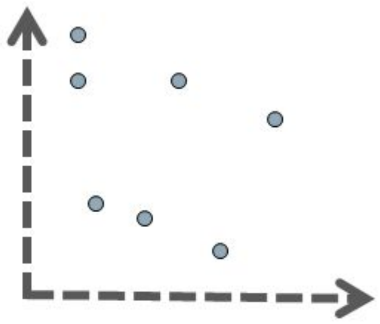
Natural Language Processing (CS5803)

Lecture 2
(Representing Text)

Recap: Drawback of TF-IDF Scheme?

- Remember?
- Solution:
  - Compact representations
  - Capturing relations between elements in consideration

Learning Representations of Words



- Represent each word with a low-dimensional vector
- Word similarity = vector similarity
- Key idea: Observe surrounding words of every word
- How do we get the vectors?
- We learn the vectors (or embeddings)
  - Representation learning
  - Word embedding

Use statistical cues

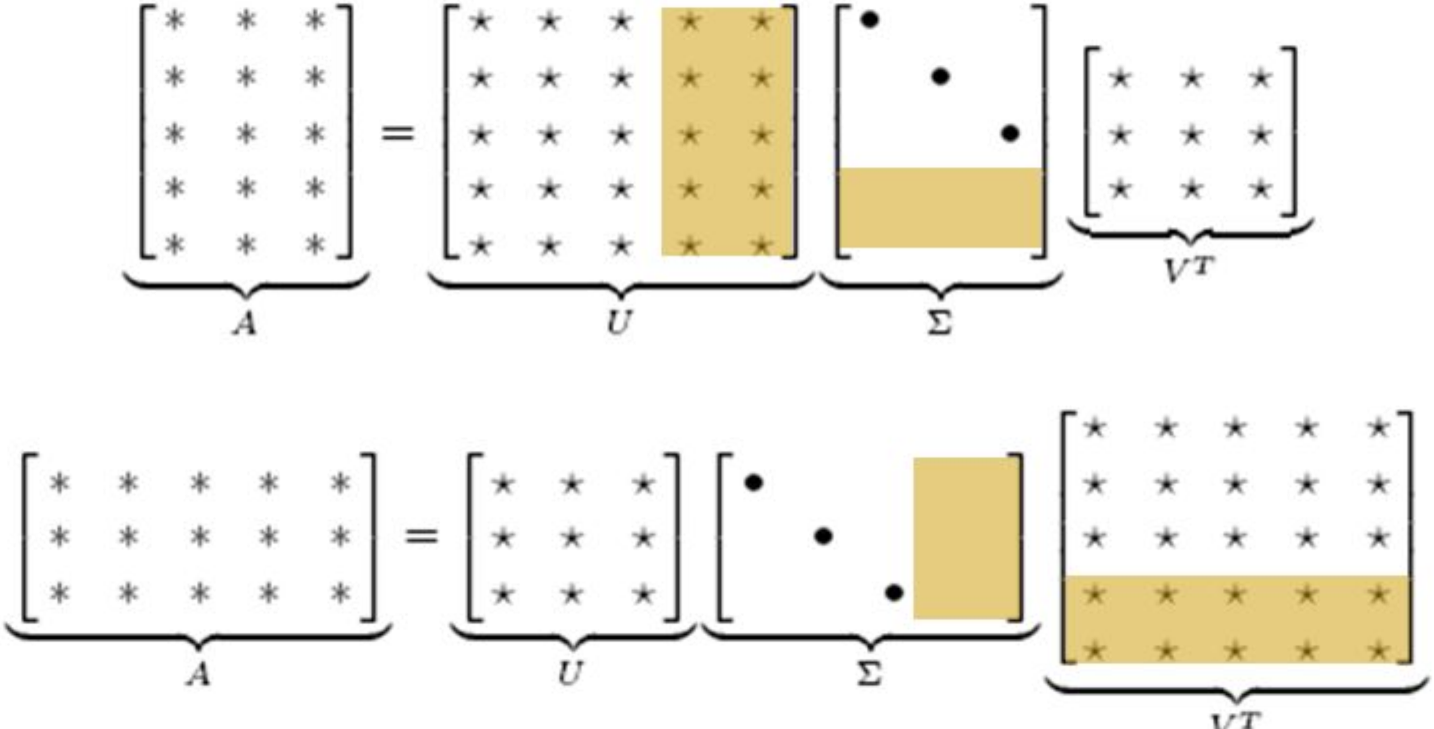
- Use Singular value Decomposition

$$A=U\Sigma V^T$$

- The columns of U are orthogonal eigenvectors of AA^T
- The columns of V are orthogonal eigenvectors of A^TA
- Eigenvalues $\lambda_1 \ldots \lambda_r$ of AA^T are the eigenvalues of A^TA

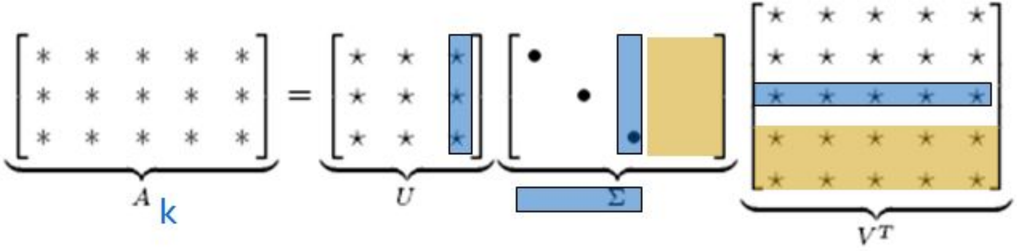
$$\sigma_i = \sqrt{\lambda_i}$$
$$\Sigma = \mathrm{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

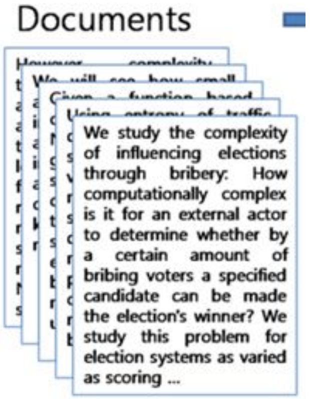
Illustration of Singular Value Decomposition



SVD for Textual Data: Summary

- A: Term-Document Incidence Matrix (Size: mxn)





- **Terms can be represented using the entries in U_k**
  - We can work with k<<m dimensions.
- **Documents can be represented using the entries in V_k**
  - We can work with k<<n dimensions.

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

SVD on BBC Dataset

Top terms per cluster:

- **Cluster 0:** mobile, phone, broadband, digital, people, technology, phones, tv, bt, said
- **Cluster 1:** show, tv, said, series, star, musical, bbc, us, film, comedy
- **Cluster 2:** economy, growth, economic, dollar, said, rate, rates, us, year, bank
- **Cluster 3:** lord, lords, said, blunkett, blair, home, government, secretary, law, house
- **Cluster 4:** games, game, software, said, microsoft, users, people, computer, search, virus
- **Cluster 5:** music, band, album, rock, song, best, chart, number, singer, said
- **Cluster 6:** said, government, would, eu, people, uk, party, minister, public, police
- **Cluster 7:** film, best, films, oscar, festival, awards, actor, award, director, actress
- **Cluster 8:** said, company, shares, firm, us, oil, market, sales, profits, bank
- **Cluster 9:** labour, election, blair, brown, party, howard, tax, chancellor, said, tory

SVD Example

Technical Memo Example

Titles:

c1: *Human* machine *interface* for Lab ABC *computer* applications

c2: A *survey* of *user* opinion of *computer system response time*

c3: The *EPS user interface* management *system*

c4: *System* and *human system* engineering testing of *EPS*

c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*

m2: The intersection *graph* of paths in *trees*

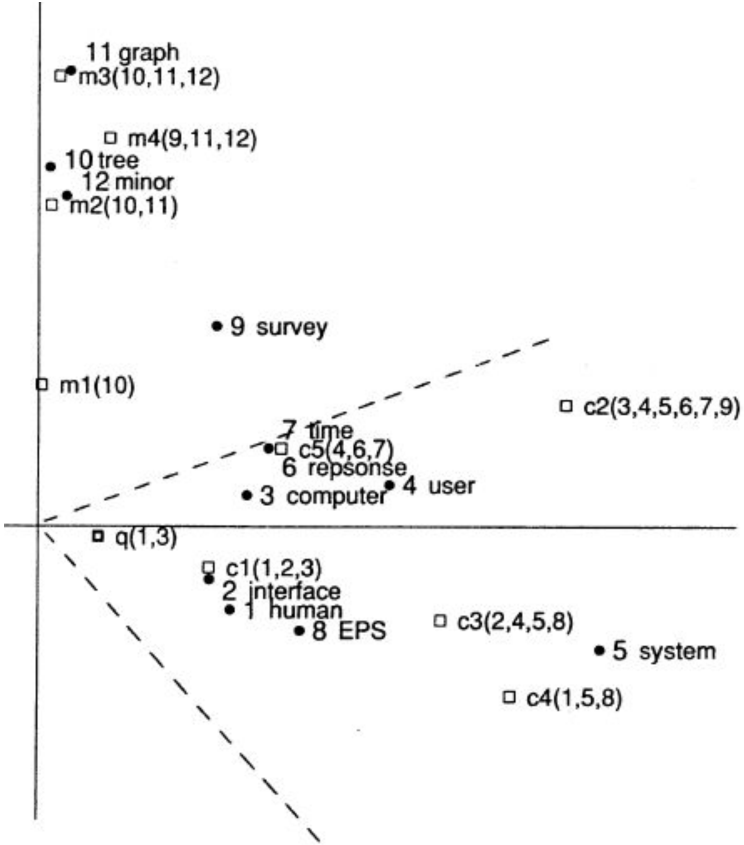
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering

m4: *Graph minors*: A *survey*

SVD Example

Terms	Documents								
	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

SVD Example



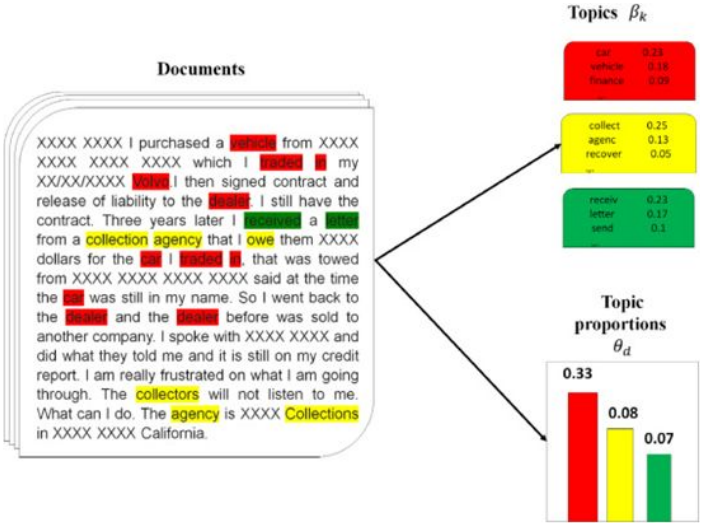
11 graph
m3(10,11,12)
m4(9,11,12)
10 tree
12 minor
m2(10,11)
9 survey
m1(10)
c2(3,4,5,6,7,9)
7 time
c5(4,6,7)
6 repsonse
4 user
3 computer
q(1,3)
c1(1,2,3)
2 interface
1 human
8 EPS
c3(2,4,5,8)
5 system
c4(1,5,8)

LDA

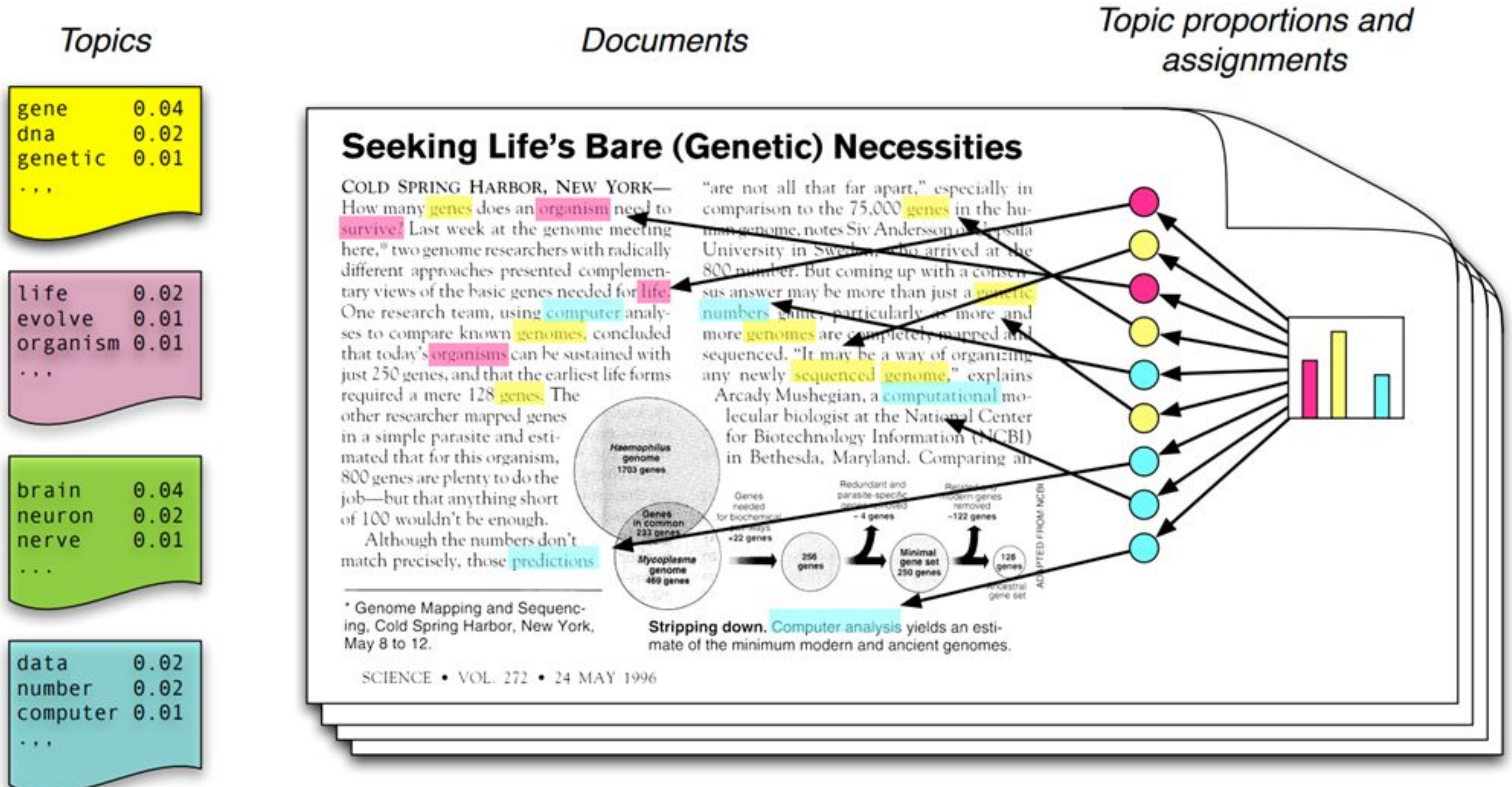
- Built on top of the notion of topics
- There are many topics
- Each document is a mixture of topics
- The topics themselves are unknown or latent

- Document is a collection of words
- Which words are present, depends on the topics in the document

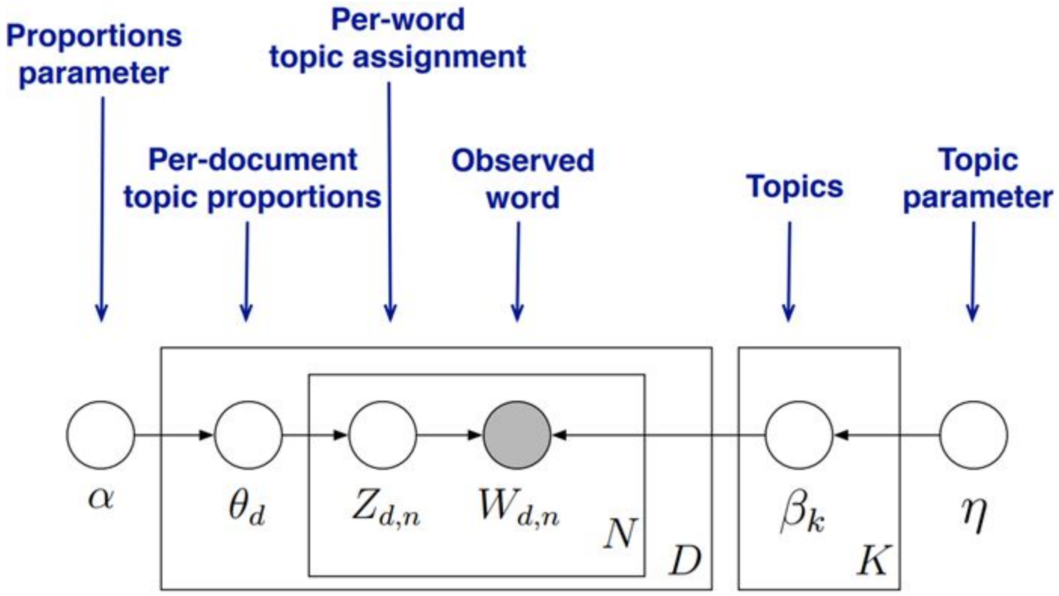
- Need a mechanism to
- Identify the topics
- Connect words with topics
- Explain the documents with these learnings



LDA

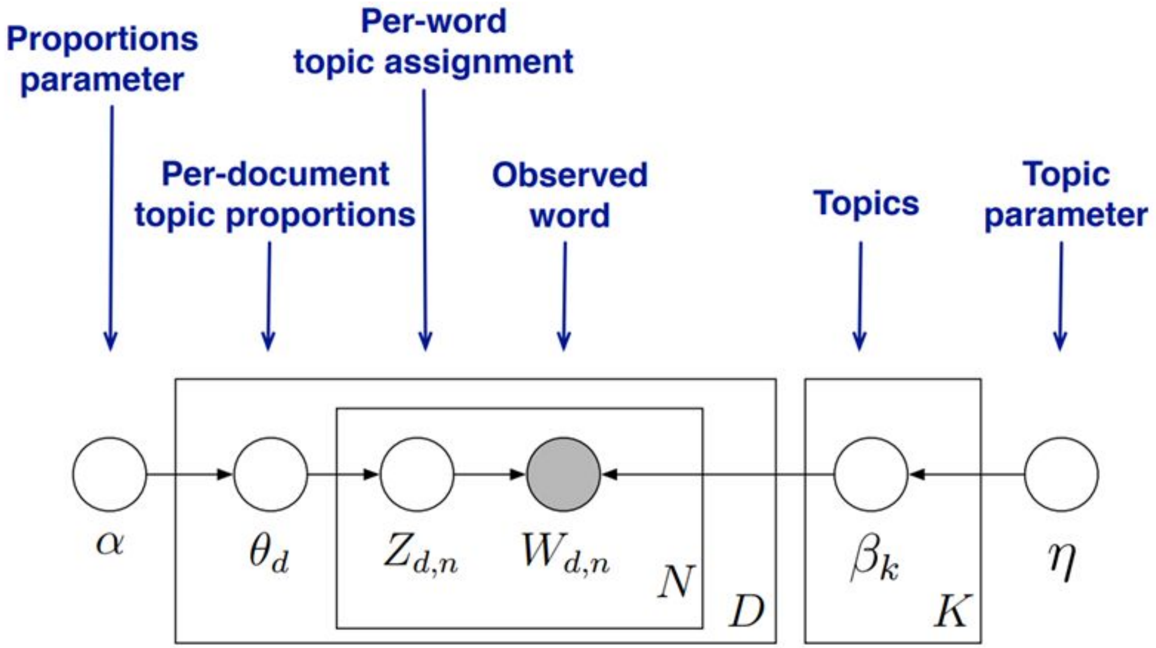

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

LDA



- α and η are hyperparameters
- θ, z and β are model parameters
- θ:: Topic distribution in the document
- z:: Per-topic word assignment
- β: Per-topic word distribution

LDA



$$p(\beta,\theta,\mathbf{z},\mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(\theta_d \mid \alpha)\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

LDA

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

```
import numpy as np
def f(theta1, theta2, theta3):
  s = np.random.dirichlet ((theta1, theta2, theta3), 5)
  for i in range(0,len(s)):
  print(i , ": %2.2f %2.2f %2.2f" % (s[i][0], s[i][1], s[i][2]))

f(10,10,10)
print("---")
```

Fore more on generating and plotting Dirichlet Distribution, please see:

- https://towardsdatascience.com/dirichlet-distribution-a82ab942a879
- https://numpy.org/doc/stable/reference/random/generated/numpy.random.dirichlet.html

Related reading: http://jrmeyer.github.io/machinelearning/2017/08/18/mle.html