# Searching Big Data with SVD: Textual Semantic Spaces, and Similar Problems

Wen-ming Ye

Developer Platform Evangelism

# Friends don’t let friends do just word count

- Understand the **math** behind Big Data
- Singular Value Decomposition can be applied to more than just textual search
- Big Data: Statistics and Pattern Search
- Big Data = Big opportunities

# Semantic Matrix Model

Basic idea:

- The meaning of a document is determined by the words that appear in it
- The meaning of a word is determined by the documents in which it appears

- d1: ***Human*** machine ***interface*** for ABC ***computer*** applications
- d2: A ***survey*** of ***user*** opinion of ***computer*** ***system*** ***response*** ***time***
- d3: The ***EPS*** ***user*** ***interface*** management ***system***
- d4: ***System*** and ***human*** ***system*** engineering testing of ***EPS***
- d5: Relation of ***user*** perceived ***response time*** to error measurement
- m1: The generation of random, binary, ordered ***trees***
- m2: The intersection ***graph*** of paths in ***trees***
- m3: ***Graph*** ***minors*** IV: Widths of ***trees*** and well-quasi-ordering
- m4: ***Graph*** ***minors***: A ***survey***

corpus

Document vector

We select some words from the corpus for demonstration purposes; most analyses will include almost all words (some use a "stop list" for high frequency words like "the").

Word vector

Word-document matrix

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

# Rebasing the Space

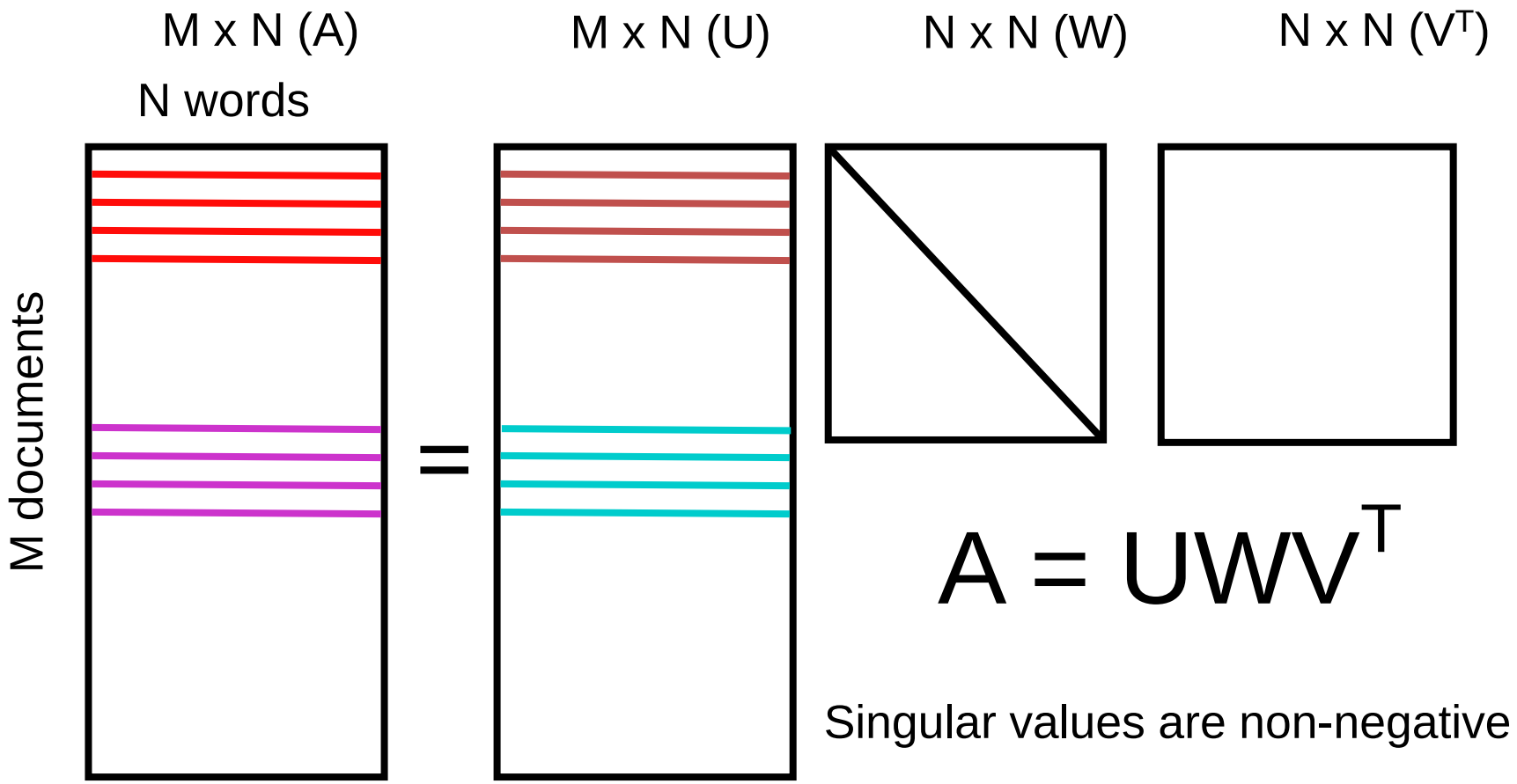
- The meaning of a document is determined by the words that appear in it
  - $m(d_i) = a_{i1}\, m(w_1) + a_{i2}\, m(w_2) + \ldots + a_{in}\, m(w_n)$
- The meaning of a word is determined by the documents in which it appears
  - $m(w_j) = a_{1j}\, m(d_1) + a_{2j}\, m(d_2) + \ldots + a_{mj}\, m(d_m)$

Although the word and document *vectors* are invariant, the *coefficients* $a_{ij}$ are transformed with each basis transformation.

**All vector-based text treatment methods aim to find a new set of coordinates that will make the relevant semantic properties of the text obvious**.

**Singular Value Decomposition** (SVD) is the basis of a number of vector-based methods like Latent Semantic Analysis (LSA) and Principal Component Analysis (PCA), Independent (ICA)

# SVD Possible for any MxN Matrix




Decomposition is unique up to the interchange of entire raws or columns

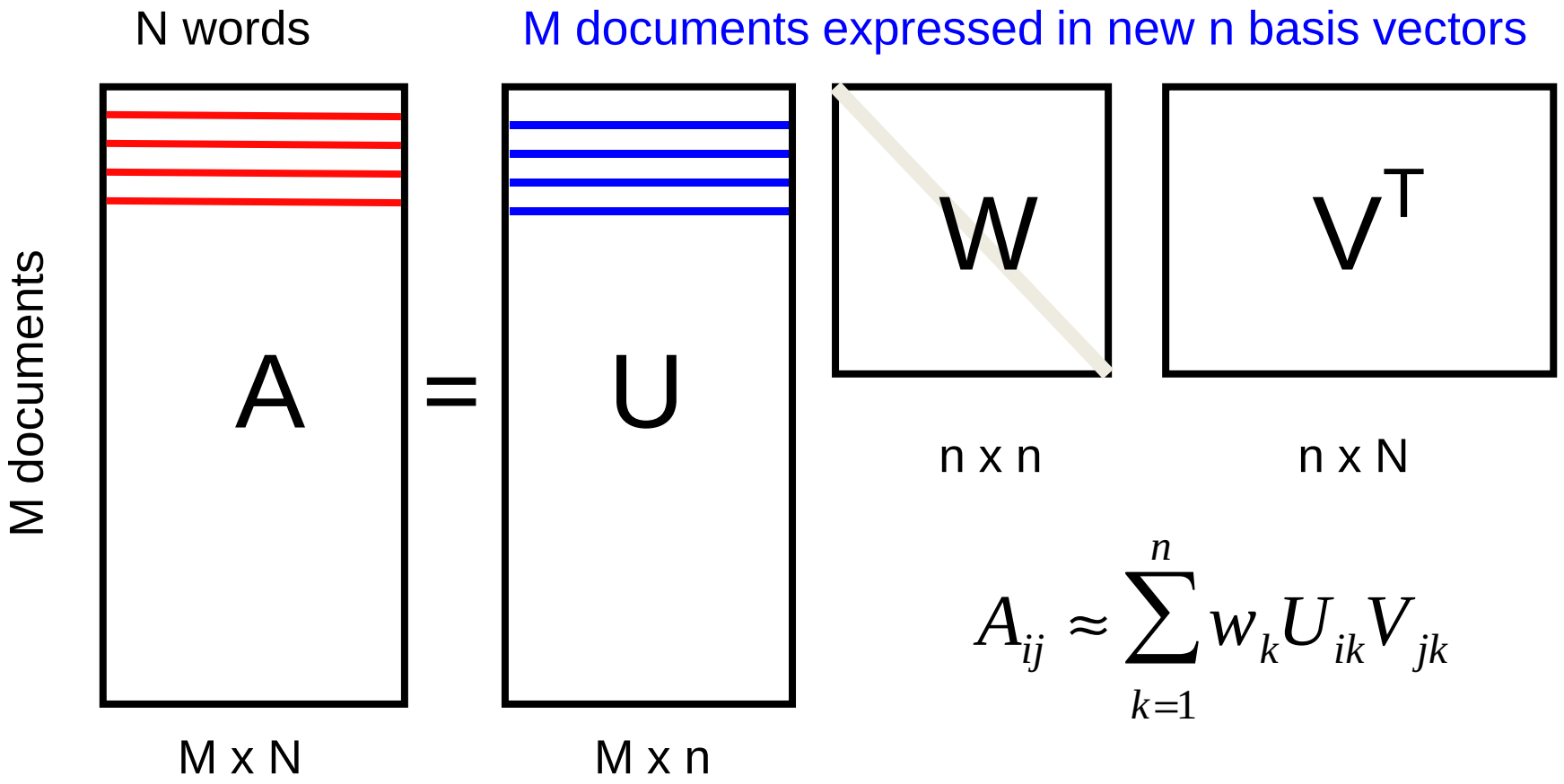
U, V are column orthonormal: $U^TU=V^TV=I$

Colums of U, V, are new basis vectors, so the new bases are orthonormal

# Dimension Reduction: Keep Only a few Singular Values

If the singular value spectrum decays quickly, A can be approximated well by only “a few” singular values

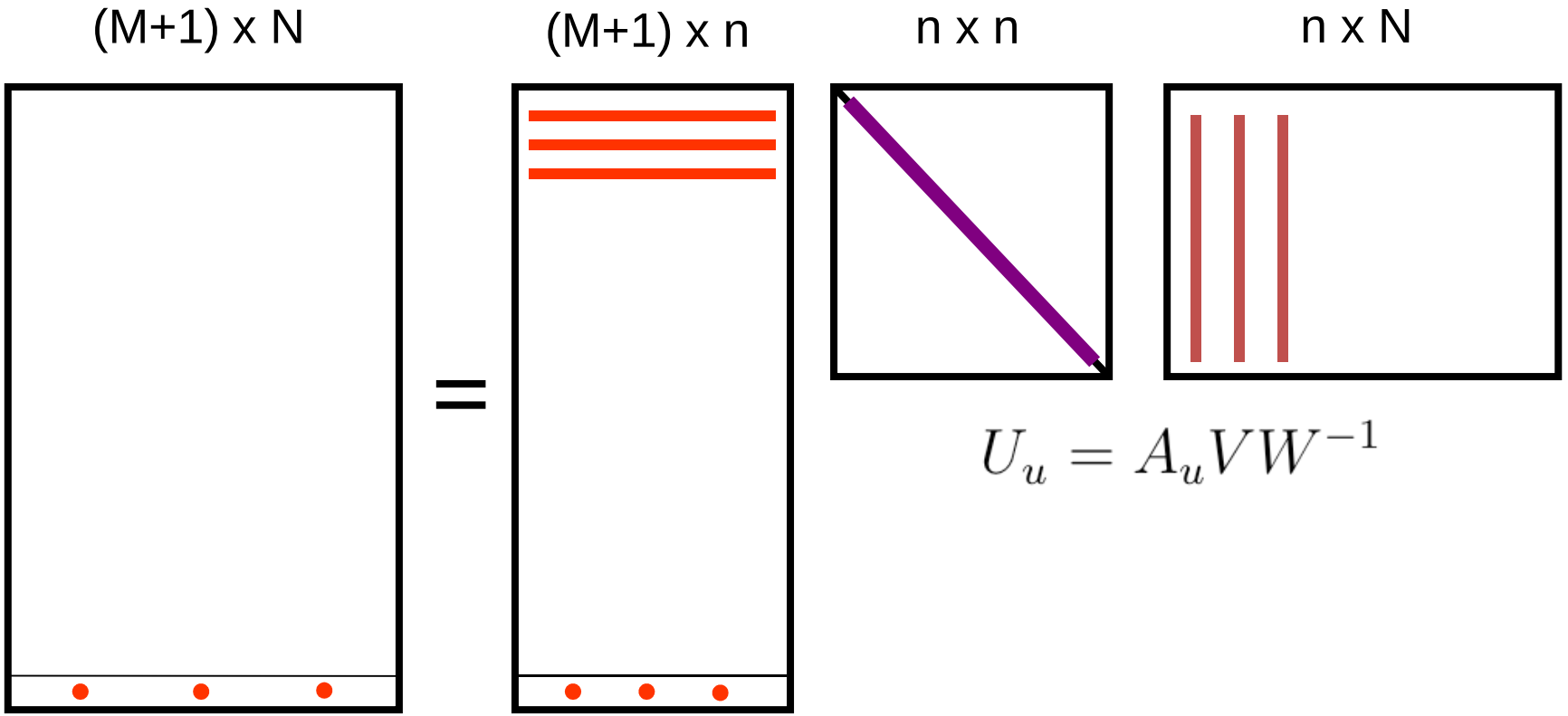
eg. in language applications M~O($10^7$), N~O($10^5$), but we keep n~300




$$A_{ij} \approx \sum_{k=1}^{n} w_k U_{ik} V_{jk}$$

# Arbitrary New Document can be Projected onto New Basis

$$
\begin{aligned}
[A : A_u] &= [U : U_u]\, WV^T \\
[A : A_u]\, V &= [U : U_u]\, WV^T V \\
[A : A_u]\, VW^{-1} &= [U : U_u]
\end{aligned}
$$




$$U_u = A_u V W^{-1}$$

# Different Similarity Measures Emphasize Different Aspects

Dot Product

$$x.y = \sum_{i=1}^{N} x_i y_i$$

Cosine

$$\cos(\theta_{xy}) = \frac{x.y}{|x||y|}$$

Euclidean

$$euclid(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

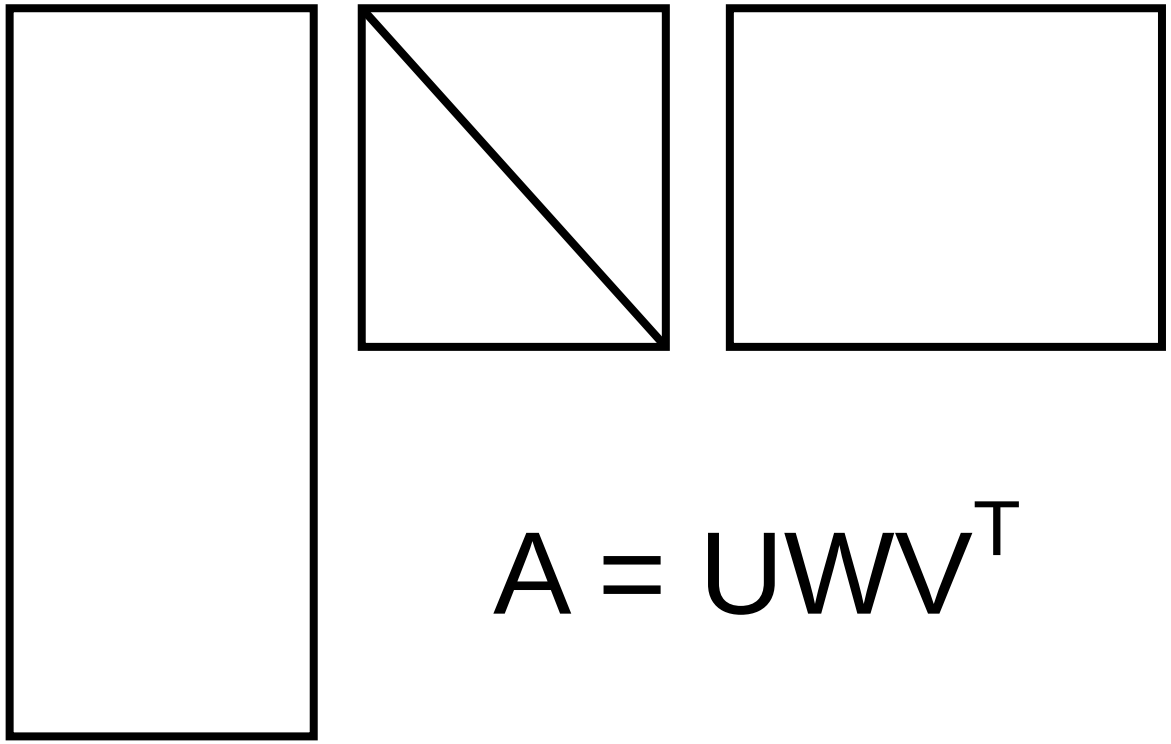
We often use inverse-cosine distances ($\cos^{-1}(q)$).
It de-emphasizes vector length in similarity measure.

# Example:
# SVD on the Word-by-Document Matrix

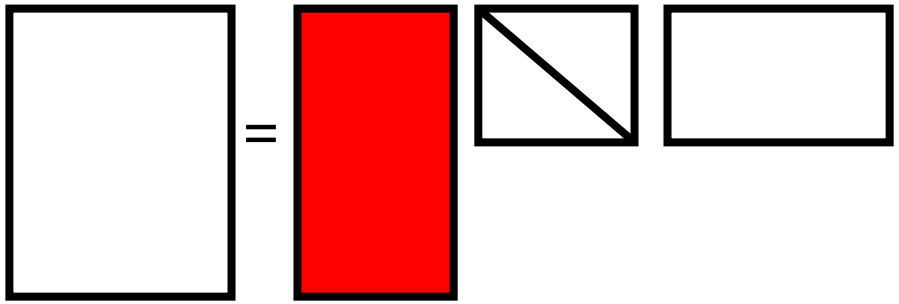
Documents

Words

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

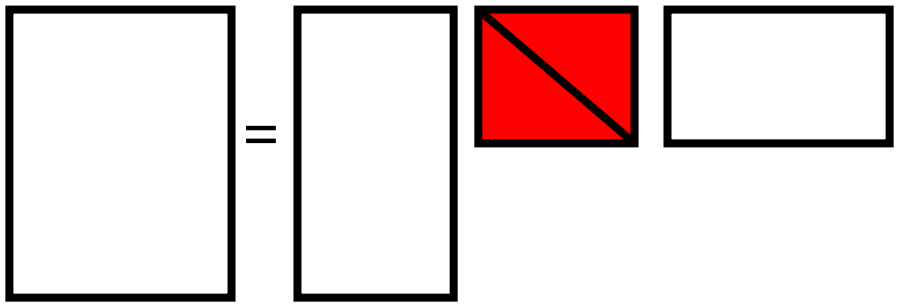


$$A = UWV^T$$



Singular value Decomposition of the words by docs matrix

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.22 | -0.11 | 0.29 | -0.41 | -0.11 | -0.34 | 0.52 | -0.06 | -0.41 |
| 0.20 | -0.07 | 0.14 | -0.55 | 0.28 | 0.50 | -0.07 | -0.01 | -0.11 |
| 0.24 | 0.04 | -0.16 | -0.59 | -0.11 | -0.25 | -0.30 | 0.06 | 0.49 |
| 0.40 | 0.06 | -0.34 | 0.10 | 0.33 | 0.38 | 0.00 | 0.00 | 0.01 |
| 0.64 | -0.17 | 0.36 | 0.33 | -0.16 | -0.21 | -0.17 | 0.03 | 0.27 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.30 | -0.14 | 0.33 | 0.19 | 0.11 | 0.27 | 0.03 | -0.02 | -0.17 |
| 0.21 | 0.27 | -0.18 | -0.03 | -0.54 | 0.08 | -0.47 | -0.04 | -0.58 |
| 0.01 | 0.49 | 0.23 | 0.03 | 0.59 | -0.39 | -0.29 | 0.25 | -0.23 |
| 0.04 | 0.62 | 0.22 | 0.00 | -0.07 | 0.11 | 0.16 | -0.68 | 0.23 |
| 0.03 | 0.45 | 0.14 | -0.01 | -0.30 | 0.28 | 0.34 | 0.68 | 0.18 |



Singular value Decomposition of the words by docs matrix

3.34

2.54

2.35

1.64

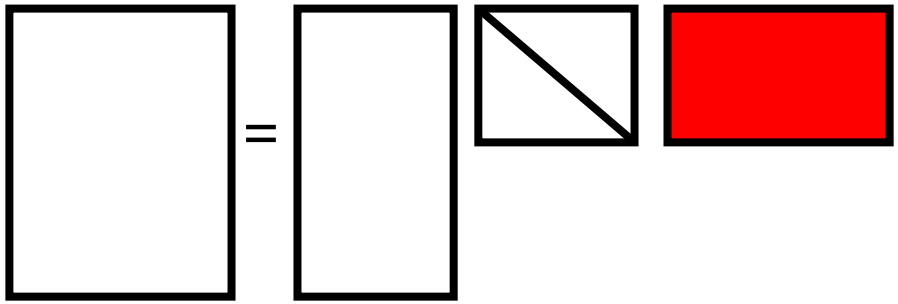
1.50

1.31

0.85

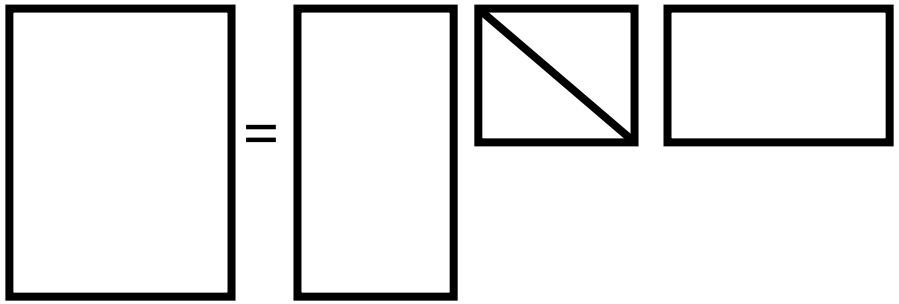
0.56

0.36



Singular value Decomposition of the words by docs matrix

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.20 | 0.61 | 0.46 | 0.54 | 0.28 | 0.00 | 0.01 | 0.02 | 0.08 |
| -0.06 | 0.17 | -0.13 | -0.23 | 0.11 | 0.19 | 0.44 | 0.62 | 0.53 |
| 0.11 | -0.50 | 0.21 | 0.57 | -0.51 | 0.10 | 0.19 | 0.25 | 0.08 |
| -0.95 | -0.03 | 0.04 | 0.27 | 0.15 | 0.02 | 0.02 | 0.01 | -0.03 |
| 0.05 | -0.21 | 0.38 | -0.21 | 0.33 | 0.39 | 0.35 | 0.15 | -0.60 |
| -0.08 | -0.26 | 0.72 | -0.37 | 0.03 | -0.30 | -0.21 | 0.00 | 0.36 |
| 0.18 | -0.43 | -0.24 | 0.26 | 0.67 | -0.34 | -0.15 | 0.25 | 0.04 |
| -0.01 | 0.05 | 0.01 | -0.02 | -0.06 | 0.45 | -0.76 | 0.45 | -0.07 |
| -0.06 | 0.24 | 0.02 | -0.08 | -0.26 | -0.62 | 0.02 | 0.52 | -0.45 |

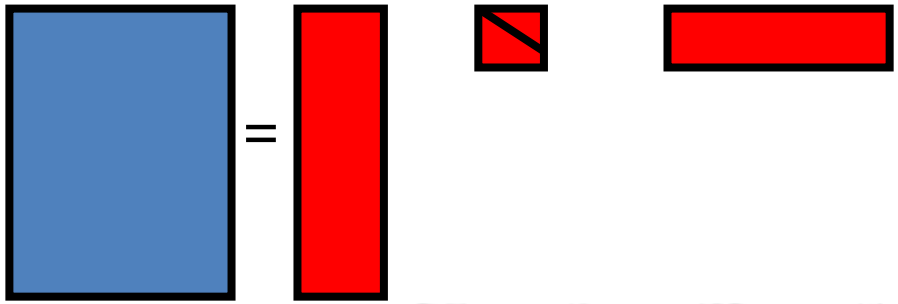


Singular value Decomposition of the words by contexts matrix

3.34

2.54

Dimensionality Reduction n=2



Singular value Decomposition of the words by contexts matrix

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 0.16 | 0.40 | 0.38 | 0.47 | 0.18 | -0.05 | -0.12 | -0.16 | -0.09 |
| interface | 0.14 | 0.37 | 0.33 | 0.40 | 0.16 | -0.03 | -0.07 | -0.10 | -0.04 |
| computer | 0.15 | 0.51 | 0.36 | 0.41 | 0.24 | 0.02 | 0.06 | 0.09 | 0.12 |
| user | 0.26 | 0.84 | 0.61 | 0.70 | 0.39 | 0.03 | 0.08 | 0.12 | 0.19 |
| system | 0.45 | 1.23 | 1.05 | 1.27 | 0.56 | -0.07 | -0.15 | -0.21 | -0.05 |
| response | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| time | 0.16 | 0.58 | 0.38 | 0.42 | 0.28 | 0.06 | 0.13 | 0.19 | 0.22 |
| EPS | 0.22 | 0.55 | 0.51 | 0.63 | 0.24 | -0.07 | -0.14 | -0.20 | -0.11 |
| survey | 0.10 | 0.53 | 0.23 | 0.21 | 0.27 | 0.14 | 0.31 | 0.44 | 0.42 |
| trees | -0.06 | 0.23 | -0.14 | -0.27 | 0.14 | 0.24 | 0.55 | 0.77 | 0.66 |
| graph | -0.06 | 0.34 | -0.15 | -0.30 | 0.20 | 0.31 | 0.69 | 0.98 | 0.85 |
| minors | -0.04 | 0.25 | -0.10 | -0.21 | 0.15 | 0.22 | 0.50 | 0.71 | 0.62 |

| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|---|---|---|
| human | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| interface | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| user | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| system | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| response | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| time | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| EPS | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| survey | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| trees | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| graph | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| minors | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

- d1: ***Human*** machine ***interface*** for ABC ***computer*** applications
- d2: A ***survey*** of ***user*** opinion of ***computer*** ***system*** ***response*** ***time***
- d3: The ***EPS*** ***user*** ***interface*** management ***system***
- d4: ***System*** and ***human*** ***system*** engineering testing of ***EPS***
- d5: Relation of ***user*** perceived ***response time*** to error measurement
- m1: The generation of random, binary, ordered ***trees***
- m2: The intersection ***graph*** of paths in ***trees***
- m3: ***Graph*** ***minors*** IV: Widths of ***trees*** and well-quasi-ordering
- m4: ***Graph*** ***minors***: A ***survey***

| | | | |
|---|---|---|---|
| Eigenvalues: | | 9 | 2 |
| human – user | : | 0 | .94 |
| human – minors: | | 0 | -.83 |

Dimensionality reduction brought out the fact that users are human, something not obvious from the text or the full matrix. This is the kind of insight the method can provide (my customers like poodles?!!...).

# Rebasing is NOT Keyword Matching

Two people  agree on best keyword 15%

| **Words:** | **cosine distances** | |
|---|---|---|
| | Keyword | LSA |
| Doctor—Doctor | 1.0 | 1.0 |
| Doctor—Physician | 0.0 | 0.8 |
| Doctor—Surgeon | 0.0 | 0.7 |

**Documents:**

| | | |
|---|---|---|
| Doctors operate on patients - Physicians do surgery | 0.0 | 0.8 |
| the radius of spheres - a circle's diameter | 0.00 | 0.55 |
| the radius of spheres - the music of spheres | 0.75 | 0.01 |

# Rebasing is NOT Co-Occurrence

Typically well over 99% of word-pairs whose similarity is induced never appear together in a paragraph.

Correlations with cosines over 10,000 random wd-wd pairs:

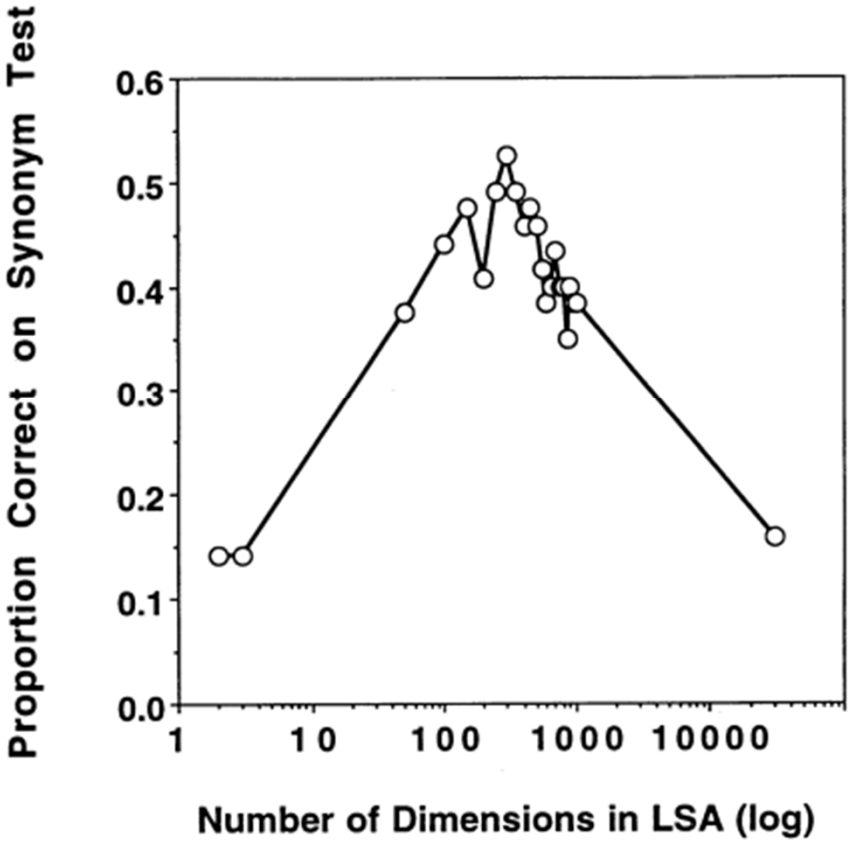
| | |
|---|---|
| Times two words co-occur in same paragraph (log both) | 0.35 |
| Times two words occur in separate paragraphs (log A only + log B only) | 0.30 |

Contingency measures:

| | |
|---|---|
| Mutual Information | 0.05 |
| Joint probability | 0.07 |

# Optimum Dimensionality Well Defined

The optimum value for n (the number of eigenvalues to keep in the approximation) is usually well defined, and is found empirically.




Notice how higher accuracy (more eigenvalues) results in *worse* performance.

This is a usual property of SVD-based methods in any field; see Gamma-ray detector discussion in next section.

# Example Application: Latent Semantic Analysis (LSA)

All methods use some form of suppression of low-impact words. This includes:

- Remove words on a high-frequency list (eg. the, a, and, etc.)
- Remove words that only appear once in the corpus
- Remove words that only appear on a single document
- etc (methods are heuristic)

LSA normalizes word counts by the **entropy** (patented).

# Example Application: Latent Semantic Analysis (LSA)

LSA is routinely used to grade long-hand essays (eg. SAT):

- Create domain semantic space
- Compute vectors for essays
- Have experts grade a small percentage of essays
- To grade an essay, compare it to the ones scored by experts

Several studies show that expert-LSA correlation is similar to expert-expert correlation:

| **Study** | **Expert-Expert** | **Expert - LSA** |
|---|---|---|
| 188 essays on human heart | 0.83 | 0.80 |
| 1205 essays on 12 different topics | 0.7 | 0.7 |
| 695 opinion essays | 0.86 | 0.86 |
| 668 argument essays | 0.87 | 0.86 |

# Some General Text-Based Applications

- Find documents based on whole document as query
  (eg. “find articles similar to this article”)
- Discover relationships between texts
  (find opinion-makers by tracing origin of an idea)
- Connect all similar paragraphs in a tech manual
  (not keyword search)
- Connect all similar paragraphs in an entire library
  (not keyword search)
- Place text into appropriate categories or taxonomies
  (eg. political articles from Democrats, or Republicans)

# SVD Can Also be Used In Non-Text Applications

- Apply whenever an object can be described by a vector.
- Dimensionality reduction can be used for noise mitigation or for bringing out relationships that are lost in the details

We will briefly discuss an example,
and then list a number of other non-text applications

# Many Inverse Problems Can Be Reduced to Searching

- Medical imaging (PET scans), face recognition
- Oil exploration (geological tomography)
- Gamma-ray detectors
- Synthetic aperture radar
- Detection of trace chemicals, Air pollutants
  - (Python Demo)

# Some Other Non-Text Applications

**Currently used applications:**

- Search for similar photos based on color distribution (patented)
- Search for similar music, based on sampling of music segments (patented)
- Predict lifespan from credit card activity (Deloitte; in the news)

**Other possibilities:**

- Predict health problems from credit card activity (similar to above)
- Predict holiday retail activity from social media texts
- Create more general indices of economic activity (eg. include social media texts)
- Xbox, Kinect dataset to analyze skeletal movement

# References

- http://blogs.msdn.com/hpctrekker/ my blog.
- http://hadooponazure.com
- **Books:**
  - Mining the Social Web
  - Data Source Handbook Pete Warden
  - Collective Intelligence *
  - Hadoop The Definitive Guide *
  - Big Data Now: Current Perspectives from O'Reilly Radar * (non-technical)