http://poloclub.gatech.edu/cse6242 CSE6242: Data & Visual Analytics

Text Analytics (Text Mining)

Duen Horng (Polo) Chau

Associate Professor, College of Computing Associate Director, MS Analytics Georgia Tech

Mahdi Roozbahani

Lecturer, Computational Science & Engineering, Georgia Tech Founder of Filio, a visual asset management platform

Partly based on materials by Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

Text is everywhere

We use documents as primary information artifact in our lives

Our access to documents has grown tremendously thanks to the Internet

- WWW: webpages, Twitter, Facebook, Wikipedia, Blogs, ...
- *Digital libraries:* Google books, ACM, IEEE, ...
- Lyrics, closed caption... (youtube)
- Police case reports
- Legislation (law)
- Reviews (products, rotten tomatoes)
- Medical reports (EHR electronic health records)
- Job descriptions

Big (Research) Questions

... in understanding and gathering information from text and document collections

- establish authorship, authenticity; plagiarism detection
- classification of genres for narratives (e.g., books, articles)
- tone classification; sentiment analysis (online reviews, twitter, social media)
- code: syntax analysis (e.g., find common bugs from students' answers)

Popular Natural Language Processing (NLP) libraries

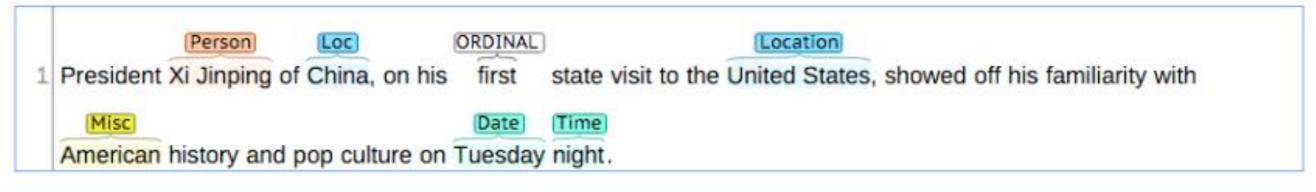
- Stanford NLP
- OpenNLP

tokenization, sentence segmentation, part-ofspeech tagging, named entity extraction, chunking, parsing

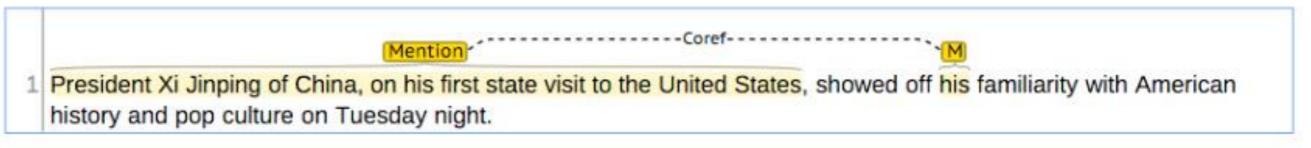
NLTK (python)

Named Entity Recognition:

Image source: https://stanfordnlp.github.io/CoreNLP/



Coreference:



Basic Dependencies:

Outline

- Preprocessing (e.g., stemming, remove stop words)
- Document representation (most common: bag-ofwords model)
- Word importance (e.g., word count, TF-IDF)
- Latent Semantic Indexing (find "concepts" among documents and words), which helps with retrieval

To learn more: CS 4650/7650 Natural Language Processing

Stemming

Reduce words to their stems (or base forms)

Words: compute, computing, computer, ...

Stem: comput

Several classes of algorithms to do this:

• Stripping suffixes, lookup-based, etc.

http://en.wikipedia.org/wiki/Stemming Stop words: http://en.wikipedia.org/wiki/Stop_words

Bag-of-words model

Represent each document as a bag of words, ignoring words' ordering. Why? For **simplicity**.

Unstructured text becomes a vector of numbers e.g., docs: "I like visualization", "I like data".

- **1**:"I"
- 2: "like"
- 3 : "data"
- 4: "visualization"
- "I like visualization" \rightarrow [1, 1, 0, 1]
- "I like data" → [1, 1, 1, 0]

TF-IDF

A word's importance score in a document, among N documents

When to use it? Everywhere you use "word count", you can likely use TF-IDF.

- TF: term frequency
- = #appearance a document
 (high, if terms appear many times in this document)
 - **IDF**: inverse document frequency
- = log(N / #document containing that term)
 (penalize "common" words appearing in almost any documents)

Final score = TF * IDF

(higher score → more "characteristic")

Vector Space Model Why?

Each document \rightarrow vector Each query \rightarrow vector

Search for documents \rightarrow find "similar" vectors Cluster documents \rightarrow cluster "similar" vectors

Latent Semantic Indexing (LSI)

Main idea

- map each document into some 'concepts'
- map each term into some 'concepts'

```
'Concept' : ~ a set of terms, with weights.
```

```
For example, DBMS_concept:

"data" (0.8),

"system" (0.5),

"retrieval" (0.6)
```

Latent Semantic Indexing (LSI) ~ pictorially (before) ~

document-term matrix

	data	system	retireval	lung	ear
doc1	1	1	1		
doc2	1	1	1		
doc3				1	1
doc4				1	1

Latent Semantic Indexing (LSI) ~ pictorially (after) ~

term-concept matrix

	database concept	medical concept
data	1	
system	1	
retrieval	1	
lung		1
ear		1

document-concept matrix

	database concept	medical concept
doc1	1	
doc2	1	
doc3		1
doc4		1

Latent Semantic Indexing (LSI)

Q: How to search, e.g., for "system"? A: find the corresponding concept(s); and the corresponding documents

	database concept	medical concept
data	1	
system	1 个	
retrieval	1	
lung		1
ear		1

	-	
	database	medical
	concept	concept
doc1	1 🔶	
doc2	1 🔶	
doc3		1
doc4		1

Latent Semantic Indexing (LSI)

Works like an automatically constructed thesaurus

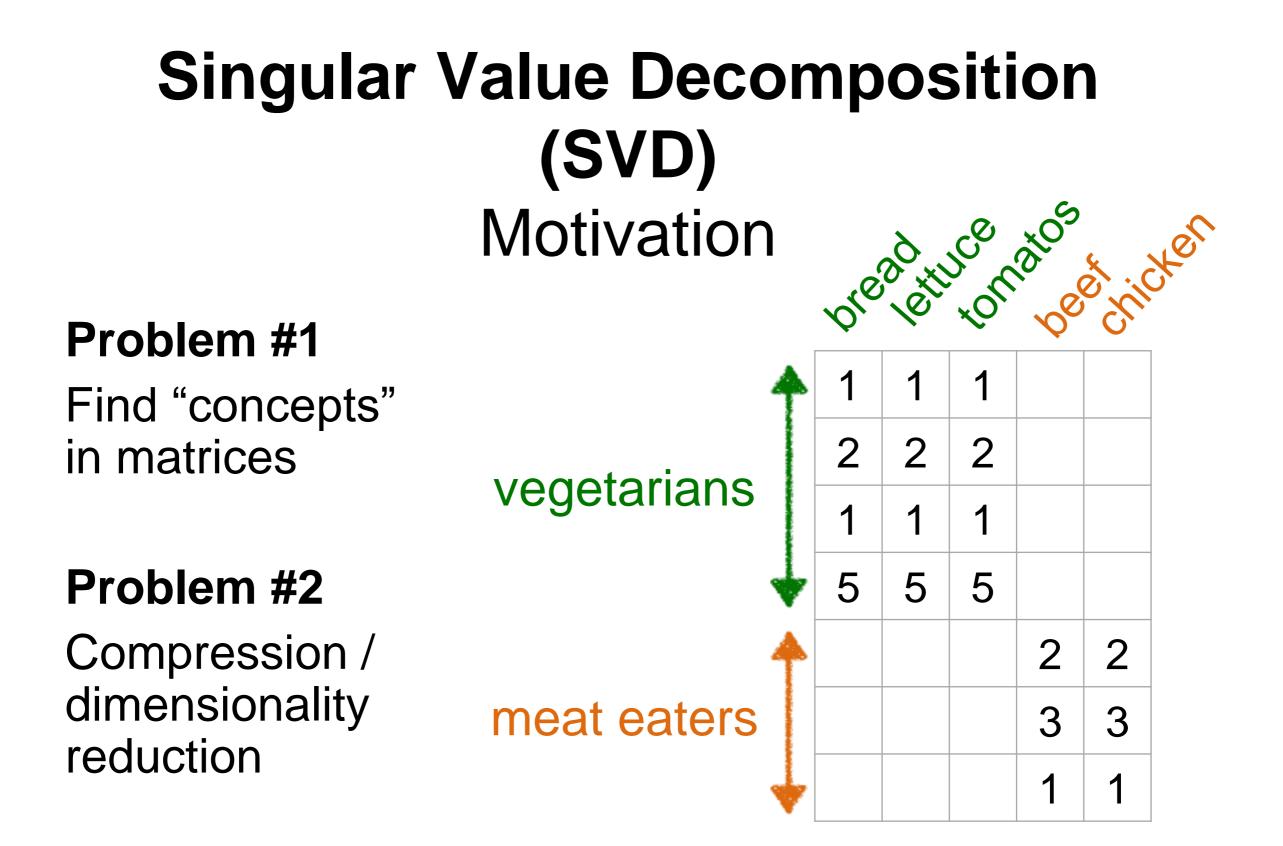
We may retrieve documents that **DON'T** have the term "system", but they contain almost everything else ("data", "retrieval")

LSI - Discussion

Great idea,

- to derive 'concepts' from documents
- to build a 'thesaurus' automatically
- to reduce dimensionality (down to few "concepts")

How does LSI work? Uses Singular Value Decomposition (SVD)

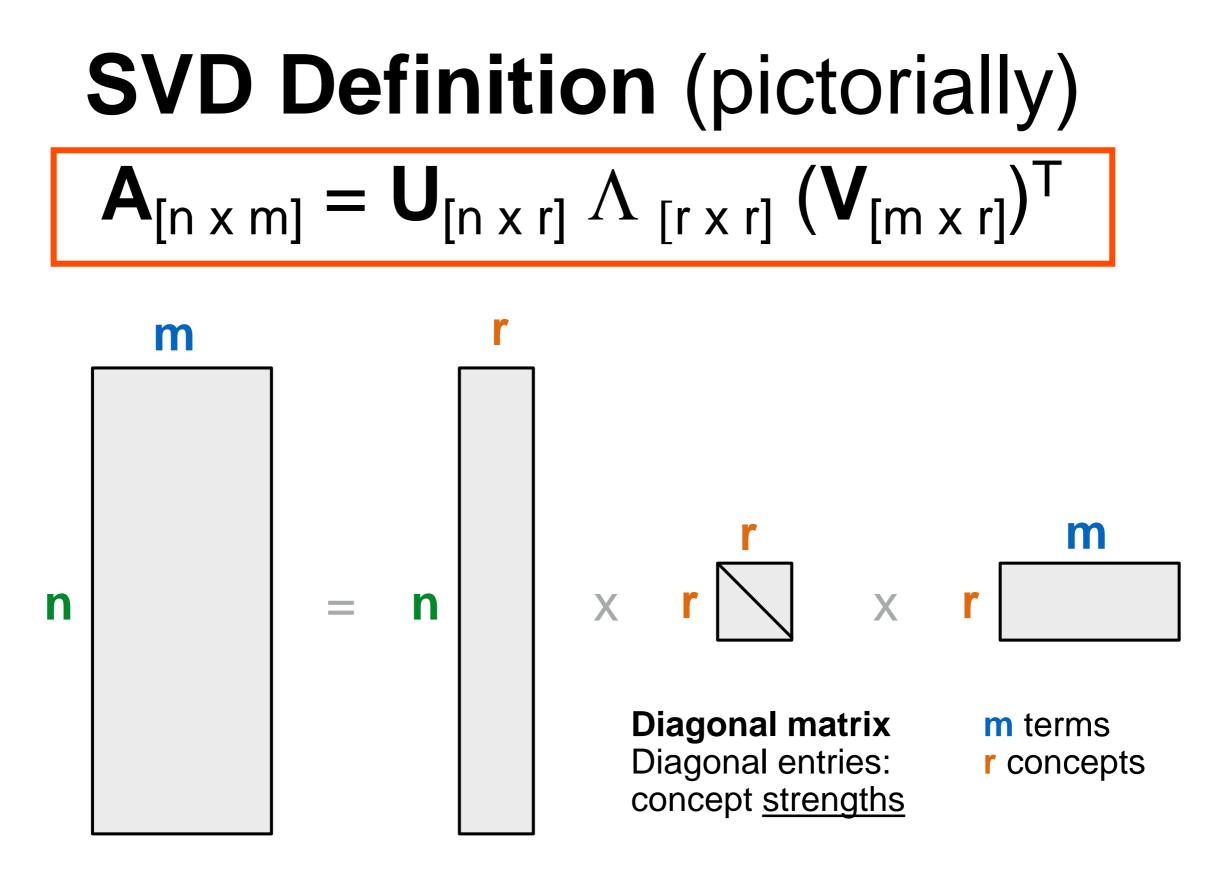


SVD is a powerful, generalizable technique.

Songs / Movies / Products

Customers

1	1	1		
2	2	2		
1	1	1		
5	5	5		
			2	2
			3	3
			1	1



n documents m terms

n documents r concepts

SVD Definition (in words)

$$\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \Lambda_{[r \times r]} (\mathbf{V}_{[m \times r]})^{\mathsf{T}}$$

A: n x m matrix

e.g., n documents, m terms

U: n x r matrix

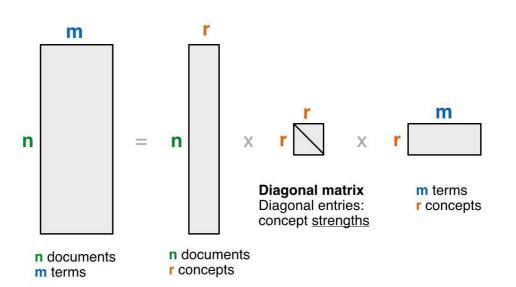
e.g., n documents, r concepts

Λ : **r x r diagonal matrix**

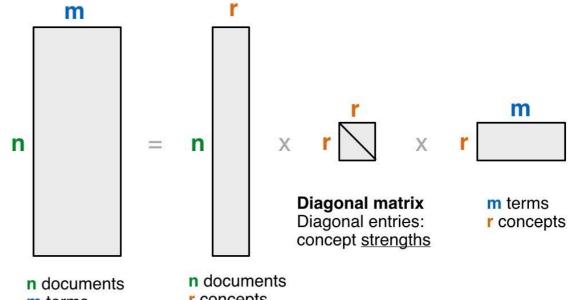
r : rank of the matrix; strength of each 'concept'

V: m x r matrix

e.g., m terms, r concepts



SVD - Properties

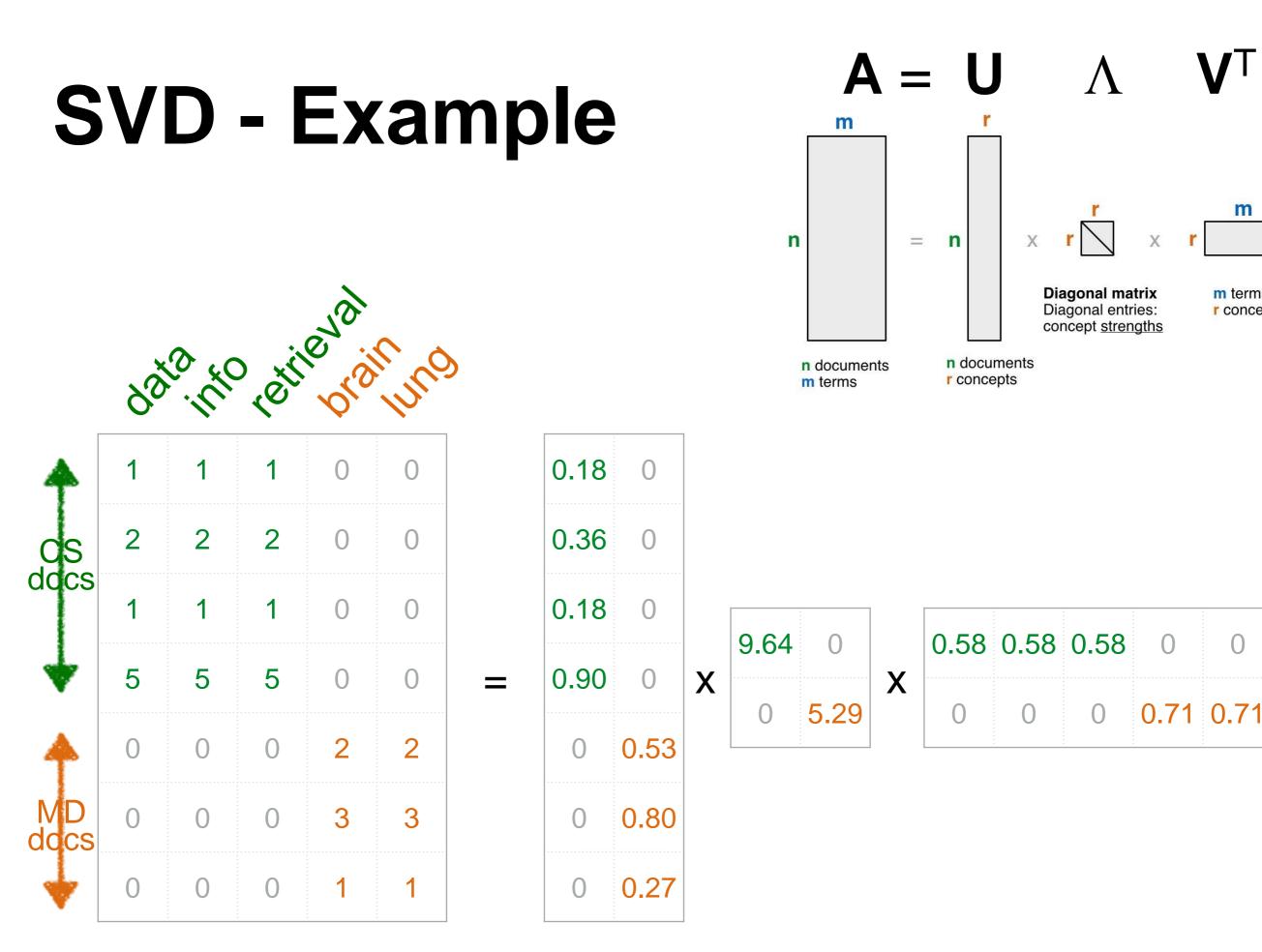


THEOREM [Press+92]:n documents
r conceptsalways possible to decompose matrix A into $A = U \land V^T$

- **U**, Λ , **V**: **unique**, most of the time
- U, V: column orthonormal

i.e., columns are unit vectors, and orthogonal to each other $U^{T} U = I$ $V^{T} V = I$ (I: identity matrix)

 Λ : diagonal matrix with non-negative diagonal entires, sorted in decreasing order



VT

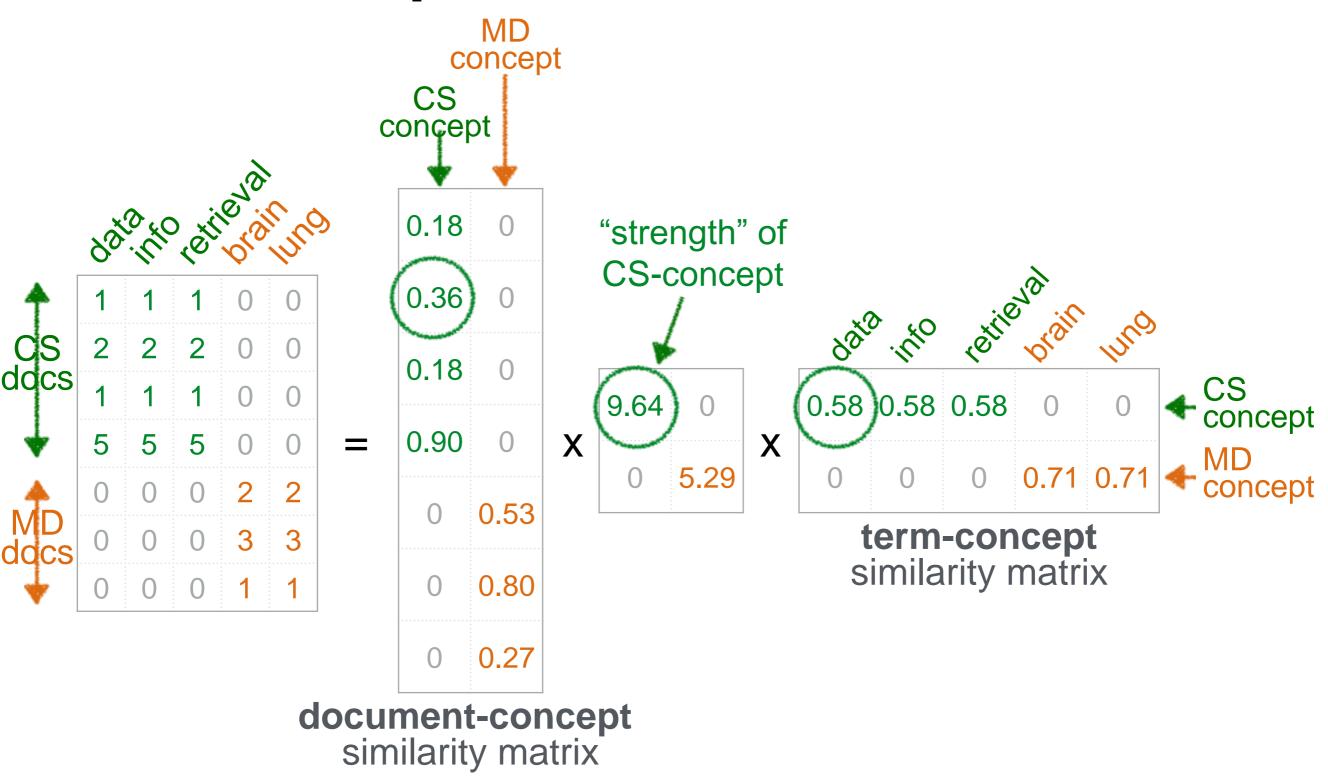
m

m terms

 $\left(\right)$

r concepts

SVD - Example



'documents', 'terms' and 'concepts':

- U: document-concept similarity matrix
- V: term-concept similarity matrix
- Λ: diagonal elements: concept "strengths"

'documents', 'terms' and 'concepts':
Q: if A is the document-to-term matrix, what is the similarity matrix A^T A ?
A:

Q: **A A**^T ? A:

'documents', 'terms' and 'concepts':

- Q: if A is the document-to-term matrix, what is the similarity matrix A^T A?
- A: term-to-term ([m x m]) similarity matrix

Q: **A A**^T ?

A: document-to-document ([n x n]) similarity matrix

SVD properties

V are the eigenvectors of the *covariance matrix* A^TA (term-to-term [m x m] similarity matrix)

$$\mathbf{A}^{\mathsf{T}}\mathbf{A} = \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}\right)^{\mathsf{I}}\left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}\right) = \mathbf{V}\boldsymbol{\Sigma}^{2}\mathbf{V}^{\mathsf{T}}$$

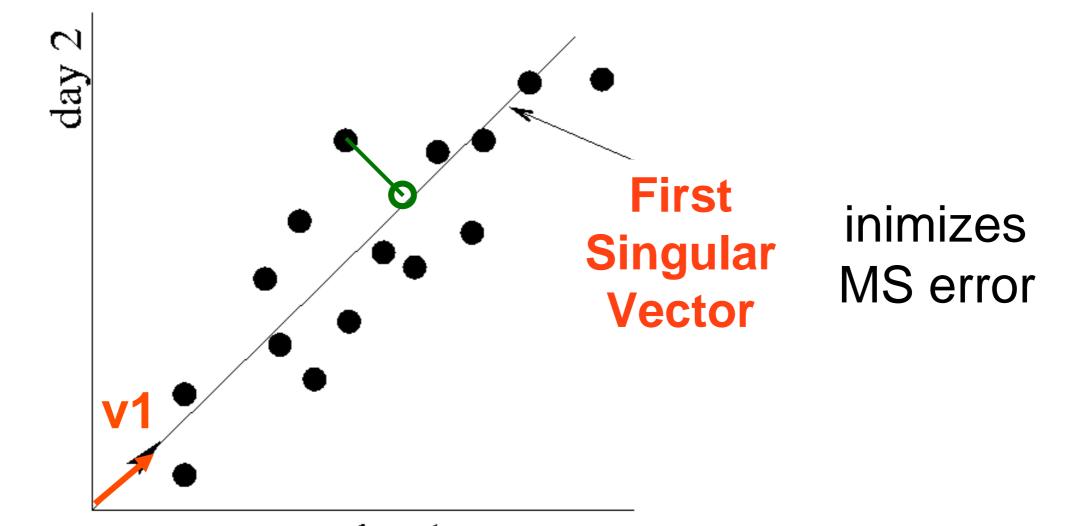
U are the eigenvectors of the Gram (inner-product) matrix **AA**^T (doc-to-doc [n x n] similarity matrix)

$$\mathbf{A}\mathbf{A}^{\mathsf{T}} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}})(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{U}\mathbf{\Sigma}^{2}\mathbf{U}^{\mathsf{T}}$$

SVD is closely related to PCA, and can be numerically more stable. For more info, see:

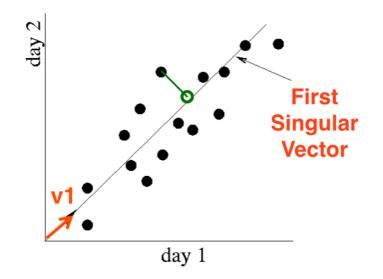
http://math.stackexchange.com/questions/3869/what-is-the-intuitive-relationship-between-svd-and-pca Ian T. Jolliffe, Principal Component Analysis (2nd ed), Springer, 2002. Gilbert Strang, Linear Algebra and Its Applications (4th ed), Brooks Cole, 2005.

SVD - Interpretation #2 Find the best axis to project on. ("best" = minimize sum of squares of projection errors)

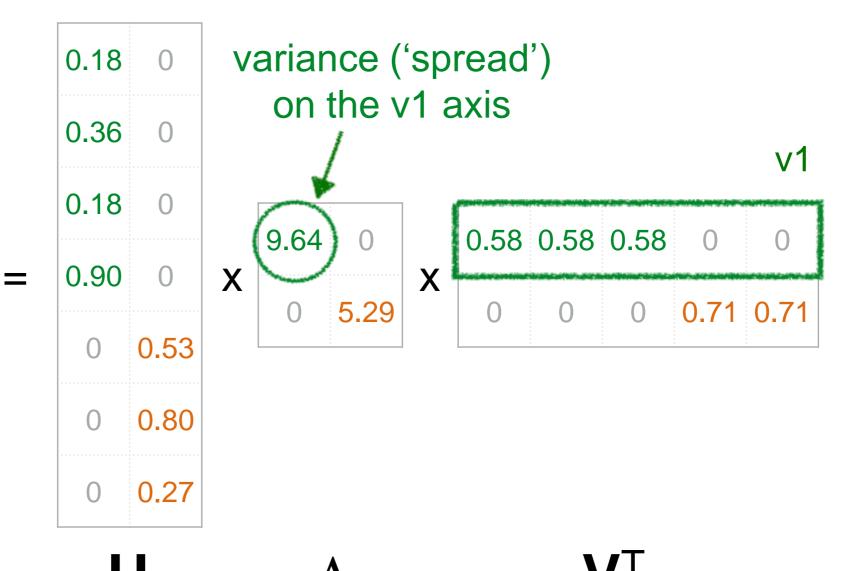


day 1 Beautiful visualization explaining PCA: http://setosa.io/ev/principal-component-analysis/

 $\mathbf{U} \ \boldsymbol{\Lambda} \$ gives the coordinates of the points in the projection axis

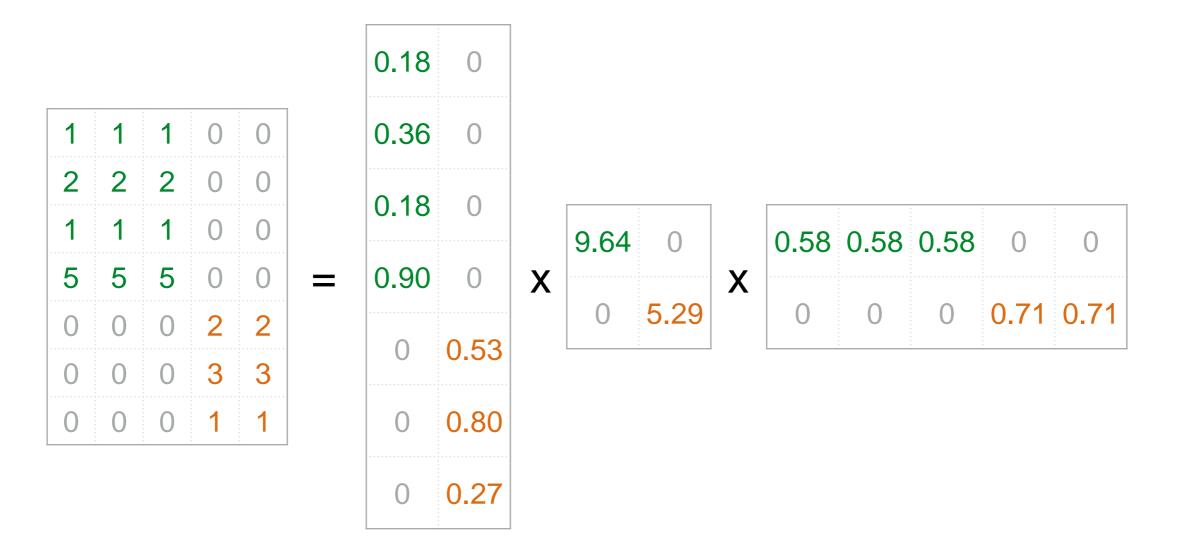


1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1



More details

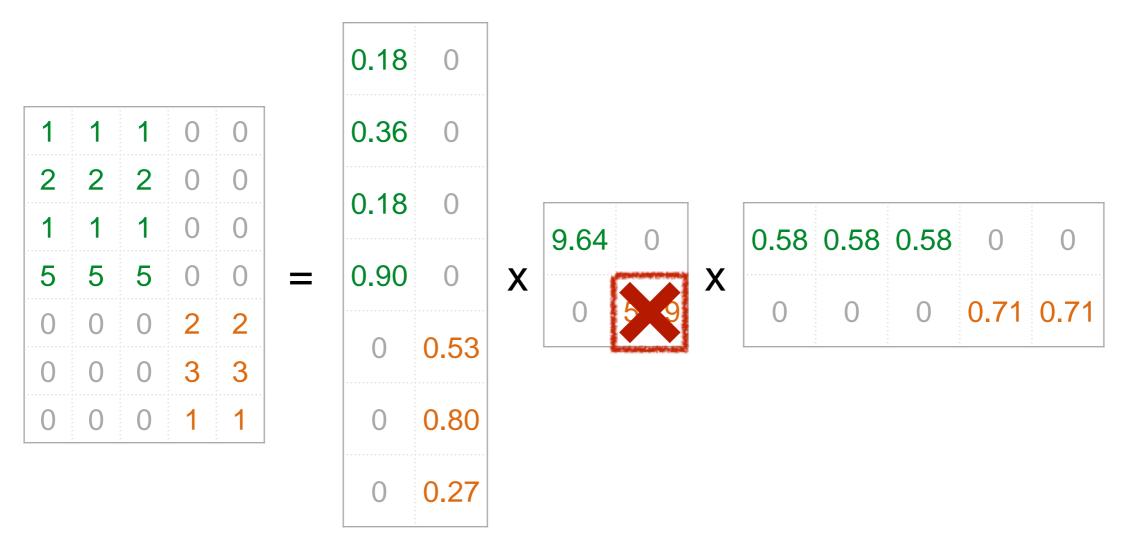
Q: how exactly is dim. reduction done?



More details

Q: how exactly is dim. reduction done?

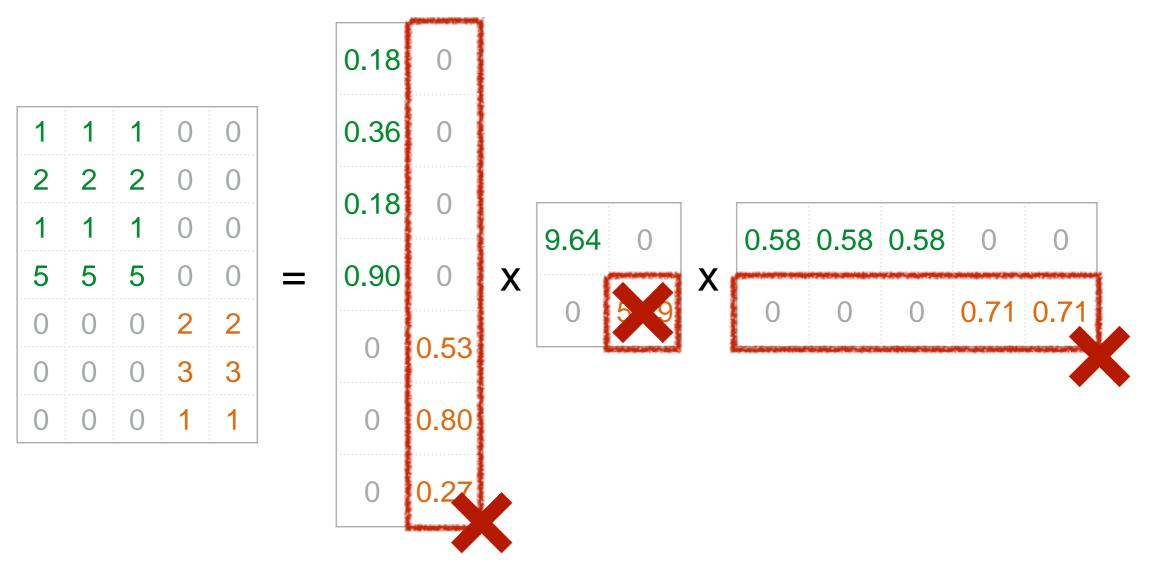
A: set the smallest singular values to zero:



More details

Q: how exactly is dim. reduction done?

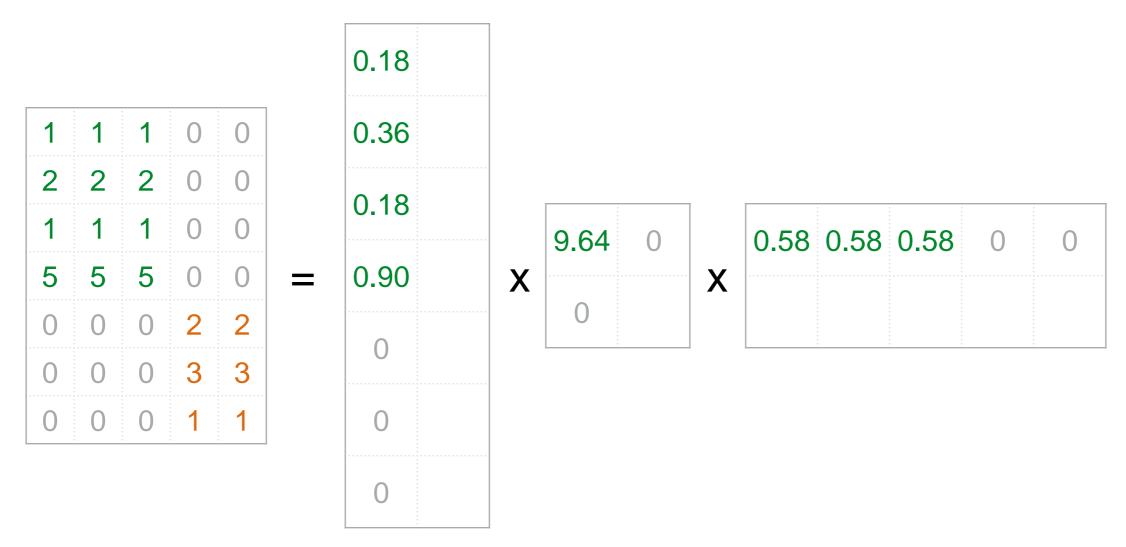
A: set the smallest singular values to zero:



More details

Q: how exactly is dim. reduction done?

A: set the smallest singular values to zero:



More details

Q: how exactly is dim. reduction done?

A: set the smallest singular values to zero:

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

0

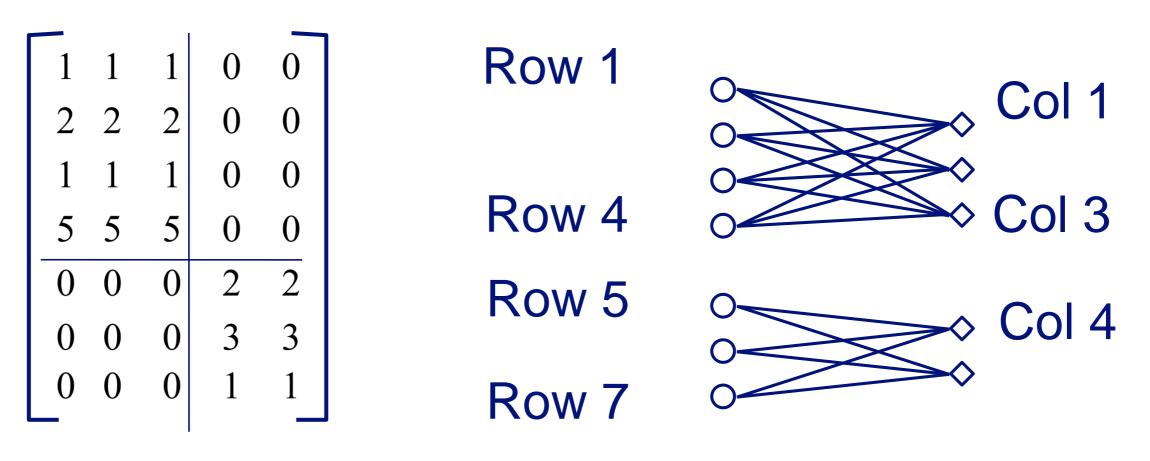
finds non-zero 'blobs' in a data matrix

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

finds non-zero 'blobs' in a data matrix

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

- finds non-zero 'blobs' in a data matrix =
- 'communities' (bi-partite cores, here)



SVD - Complexity

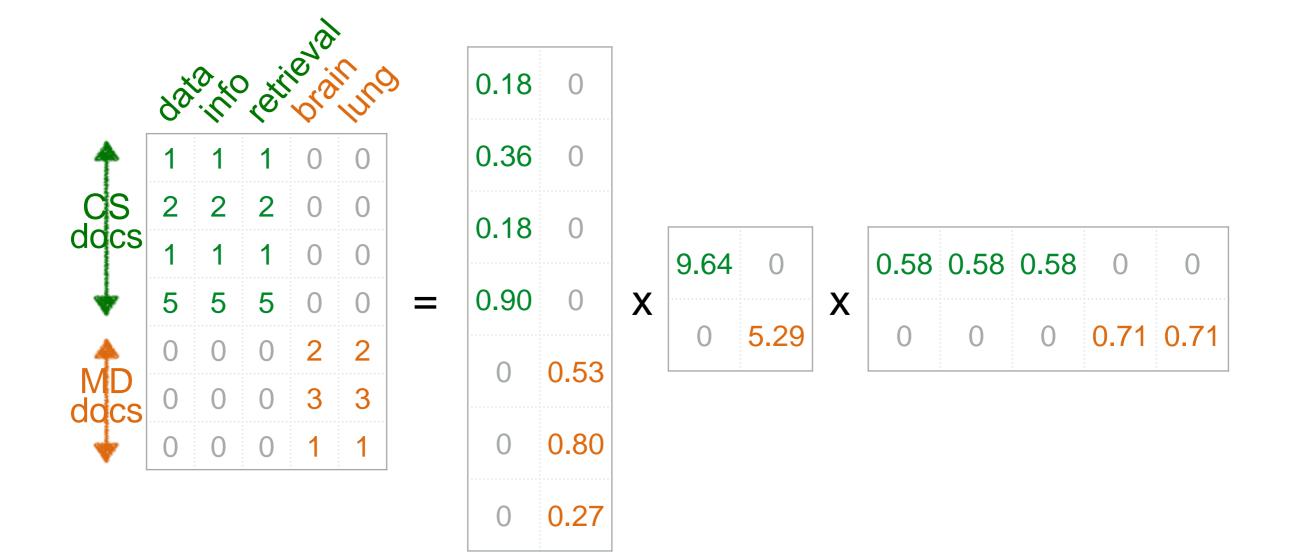
O(n*m*m) or O(n*n*m) (whichever is less)

Faster version, if just want singular values or if we want first *k* singular vectors or if the matrix is sparse [Berry]

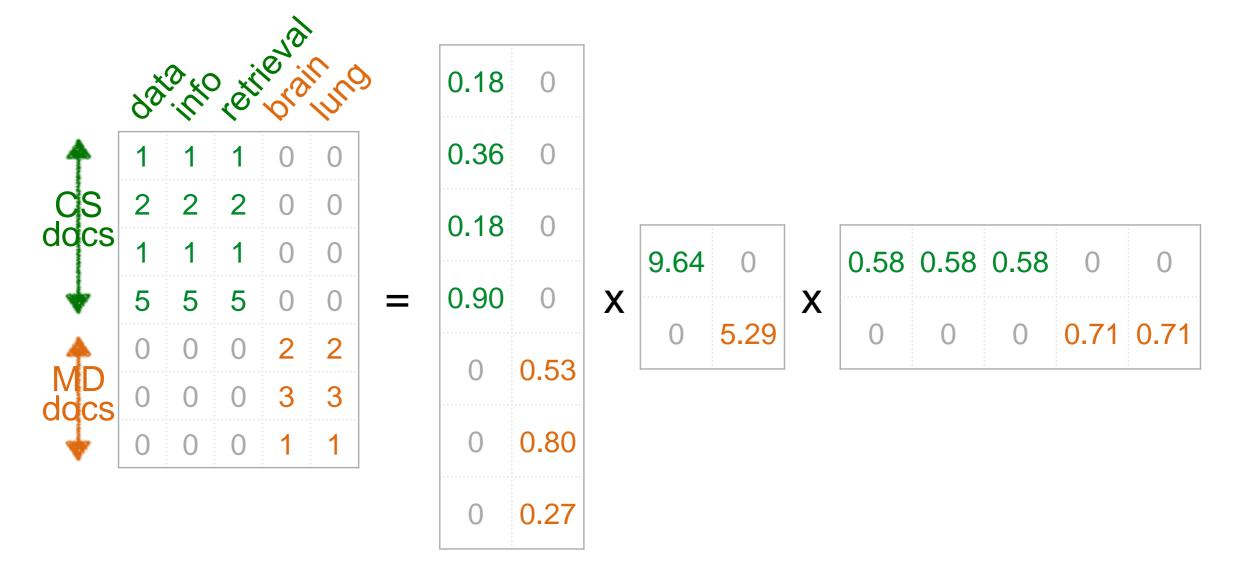
No need to write your own!

Available in most linear algebra packages (LINPACK, matlab, Splus/R, mathematica ...)

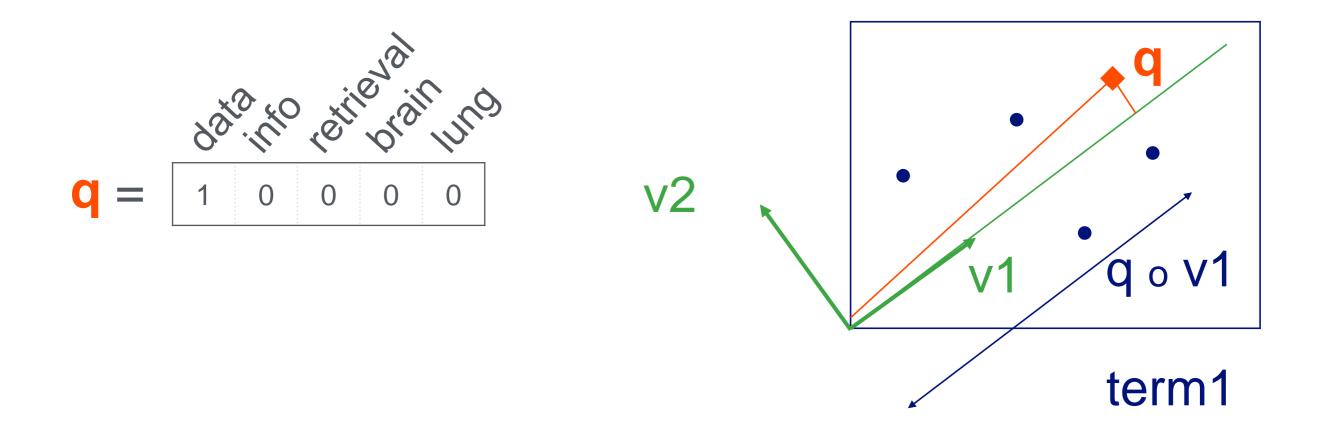
For example, how to find documents with 'data'?



For example, how to find documents with 'data'? A: map query vectors into 'concept space' – how?

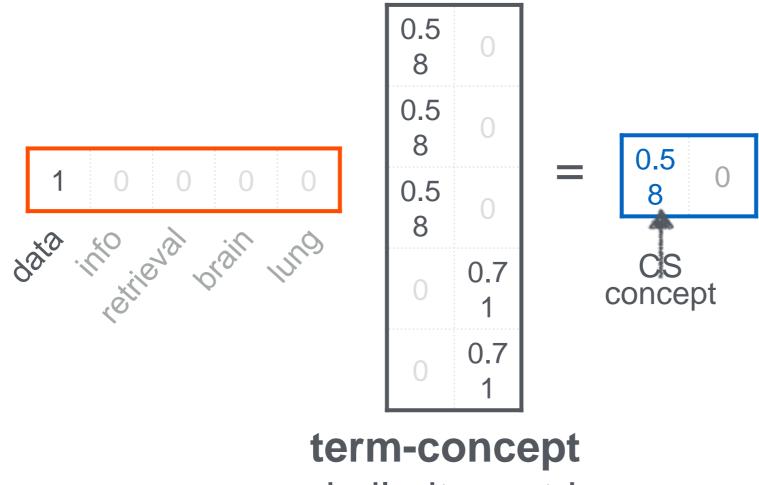


For example, how to find documents with 'data'? A: map query vectors into 'concept space', using inner product (cosine similarity) with each 'concept' vector v_i



Case Study How to do queries with LSI? Compactly, we have:

 $\mathbf{q} \mathbf{V} = \mathbf{q}_{\text{concept}}$



similarity matrix

Case Study How would the document ('information', 'retrieval') be handled?

Case Study How would the document ('information', 'retrieval') be handled?

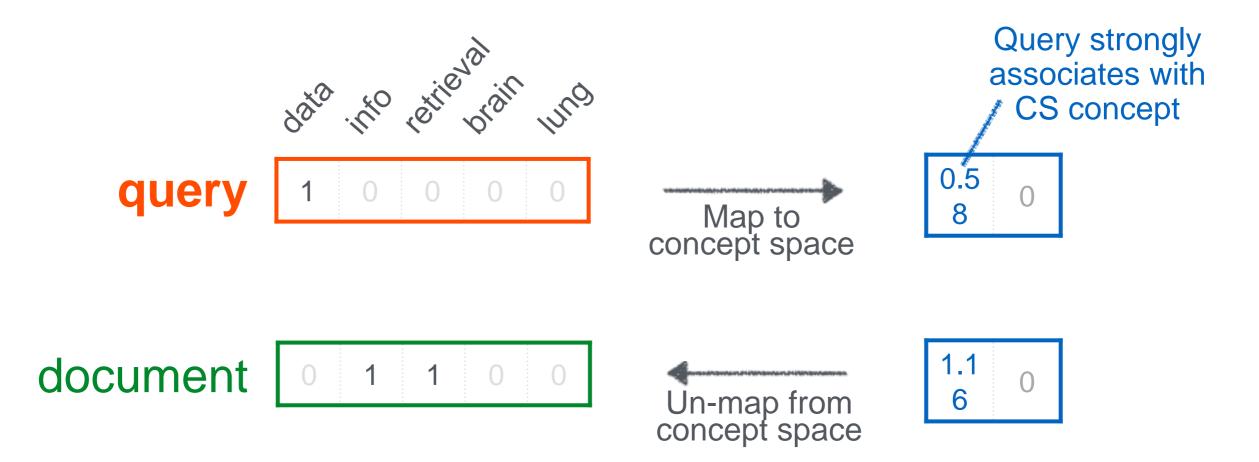
 $d V = d_{concept}$

SAME!

0.5 8 0.5 8 1.1 1 0.5 6 8 data into reval brain uno 0.7 concept 1 0.7 term-concept similarity matrix

Case Study Observation

Document ('information', 'retrieval') will be retrieved by **query** ('data'), even though it does not contain 'data'!!

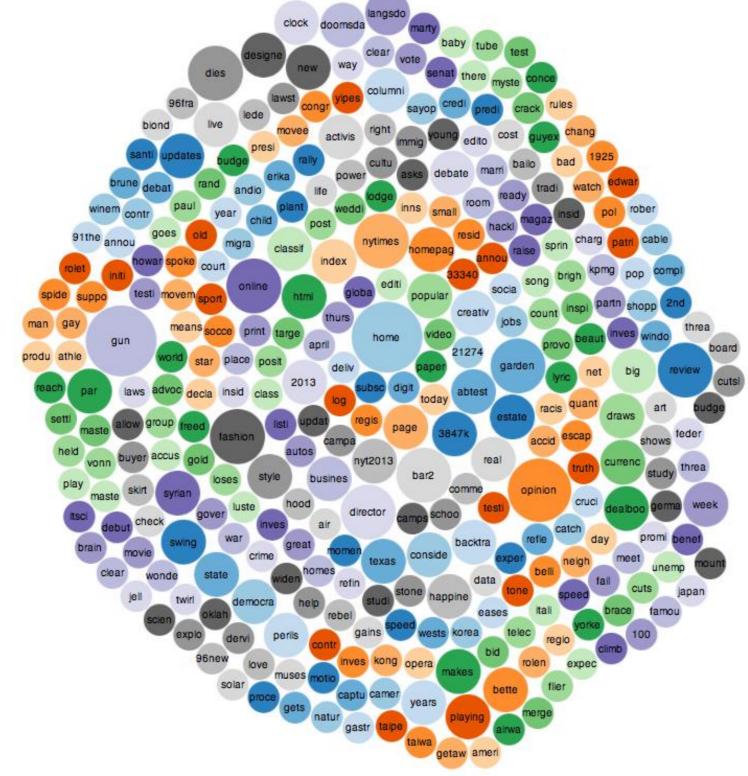


Switch Gear to Text Visualization

Word/Tag Cloud (still popular?)



Word Counts (words as bubbles)



http://www.infocaptor.com/bubble-my-page

Word Tree

word tree

We

📄 reverse tree 📄 one phrase per line

Shift-click to make that word the root.

We

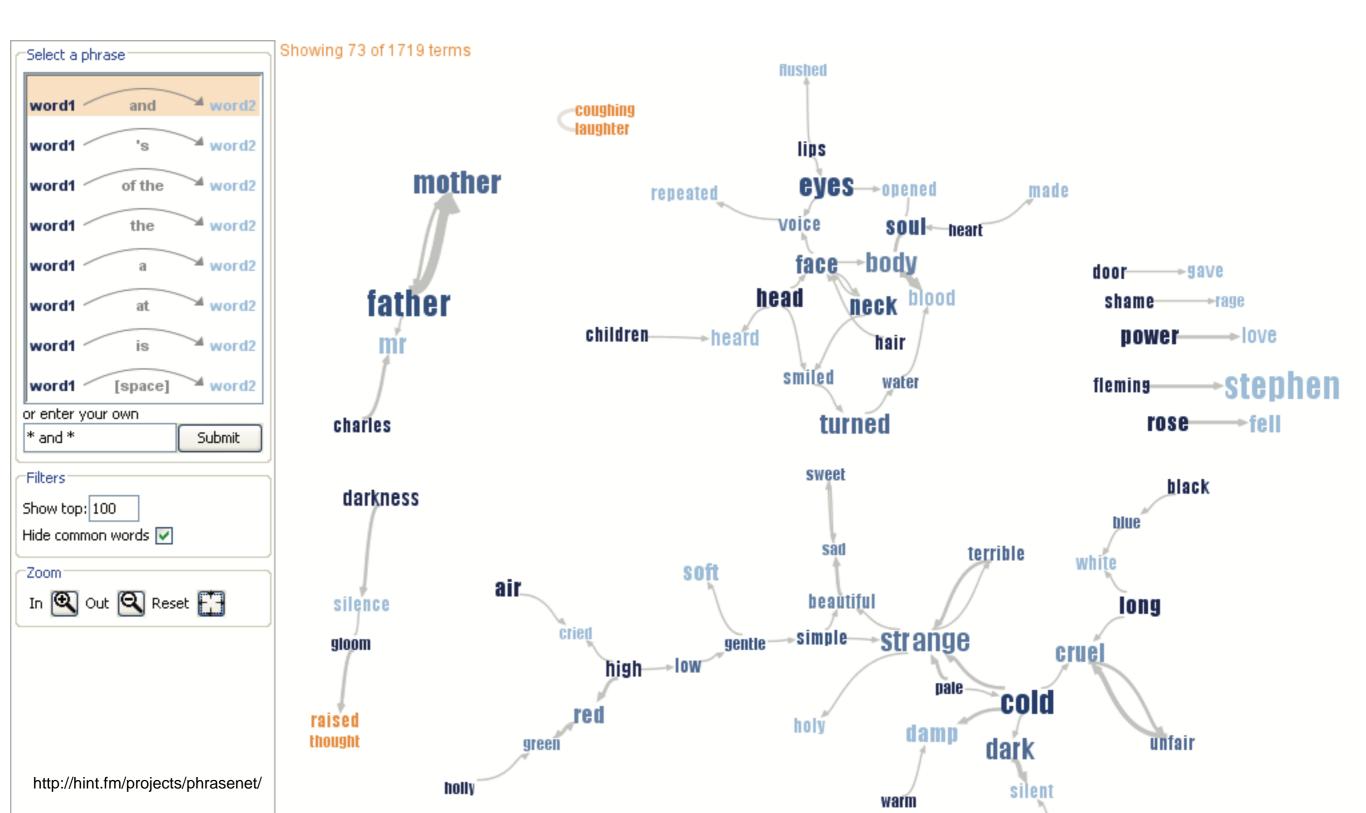
substitute spectacle for we must act knowing that our work will be imperfect. We must act, knowing that today's victories will be only partial, and th politics, or treat nameact knowing that today's victories will be only partial, and that it will be up to those who stand here in four years, and forty years knowing that our work will be imperfect. We must act, knowing that today's victories will be only partial, and that it will be up to those wh calling as reasoned debate. do these things together, as one nation, and one people. This generation of Americans has been tested by crises that steeled our resolve and proved our resilience. harness new ideas and technology to remake our government, revamp our tax code, reform our schools, and empower our citizens with the skills they need to wor We must act, we must act must make the hard choices to reduce the cost of health care and the size of our deficit. But we reject the belief that America must choose between caring for the generat lead it. We cannot cede to other nations the technology that will power new jobs and new industries - we must claim its promise. knowing that our work will claim its promise. That's how we will maintain our economic vitality and our national treasure - our forests and waterways; our croplands and snowcapped per carry those lessons into this time as well. We will defend our people and uphold our values through strength of arms and rule of law. be imperfect. We must act, be a source of hope to the poor, the sick, the marginalized, the victims of prejudice - not out of mere charity, but because peace in our time requires the constant as faithfully execute that pledge during the duration of our service. But the words I spoke today are not so different from the oath that is taken each time a soldier sig knowing that today's seize it - so long as we seize it together. For we, the people, understand that our country cannot succeed when a shrinking few do very well and a growing many barely m respond to the threat of climate change, knowing that the failure to do so would betray our children and future generations. victories will be only maintain our economic vitality and our national treasure - our forests and waterways; our croplands and snowcapped peaks. That is how we will preserve our planet, or will preserve our planet, commanded to our care by God. That's what will lend meaning to the creed our fathers once declared. partial, and that it will be defend our people and uphold our values through strength of arms and rule of law. We will show the courage to try and resolve our differences with other nations peaced show the courage to try and resolve our differences with other nations peacefully - not because we are naive about the dangers we face, but because engagement can mor up to those who stand here renew those institutions that extend our capacity to manage crisis abroad, for no one has a greater stake in a peaceful world than its most powerful nation. support democracy from Asia to Africa; from the Americas to the Middle East, because our interests and our conscience compel us to act on behalf of those who long for in four years, and forty made for this moment, and we will seize it - so long as we seize it together. For we, the people, understand that our country cannot succeed when a shrinking few do very we true to our creed when a little girl born into the bleakest poverty knows that she has the same chance to succeed as anybody else, because she is an American, she is free, and years, and four hundred are also heirs to those who won the peace and not just the war, who turned sworn enemies into the surest of friends, and we must carry those lessons into this time as well. naïve about the dangers we face, but because engagement can more durably lift suspicion and fear. America will remain the anchor of strong alliances in every corner of th years hence to advance the truly created equal, then surely the love we commit to one another must be equal as well. Our journey is not complete until no citizen is forced to wait for hours to exercise t every citizen deserves a basic measure of security and dignity. We must timeless spirit once our obligations as Americans are not just to ourselves, but to all poster , the people, enduring security and lasting peace do not require perpetual war. Our conferred to us in a spare understand that our country cannot succeed when a shrinking few do very well and a growing many barely make it. declare today that the most evident of truths - that all of us are created equal - is the star that guides us still; just as it guided our forebeau Philadelphia hall. . Through blood drawn by lash and blood drawn by sword, we learned that no union founded on the principles of liberty and equality could survive half-slave and half never relinquished our skepticism of central authority, nor have we succumbed to the fiction that all society's ills can be cured through government alone. have always understood that when times change, so must we; that fidelity to our founding principles requires new responses to new challenges; that preserving our individ 📕 My fellow Americans, the lost, know too well the price that is paid for liberty. The knowledge of their sacrifice will keep us forever vigilant against those who would do us harm. cede to other nations the technology that will power new jobs and new industries - we must claim its promise. That's how we will maintain our economic vita walk alone; to hear a King proclaim that our individual freedom is inextricably bound to the freedom of every soul on Earth. oath I have sworn before you cannot afford delay. We cannot mistake absolutism for principle, or substitute spectacle for politics, or treat name-calling as reasoned debate. mistake absolutism for principle, or substitute spectacle for politics, or treat name-calling as reasoned debate. We must act, we must act knowing that our we today, like the one recited define liberty in exactly the same way, or follow the same precise path to happiness. Progress does not compel us to settle centuries-long debates about the role of government for all make to the flag that waves above and that fills our hearts with pride. They are the words of citizens, and they represent our greatest hope. by others who serve in this gather to inaugurate a president, we bear witness to the enduring strength of our Constitution. We affirm the promise of our democracy. bear witness to the enduring strength of our Constitution. We affirm the promise of our democracy. We recall that what binds this nation together is not the colors of our skin or the tenets of our fa Capitol, was an oath to God affirm the promise of our democracy. We recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names. and country, not party or recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names. hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable rights, that among these are Life, Liberty, and the pursuit of Happ faction - and we must continue a never-ending journey, to bridge the meaning of those words with the realities of our time. For history tells us that while these truths may be self-evident, they have never been self-executir learned that no union founded on the principles of liberty and equality could survive half-slave and half-free. We made ourselves anew, and vowed to move forward together. made ourselves anew, and vowed to move forward together. Together, we determined that a modern economy requires railroads and highways to speed travel and commerce; schools and colleges faithfully execute that determined that a modern economy requires railroads and highways to speed travel and commerce; schools and colleges to train our workers. pledge during the duration discovered that a free market only thrives when there are rules to ensure competition and fair play. Together, we resolved that a great nation must care for the vulnerable, and protect its people from resolved that a great nation must care for the vulnerable, and protect its people from life's worst hazards and misfortune. of our service. But the succumbed to the fiction that all society's ills can be cured through government alone. Our celebration of initiative and enterprise; our insistence on hard work and personal responsibility, these ar that fidelity to our founding principles requires new responses to new challenges; that preserving our individual freedoms ultimately requires collective action. words I spoke today are not Il need to equip our children for the future, or build the roads and networks and research labs that will bring new jobs and businesses to our shores. possess all the qualities that this world without boundaries demands: youth and drive; diversity and openness; an endless capacity for risk and a gift for reinvention. so different from the oath seize it together. For we, the people, understand that our country cannot succeed when a shrinking few do very well and a growing many barely make it. believe that America's prosperity must rest upon the broad shoulders of a rising middle class. We know that America thrives when every person can find independence and pride in their work; whe that is taken each time a know that America thrives when every person can find independence and pride in their work; when the wages of honest labor liberate families from the brink of hardship. understand that outworn programs are inadequate to the needs of our time. We must harness new ideas and technology to remake our government, revamp our tax code, reform our schools, and er soldier signs up for duty, reject the belief that America must choose between caring for the generation that built this country and investing in the generation that will build its future. remember the lessons of our past, when twilight years were spent in poverty, and parents of a child with a disability had nowhere to turn. or an immigrant realizes her do not believe that in this country, freedom is reserved for the lucky, or happiness for the few. We recognize that no matter how responsibly we live our lives, any one of us, at any time, may face a jo recognize that no matter how responsibly we live our lives, any one of us, at any time, may face a job loss, or a sudden illness, or a home swept away in a terrible storm. dream. My oath is not so live our lives, any one of us, at any time, may face a job loss, or a sudden illness, or a home swept away in a terrible storm. make to each other - through Medicare, and Medicaid, and Social Security - these things do not sap our initiative; they strengthen us.

http://www.jasondavies.com/wordtree/

132

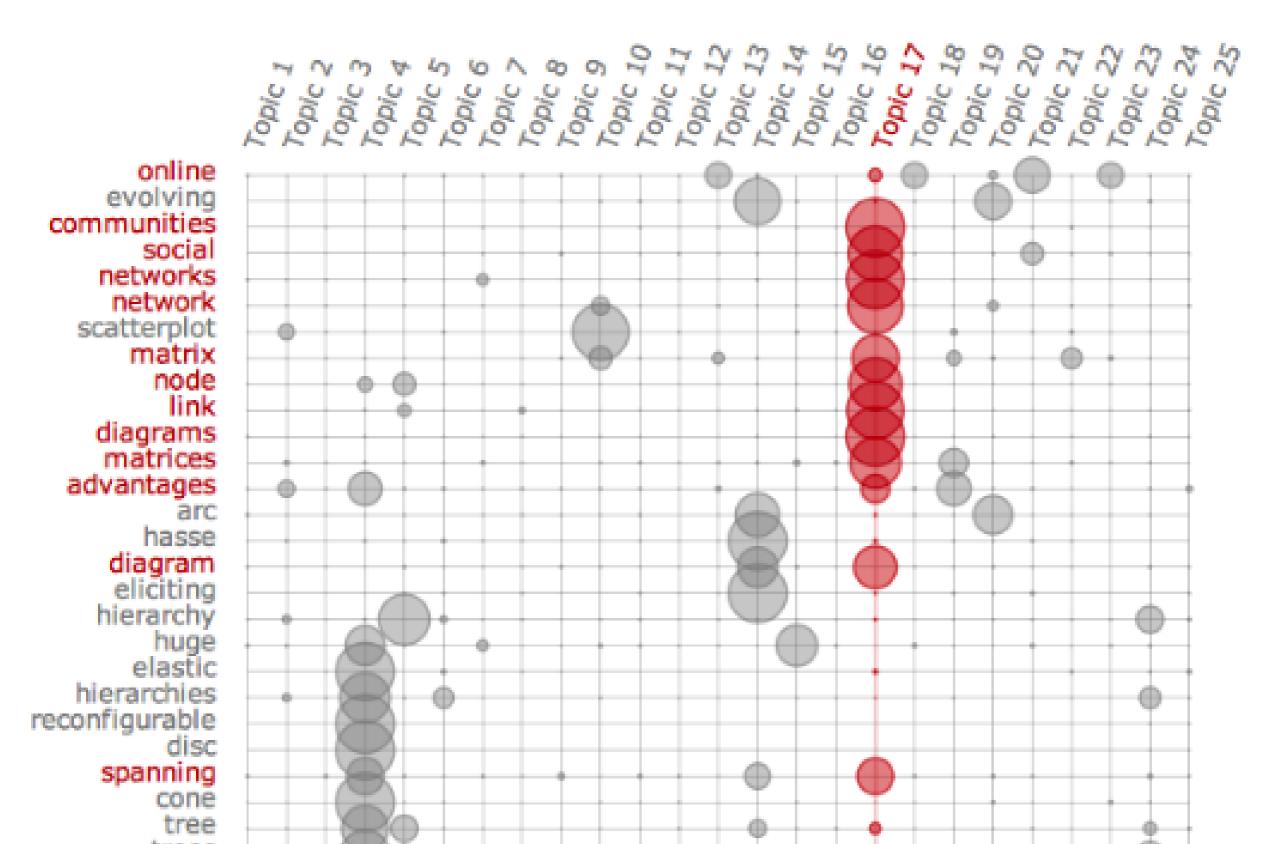
Phrase Net

Visualize pairs of words satisfying a pattern ("X and Y")



Termite: Topic Model Visualization

http://vis.stanford.edu/papers/termite



Termite: Topic Model Visualization.

