

# CS6220: DATA MINING TECHNIQUES

## Text Data: Topic Models

---

**Instructor: Yizhou Sun**

[yzsun@ccs.neu.edu](mailto:yzsun@ccs.neu.edu)


October 26, 2015

# Methods to Learn

	Matrix Data	Text Data	Set Data	Sequence Data	Time Series	Graph & Network	Images
Classification	Decision Tree; Naïve Bayes; Logistic Regression SVM; kNN			HMM		Label Propagation*	Neural Network
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means*	<b>PLSA</b>				SCAN*; Spectral Clustering*	
Frequent Pattern Mining			Apriori; FP-growth	GSP; PrefixSpan			
Prediction	Linear Regression				Autoregression		
Similarity Search					DTW	P-PageRank	
Ranking						PageRank	

# Text Data: Topic Models

---

- Text Data and Topic Models 
- Probabilistic Latent Semantic Analysis
- Summary

# Text Data

- Word/term
- Document
  - A sequence of words
- Corpus
  - A collection of documents

### Cancer effort honored

Handsworth memorial award set up

By Anna Marie D'Angelo

News Reporter

MELINDA Hathaway, sponsor for her Internet Web page on cancer, will be honored by her school with a memorial award.

Hathaway, 33, died in April after a 2 1/2 year battle with Aden's Tumor, a rare soft tissue cancer that started in her spine.

Hathaway was originally featured in a News story in April about her popular Internet home page on the World Wide Web.

Hathaway used the Web site to share her experiences with cancer and raise about her family.

More than 4,000 hits were made to Melinda's home page in less than two months.

In July more than 20,000 hits have been made since the home page was launched in February.

Things that Melinda Hathaway wrote.

Melinda Hathaway would have been a page on July 10th and she would have been a similar experience.

Her work is shared with her son's looking for



MELINDA Hathaway showed her positive outlook on life and her cancer experiences with kids on the Internet. Hathaway died last month.

### LIFESTYLE

## The geek grew up

Now Amanda Welliver helps teens bully-proof themselves

By Jeff Labadie

News Reporter

Amanda Welliver, a former geek, is now a successful businesswoman and author of a book on how to deal with bullies.

Welliver, 30, is a former student of the University of Wisconsin-Madison. She was a member of the Phi Kappa Phi Honor Society and the Phi Kappa Phi Honor Society.

Welliver's book, "The Geek's Guide to Life: How to Deal with Bullies, How to Find Love, and How to Succeed in Business," is a guide for young adults who were once geeks.



News, Thursday, March 16, 2012

### Raven in his 80s set to see again after op

Tarquin's blind, has only one eye but has mate 70 years his junior

By Steve Bortol

A BIRD named "Tarquin" is set to have an operation within days to correct the defect in his beak that has prevented him from seeing for the past 10 years.

Tarquin, a 75-year-old male, was found in a cage at a veterinary hospital in Madison, Wis., in 2001. He was blind and had a severely deformed beak.

Dr. Bortol, who treated Tarquin, said the bird is "a real character" and is "a real joy to work with."

Tarquin's mate, a 4-year-old female named "Lily," is a "real beauty" and is "a real joy to work with."



### It's man vs burger

Monster meal deals hard to digest

By Matt Miller

Man vs. Monster: A man who ate a 100-calorie burger in 10 seconds.

The man, who is a professional eater, ate the burger in 10 seconds.



### INDIANAPOLIS NEWS

MINOR EVENING, JULY 18, 1912

## HUNDREDS IN STATE SEE 'FLYING SAUCERS'

Franklin 'Dogfight' Alerts State Troopers

State Troopers were alerted to sightings of "flying saucers" in Franklin County, Ind., last night.

The sightings appeared to have a "dogfight" character and resembled a "dogfight" between two aircraft.



# Represent a Document

- Most common way: Bag-of-Words
  - Ignore the order of words
  - keep the count

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*



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

# More Details

---

- Represent the doc as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it)
  - Number of words is huge
  - Select and use a smaller set of words that are of interest
  - E.g. uninteresting words: 'and', 'the', 'at', 'is', etc. These are called stop-words
  - Stemming: remove endings. E.g. 'learn', 'learning', 'learnable', 'learned' could be substituted by the single stem 'learn'
  - Other simplifications can also be invented and used
  - The set of different remaining words is called dictionary or vocabulary. Fix an ordering of the terms in the dictionary so that you can operate them by their index.
  - Can be extended to bi-gram, tri-gram, or so



# Topic Models

- Topic modeling
  - Get topics automatically from a corpus
  - Assign documents to topics automatically
- Most frequently used topic models
  - pLSA
  - LDA


“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

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.



# Text Data: Topic Models

---

- Text Data and Topic Models
- Probabilistic Latent Semantic Analysis 
- Summary

# Notations

---

- Word, document, topic
  - $w, d, z$
- Word count in document
  - $c(w, d)$
- Word distribution for each topic ( $\beta_z$ )
  - $\beta_{zw}: p(w|z)$
- Topic distribution for each document ( $\theta_d$ )
  - $\theta_{dz}: p(z|d)$  (Yes, fuzzy clustering)

# Review of Multinomial Distribution

---

- Select  $n$  data points from  $K$  categories, each with probability  $p_k$ 
  - $n$  trials of independent categorical distribution
  - E.g., get 1-6 from a dice with  $1/6$
- When  $K=2$ , binomial distribution
  - $n$  trials of independent Bernoulli distribution
  - E.g., flip a coin to get heads or tails



# Generative Model for pLSA

- Describe how a document is generated probabilistically

- For each position in  $d$ ,  $n = 1, \dots, N_d$

- Generate the topic for the position as

$$z_n \sim \text{mult}(\cdot | \theta_d), \text{ i.e., } p(z_n = k) = \theta_{dk}$$

(Note, 1 trial multinomial, i.e., categorical distribution)

- Generate the word for the position as

$$w_n \sim \text{mult}(\cdot | \beta_{z_n}), \text{ i.e., } p(w_n = w) = \beta_{z_n w}$$

“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

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.

# The Likelihood Function for a Corpus

- Probability of a word

$$p(w|d) = \sum_k p(w, z = k|d) = \sum_k p(w|z = k)p(z = k|d) = \sum_k \beta_{kw} \theta_{dk}$$

- Likelihood of a corpus

$$\begin{aligned} & \prod_{d=1} P(w_1, \dots, w_{N_d}, d | \theta, \beta, \pi) \\ &= \prod_{d=1} P(d) \left\{ \prod_{n=1}^{N_d} \left( \sum_k P(z_n = k | d, \theta_d) P(w_n | \beta_k) \right) \right\} \\ &= \prod_{d=1} \pi_d \left\{ \prod_{n=1}^{N_d} \left( \sum_k \theta_{dk} \beta_{kw_n} \right) \right\} \end{aligned}$$

*$\pi_d$  is usually considered as uniform: i.e.,  $1/N_d$*

# Re-arrange the Likelihood Function

---

- Group the same word from different positions together

$$\max \log L = \sum_{dw} c(w, d) \log \sum_z \theta_{dz} \beta_{zw}$$

$$s. t. \sum_z \theta_{dz} = 1 \text{ and } \sum_w \beta_{zw} = 1$$

# Optimization: EM Algorithm

- Repeat until converge

- E-step: for each word in each document, calculate is conditional probability belonging to each topic

$$p(z|w, d) \propto p(w|z, d)p(z|d) = \beta_{zw}\theta_{dz} \text{ (i. e., } p(z|w, d) = \frac{\beta_{zw}\theta_{dz}}{\sum_{z'} \beta_{z'w}\theta_{dz'}})$$


- M-step: given the conditional distribution, find the parameters that can maximize the expected likelihood

$$\beta_{zw} \propto \sum_d p(z|w, d)c(w, d) \text{ (i. e., } \beta_{zw} = \frac{\sum_d p(z|w, d)c(w, d)}{\sum_{w', d} p(z|w', d)c(w', d)})$$

$$\theta_{dz} \propto \sum_w p(z|w, d)c(w, d) \text{ (i. e., } \theta_{dz} = \frac{\sum_w p(z|w, d)c(w, d)}{N_d})$$

# Text Data: Topic Models

---

- Text Data and Topic Models
- Probabilistic Latent Semantic Analysis
- Summary 



# Summary

---

- Basic Concepts
  - Word/term, document, corpus, topic
  - How to represent a document
- pLSA
  - Generative model
  - Likelihood function
  - EM algorithm