an approximation of \mathbf{t} that can be measured by cosine similarity, etc. Consequently, the clustering of terms.

Homework 9 Calculate the cosine similarity of terms, in Example 1, in $\text{span}(\mathbf{V}_2)$ and explain your result.

Homework 10 Design an algorithm of term training and explain why it improves the performance of text clustering.

^aThat is, the row vector of TD matrix.

Anathema of Dimension

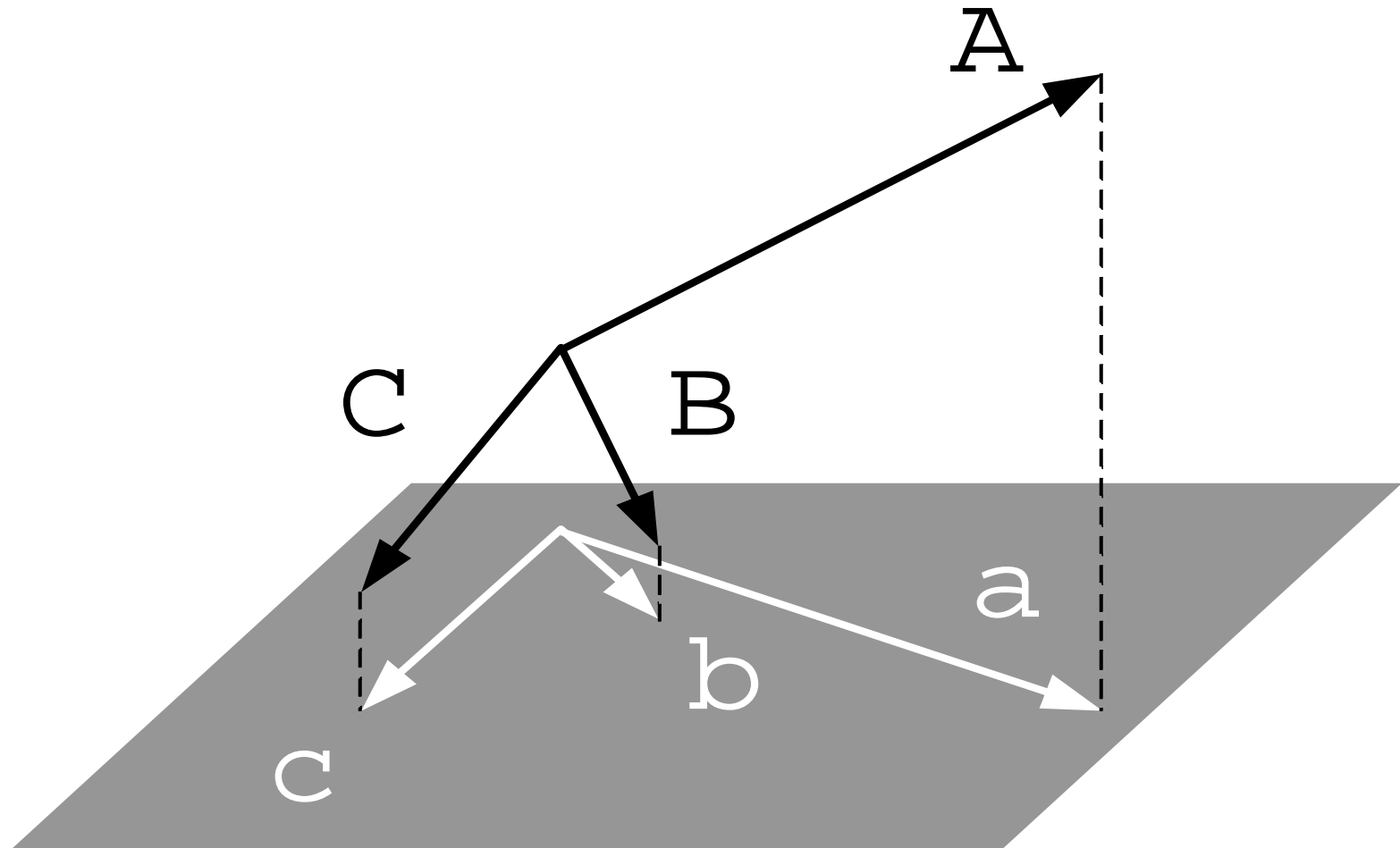


Figure 7: Vectors projected on a space

Free Choices of SVD Method

1. Choice of terms (Uniformly distributed terms and infrequent terms should be wiped. Verbs and adjectives are considered.)^a
2. Choice of k (Now, only by the performance of experiment. The usual dimension is $50 \sim 150$.)
3. Entity of TD matrix:
 - count of term
 - frequency of term
 - Boolean value of term

Or, term \rightarrow concept firstly, then count, frequency or Boolean value of concept.

^aThe orthogonality can be explained by independency.

Conclusion

1. How SVD works in IR and text clustering?
2. How to utilize the semantic knowledge base?
3. The open problems of SVD method
 - Choosing the efficient set of terms
 - Choosing the best k
 - The statistical method associated with SVD
 - (a) Nonparametric modification of TD matrix
$$\mathbf{A} = (a_{ij})_{m \times n}, \text{ or}$$
 - (b) Bayesian method ([1])

Homework 11 Further reading of [2] for SVD updating in real-time, which is related to the theory of Matrix Computation ([4]).

References

1. J. O. Berger (1985), *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.
2. M. W. Berry S. T. Dumais and G. W. O'brien (1995), *Using Linear Algebra for Intelligent Information Retrieval*. In SIAM Review, Vol. 37, No. 4, pp573-595.
3. P. J. Bickel and K. A. Doksum (2001), *Mathematical Statistics — Basic Ideas and Selected Topics* (2nd Ed). Prentice-Hall, Inc.
4. G. H. Golub and C. F. van Loan (1996), *Matrix Computation*. The John Hopkins University Press.
5. R. A. Johnson and D. W. Wichern (2003), *Applied Multivariate Statistical Analysis* (5th Ed). Pearson Education, Inc.
6. C. D. Manning and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*. The MIT Press.



**Thank you
for your attention!**