CSE 190 – Lecture 14

Data Mining and Predictive Analytics

Dimensionality-reduction approaches to document representation – part 2

Assignment 1!

Task 1:

UCSDCSE
Computer Science and Engineering

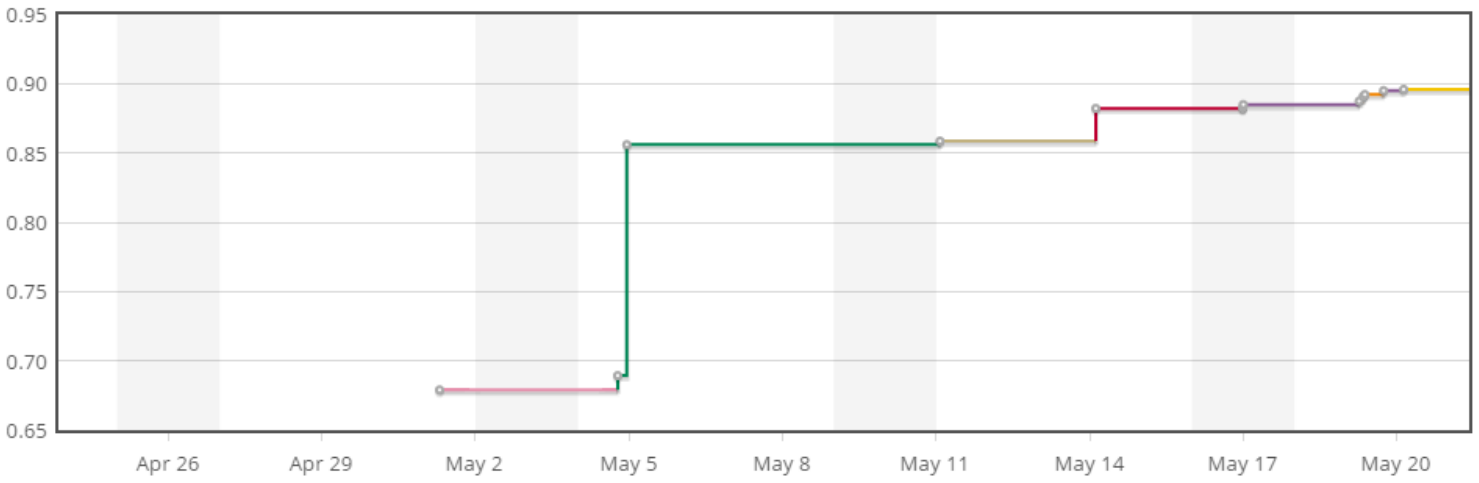
Completed • Knowledge • 117 teams

CSE 190 -- Assignment 1 -- Task 1 -- Purchase Prediction

Thu 23 Apr 2015 – Thu 21 May 2015 (9 hours ago)

Dashboard

Private Leaderboard - CSE 190 -- Assignment 1 -- Task 1 -- Purchase Prediction



This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts?
Let us know.

#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑3	zxcv1234	0.89588	46	Thu, 21 May 2015 05:58:12 (-26.7h)
2	↑3	Fedora The Explora	0.89568	133	Thu, 21 May 2015 07:37:45

Assignment 1!

Task 2:

UCSDCSE Computer Science and Engineering

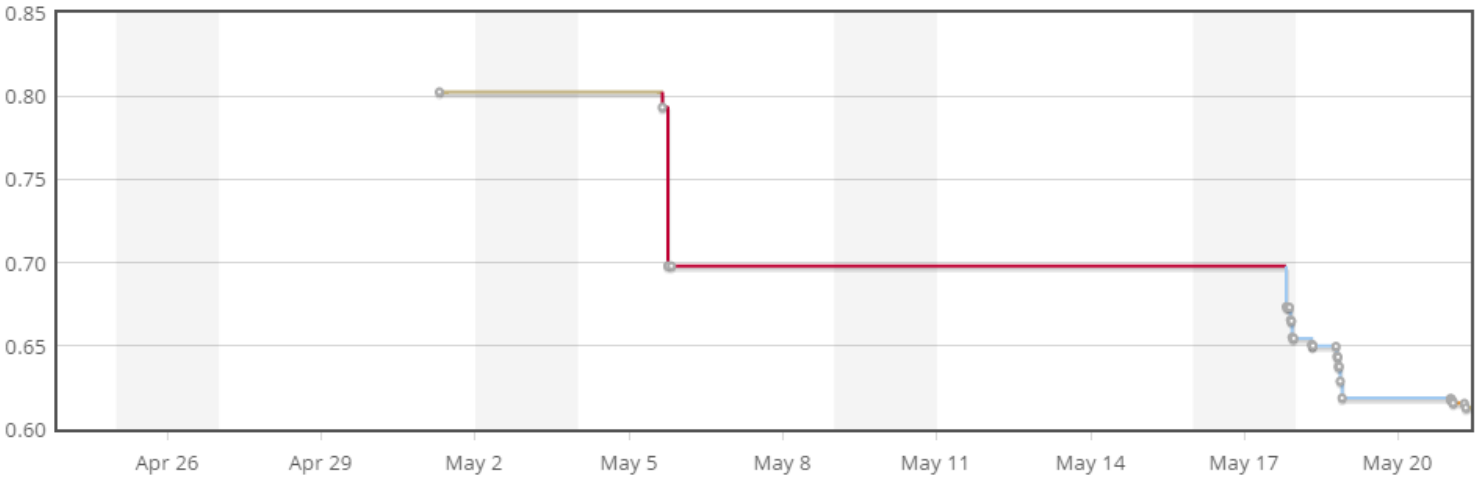
Completed • Knowledge • 116 teams

CSE 190 -- Assignment 1 -- Task 2 -- Helpfulness Prediction

Thu 23 Apr 2015 – Thu 21 May 2015 (9 hours ago)

Dashboard

Private Leaderboard - CSE 190 -- Assignment 1 -- Task 2 -- Helpfulness Prediction



This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts? Let us know.

#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑8	+_+	0.61576	20	Thu, 21 May 2015 07:00:13 (-5.3h)
2	↑9	┬─┬ノ(' - 'ノ)	0.61576	72	Thu, 21 May 2015 08:24:35 (-2.7h)

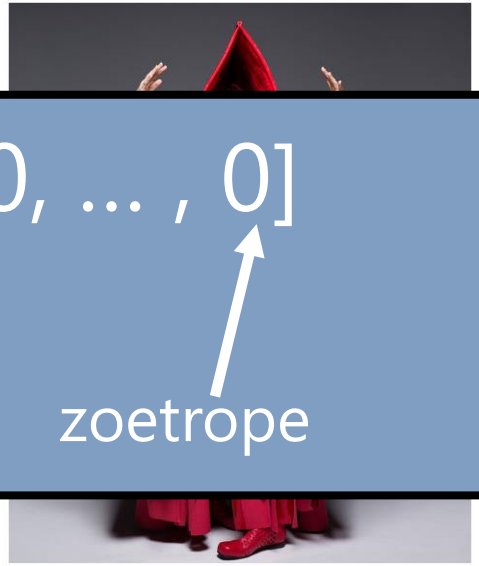
Previously – text mining

Bag-of-Words models

The Peculiar Genius of Bjork

CULTURE | BY EMILY WITT | JANUARY 23, 2015 11:30 AM

Solo musician or master collaborator? For her new album, Bjork has merged the two sides of her artistry to create a new experience of music — again.



F_text = [150, 0, 0, 0, 0, 0, … , 0]

a

aardvark

zoetrope

musician, who creates her music in an emotional cocoon, tinkering with technologies, concepts and feelings; and Bjork the producer and curator, who seeks out

Previously – text mining

Inference!

Problem 1: Sentiment analysis

Let's build a predictor of the form:

$$f(\text{text}) \rightarrow \text{rating}$$

using a model based on linear regression:

$$\text{rating} \simeq \alpha + \sum_{w \in \text{text}} \text{count}(w) \cdot \theta_w$$

Code: http://jmcauley.ucsd.edu/cse190/code/week6.py

Dimensionality reduction

How can we find **low-dimensional structure** in documents?

What we would like:

87 of 102 people found the following review helpful

★★★★★ **You keep what you kill**, December 27, 2004

By **Schtinky "Schtinky"** (Washington State) - See all my reviews

VINE™ VOICE

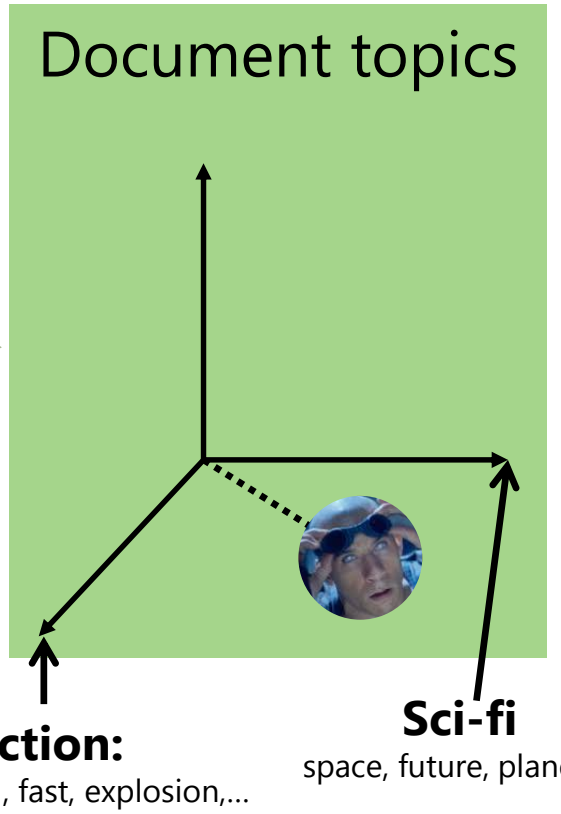
This review is from: **The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)**

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from `Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to `Pitch Black' fans like myself.

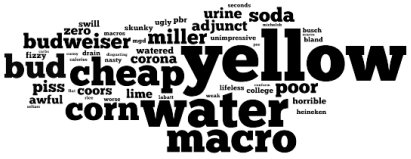
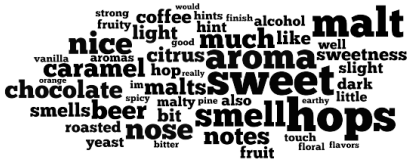
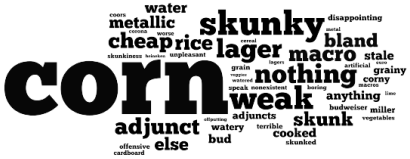
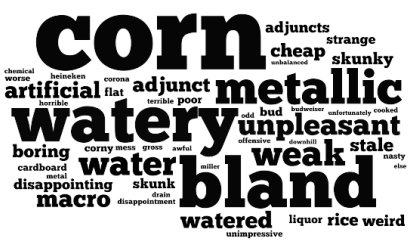
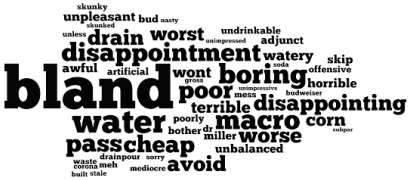
First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

topic model



Aspects of opinions

	Aspect words θ_k	Sentiment words (2-star) $\phi_{k,2}$	Sentiment words (5-star) $\phi_{k,5}$
Feel			
Look			
Smell			
Taste			
Overall impression			

Dimensionality reduction approaches to text

In the case study we just saw, the dimensions were **given** to us – we just had to find the topics corresponding to them

What can we do to find the dimensions **automatically?**

Singular-value decomposition

Recall (from weeks 3&5)

$$R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix}$$

(e.g.) matrix of ratings

(square roots of) eigenvalues of RR^T

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

eigenvectors of RR^T

eigenvectors of $R^T R$

Singular-value decomposition

Taking the eigenvectors corresponding to the top-K eigenvalues is then the "best" rank-K approximation

$$R = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix}$$

$$R \simeq U^{(k)} \Sigma^{(k)} V^{(k)T}$$

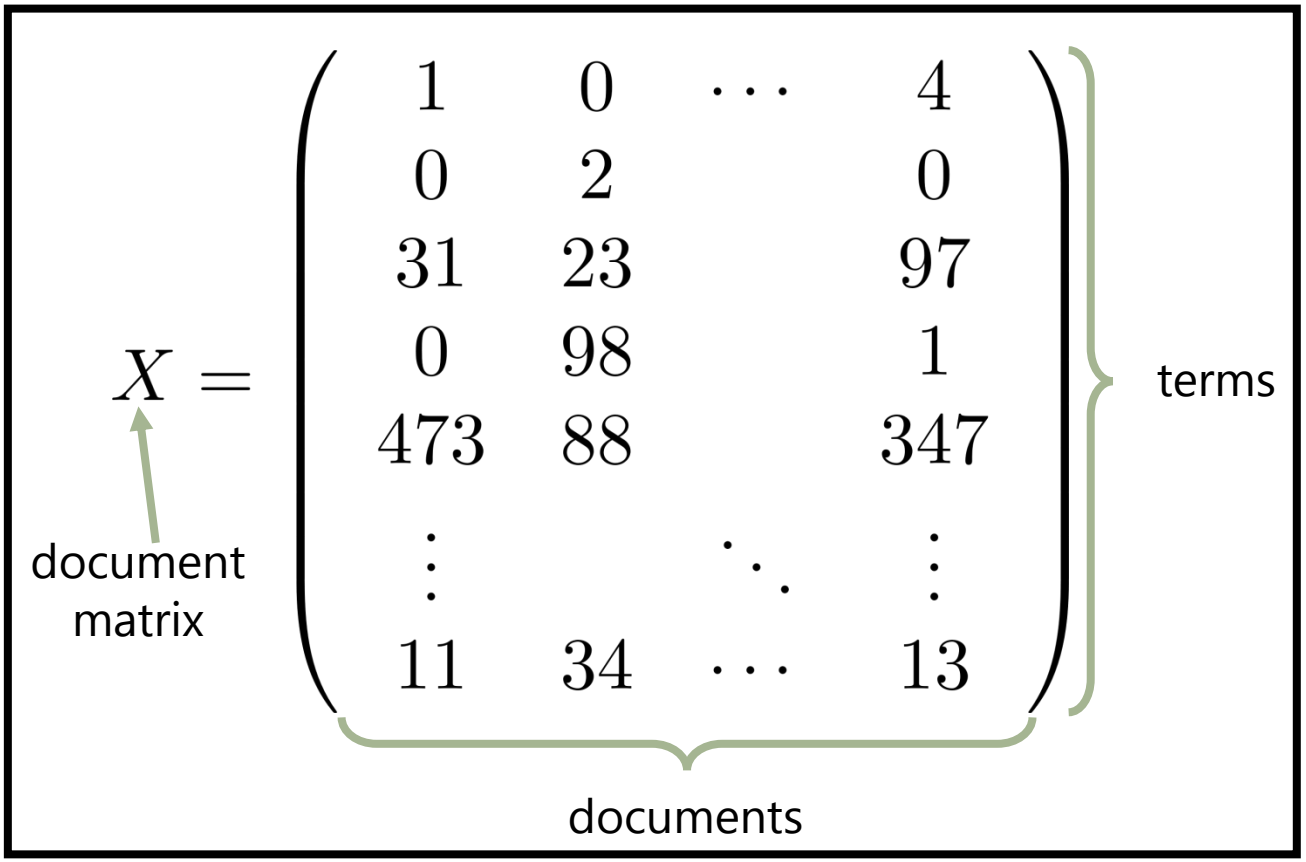
(square roots of top k) eigenvalues of RR^T

(top k) eigenvectors of RR^T

(top k) eigenvectors of $R^T R$

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?



X is a *TxD* matrix whose **columns** are bag-of-words representations of our documents

T = dictionary size
D = number of documents

Singular-value decomposition

What happens when we apply this to a matrix encoding our documents?

$X^T X$ is a *D*x*D* matrix.

$U^{(k)}\Sigma^{(k)}$ is a low-rank approximation of each **document**

eigenvectors of $X^T X$

XX^T is a *T*x*T* matrix.

$V^{(k)}\Sigma^{(k)}$ is a low-rank approximation of each **term**

eigenvectors of XX^T

Singular-value decomposition

Using our low rank representation of each **document** we can…

- Compare two documents by their low dimensional representations (e.g. by cosine similarity)
- To retrieve a document (by first projecting the query into the low-dimensional document space)
- Cluster similar documents according to their low-dimensional representations
- Use the low-dimensional representation as features for some other predictive task

Singular-value decomposition

Using our low rank representation of each **word** we can…

- Identify potential synonyms – if two words have similar low-dimensional representations then they should have similar "roles" in documents and are potentially synonyms of each other
- This idea can even be applied across languages, where similar terms in different languages ought to have similar representations in parallel corpora of translated documents

Singular-value decomposition

This approach is called **latent semantic analysis**

- In practice, computing eigenvectors for matrices of the sizes in question is not practical – neither for XX^T nor X^TX (they won't even fit in memory!)
- Instead one needs to resort to some approximation of the SVD, e.g. a method based on stochastic gradient descent that never requires us to compute XX^T or X^TX directly (much as we did when approximating rating matrices with low-rank terms)

Probabilistic modeling of documents

Finally, can we represent documents in terms of the topics they describe?

What we would like:

87 of 102 people found the following review helpful

★★★★★ **You keep what you kill**, December 27, 2004

By **Schtinky "Schtinky"** (Washington State) - See all my reviews

VINE™ VOICE

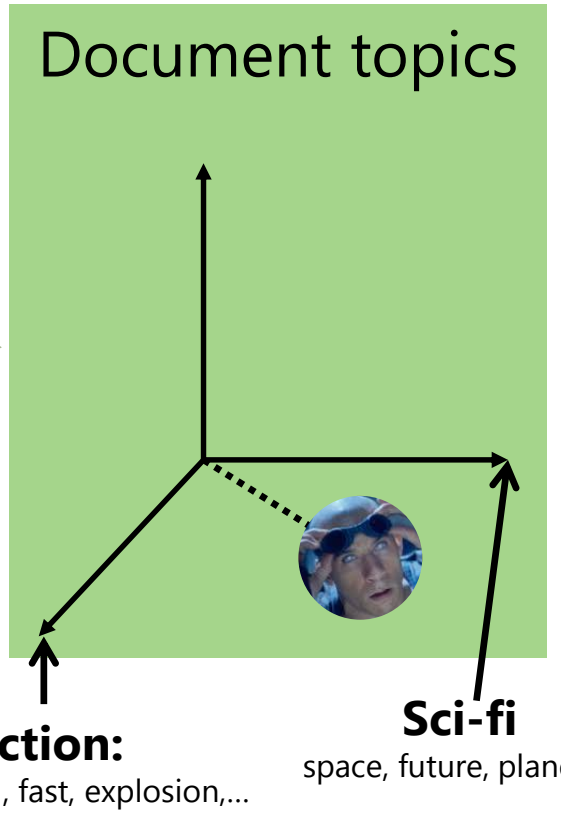
This review is from: **The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)**

Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from `Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to `Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of "The Chronicles of Riddick")

topic model

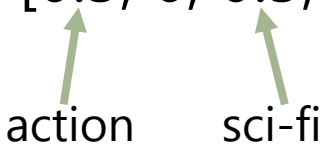


Probabilistic modeling of documents

Finally, can we represent documents in terms of the topics they describe?

- We'd like each document to be a **mixture over topics**

(e.g. if movies have topics like "action", "comedy", "sci-fi", and "romance", then reviews of action/sci-fis might have representations like [0.5, 0, 0.5, 0])

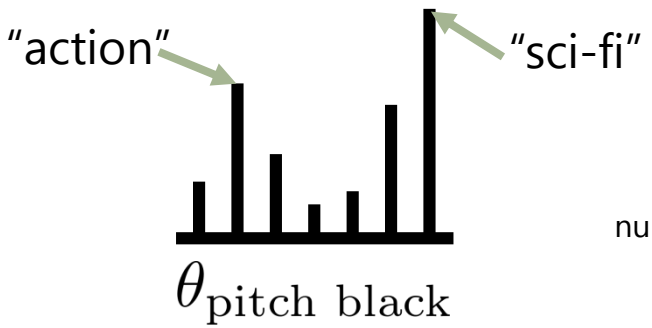


- Next we'd like each topic to be a **mixture over words**

(e.g. a topic like "action" would have high weights for words like "fast", "loud", "explosion" and low weights for words like "funny", "romance", and "family")

Latent Dirichlet Allocation

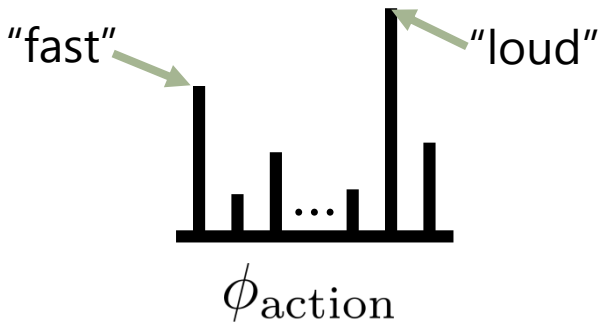
Both of these can be represented by **multinomial distributions**



Each document has a **topic distribution** which is a mixture over the topics it discusses

number of topics

$$\theta_d \in \Delta^K \text{ i.e., } \forall_d \sum_k \theta_{d,k} = 1$$



Each topic has a **word distribution** which is a mixture over the words it discusses

number of words

$$\phi_k \in \Delta^D \text{ i.e., } \forall_k \sum_w \phi_{k,w} = 1$$

Latent Dirichlet Allocation

LDA assumes the following "process" that generates the words in a document

(suppose we already know the topic distributions and word distributions)

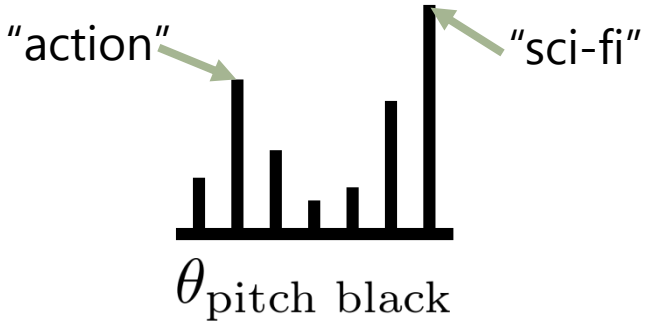
```
for j = 1 .. length of document:
  sample a topic for the word:
    z_dj ← \theta_d
  sample a word from the topic:
    w_j ← \phi_{z_dj}
```

Since each word is sampled independently, the output of this process is a **bag of words**

Latent Dirichlet Allocation

LDA assumes the following "process" that generates the words in a document

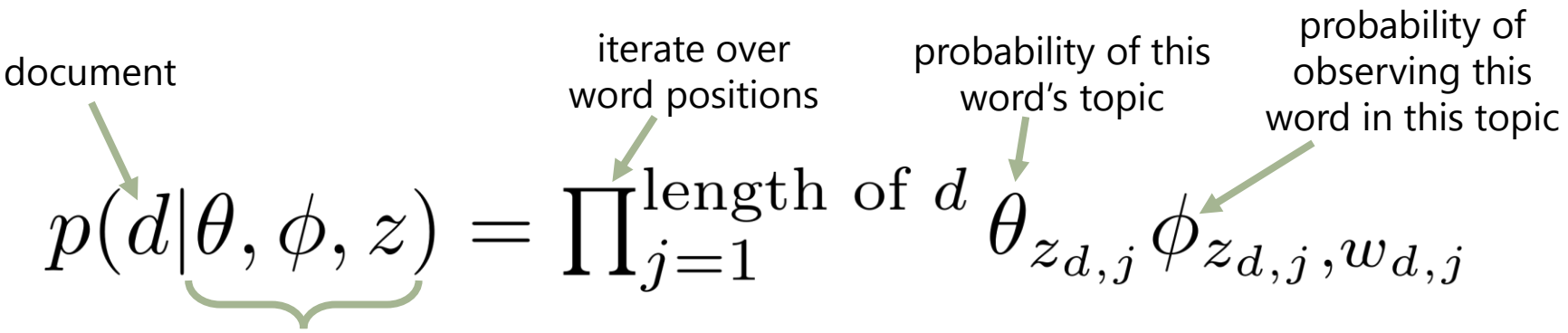
e.g. generate a likely review for pitch black:



j	Sample a topic	Sample a word
1	$z_{d1} = 2$	"explosion"
2	z_{d2}=7	"space"
3	z_{d3}=2	"bang"
4	z_{d4}=7	"future"
5	z_{d5}=7	"planet"
6	z_{d6}=6	"acting"
7	z_{d7}=2	"explosion"

Latent Dirichlet Allocation

Under this model, we can estimate the probability of a particular bag-of-words appearing with a particular topic and word distribution



Problem: we need to estimate all this stuff before we can compute this probability!

Latent Dirichlet Allocation

We need to estimate the topics (theta), the word distributions (phi) **and** the topic assignments (z, latent variables) that explain the observations (the words in the document)

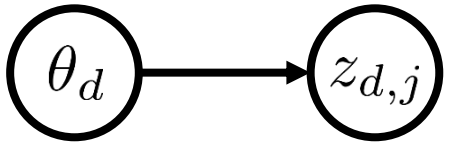
We can write down the dependencies between these variables using a (big!) **graphical model**

Latent Dirichlet Allocation

For every single word we have an edge like:

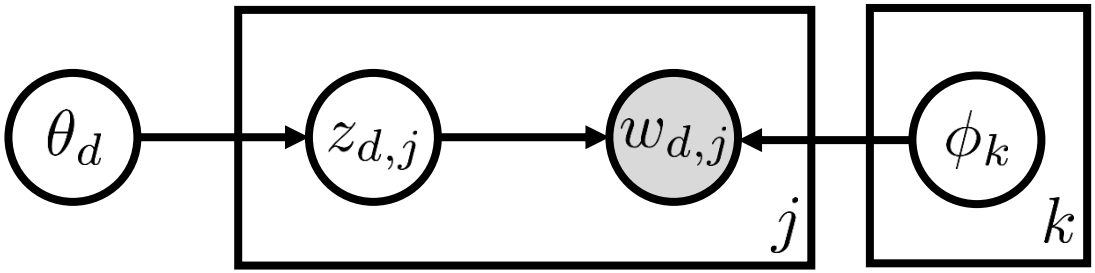
$$p(w|z) = \phi_{z,w}$$

and an edge like:



$$p(z|\theta) = \theta_z$$

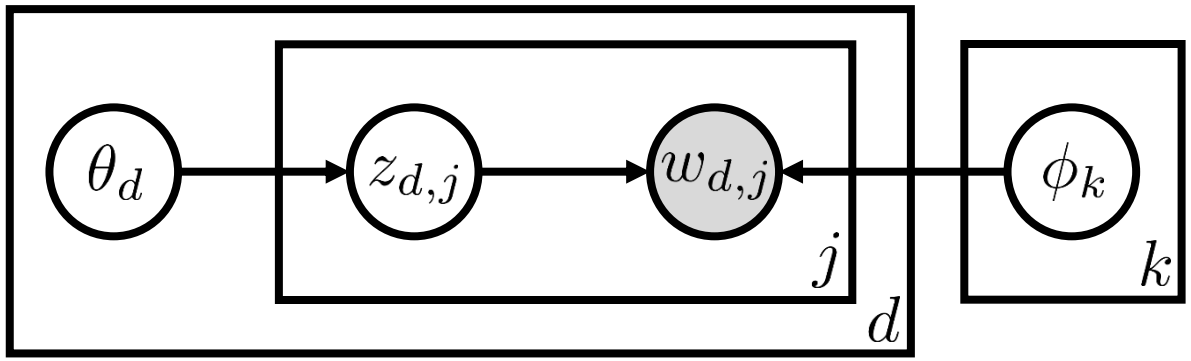
for convenience we draw this like:



(this is called "plate notation")

Latent Dirichlet Allocation

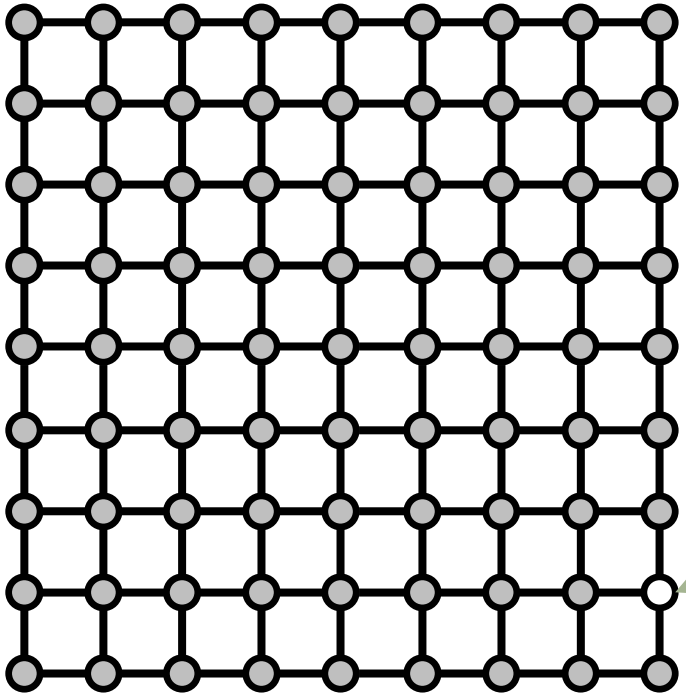
And we have a copy of this for every document!



Finally we have to estimate the parameters of this (rather large) model

Gibbs Sampling

Modeling fitting is traditionally done by **Gibbs Sampling.** This is a very simple procedure that works as follows:



1. Start with some initial values of the parameters
2. For each variable (according to some schedule), condition on its neighbors
3. **Sample** a new value for that variable (*y*) according to p(y|neighbors)
4. Repeat until you get bored

Gibbs Sampling

Modeling fitting is traditionally done by **Gibbs Sampling.** This is a very simple procedure that works as follows:

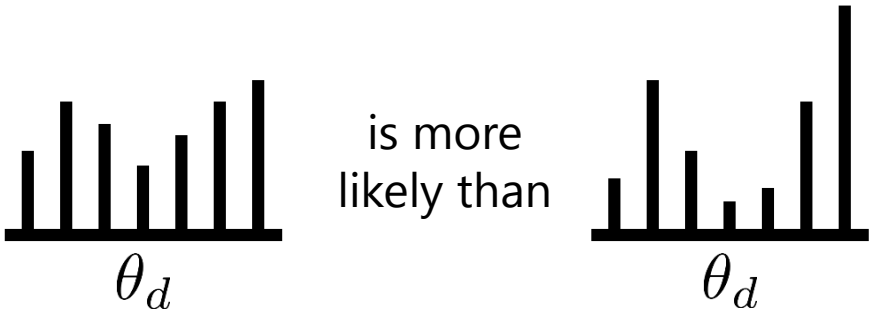
Gibbs Sampling has useful theoretical properties, most critically that the probability of a variable occupying a particular state (over a sequence of samples) is equal to the true marginal distribution, so we can (eventually) estimate the unknowns (theta, phi, and z) in this way

Gibbs Sampling

What about regularization?

How should we go about fitting topic distributions for documents with few words, or word distributions of topics that rarely occur?

- Much as we do with a regularizer, we'd like to penalize the deviation from uniformity
- That is, we'd like to penalize \theta and \phi for being too non-uniform



Gibbs Sampling

Since we have a probabilistic model, we want to be able to write down our regularizer as a **probability** of observing certain values for our parameters

$$p(\theta_d) = ? \quad p(\phi_k) = ?$$

- We want the probability to be higher for \theta and \phi closer to uniform
  - This property is captured by a **Dirichlet distribution**

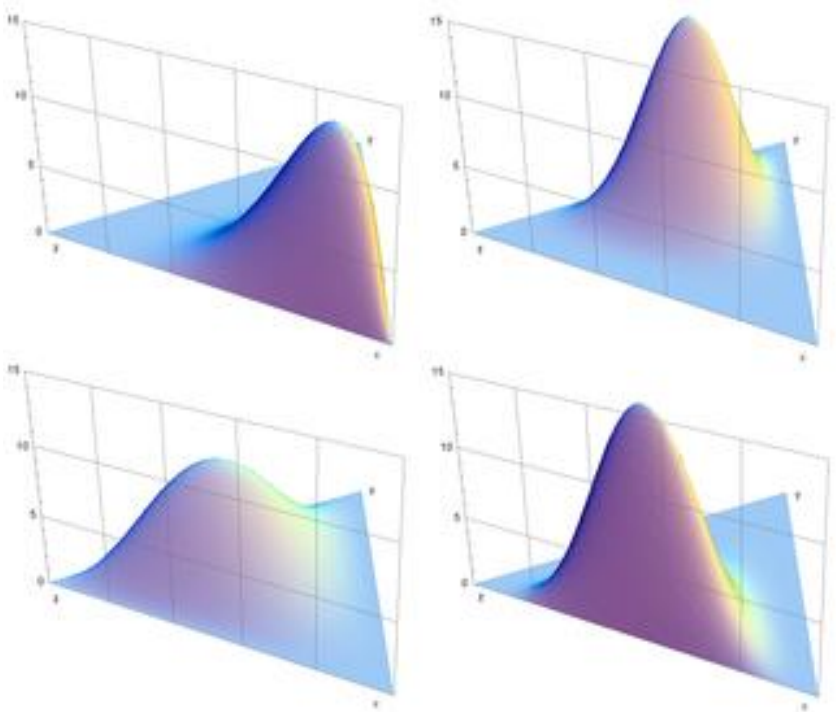
Dirichlet distribution

A Dichlet distribution "generates" multinomial distributions. That is, it's support is the set of points that lie on a simplex (i.e., positive values that add to 1)

concentration parameters

p.d.f.: $\frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

beta function



Visualization of a three-dimensional dirichlet distribution (from wiki)

Dirichlet distribution

The concentration parameters \alpha encode our prior probability of certain topics having higher likelihood than others

concentration parameters

p.d.f.: $\frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

beta function

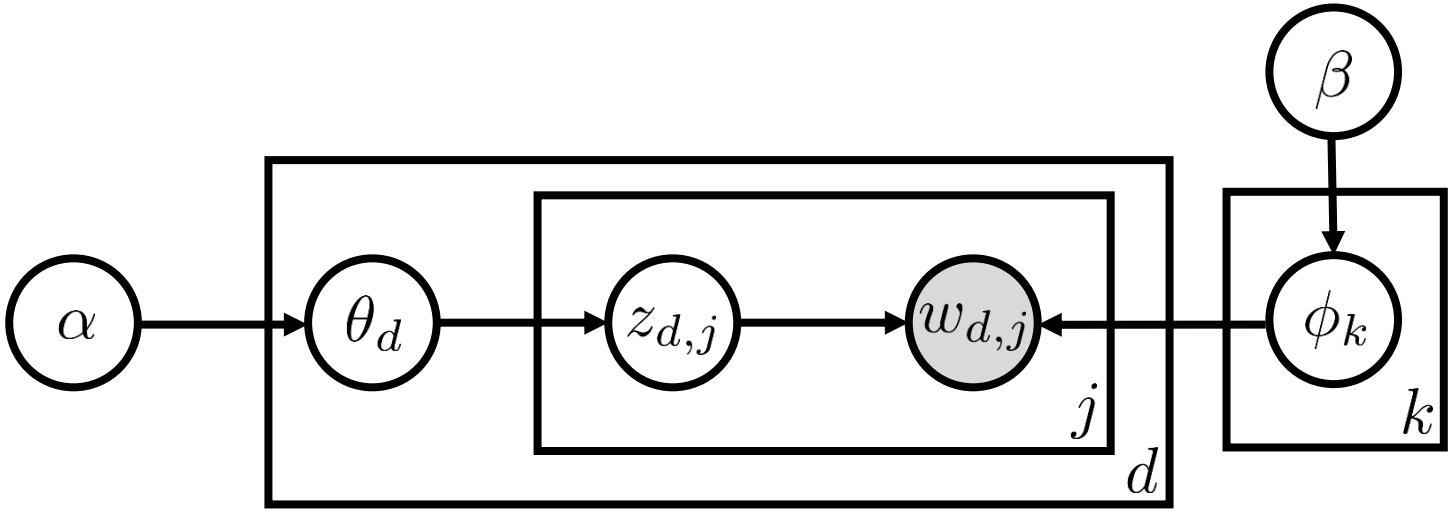
- In the most typical case, we want to penalize deviation from uniformity, in which case \alpha is a uniform vector
- In this case the expression simplifies to the **symmetric** Dirchlet distribution:

p.d.f.: $\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} x_i^{\alpha - 1}$

gamma function

Latent Dirichlet Allocation

These two parameters now just become additional unknowns in the model:



- The larger the values of alpha/beta, the more we penalize deviation from uniformity
- Usually we'll set these parameters by grid search, just as we do when choosing other regularization parameters

Latent Dirichlet Allocation

E.g. some topics discovered from an Associated Press corpus

labels are determined manually

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Latent Dirichlet Allocation

And the topics most likely to have generated each word in a document

labels are determined manually

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

From http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

Latent Dirichlet Allocation

Many many many extensions of Latent Dirichlet Allocation have been proposed:

- To handle temporally evolving data:

"Topics over time: a non-Markov continuous-time model of topical trends" (Wang & McCallum, 2006)
http://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf

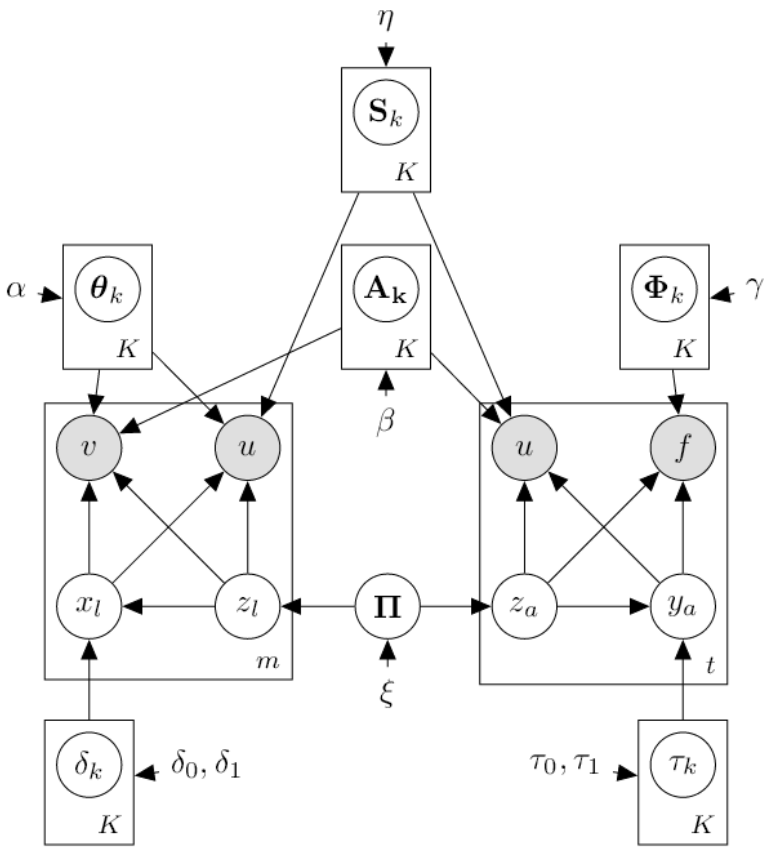
- To handle **relational** data:

"Block-LDA: Jointly modeling entity-annotated text and entity-entity links" (Balasubramanyan & Cohen, 2011)
http://www.cs.cmu.edu/~wcohen/postscript/sdm-2011-sub.pdf
"Relational topic models for document networks" (Chang & Blei, 2009)
https://www.cs.princeton.edu/~blei/papers/ChangBlei2009.pdf
"Topic-link LDA: joint models of topic and author community" (Liu, Nicelescu-Mizil, & Gryc, 2009)
http://www.niculescu-mizil.org/papers/Link-LDA2.crc.pdf

Latent Dirichlet Allocation

Many many many extensions of Latent Dirichlet Allocation have been proposed:

"WTFW" model (Barbieri, Bonch, & Manco, 2014), a model for relational documents



Latent Dirichlet Allocation

Many many many extensions of Latent Dirichlet Allocation have been proposed:

- To handle user opinions & rating data

Case study!

Summary

Text mining

Using **text** to solve predictive tasks

- Representing documents using bags-of-words and TF-IDF weighted vectors
- Stemming & stopwords
- Sentiment analysis and classification

Dimensionality reduction approaches:

- Latent Semantic Analysis
- Latent Dirichlet Allocation

Questions?

Further reading:

- Latent semantic analysis

"An introduction to Latent Semantic Analysis" (Landauer, Foltz, & Laham, 1998)
http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

- LDA

"Latent Dirichlet Allocation" (Blei, Ng, & Jordan, 2003)
http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf

- Plate notation

http://en.wikipedia.org/wiki/Plate_notation
"Operations for Learning with Graphical Models" (Buntine, 1994)
http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume2/buntine94a.pdf