Visually Do Statistics

(1)What is the advantage of studying statistics ?
(2)Visualize Singular Value Decomposition

2019/02/26

Prof. Yukari SHIROTA (Gakushuin University)
Prof. Basabi Chakraborty (Iwate Prefectural University)

Today I would like to talk about why we need to study statistics. If you learn statistics, you will be able to take more scientific approaches when you handle data around you such as experimental data and stock prices. In this talk, using frequently done mistakes, I will show you how we should answer a question using the knowledge of statistics. Then, let me move to the first mistake.

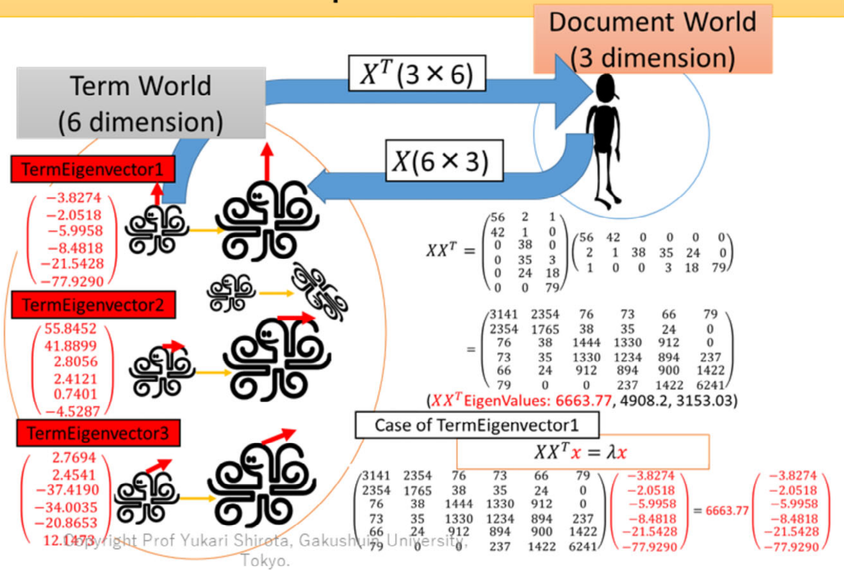
Visualization of Latent Semantic Analysis

http://www.iaiai.org/journals/index.php/IEE/article/view/70

Let me move on to the next topic, Latent Semantic Analysis.

SVD (Singular Value Decomposition) can be used for non square matrixes

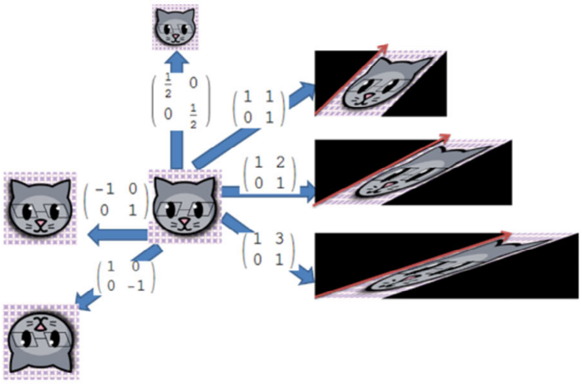
- Latent Semantic Analysis



Latent Semantic Analysis is used for topic extraction in text mining. The mathematics used here is Singular Value Decomposition. SVD can extract eigenvectors and eigenvalues from the given matrix. In the event, the shape of the matrix X is rectangle, not square. In my talk, finally you can understand the above illustration. Let's start the explanation.

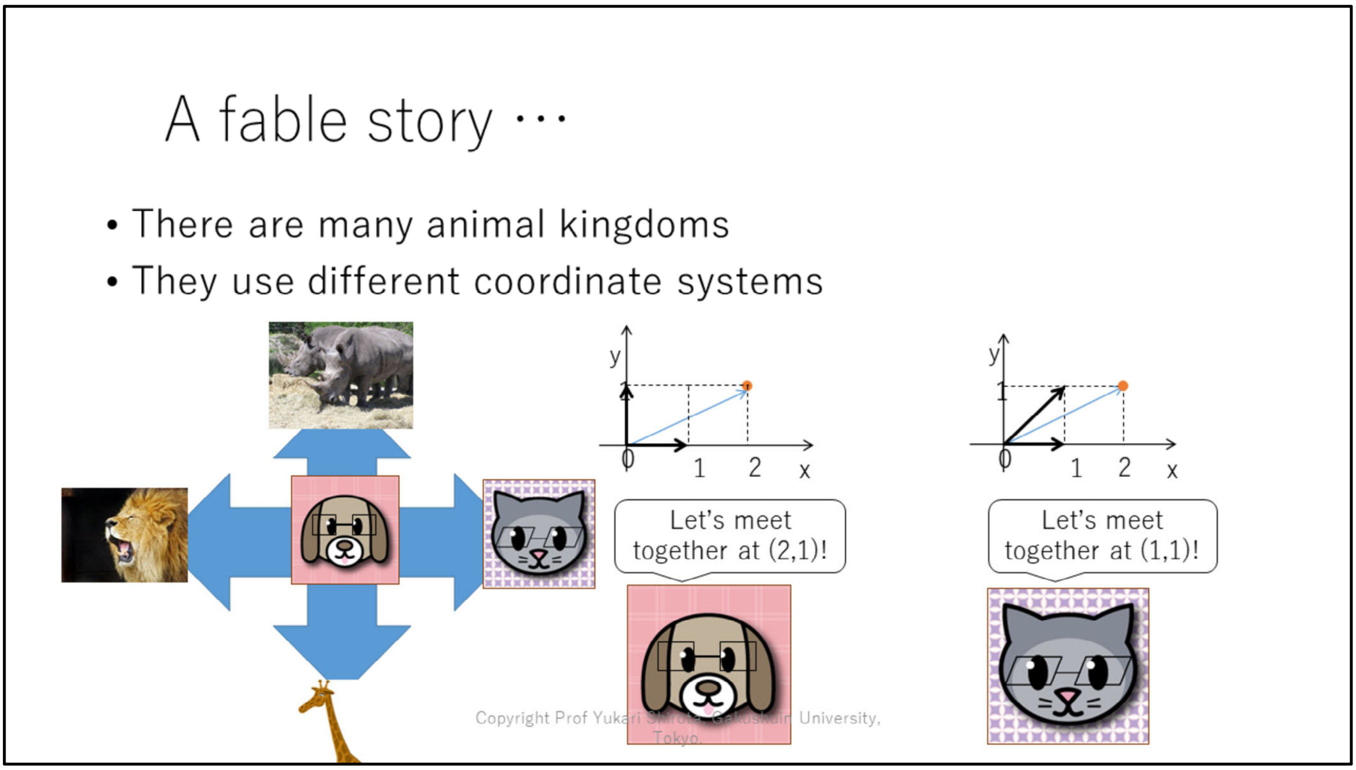
Review: Eigenvalues

- In advance, students should review the concept of eigenvalues
- A fable story for understanding eigenvalues "Enlargement Factors of the Magnification Machine are Eigenvalues" by Shirota

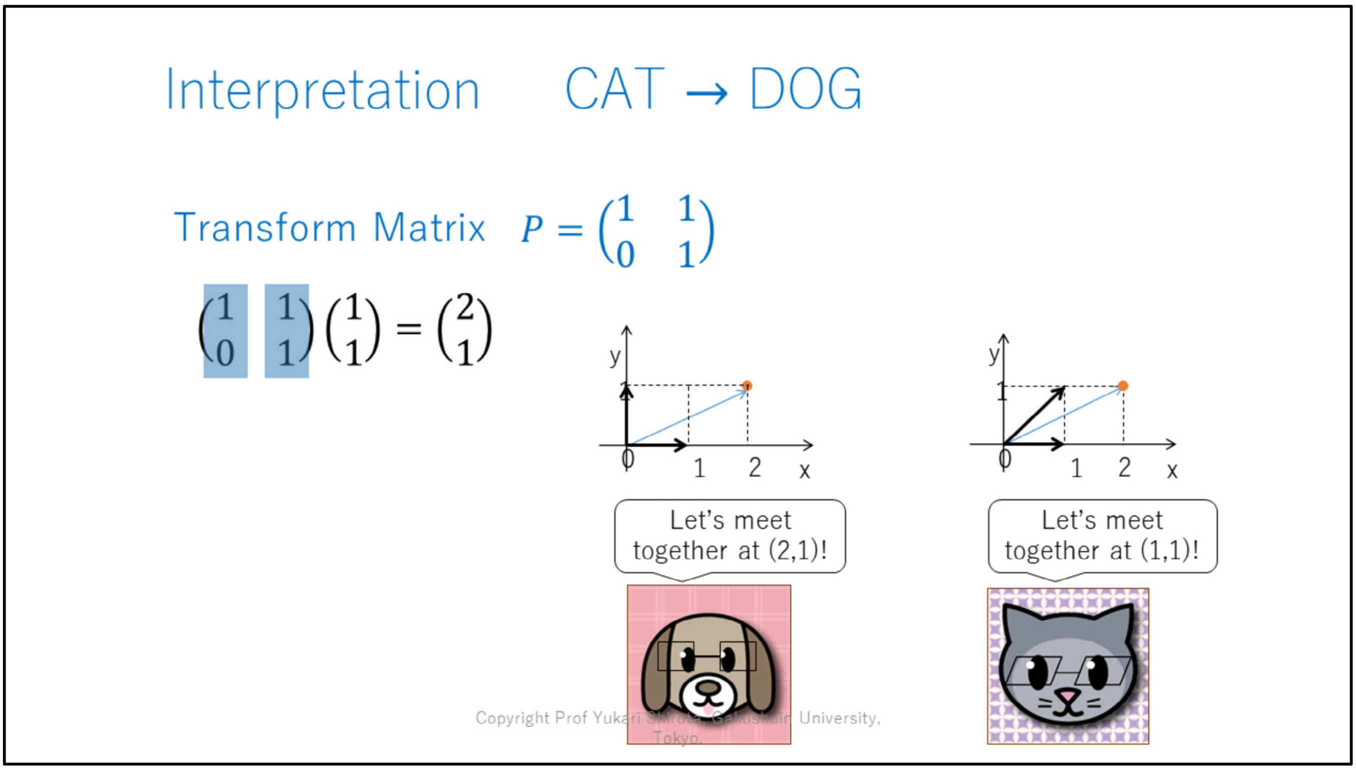


http://www-cc.gakushuin.ac.jp/~20010570/mathABC/SELECTED/

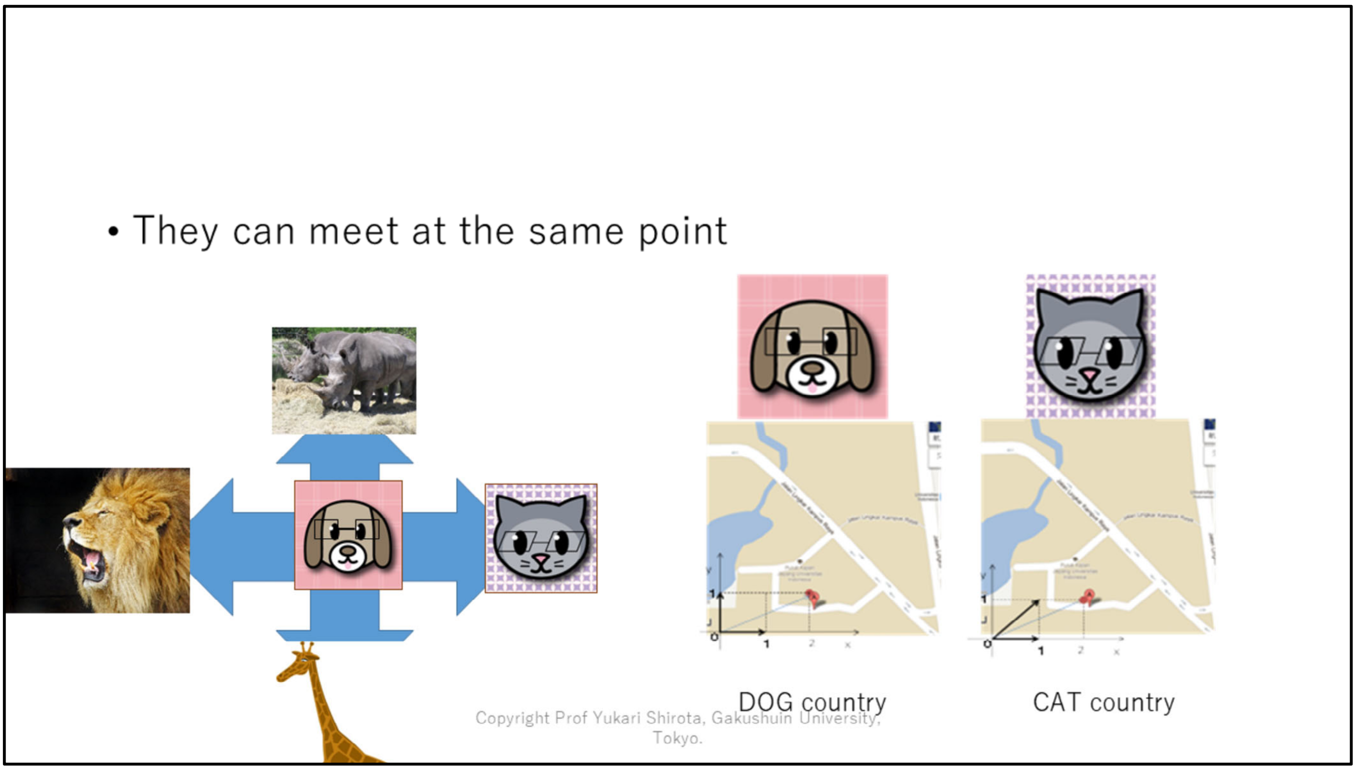
Shortly review eigenvalues here.



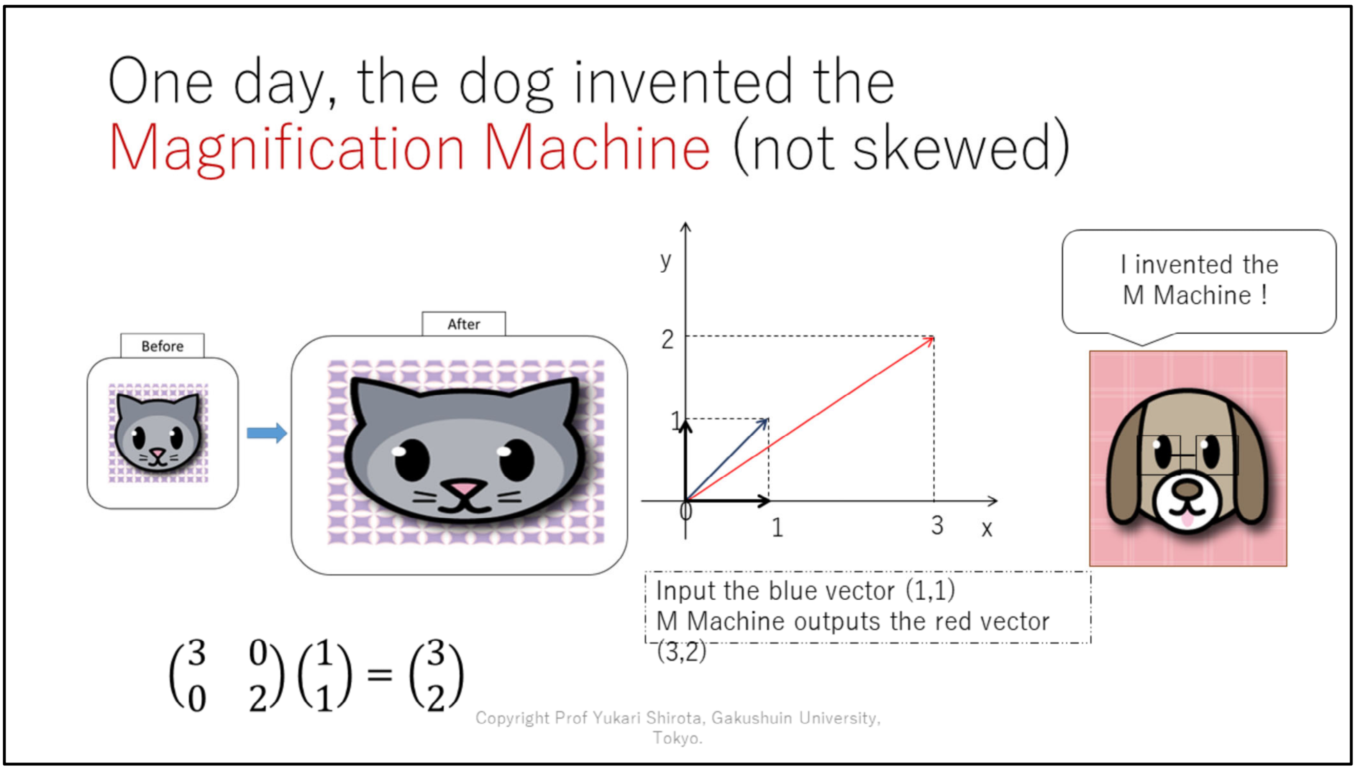
There are many animal kingdoms. Suppose that they live in a two dimensional world.
They use different coordinate systems, namely they have different base vectors.
So the same position is expressed up to their coordinate systems.



For example, let's consider the dog and cat countries.
To translate the cat's country position expression to the dog's one, the transform matrix is required.



Anywhere around the world can be expressed differently.



One day, the genius dog invented the magnification machine. That is the special matrix.
Any vector can be magnified without skewing.
But when the cat who has moved to the dog country tried to use this machine, the machine could not work well.
The input photo has been skewed.
What is the reason of the trouble.
The answer is the difference of their coordinate systems.

To import the M Machine to other countries, an interpretation is needed

- Input the image expressed by CAT coordinates
  ↓
- P Transform from CAT coordinates to DOG coordinates
  ↓
- A Magnify the image
  ↓
- P^{-1} Transform from DOG coordinates to CAT coordinates
  ↓
- Output the image expressed by CAT coordinates

Some interpretation must been needed before the magnification.
The machine is based on the dog's country base vectors. So
In advance we have to change the transform to the cat's country expression to the dog's one.
Then we should have done the magnification.

The eigenvalue is the same value after any transformation

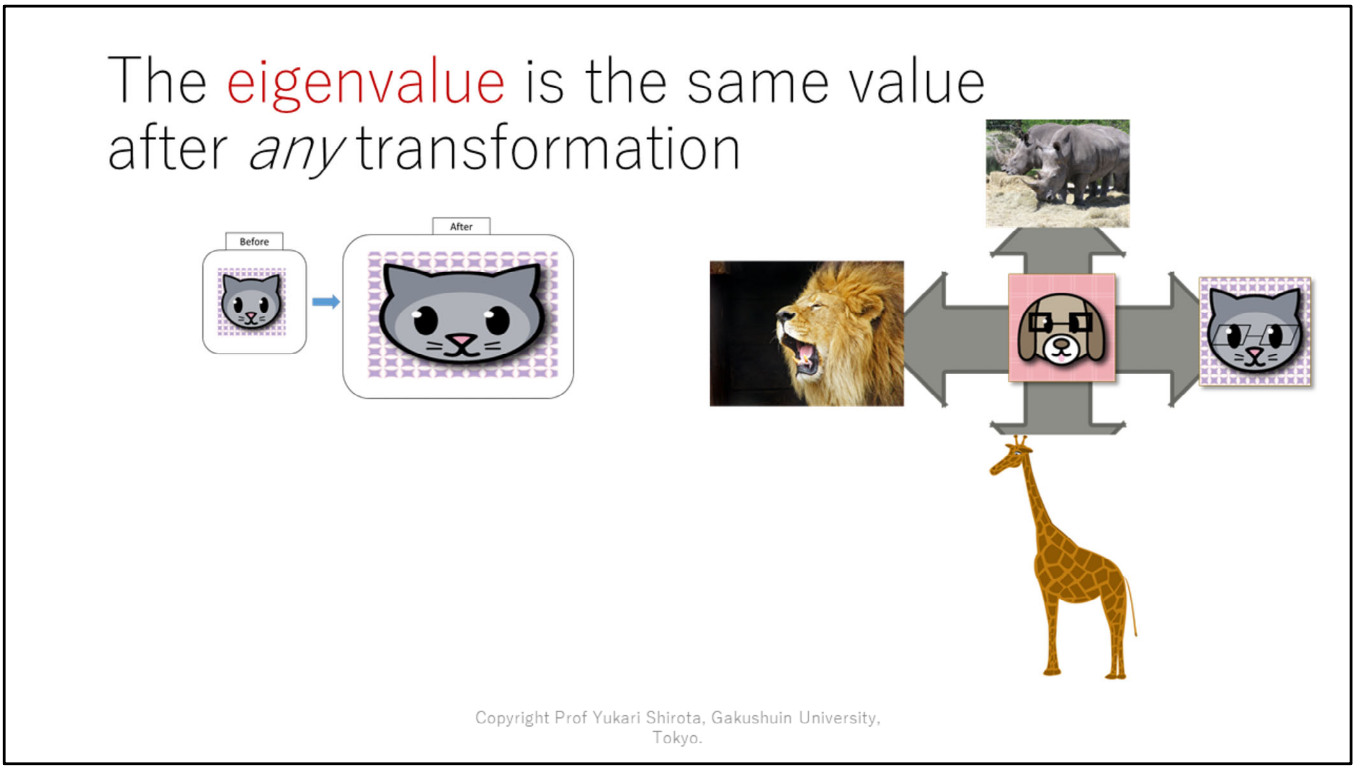
DOG $\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$ → Though the expressions are different, the same point.

CAT $\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ ↗

$\begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}$ is the cat's country's 3 × 2 magnification machine

The eigenvalues are 3 and 2 which are the same as the magnification factors.

The magnification machine for the cat country becomes a 3, 1, 0, 2 matrix.
In addition, the eigenvalues of the matrix are 2 and 3, which are the same as the initial magnification factors.
The magnification machine can keep the magnification factors even if the interpretation has been conducted like this.
This is called invariance of eigenvalues.
Eigenvectors change but the eigenvalues do not change.



That is the concept of eigenvalues. They keep the invariance.
This is the end of the explanation of the eigenvalues.

LSA uses SVD

- LSA is widely used in text-mining.
- Make a term-document matrix, counting the occurrences of each term in each document
- Ex: 3 documents and 6 terms
  - Transformation from document-world to term-world
  - Transformation from term-world to document-world

$$\begin{pmatrix} 56 & 2 & 1 \\ 42 & 1 & 0 \\ 0 & 38 & 0 \\ 0 & 35 & 3 \\ 0 & 24 & 18 \\ 0 & 0 & 79 \end{pmatrix}$$

transpose

$$\begin{pmatrix} 56 & 42 & 0 & 0 & 0 & 0 \\ 2 & 1 & 38 & 35 & 24 & 0 \\ 1 & 0 & 0 & 3 & 18 & 79 \end{pmatrix}$$

Let's be back to the LSA.
In the topic extraction, firstly, the term-document matrix is given. In the above case, there are 3 documents and the 6 significant terms. Each figure in the matrix shows the number of appearances of each term in each document. The transpose of the given matrix has the size of 3 lines times 6 columns.

Singular value Decomposition

The given data is only the term-document matrix.

6 Keywords By TFIDF

3 Documents

	Document #1	Document #2	Document #3
affection	56	2	1
passion	42	1	0
expectation	0	38	0
desire	0	35	3
promise	0	24	18
believe	0	0	79

Unknown What intrinsic concepts exist ? -> Latent Semantics

Suppose that there are 3 documents as shown here. The significant terms selected in advance are supposed to be affection, passion, expectation, desire, promise, and believe. In Document #1, we found 56 time the term "affection". Here I will ask you a question ? Latent Semantic Analysis is supposed to be an analysis for intrinsic concepts which are called Latent Semantics. In these documents, what intrinsic concepts exist ? Only seeing the given matrix enables us to think there are just 3 latent semantics.

Singular Value Decomposition

The given data is only the term-document matrix.

6 Keywords By TFIDF | 3 Documents

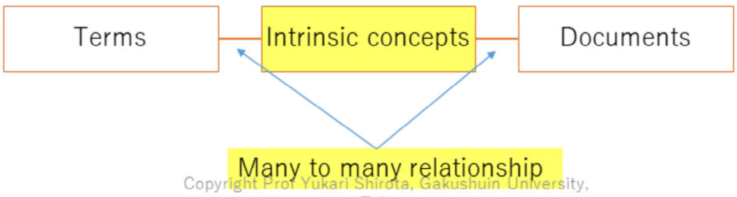
	Document #1	Document #2	Document #3
affection	56	2	1
passion	42	1	0
expectation	0	38	0
desire	0	35	3
promise	0	24	18
believe	0	0	79

Unknown	This group latent semantic may be [Love]	This group latent semantic may be [Hope]	This group latent semantic may be [Trust]

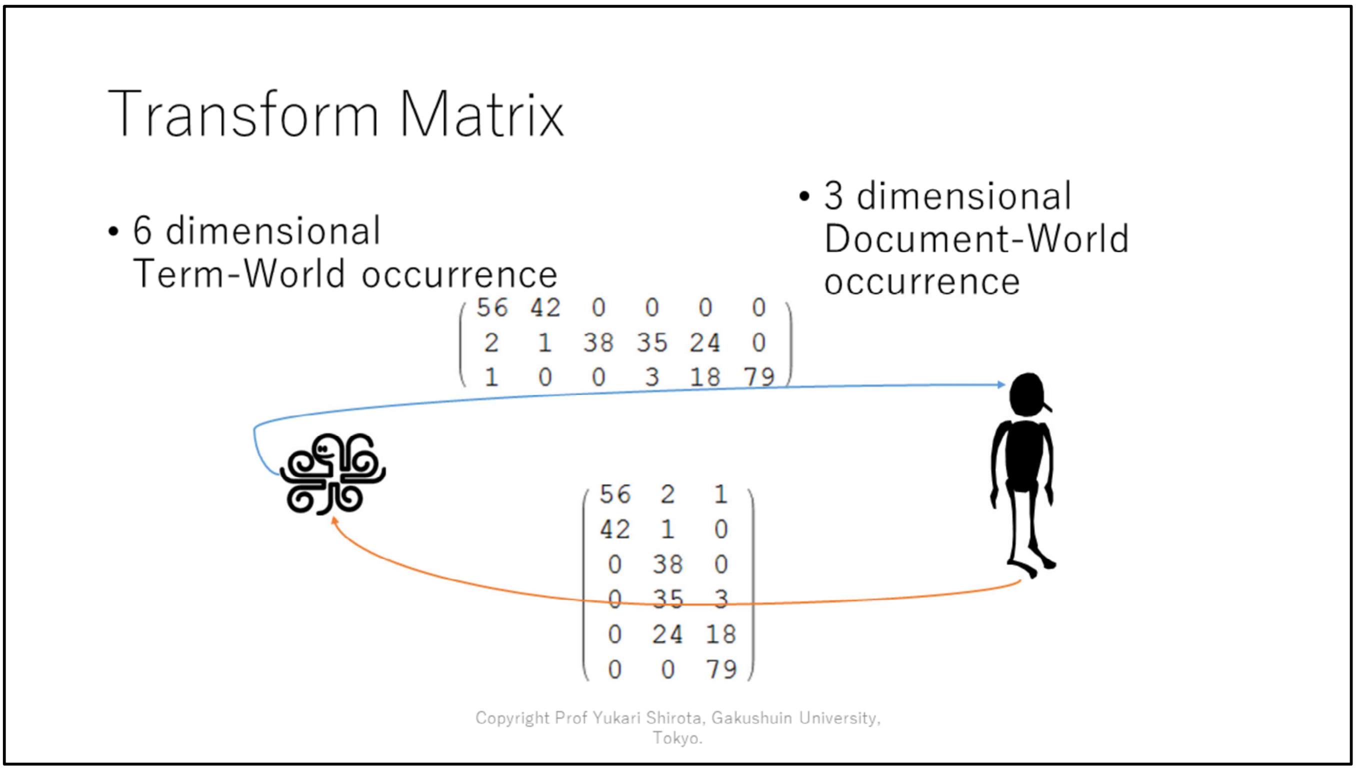
Can you name the three latent semantic concepts ? We have to select the concept to associate the terms to the concept. I named the first one as LOVE, the second one as HOPE, and the last one as TRUST. The terms affection and passion are associated to LOVE. The terms expectation, desire, and promise are associated to HOPE. The terms promise and believe are associated to TRUST.

LSA

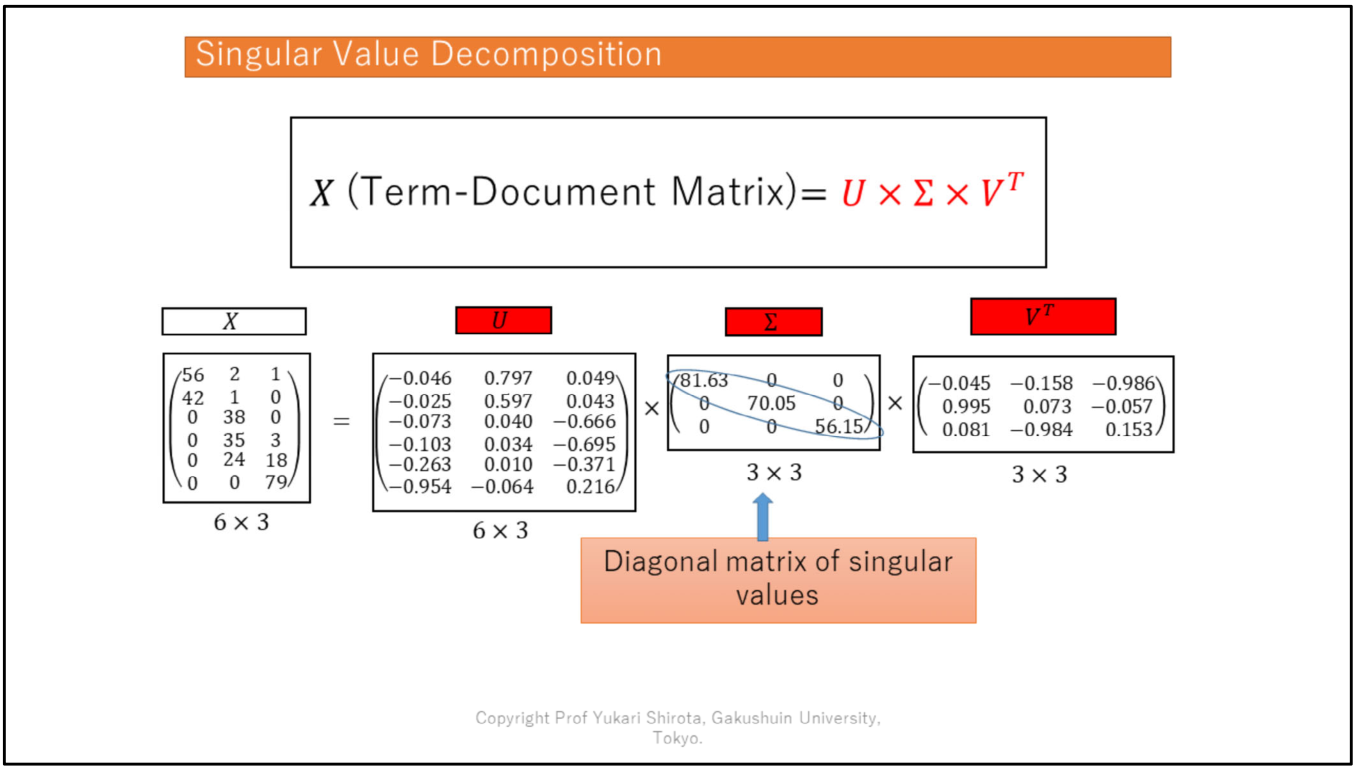
- From the relationship between terms and documents, we want to find intrinsic concepts on the given term-document matrix.
- For example, the intrinsic concepts are love, hope, and trust.
- The latent semantics corresponds to the eigenvector. The eigenvalue shows the amplitude of the latent semantics.



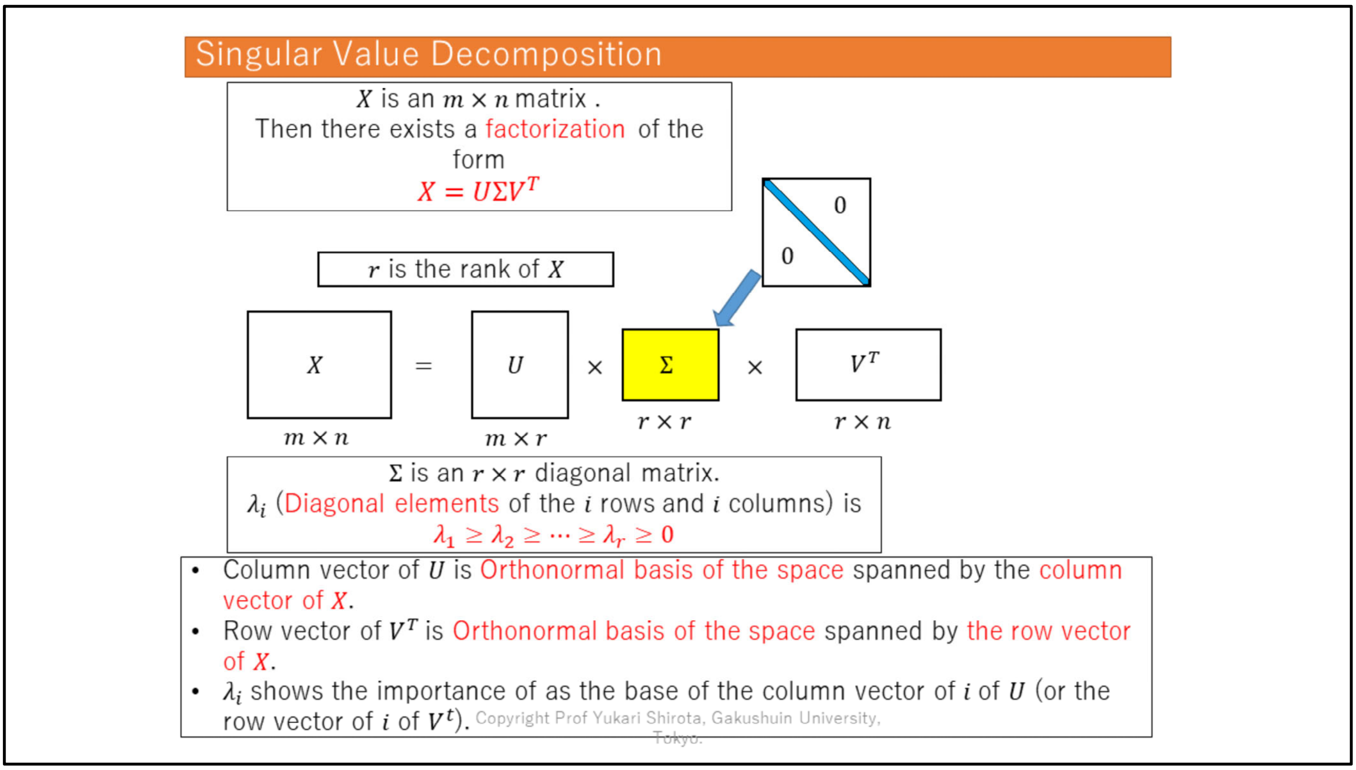
Suppose that there are three concepts in the three documents; they are LOVE, HOPE, and TRUST. How can we express the concepts ? If you use terms, you can express the concept using terms. For example, LOVE is such a concept which included the meaning of affection and passion. On the other hand, if you use documents, you can express the concept using the documents. You will be able to express the concepts using the ratio of the given documents. For example, I would like to read a comic which consists of 50% “The Rose of Versailles”, 30% “NARUTO”, and 20% “ONE PIECE”.



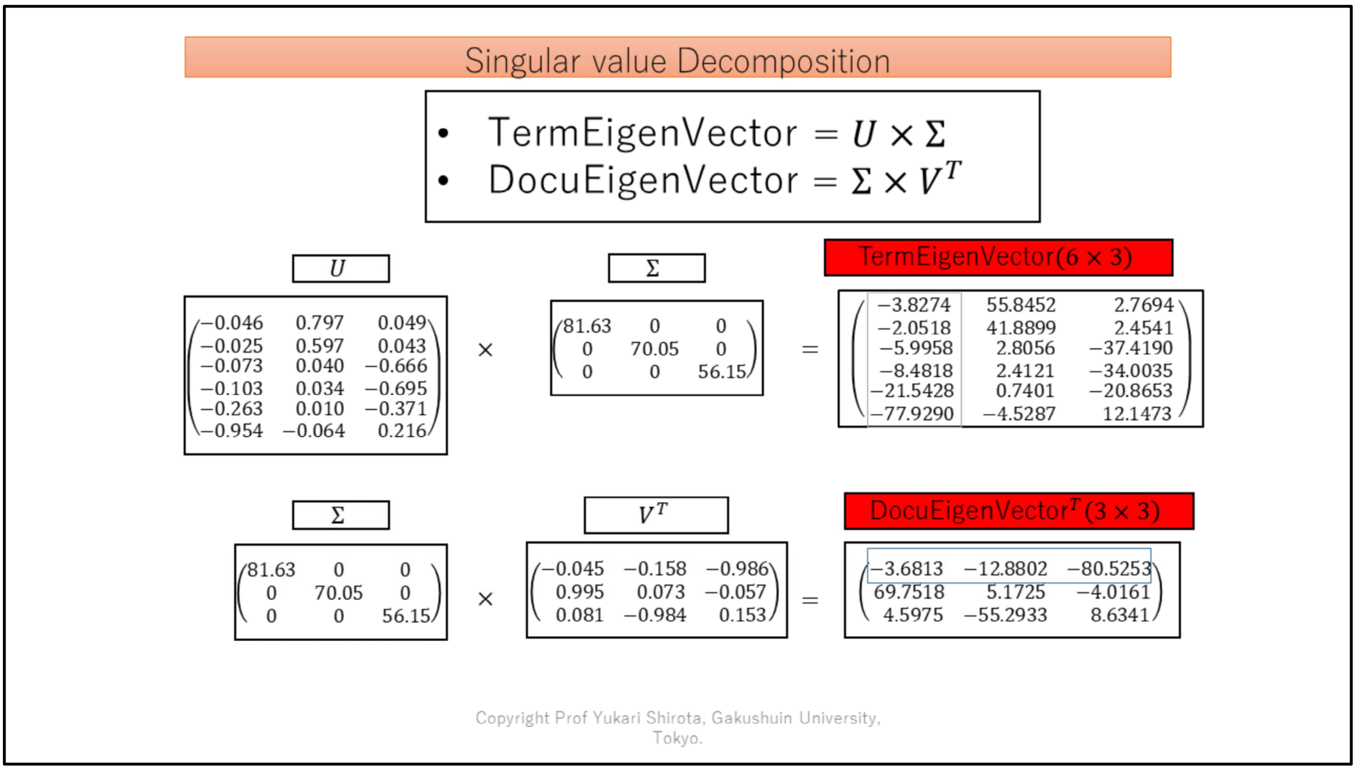
Let me here think about the two worlds. They are a term-world and a document-world. In the given case, the term-world is a 6 dimensional world, because the selected terms are 6. On the other hand, the document-world is a 3 dimensional world, because the number of the given documents are 3. The given matrix named X expresses the relationship from the 6 dimension world to the 3 dimension world. On the other hand, the transpose of X expresses the relationship from the 3 dimension world to the 6 dimension world.



Let me explain the Singular Value Decomposition. Input the matrix X, the process SVD outputs three matrices, U and sigma, and a transpose of V. The central matrix named sigma is a diagonal matrix singular values. The three figures are eigenvalues of matrix X. They are 81.63, 70.05, and 56.15. The value shows the impact or amplitude of each intrinsic concept such as LOVE.



This is an explanation of SVD which you would often see in a textbook.



From SVD, we can get 2 kinds of eigenvectors. The first one is U times sigma. Another is sigma times transpose of V. I will call them TermEigenVector and DocuEigenVector. The names are just in my talk and the names are not formal terms in mathematics. The number of TermEigenVectors is 3. Among them, let's see the framed column vector. The element values are -3.8, -2.0, -5.9, -8.4, -21.5, and -77.9 approximately. Which ones have a large absolute value ? The 5th and 6th elements have large absolute values. The corresponding terms are promise and believe. So we understand the concept is TRUST. Similar to this, we found the second column shows LOVE and the third column shows HOPE. Next let me move on to the transpose of DocuEigenVectors. Here the first line vector shows TRUST. The second line shows LOVE. The third line shows HOPE.

Invariance of Eigenvalues

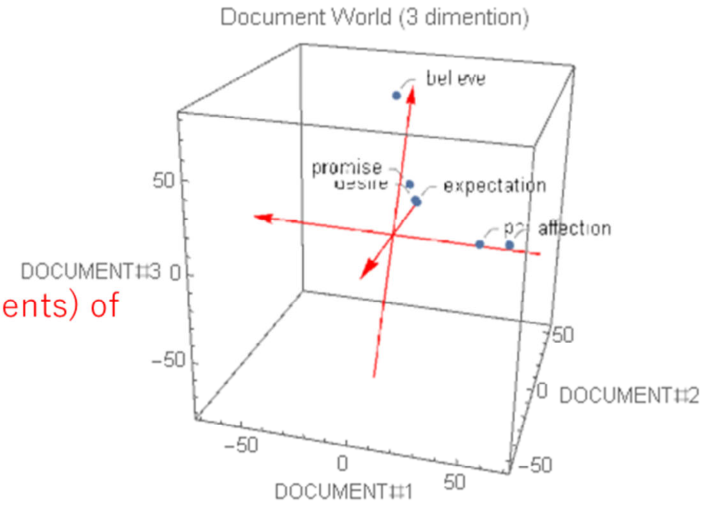
$$\begin{pmatrix} 81.63 & 0 & 0 \\ 0 & 70.05 & 0 \\ 0 & 0 & 56.15 \end{pmatrix}$$

If there are three intrinsic concepts in the documents.
The concept can be expressed both in the term distribution and the document distribution.
ASPECT 1: Blend ratio among 6 keywords
ASPECT 2: Blend ration among three documents
（＃１：８０％、＃２：１５％、＃３：５％）
The eigenvalue corresponds to the impact factor.

Let's here conclude the discussion so far. The intrinsic semantics can be expressed in the two approaches. The first one is a ratio of 6 terms. The second one is a ratio of three documents. For example, instead of saying 80% LOVE, 15% HOPE, and 5% TRUST, you can say 80% Document #1, 15% Document #2, and 5% Document 5%. In addition, eigenvalues show amplitude of each concept in the given data. In both worlds, the eigenvalues are the same.

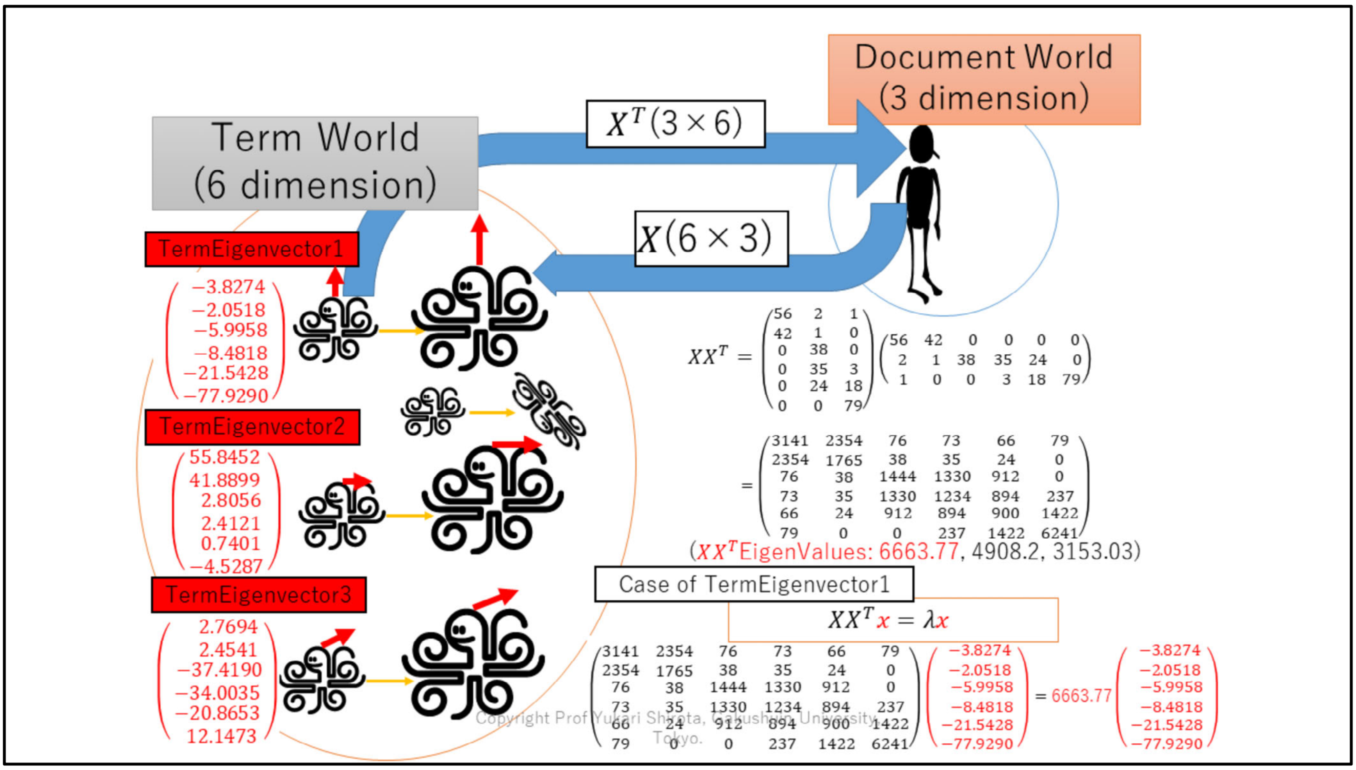
“affection” is transformed by X^T

$$\begin{pmatrix} 56 & 42 & 0 & 0 & 0 & 0 \\ 2 & 1 & 38 & 35 & 24 & 0 \\ 1 & 0 & 0 & 3 & 18 & 79 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 56 \\ 2 \\ 1 \end{pmatrix}$$

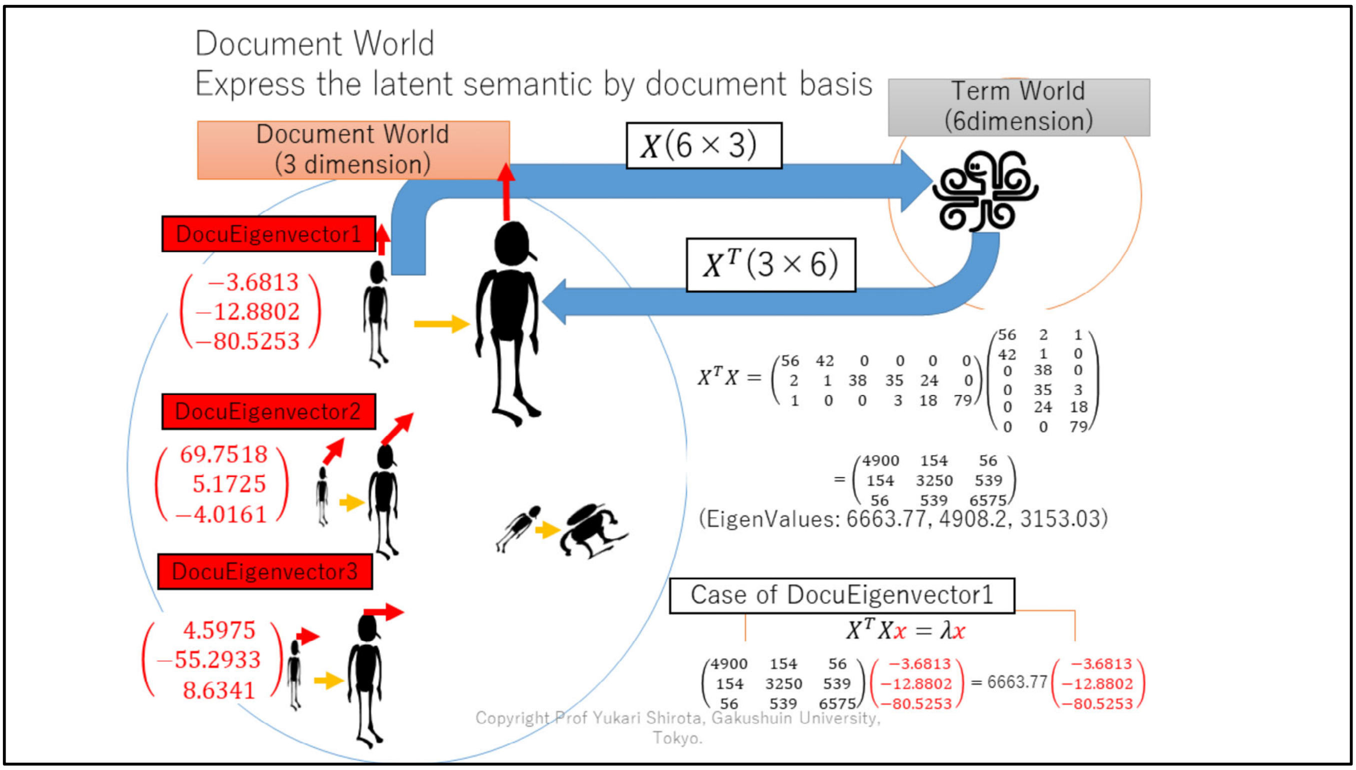


3 Eigenvectors (Principal Components) of the Document World

Let’s consider the transformation between the term world and the document world.
The term affection is expressed in the term world as (1, 0, 0, 0, 0, 0).
The tern affection is moved to (56,2,1) in the document world.

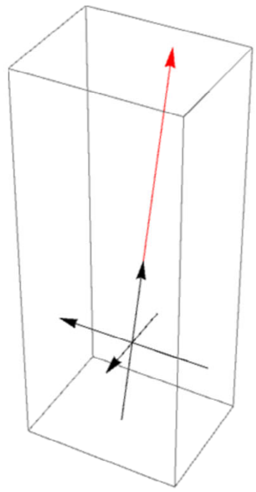


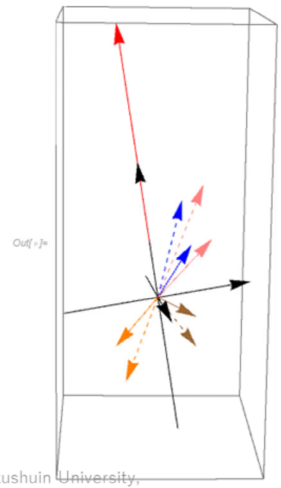
Visually I will explain the invariance of eigenvalues. Let's see this illustration. In the left side, there is a term-world with 6 dimensions. The term-world occurrence is supposed to be transformed to the document world using the transpose of X. Namely, we shall calculate the transpose of X times the occurrence vector. And then, we shall consequently transform the transformed occurrence back to the term-world, using X. Any occurrence term vector has been skewed after the 2 transformations. However, only 3 TermEigenVectors have been enlarged, not skewed. The enlargement factors are approximately 6663.77, 4908.2, and 3153.03. The figures are squares of eigenvalues which are approximately 81.63, 70.05, and 56.15. This is because the round trip make then squared.



Let me try the same operation from the document-world. The only three DocuEigenVectors can keep the same shape after the round trip transformations by X times the transpose of X. The DocuEigenVector showing LOVE is transformed to the TermEigenVector showing LOVE. Let's consider the given data of SVD problem. The given data was only matrix X. The matrix X includes the intrinsic semantic eigenvectors which can be expressed in both document-world and term-world.
SVD is also used in Random Matrix Theory in Econophysics. Random Matrix Theory is a time-series data analysis method of stock price fluctuations. SVD is used in many application fields.

Only 3 eigenvectors (Principal Components) of the Document World keep the same direction after the X and X^T transformations





Conclusion of LSA and SVD

- LSA uses SVD
- LSA extracts the intrinsic semantics.
- The intrinsic semantics is the eigenvectors which can be found by the SVD.
- The eigenvalue expresses the amplitude of the eigenvector.