

Clustering the Tagged Web



Daniel Ramage, Paul Heymann,
Christopher D. Manning, Hector Garcia-Molina

Stanford University

WSDM 2009

Images from del.icio.us, lbaumann.com, www.hometrainingtools.com

Web document text



QUICK ORDER

FREE CATALOG

ORDER FORMS

TEACHING TIPS

SCIENCE PROJECTS

NEWSLETTERS

ONLINE STORE

Microscopes & Accessories
Life Science & Biology
Earth & Space Science
Chemistry
Physical Science & Physics
Technology
General Science
Science Books
Science by Grade Level
Science Curriculum
Science Kits for Curriculum

MORE SHOPPING

Nature Backpack Kits
Science Kits
Great Gift Ideas
Bestsellers
New Products
Monthly Specials



Welcome! Looking for hands-on science ideas? Try these:

- ▶ [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- ▶ [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- ▶ [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.



Chemistry Supplies:

Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts,

molecular models, digital balances, and more. Or make it



Dissection Supplies

Find everything you need to dissect frogs, cow eyes, and more.

▶ [Read more](#)



Study Bacteria
Experiment with

HACKER SAFE

TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- ▶ World of Germs
\$19.95

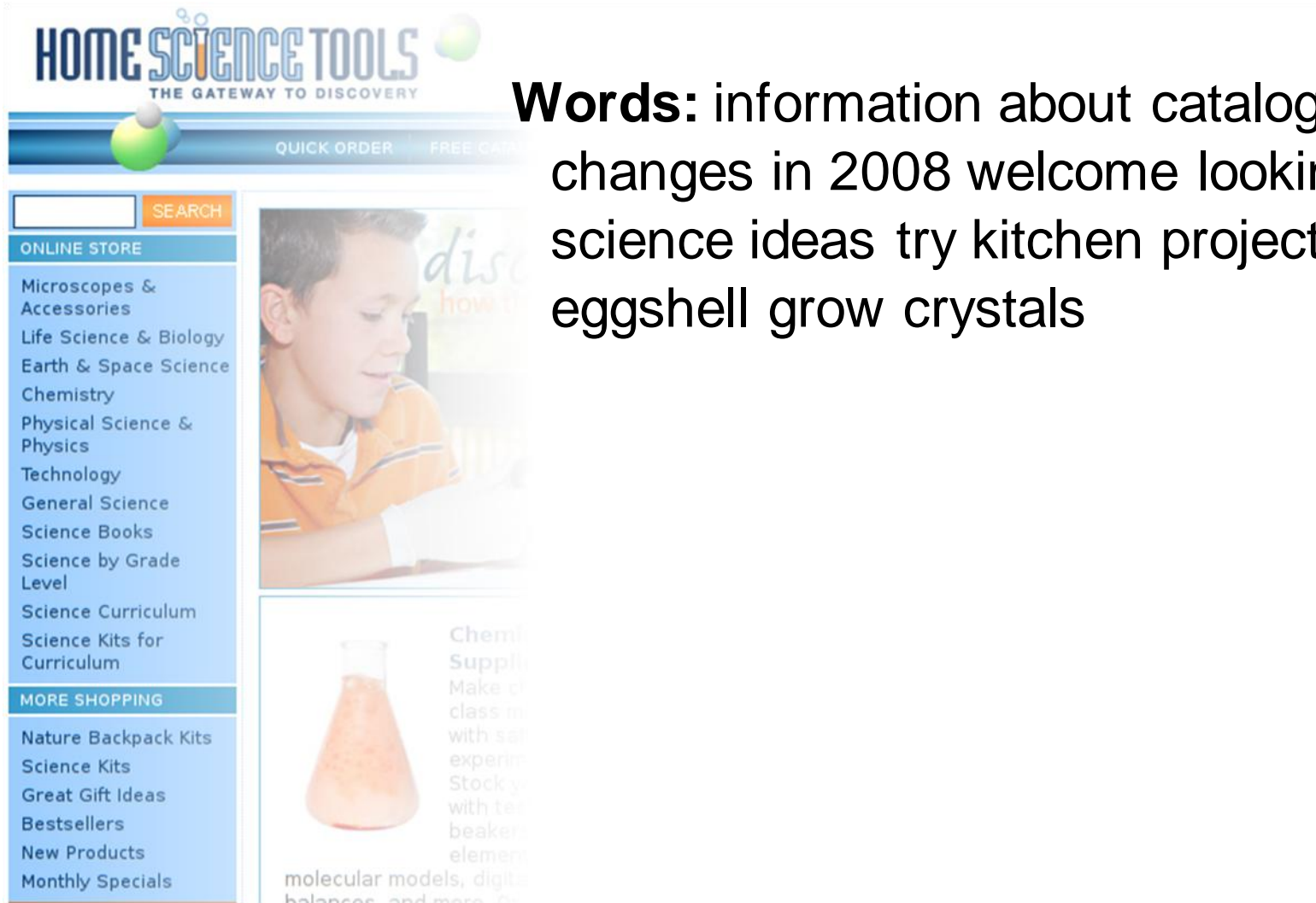


- ▶ Chemlab 1100 Chemistry Kit
\$30.95



- ▶ Kids LED Cordless Microscope

Web document text



HOME SCIENCE TOOLS
THE GATEWAY TO DISCOVERY

QUICK ORDER FREE CASH

SEARCH

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

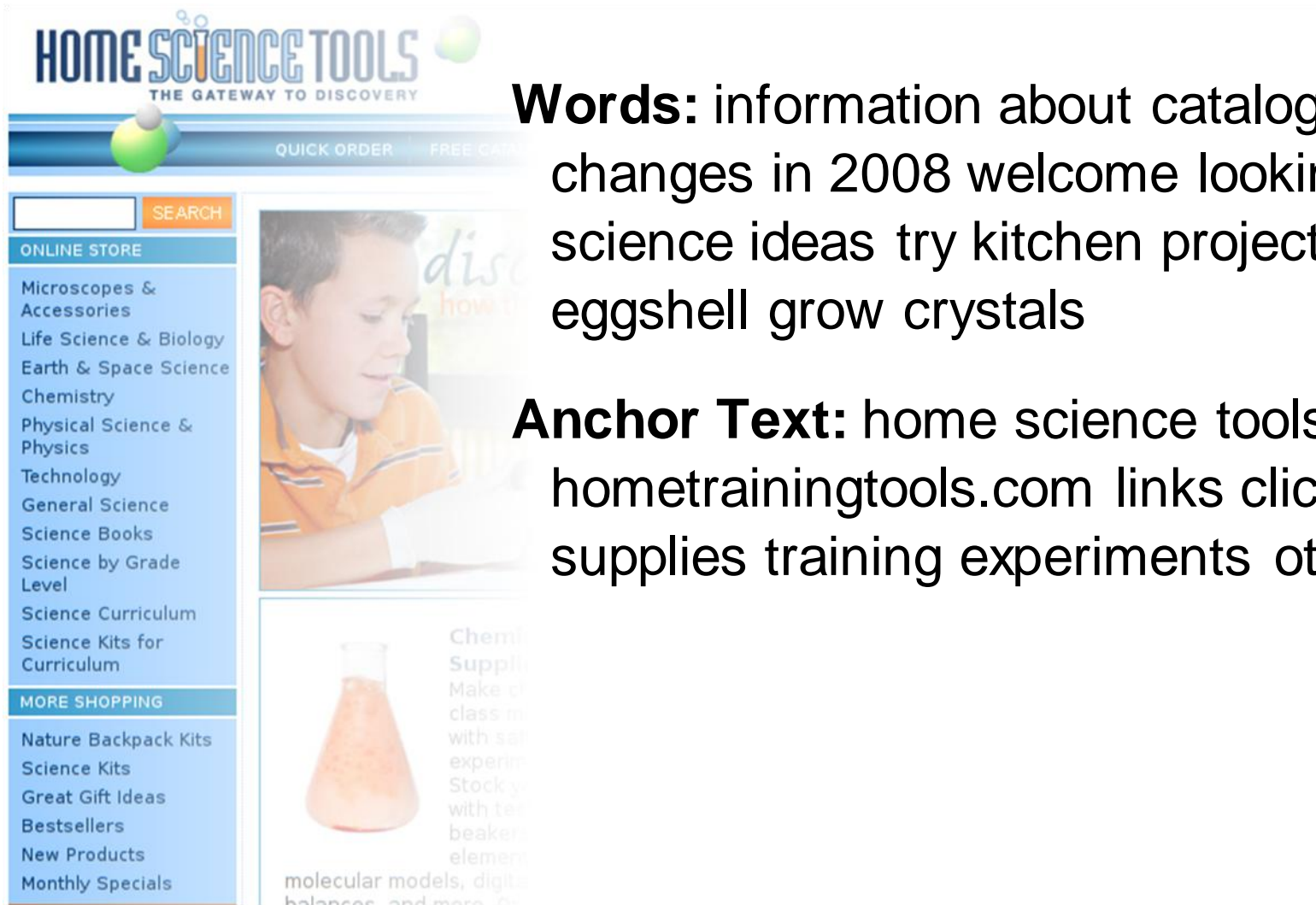
- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

disc
how to

Chemical Supplies
Make it
class m
with saf
experim
Stock y
with tes
beakers
elemen
molecul
balance
and more

Words: information about catalog pricing changes in 2008 welcome looking hands-on science ideas try kitchen projects dissolve eggshell grow crystals

Web document text



Words: information about catalog pricing changes in 2008 welcome looking hands-on science ideas try kitchen projects dissolve eggshell grow crystals

Anchor Text: home science tools hometrainingtools.com links click follow supplies training experiments other pages

Web document text



Words: information about catalog pricing changes in 2008 welcome looking hands-on science ideas try kitchen projects dissolve eggshell grow crystals

Anchor Text: home science tools
hometrainingtools.com links click follow supplies training experiments other pages

Tags: science homeschool education shopping curriculum homeschooling experiments tools chemistry supplies

Why tags? – del.icio.us

The most popular bookmarks on Delicious right now

[See more Popular bookmarks](#) 

New bo



[WordPress Resources: The Ultimate Collection » DivitoDesign](#) SAVE

136

[wordpress](#) [resources](#) [tips](#) [themes](#) [tutorials](#)



[7 Cool Things to Do With Linux | davehayes.org](#) SAVE

136

[linux](#) [tips](#) [utilities](#) [security](#) [mediacenter](#)



[Embedr - Create Video Playlists and Embed Them Anywhere](#) SAVE

137

[video](#) [embed](#) [youtube](#) [playlist](#) [tools](#)



[50 Useful JavaScript Tools | Developer's Toolbox | Smashing Magazine](#) SAVE

228

[javascript](#) [tools](#) [webdesign](#) [ajax](#) [jquery](#)



[「クックパッド」の裏側にいった](#) SAVE

128

[cookpad](#) [rails](#) [server](#) [ruby](#) [development](#)

Why tags? – del.icio.us

The most popular bookmarks on Delicious right now

[See more Popular bookmarks](#) →

New bo



[WordPress Resources: The Ultimate Collection » DivitoDesign](#) SAVE

136

wordpress < resources < tips < themes < tutorials



[7 Cool Things to Do With Linux | davehayes.org](#) SAVE

136

linux < tips < utilities < security < mediacenter



[Embedr - Create Video Playlists and Embed Them Anywhere](#) SAVE

137

video < embed < youtube < playlist < tools

≈120,000 posts / day

12-75 million (≈ 10^7 – 10^8) unique URLs

(versus 10^9 – 10^{11} total URLs)

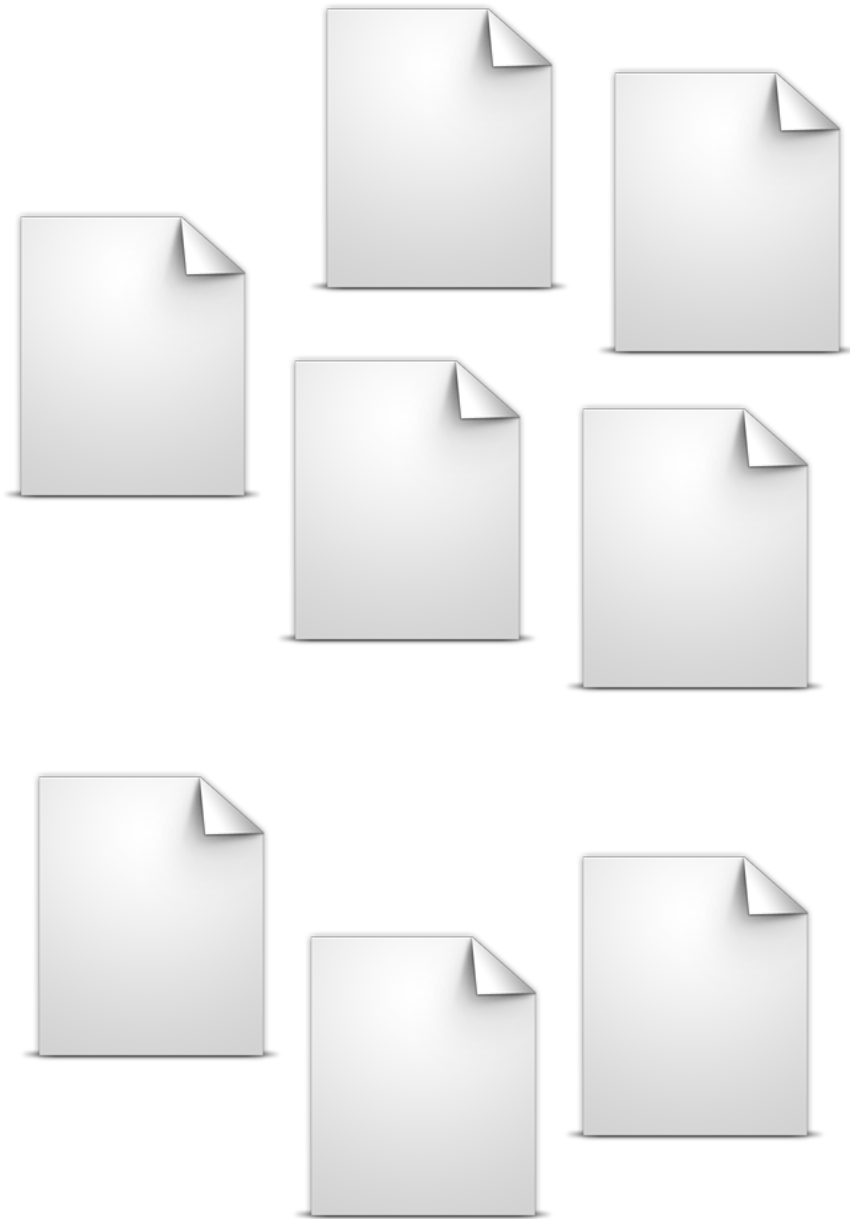
Disproportionately the web's most useful URLs

(and those URLs have many tags)

Using tags to understand the web

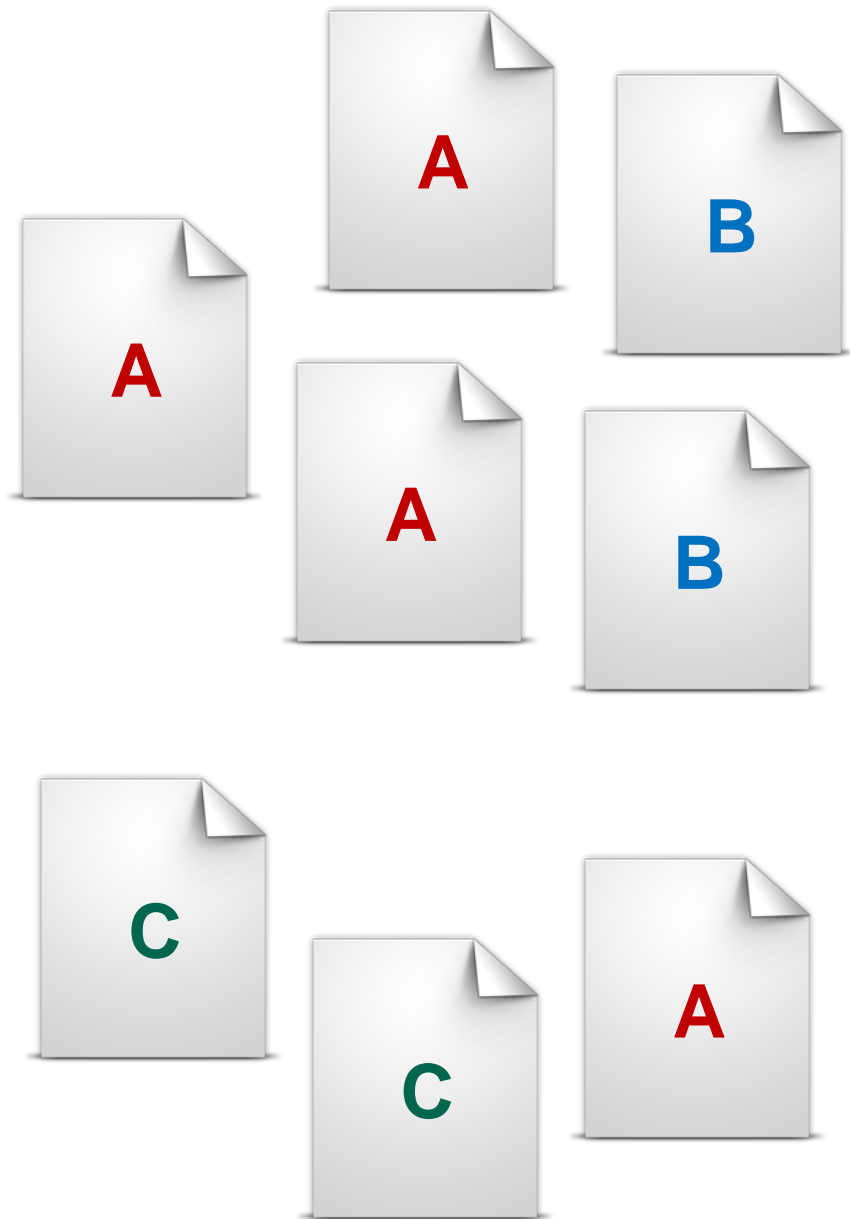
- The web is large and growing: anything that helps us understand high level structure is useful
- Tags encode semantically meaningful labels
- Tags cover much of the web's best content
- How can we use tags to provide high-level insight?

Web page clustering task



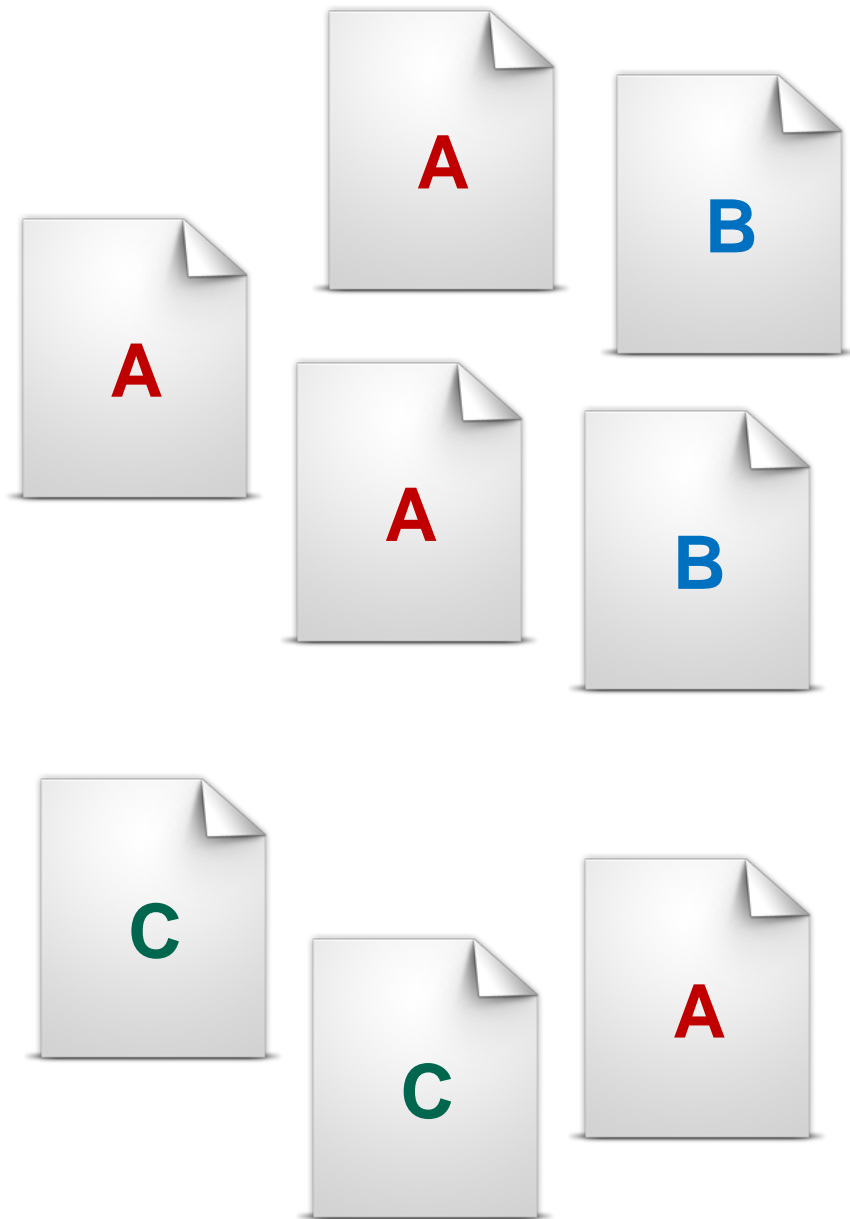
- Given a collection of web pages

Web page clustering task



- Given a collection of web pages
- Assign each page to a cluster, maximizing similarity within clusters

Web page clustering task



- Given a collection of web pages
- Assign each page to a cluster, maximizing similarity within clusters
- Applications: improved user interfaces, collection clustering, search result diversity, language-model based retrieval

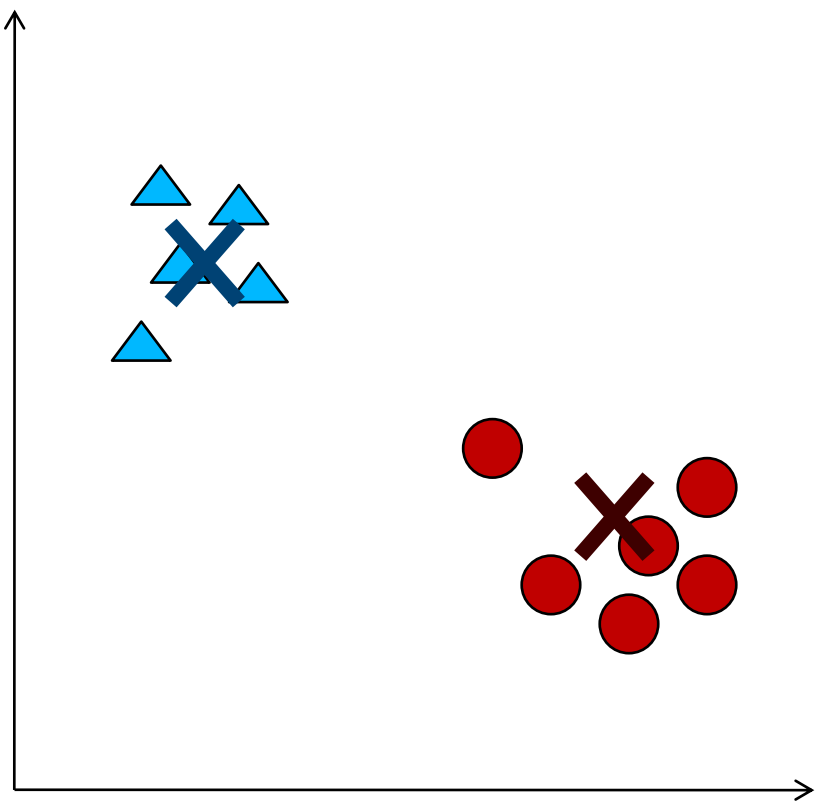
Structure of this talk

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means			
	Generative Model: MM-LDA			

Models: K-means and MM-LDA

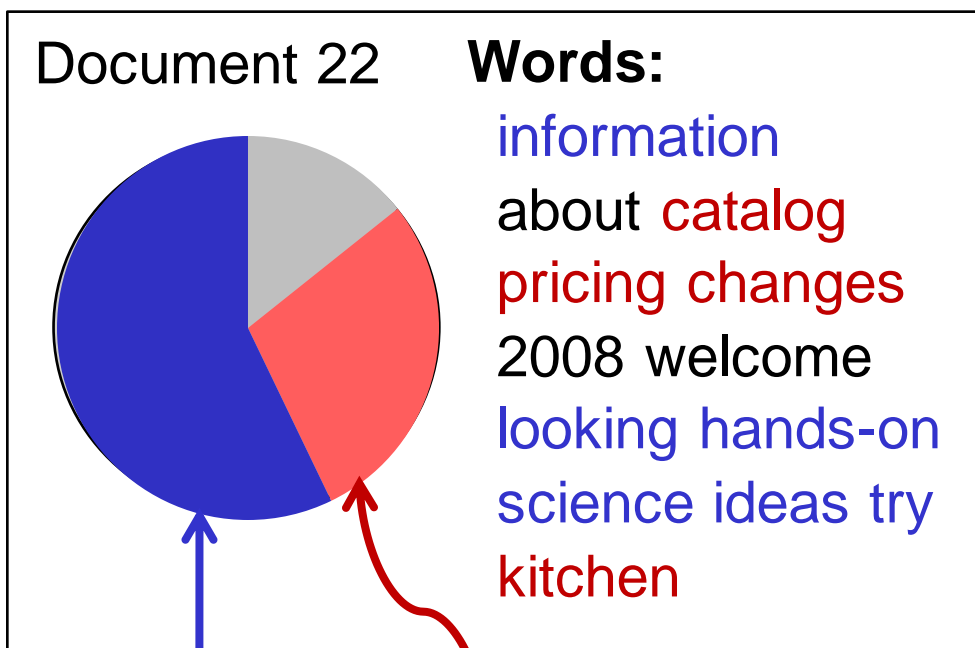
		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means			
	Generative Model: MM-LDA			

Model 1: K-means clustering



- K-means assumes the standard **Vector Space Model**: documents are Euclidean normalized real-valued vectors
- Algorithm: iteratively
 - Re-assign documents to closest cluster centroid
 - Update cluster centroids from document assignments

Model 2: Latent Dirichlet Allocation



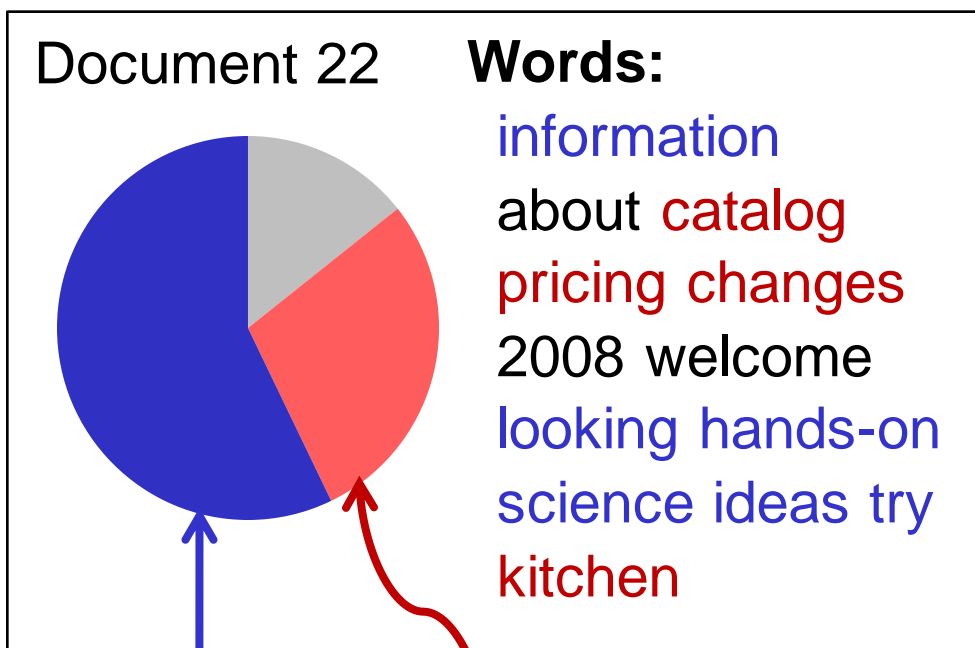
- LDA assumes each document's words generated by some topic's word distribution

Topic 5
science
experiment
learning
ideas
practice
information

Topic 12
catalog
shopping
buy
Internet
checkout
cart

...

Model 2: Latent Dirichlet Allocation



Topic 5
science
experiment
learning
ideas
practice
information

Topic 12
catalog
shopping
buy
Internet
checkout
cart

...

- LDA assumes each document's words generated by some topic's word distribution
- Paired with an inference mechanism (Gibbs sampling), learns per-document distributions over topics, per-topic distributions over words

Features: words, anchors, and tags

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means			
	Generative Model: MM-LDA			

Combining features

Feature Combination

Feature Space Size

Words

Words

Anchors

Anchors

Tags

Tags

Combining features

Feature Combination

Feature Space Size

Words

Words

Anchors

Anchors

Tags

Tags

Tags as Words

Tags as Words

Anchors as Words

Tags & Anchors as Words

Combining features

Feature Combination

Feature Space Size

Words

Words

Anchors

Anchors

Tags

Tags

Tags as Words

Tags as Words

Tags as New Words

Words

Tags

Words

Anchors

Words

Tags

Anchors

Combining features

Feature Combination

Feature Space Size

Words

Words

Anchors

Anchors

Tags

Tags

Tags as Words

Tags as Words

Tags as New Words

Words

Tags

Words

Anchors

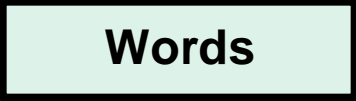
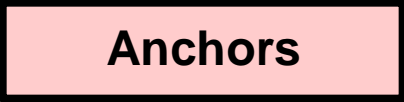



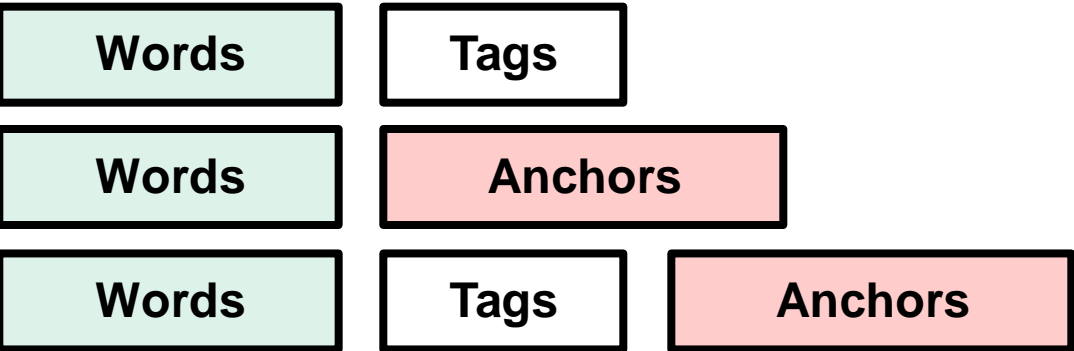
Words

Tags

Anchors

Simple feature space modifications for existing models

Combining features

Feature Combination	Feature Space Size
Words	
Anchors	
Tags	
Tags as Words	
Tags as New Words	
Words + Tags	

Combining features

Feature Combination	Feature Space Size
Words	Words
Anchors	Anchors
Tags	Tags
Tags as Words	Tags as Words
Tags as New Words	Words Tags
Words + Tags	K-means: normalize feature input vectors independently LDA : multiple parallel sets of observations via MM-LDA

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	1. Combining words and tags in the VSM		
	Generative Model: MM-LDA			

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	2. Comparing models, at multiple levels of specificity		
	Generative Model: MM-LDA			

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	3. Do words and tags complement or substitute for anchor text?		
	Generative Model: MM-LDA			

Experimental Setup

- Construct surrogate “gold standard” clustering using Open Directory Project
- Reflects a (problematic) consensus clustering, with known number of clusters

ODP Category	# Documents	Top Tags
Computers	5361	web css tools software programming
Health	434	parenting medicine healthcare medical
Reference	1325	education reference time research dictionary

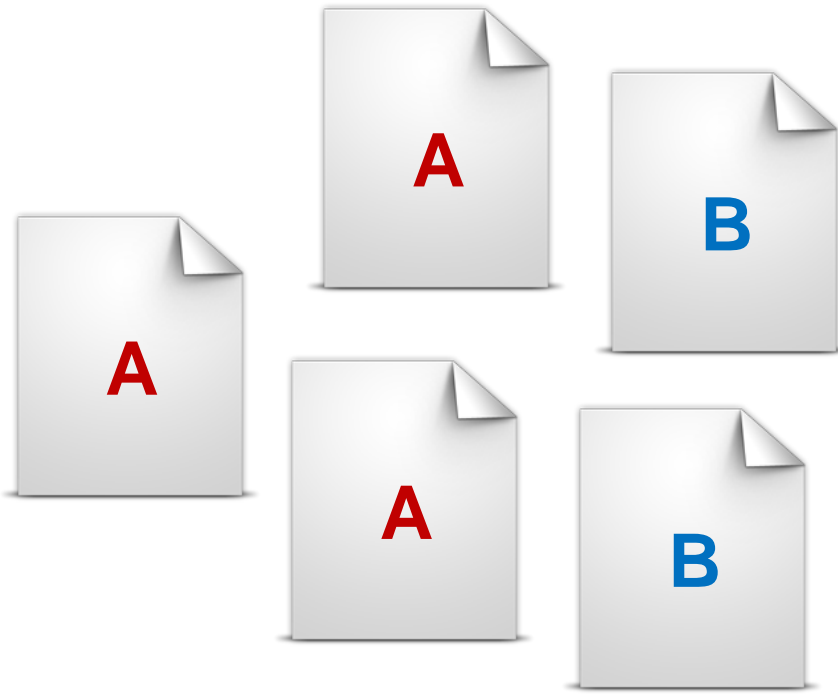
Experimental Setup

- Score predicted clusterings with ODP, but *not* trying to predict ODP
- Useful for relative system performance

ODP Category	# Documents	Top Tags
Computers	5361	web css tools software programming
Health	434	parenting medicine healthcare medical
Reference	1325	education reference time research dictionary

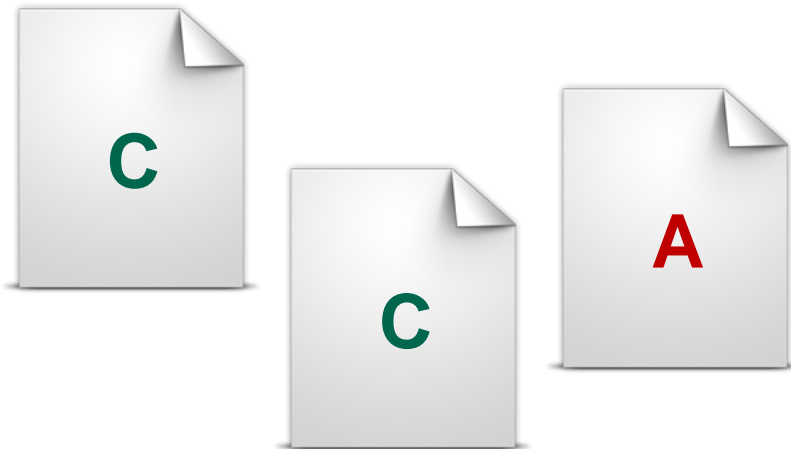
Evaluation: Cluster F1

Reference



Intuition: balance
pairwise precision
(place only similar documents together)
with

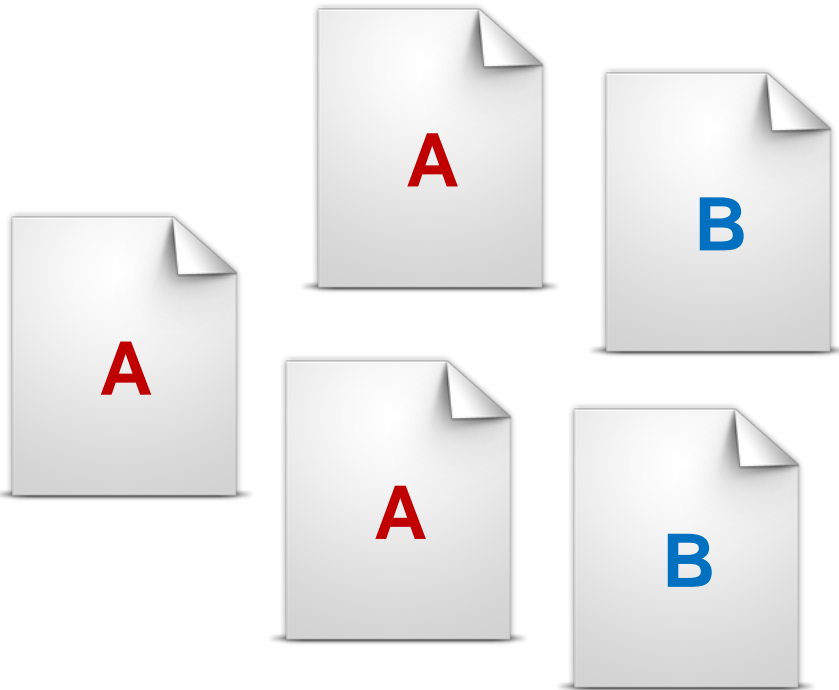
Health



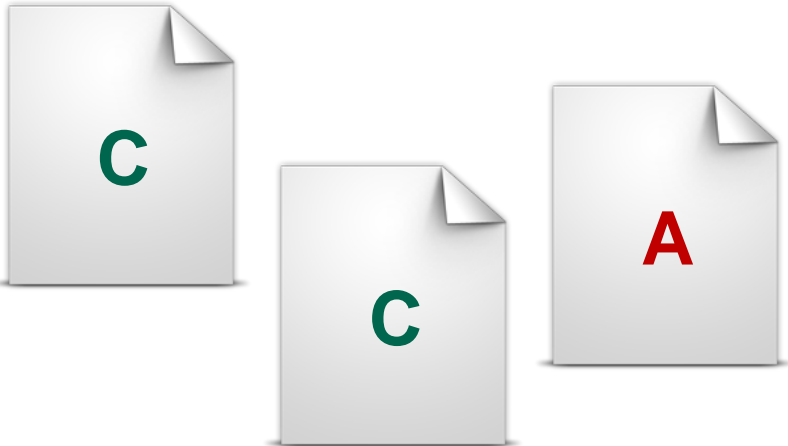
pairwise recall (keep
all similar documents
together)

Evaluation: Cluster F1

Reference

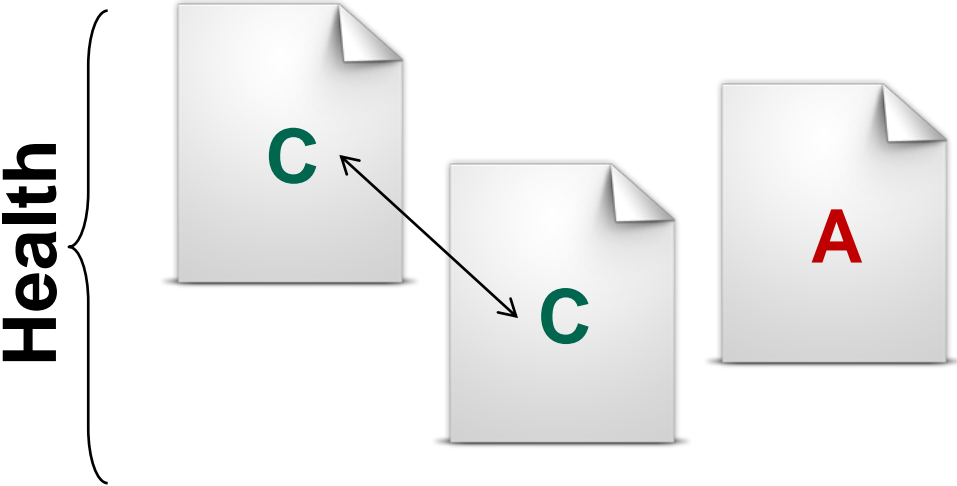
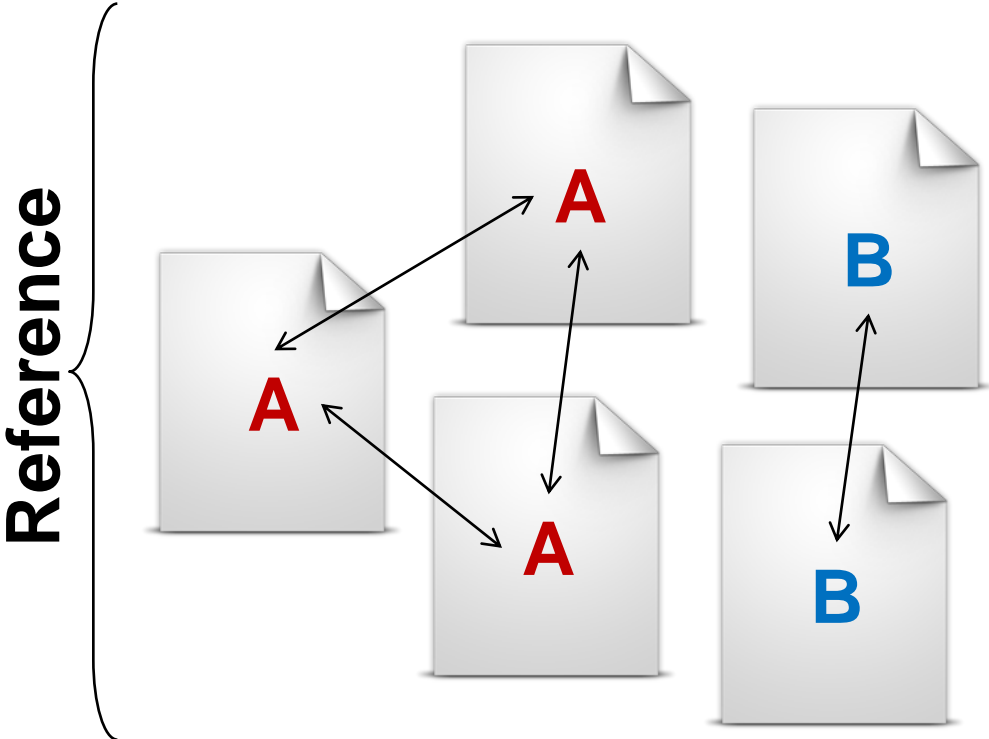


Health



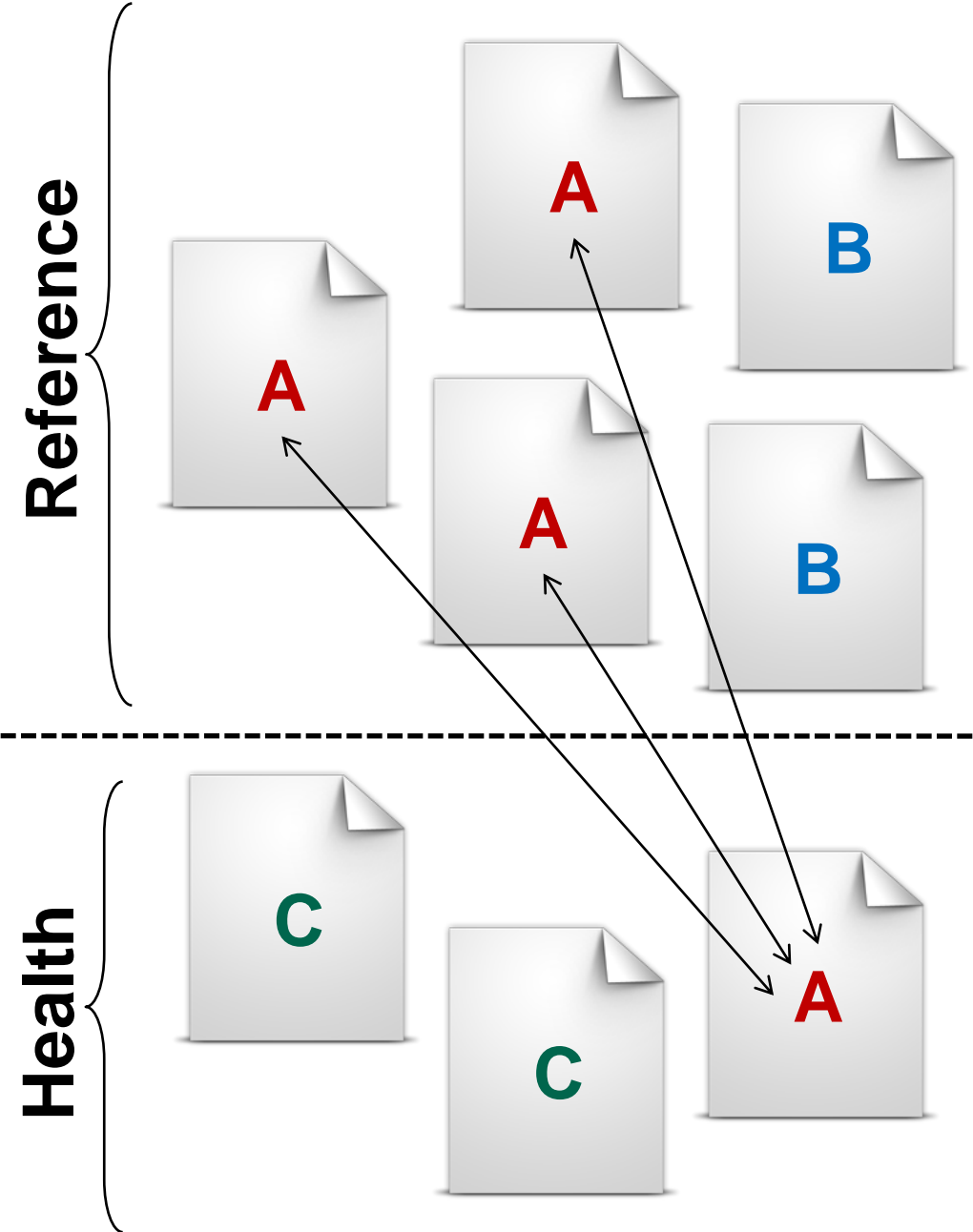
	Same Label	Different Label
Same Cluster		
Different Cluster		

Evaluation: Cluster F1



	Same Label	Different Label
Same Cluster	5	
Different Cluster		

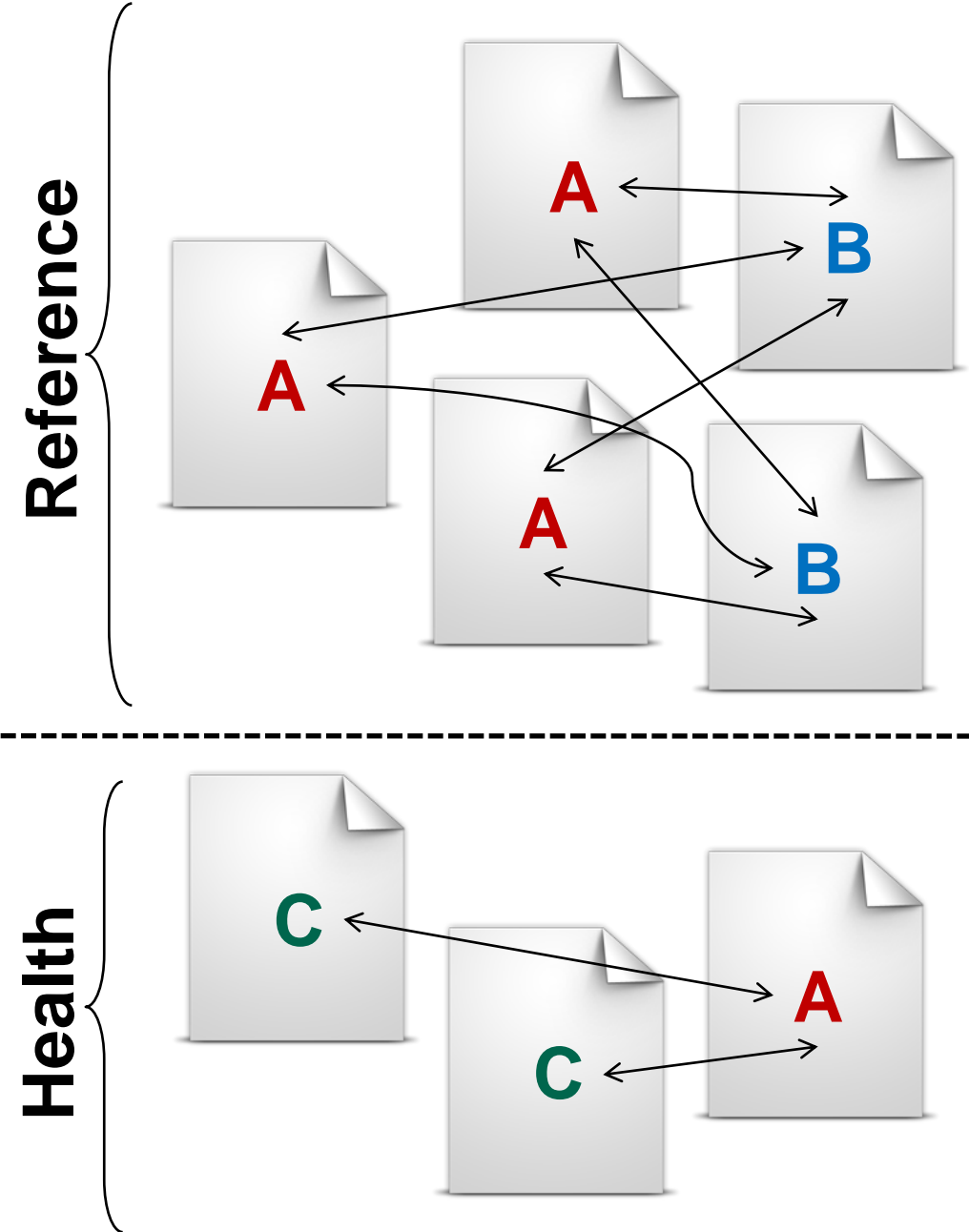
Evaluation: Cluster F1



	Same Label	Different Label
Same Cluster	5	3
Different Cluster		

Cluster Precision: 5/8

Evaluation: Cluster F1

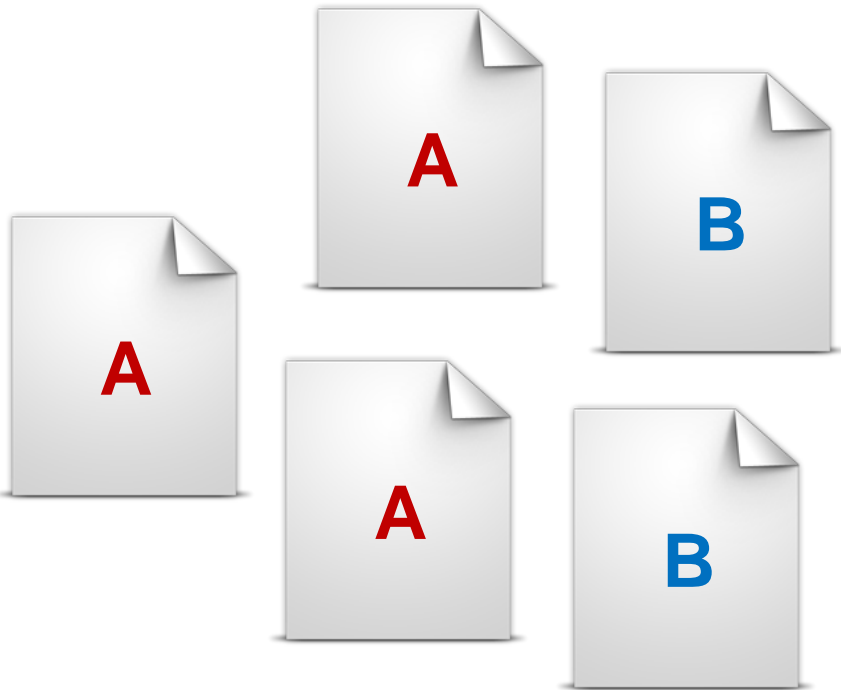


	Same Label	Different Label
Same Cluster	5	3
Different Cluster	8	

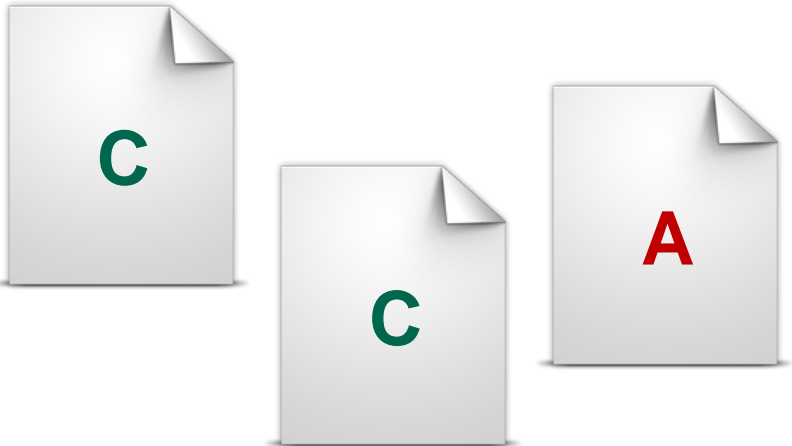
Cluster Precision: 5/8
Cluster Recall: 5/13

Evaluation: Cluster F1

Reference



Health



	Same Label	Different Label
Same Cluster	5	3
Different Cluster	8	

Cluster Precision: 5/8
Cluster Recall: 5/13
Cluster F1: **.476**

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	1. Combining words and tags in the VSM		
	Generative Model: MM-LDA			

Result: normalize words and tags independently in the Vector Space Model

Features			K-means
Words	Words		.139
Tags	Tags		.219
Words+Tags	Words	Tags	.225

Possible utility for other applications of the VSM

Result: normalize words and tags independently in the Vector Space Model

Features		K-means
Words	Words	.139
Tags	Tags	.219
Words+Tags	Words Tags	.225
Tags as Words (x1)	Tags as Words	.158
Tags as Words (x2)	Tags as Words	.176
Tags as New Words	Words Tags	.154

Possible utility for other applications of the VSM

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	2. Comparing models, at multiple levels of specificity		
	Generative Model: MM-LDA			

Result: MM-LDA outperforms K-means on top-level ODP categories

Features		K-means	(MM-)LDA
Words	Words	.139	.260
Tags	Tags	.219	.270
Words+Tags	Words	Tags	.307

Tagging at multiple basic levels

People use tags to help find the same page later, often at a “natural” level of specificity

Programming/**Languages** (1094 documents)

- Java PHP Python C++
JavaScript Perl Lisp
Ruby C

Society/**Social Sciences** (1590 documents)

- Issues, Religion &
Spirituality, People,
Politics, History, Law,
Philosophy

Tagging at multiple basic levels

People use tags to help find the same page later, often at a “natural” level of specificity

Programming/**Languages**
(1094 documents)

– Java PHP Python C++

Java Script Perl Lisp

Ruby C *java* applies to 73% of

Programming/Java pages

but *software* applies to only 21% of

Top/Computer pages

Society/**Social Sciences**
(1590 documents)

– Issues, Religion &

Spirituality, People,

Politics, History, Law,

Philosophy

Result: Sometimes, tags tell you more about cluster membership than words do

	Features	K-means	(MM-)LDA
Programming Languages	Words	.189	.288
	Tags	.567	.463
	Words+Tags	.556	.297
Social Sciences	Words	.196	.300
	Tags	.307	.310
	Words+Tags	.308	.302

- Tags are very discriminating in subcategories
- K-means wins when the feature space is cleaner

Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means	3. Do words and tags complement or substitute for anchor text?		
	Generative Model: MM-LDA			

Result: Tags complement anchor text

Features	K-means	(MM-)LDA
Words	.139	.260
Words+Anchors	.128	.248
Words+Anchors+Tags	.224	.306

Anchors can depress performance, but adding tags brings to within delta of Words+Tags.

Conclusions

- Tags add real value when high-level semantic information is needed
- Tags act differently than words, anchor text
- At the right level of specificity, tags describe pages better than anything else
- Treat tags and words as separate information channels to maximize utility

Thanks! Questions?

Backup material

Result: Tags complement anchor text

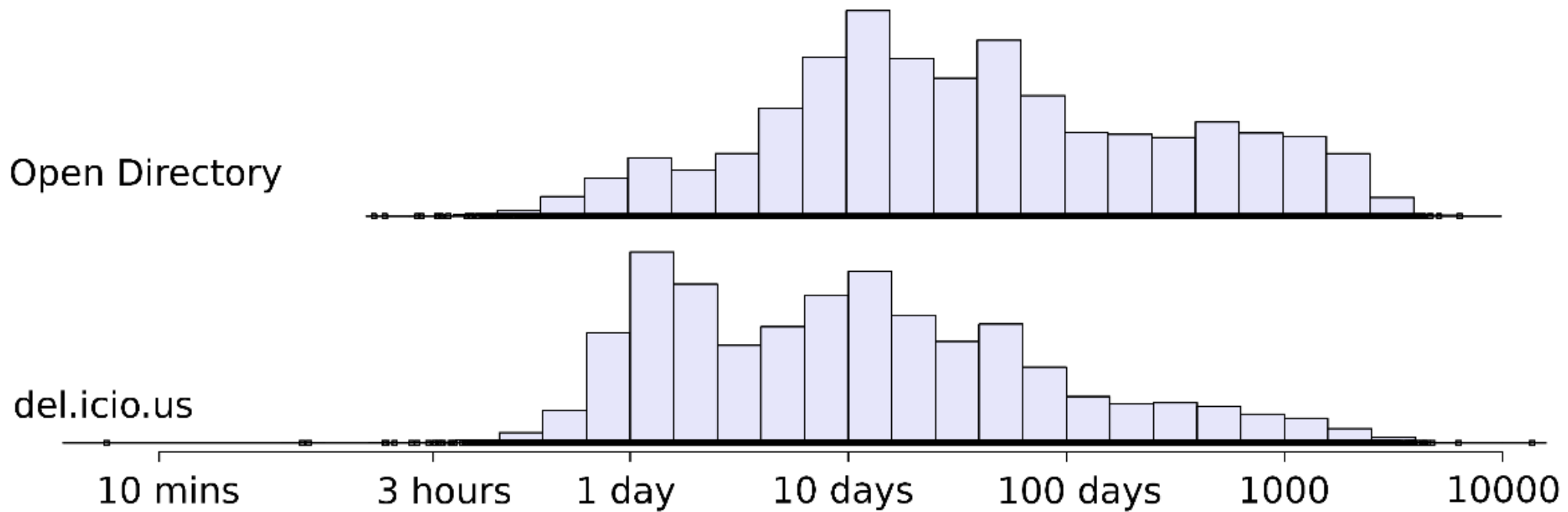
	(MM-)LDA	K-means
Words	.260	.139
Anchors as Words	.270	.120
(Anchors as Words)+Tags	.281	.214
Words+Anchors	.248	.128
Words+Anchors+Tags	.306	.224

Anchor text acts as annotations from another web author. Noisier than words and tags, but can be usefully integrated into a joint model.

Future directions

- More targeted graphical models
 - Individual users with individual vocabularies
 - Time series
- Direct evaluation in retrieval / browsing
- More types of annotated documents
 - Product reviews; academic papers; blog posts

Content age: ODP versus del.icio.us



57% of Tag Crawl data initially indexed by Google

Clustering (flat, parametric)

- **Input**

- Number of clusters K
- Set of documents: <words,tags,anchors>

- **Output**

- Assignment of documents to clusters

- **Evaluation**

- Comparison to a gold standard

Outline

- The tagged web
- Dataset and methodology
- Clustering with tags and words
 - K-Means in tag-augmented vector space
 - Multi-Multinomial LDA
- **Experiments**
- Discussion

Outline

- The tagged web
- Dataset and methodology
- Clustering with tags and words
 - K-Means in tag-augmented vector space
 - Multi-Multinomial LDA
- Experiments
- **Discussion**

Outline

- The tagged web
- **Dataset and methodology**
- Clustering with tags and words
 - K-Means in Tag-Augmented Vector Space
 - Multi-Multinomial LDA
- Experiments
- Discussion

Automatic cluster evaluation

- Pick a slice of ODP with k subtrees
- Cluster relevant documents into k sets
- Compare inferred assignments to ODP labeling

Automatic cluster evaluation

- Pick a slice of ODP with k subtrees
- Cluster relevant documents into k sets
- Compare inferred assignments to ODP labeling

Advantages

- Scalable, automatic, reflects “consensus” clustering

Drawbacks

- May not translate to performance gains in task
- Does not address choosing best k

F-measure of cluster quality

# Pairs of Examples	Same cluster	Different cluster
Same ODP Category	A	C
Different ODP Category	B	D

$$P = \frac{A}{A + B}$$

$$R = \frac{A}{A + C}$$

$$F_1 = \frac{2 P R}{P + R}$$

A tagged document

Tags

curriculum
education(2)
homeschool
imported
learning
science(4)
shopping
slinky
teachers
teaching
tools

HOME SCIENCE TOOLS
THE GATEWAY TO DISCOVERY

CART | MY ACCOUNT | WISH LIST

Information about catalog pricing changes in 2008. -800-860-6272

QUICK ORDER | FREE CATALOG | ORDER FORMS | TEACHING TIPS | SCIENCE PROJECTS | NEWSLETTERS

SEARCH

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

discover
how things work

Welcome! Looking for hands-on science ideas? Try these:

- [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.

Chemistry Supplies:
Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts, molecular models, digital balances, and more. Or make it

Dissection Supplies
Find everything you need to dissect frogs, cow eyes, and more.

[Read more](#)

Study Bacteria
Experiment with

HACKER SAFE
TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- [World of Germs](#)
\$19.95
- [Chemlab 1100 Chemistry Kit](#)
\$30.95
- [Kids LED Cordless Microscope](#)

ODP Label: Top/Reference

Top/Reference/Education/K_through_12/Home_Schooling/Curriculum/Science

MM-LDA implementation

- Collapsed Gibbs-sampler with hard assignments
 - Repeatedly samples new z for each word
 - Usually converges within several dozen passes
 - Could be parallelized
- Runtime:
 - 22 min (MM-LDA) versus 6 min (K-means) on 2000 documents

K-means generated clusters

tags

linux security php opensource vpn unix
games go game sports firefox gaming
music research finance audio mp3 lyrics
news business newspaper politics media magazine
politics activism travel movies law government

words

linux ircd php beware kernel exe
dmg munsey ballparks suppes racer game
music research redirect nottingham meta laboratory
v business leadership d news j
aquaculture terrapass geothermal anarchist wwoof cpssc

MM-LDA generated clusters

tags

web2.0 tools online editor photo office
guitar scanner chemistry military earthquake groupware
health medical medicine healthcare process gardening
bible christian space astronomy religion christianity
politics activism environment copyright law government

words

icons uml powerpoint lucid dreams dreaming
grub outlook bittorrent rendering recovery boot
exe health openpkg okino dll polytrans
gaelic bible nt bone scottish english
war shall power prisoners their article

K-means term weighting

	tf	tf-idf
Words	.131	.152
Tags	.201	.154
Words+Tags	.209	.168

$$\vec{d}_j^{(i)} = \begin{cases} f(w^{(i)}, W, j) & \text{if } 1 \leq j \leq |W| \\ f(t^{(i)}, T, j - |W|) & \text{otherwise} \end{cases}$$

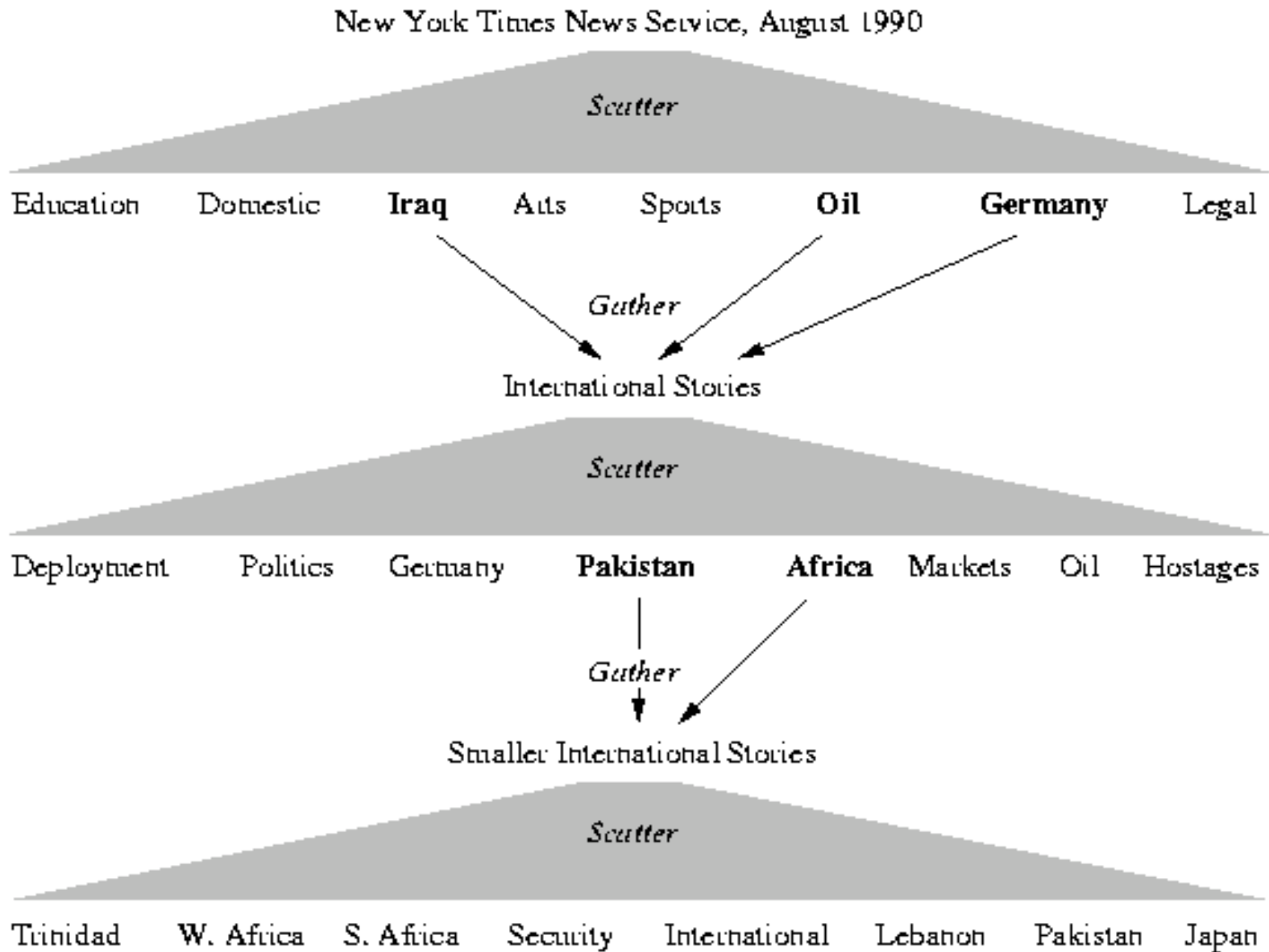
$$f_{tf}(\vec{w}^{(i)}, W_j) = \frac{1}{2N_w} \sum_{k=1}^N I[\vec{w}_k^{(i)} = W_j]$$

$$f_{tfidf}(\vec{w}^{(i)}, W_j) \propto \log f_{tf}(\cdot) \log \frac{D}{\sum_{l=1}^D I[W_j \in \vec{w}^{(l)}]}$$

Impact

- Social bookmarking is big and getting bigger
- Tags hold promise of specific, relevant indexing vocabulary for the web
 - Not quite full-text indexing
 - Not quite controlled pre-coordinate indexing
- Tagging data improves web clustering performance, which promises better IR
 - How else will tagging impact IR?

Scatter/Gather [Cutting et al 1992]



Stanford tag crawl dataset

Heymann, et. al 2008

Bookmarks/Posts

paul: news, uk → bbc.co.uk
08:33:25

mary: recipes, food → food.com
08:33:23

dave: tv, cnn, news → cnn.com
08:33:21

Triples

(paul, news, bbc.co.uk)
(paul, uk, bbc.co.uk)

(mary, recipes, food.com)
(mary, food, food.com)

(dave, tv, cnn.com)
(dave, cnn, cnn.com)
(dave, news, cnn.com)

Stanford tag crawl dataset

Back Link Text

... He is also a CNN Contributor, appearing on a variety of shows, including The Situation Room, Anderson Cooper 360, Lou Dobbs Tonight, and many others...



Page Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Forward Link Text

CNN.com is among the world's leaders in online news and information delivery. Staffed 24 hours, seven days a week by a dedicated staff in CNN's world headquarters in Atlanta, Georgia, ...



Tags

news cnn
daily media

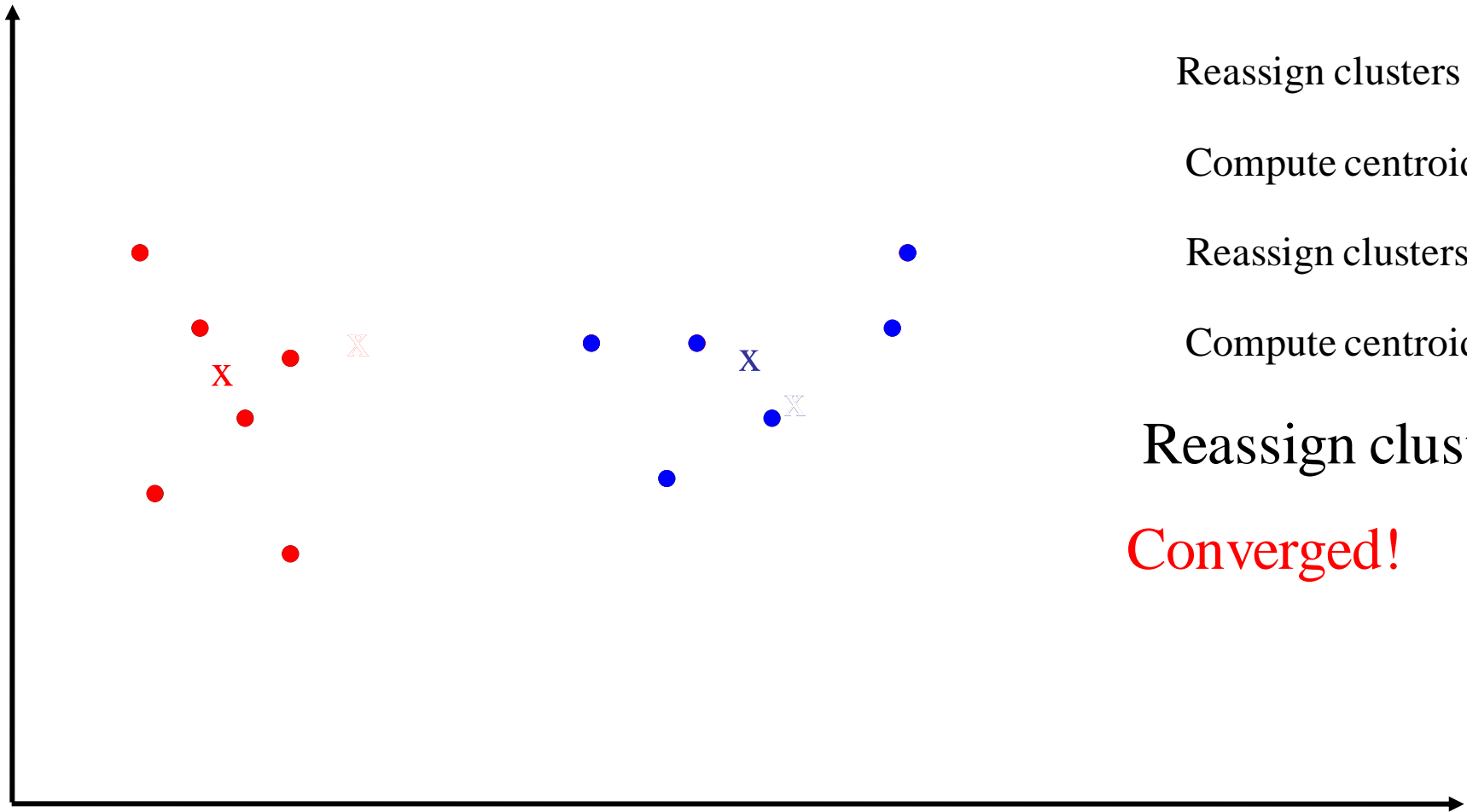
K-means [CS276]

- Assumes documents are real-valued vectors
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, c :

$$\mu_c = \frac{1}{|c|} \sum_{x \in c} x$$

- Reassignment of instances to clusters is based on distances to the current cluster centroids

K-means example ($K=2$) [CS276]



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

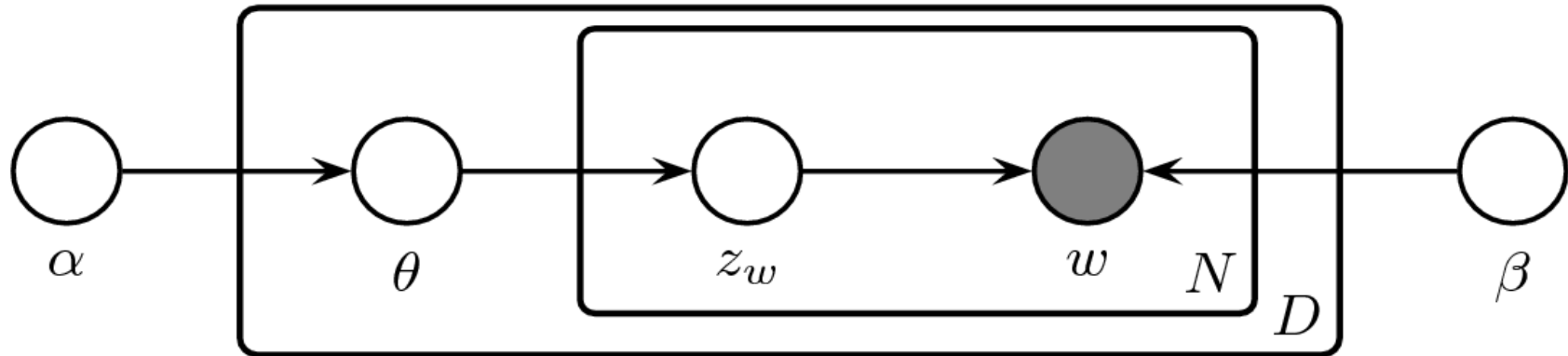
Converged!

MM-LDA outperforms K-means

On top-level ODP categories

	LDA	K-means
Words	0.260	.139
Tags	0.270	.219
Words+Tags	0.307	.225

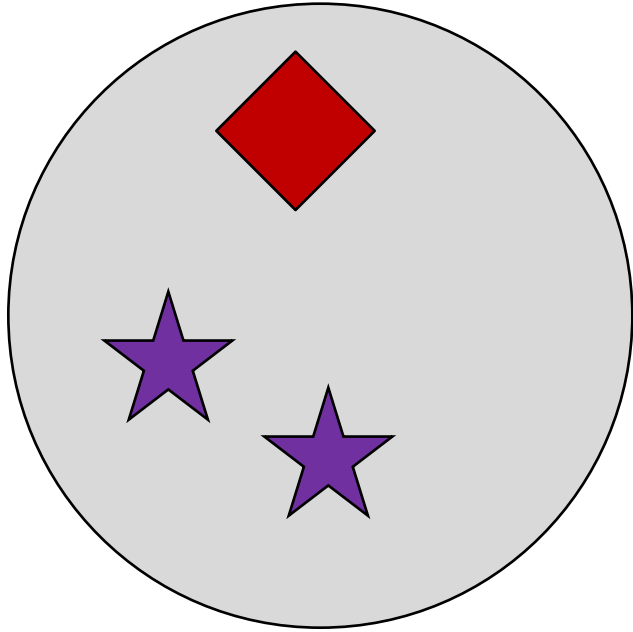
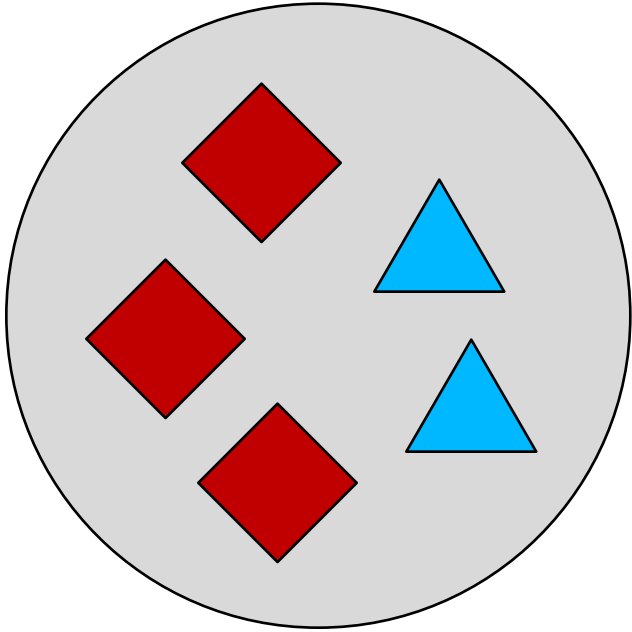
Latent Dirichlet Allocation (LDA)



- D – number of documents
- N – number of words in document
- $alpha$ – symmetric Dirichlet prior
- $theta$ – per document topic multinomial
- z_w – per word topic assignment
- w – word observation
- $beta$ – per topic word multinomial

MM-LDA Properties

- Natural extension of LDA
- Jointly models multiple types of observations
 - Similar to Blei et al.'s GM-LDA for images with captions
- Words and tags counted independently, contribute jointly to document topic model



A web document collection

Stanford Tag Crawl Dataset:

One month of del.icio.us posts in May/June 2007



Most web pages come with words

HOME SCIENCE TOOLS
THE GATEWAY TO DISCOVERY

CART | MY ACCOUNT | WISH LIST

Information about catalog pricing changes in 2008. -800-860-6272

QUICK ORDER | FREE CATALOG | ORDER FORMS | TEACHING TIPS | SCIENCE PROJECTS | NEWSLETTERS

SEARCH

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

discover
how things work

Welcome! Looking for hands-on science ideas? Try these:

- [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.

HACKER SAFE
TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- [World of Germs](#)
\$19.95
- [Chemlab 1100 Chemistry Kit](#)
\$30.95
- [Kids LED Cordless Microscope](#)

Chemistry Supplies:
Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts, molecular models, digital balances, and more. Or make it

Dissection Supplies
Find everything you need to dissect frogs, cow eyes, and more.
[Read more](#)

Study Bacteria
Experiment with

Words: welcome looking hands-on science ideas try kitchen projects dissolve eggshell grow crystals ...

Words can be used to cluster



Text surrounds links from other pages



The screenshot shows the Home Science Tools website. The logo at the top left reads "HOME SCIENCE TOOLS THE GATEWAY TO DISCOVERY". Navigation links include "QUICK ORDER", "FREE CATALOG", "ORDER FORMS", "TEACHING TIPS", "SCIENCE PROJECTS", and "NEWSLETTERS". A search bar is located in the top left. A sidebar on the left lists categories under "ONLINE STORE" and "MORE SHOPPING". The main content area features a "Welcome!" message, a "HACKER SAFE" badge, and several product recommendations with images and descriptions.

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

Discover how things work

Welcome! Looking for hands-on science ideas? Try these:

- [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.

HACKER SAFE
TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- [World of Germs](#)
\$19.95
- [Chemlab 1100 Chemistry Kit](#)
\$30.95
- [Kids LED Cordless Microscope](#)

Chemistry Supplies:
Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts, molecular models, digital balances, and more. Or make it

Dissection Supplies
Find everything you need to dissect frogs, cow eyes, and more.
[Read more](#)

Study Bacteria
Experiment with

Anchor Text:

tools home science links click buy supplies experiments ...

Words: welcome looking hands-on science ideas try kitchen projects dissolve eggshell grow crystals ...

Social bookmarking websites add tags



HOME SCIENCE TOOLS
THE GATEWAY TO DISCOVERY

Information about catalog pricing changes in 2008. -800-860-6272

QUICK ORDER | FREE CATALOG | ORDER FORMS | TEACHING TIPS | SCIENCE PROJECTS | NEWSLETTERS

SEARCH

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

discover
how things work

Welcome! Looking for hands-on science ideas? Try these:

- [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.

HACKER SAFE
TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- [World of Germs](#)
\$19.95
- [Chemlab 1100 Chemistry Kit](#)
\$30.95
- [Kids LED Cordless Microscope](#)

Chemistry Supplies:
Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts, molecular models, digital balances, and more. Or make it

Dissection Supplies
Find everything you need to dissect frogs, cow eyes, and more.
[Read more](#)

Study Bacteria
Experiment with

Tags:
curriculum
education
homeschool
imported
learning
science
shopping
slinky
teachers
teaching
tools

Anchor Text:

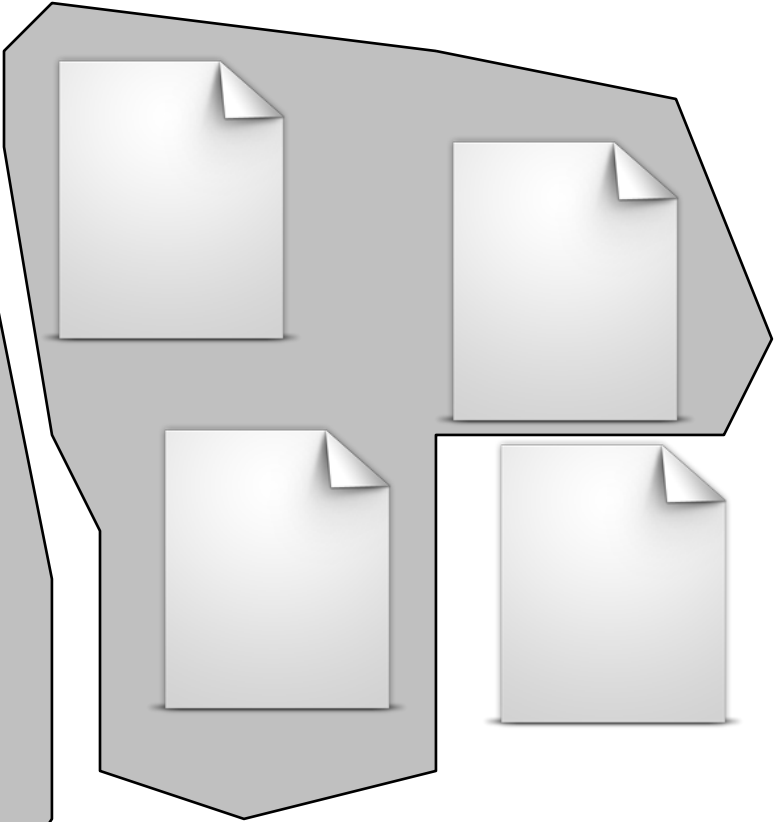
tools home science links click buy
supplies experiments ...

Words: welcome looking hands-on
science ideas try kitchen projects
dissolve eggshell grow crystals ...

How do we use words, anchor text, and tags together to most improve clustering?



The screenshot shows the Home Science Tools website. The header includes the logo "HOME SCIENCE TOOLS THE GATEWAY TO DISCOVERY" and navigation links like "CART | MY ACCOUNT | WISH LIST". A search bar is present. The main content area features a "discover how things work" banner with a child at a microscope. Below this are several product and category tiles: "Chemistry Supplies" (Make chemistry class memorable...), "Dissection Supplies" (Find everything you need to dissect...), "Study Bacteria" (Experiment with...), "Kitchen Science Projects" (Dissolve an eggshell...), "Science Fair Projects" (Ideas for a biology, chemistry, or physics project...), and "Make a Bouncy Ball" (Discover how polymers help a homemade ball bounce...). A "YOU MIGHT LIKE:" section lists items like "World of Germs" (\$19.95), "Chemlab 1100 Chemistry Kit" (\$30.95), and "Kids LED Cordless Microscope". The website is surrounded by a grey octagonal frame with three white document icons at the bottom.

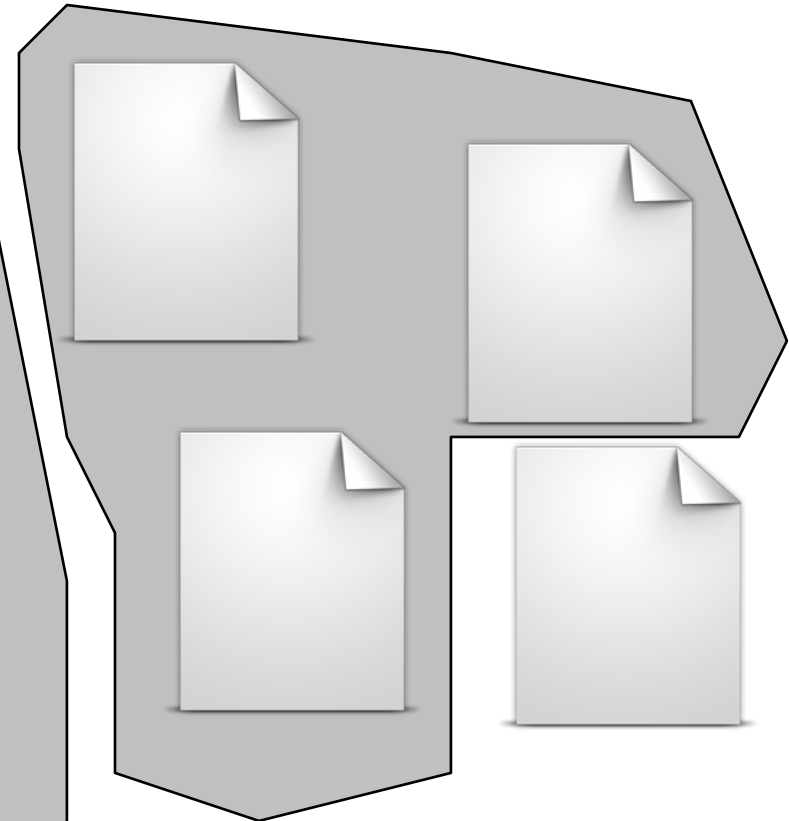


A grey octagonal frame containing five white document icons with folded corners, arranged in a pattern that suggests a collection of related documents or a cluster of information.

How do we test if clustering improves?



The screenshot shows the Home Science Tools website. The header includes the logo "HOME SCIENCE TOOLS THE GATEWAY TO DISCOVERY" and navigation links like "CART | MY ACCOUNT | WISH LIST". A search bar is present. The main content area features a "discover" banner with a child's photo, a "Welcome! Looking for hands-on science ideas? Try these:" section with three project ideas, and several product categories like "Chemistry Supplies", "Dissection Supplies", and "Study Bacteria". A "YOU MIGHT LIKE:" section lists items like "World of Germs" and "Chemlab 1100 Chemistry Kit". Five document icons are overlaid on the page: one at the top left, one at the top right, one at the bottom left, one at the bottom center, and one at the bottom right.



A large document icon with four smaller document icons overlaid on it. The smaller icons are positioned at the top left, top right, bottom left, and bottom right of the larger icon, representing a hierarchical or nested structure.

Many pages have a “gold standard” label

Reference/Education

HOME SCIENCE TOOLS
THE GATEWAY TO DISCOVERY

CART | MY ACCOUNT | WISH LIST

Information about catalog pricing changes in 2008. -800-860-6272

QUICK ORDER | FREE CATALOG | ORDER FORMS | TEACHING TIPS | SCIENCE PROJECTS | NEWSLETTERS

SEARCH

ONLINE STORE

- Microscopes & Accessories
- Life Science & Biology
- Earth & Space Science
- Chemistry
- Physical Science & Physics
- Technology
- General Science
- Science Books
- Science by Grade Level
- Science Curriculum
- Science Kits for Curriculum

MORE SHOPPING

- Nature Backpack Kits
- Science Kits
- Great Gift Ideas
- Bestsellers
- New Products
- Monthly Specials

discover
how things work

Welcome! Looking for hands-on science ideas? Try these:

- [Kitchen Science Projects](#) Dissolve an eggshell, grow your own crystals...
- [Science Fair Projects](#) Ideas for a biology, chemistry, or physics project.
- [Make a Bouncy Ball](#) Discover how polymers help a homemade ball bounce.

Chemistry Supplies:
Make chemistry class memorable with safe & fun experiments! Stock your lab with test tubes, beakers, elements charts, molecular models, digital balances, and more. Or make it

Dissection Supplies
Find everything you need to dissect frogs, cow eyes, and more.
[Read more](#)

Study Bacteria
Experiment with

HACKER SAFE
TESTED DAILY 20-FEB

YOU MIGHT LIKE:

- [World of Germs](#)
\$19.95
- [Chemlab 1100 Chemistry Kit](#)
\$30.95
- [Kids LED Cordless Microscope](#)

Tags:
curriculum
education
homeschool
imported
learning
science
shopping
slinky
teachers
teaching
tools

Anchor Text:

tools home science links click buy
supplies experiments ...

Words:

welcome looking hands-on
science ideas try kitchen projects
dissolve eggshell grow crystals ...

Open Directory Project

 open directory project In partnership with
AOL search

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

 [advanced](#)

Arts

[Movies](#), [Television](#), [Music](#)...

Games

[Video Games](#), [RPGs](#), [Gambling](#)...

Kids and Teens

[Arts](#), [School Time](#), [Teen Life](#)...

Reference

[Maps](#), [Education](#), [Libraries](#)...

Shopping

[Clothing](#), [Food](#), [Gifts](#)...

World

[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#), [Русский](#), [Svenska](#)...

Business

[Jobs](#), [Real Estate](#), [Investing](#)...

Health

[Fitness](#), [Medicine](#), [Alternative](#)...

News

[Media](#), [Newspapers](#), [Weather](#)...

Regional

[US](#), [Canada](#), [UK](#), [Europe](#)...

Society

[People](#), [Religion](#), [Issues](#)...

Computers

[Internet](#), [Software](#), [Hardware](#)...

Home

[Family](#), [Consumers](#), [Cooking](#)...

Recreation

[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...

Science

[Biology](#), [Psychology](#), [Physics](#)...

Sports

[Baseball](#), [Soccer](#), [Basketball](#)...

[Become an Editor](#) Help build the largest human-edited directory of the web



Copyright © 1998-2008 Netscape

4,583,933 sites - 78,876 editors - over 590,000 categories

Cluster evaluation

Reference



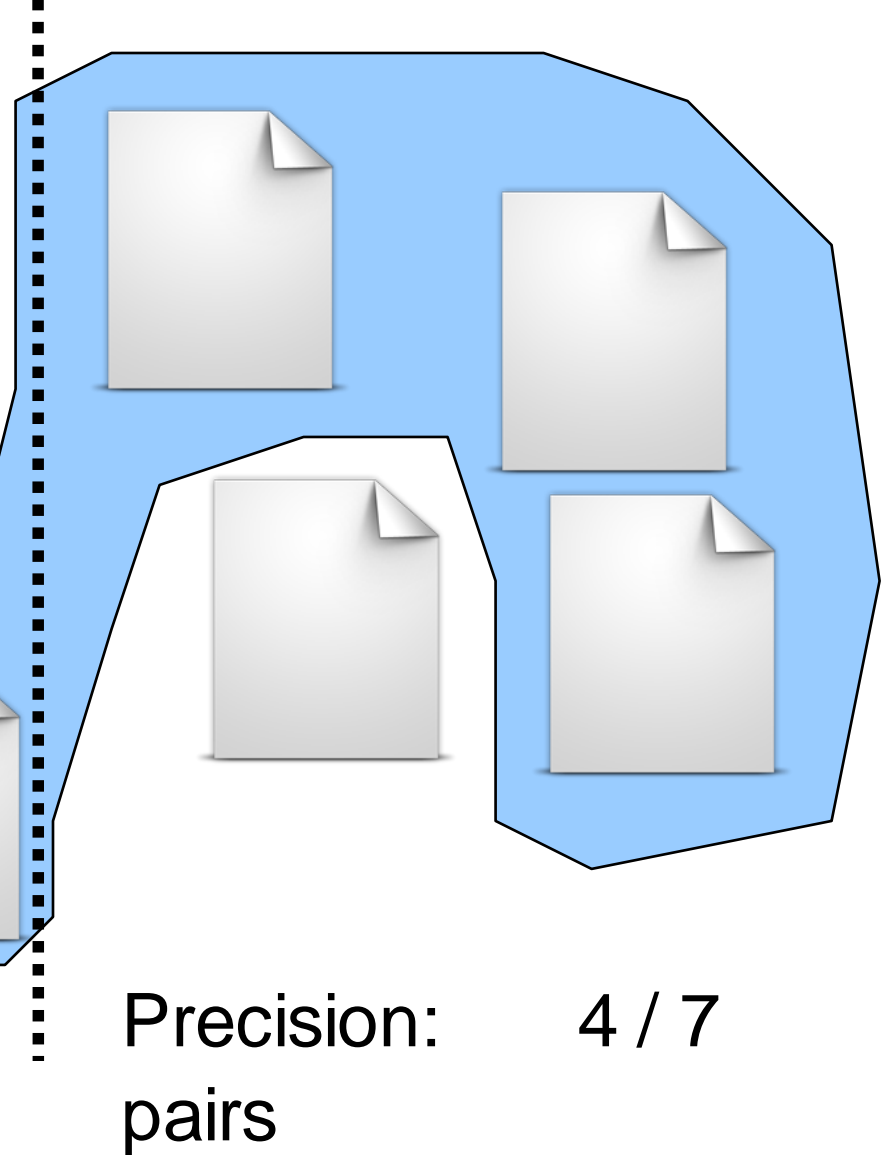
Arts



Cluster evaluation

Reference

Arts



Precision: 4 / 7

pairs

Recall: 4 / 12

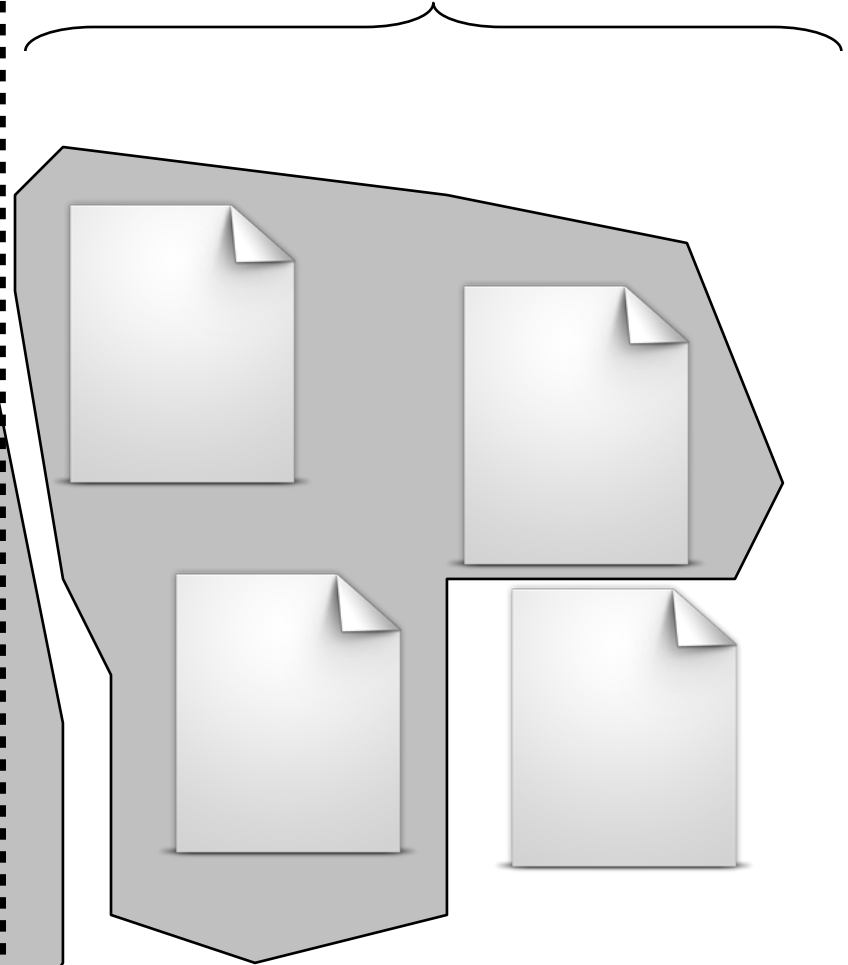
Cluster evaluation

Reference

Arts



The Reference cluster contains a screenshot of the Home Science Tools website. The website header includes the logo "HOME SCIENCE TOOLS THE GATEWAY TO DISCOVERY" and navigation links for "QUICK ORDER", "FREE CATALOG", "ORDER FORMS", "TEACHING TIPS", "SCIENCE PROJECTS", and "NEWSLETTERS". The main content area features a "discover how things work" banner with a child at a microscope, a "Welcome! Looking for hands-on science ideas? Try these:" section with three project ideas, a "HACKER SAFE" badge, and a "YOU MIGHT LIKE:" section with product recommendations like "World of Germs" and "Chemlab 1100 Chemistry Kit". Below the website screenshot are three document icons representing related content.



The Arts cluster contains five document icons representing related content.

Precision:

9 / 9 pairs

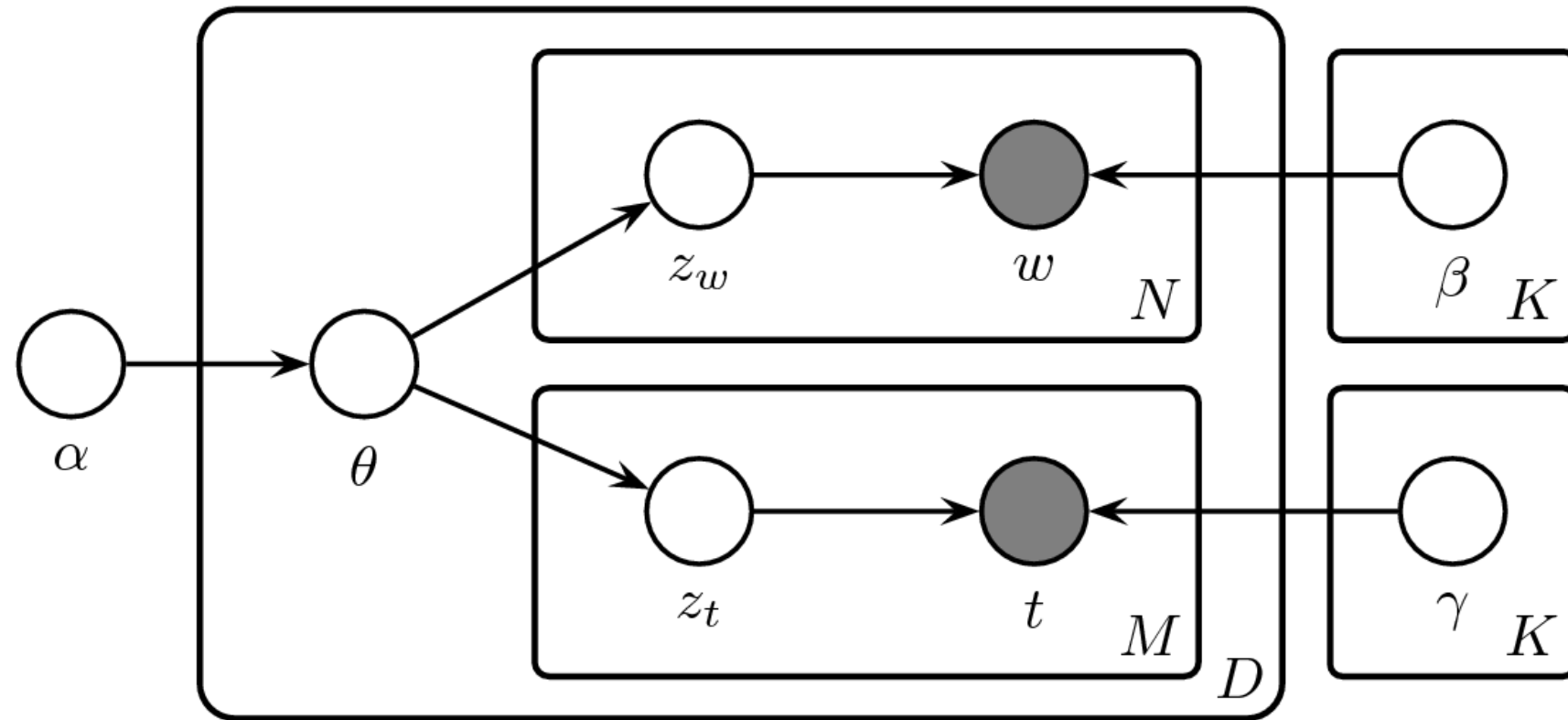
Recall:

9 / 12 pairs

Outline

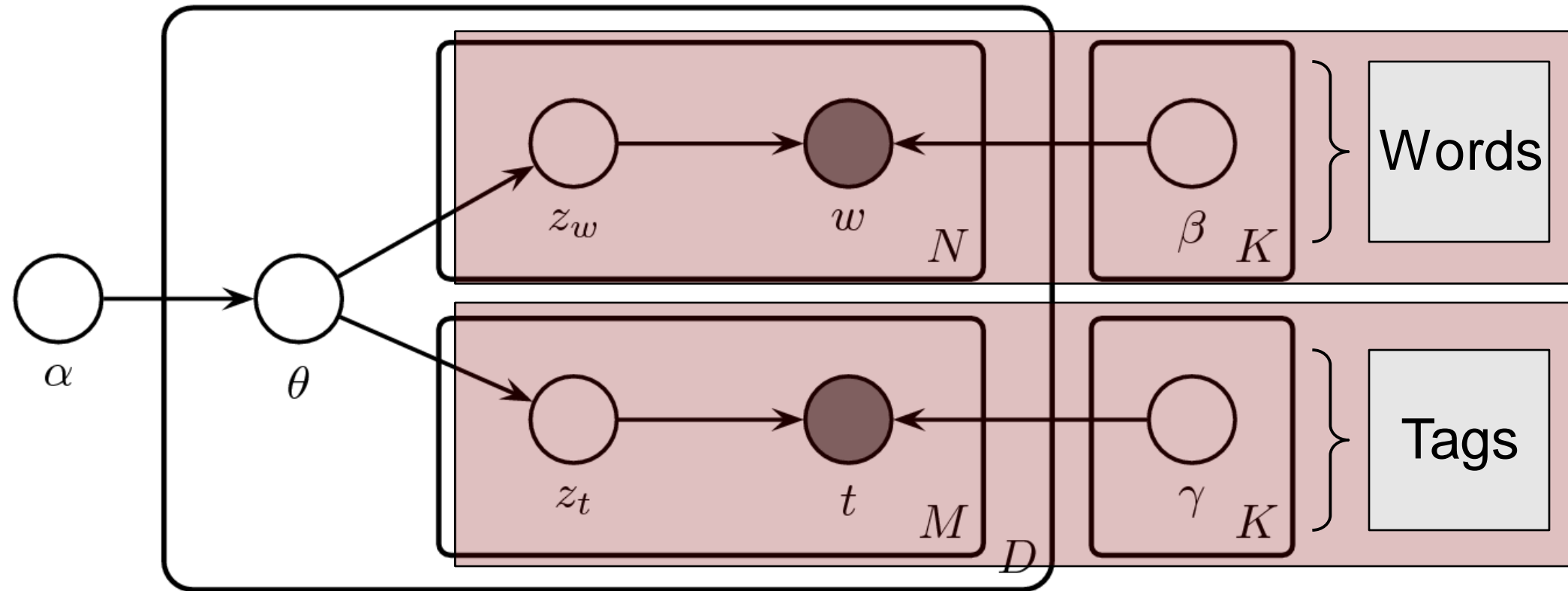
- × The tagged web
- × Dataset and methodology
- **Algorithms for clustering with tags and words**
 - K-Means in tag-augmented vector space
 - Multi-Multinomial LDA
- **Results**
 - Tag and word normalization
 - Clustering at varying levels of specificity
 - Incorporating anchor text

Multi-Multinomial LDA (MM-LDA)



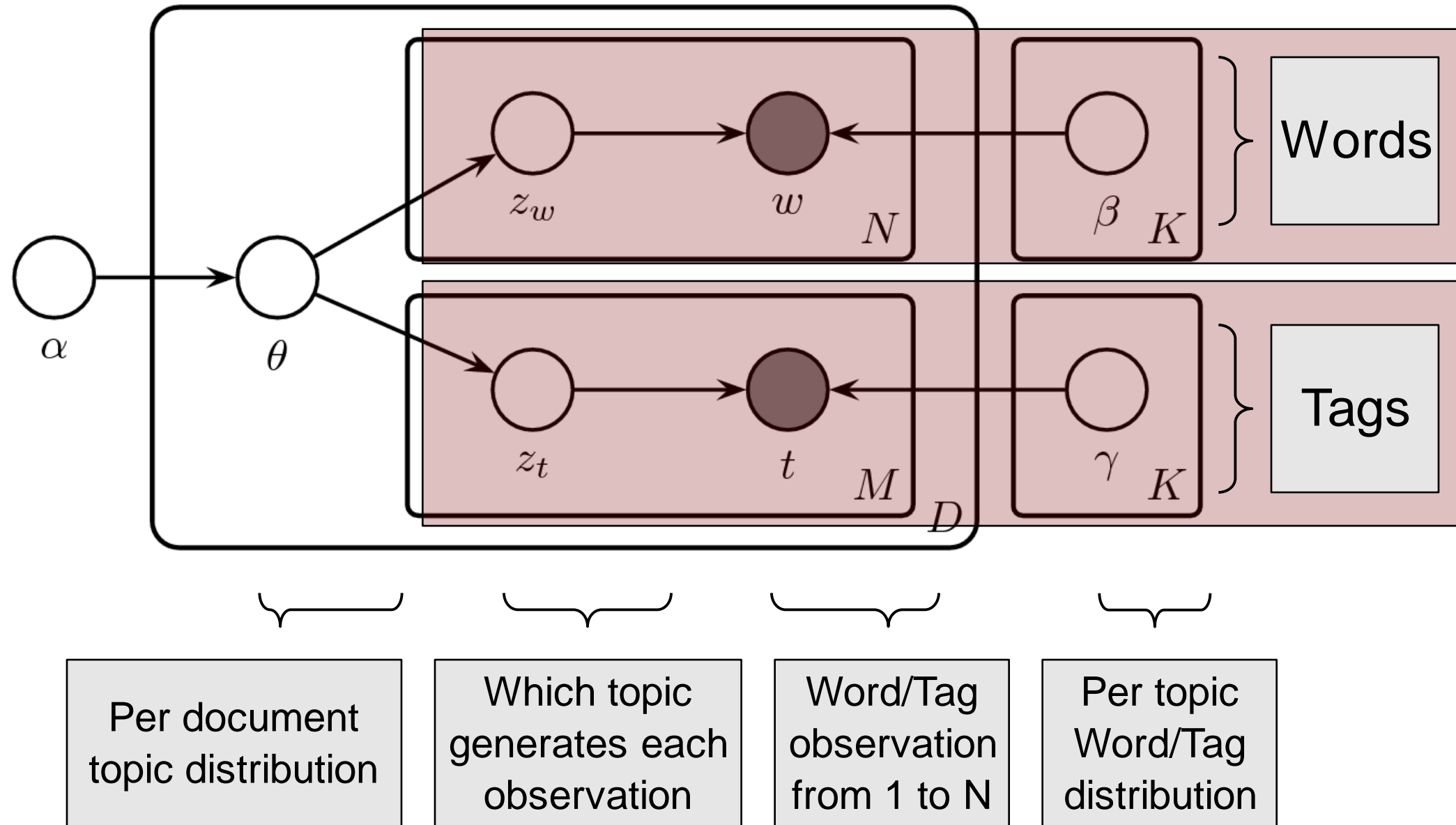
Extends *Latent Dirichlet Allocation*: Words and tags (and anchors, etc.) are counted independently, contribute jointly to topic probabilities

Multi-Multinomial LDA (MM-LDA)



Extends *Latent Dirichlet Allocation*: Words and tags (and anchors, etc.) are counted independently, contribute jointly to topic probabilities

Multi-Multinomial LDA (MM-LDA)

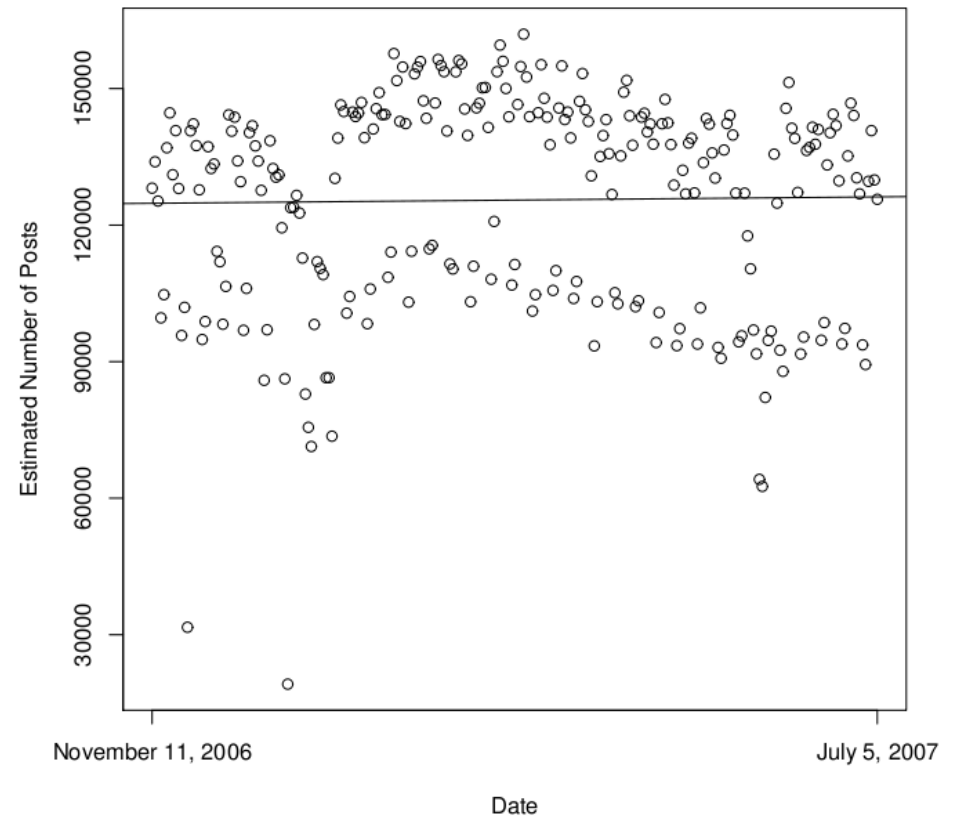
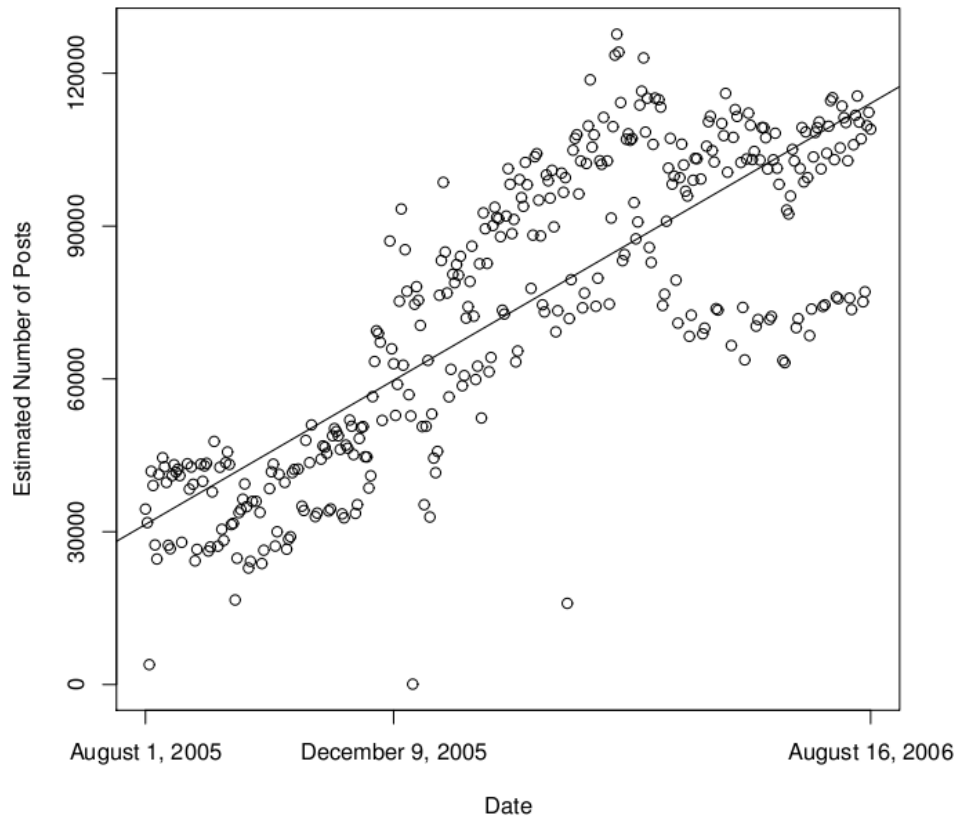


The tagged web (Heymann, et al., WSDM 2008)

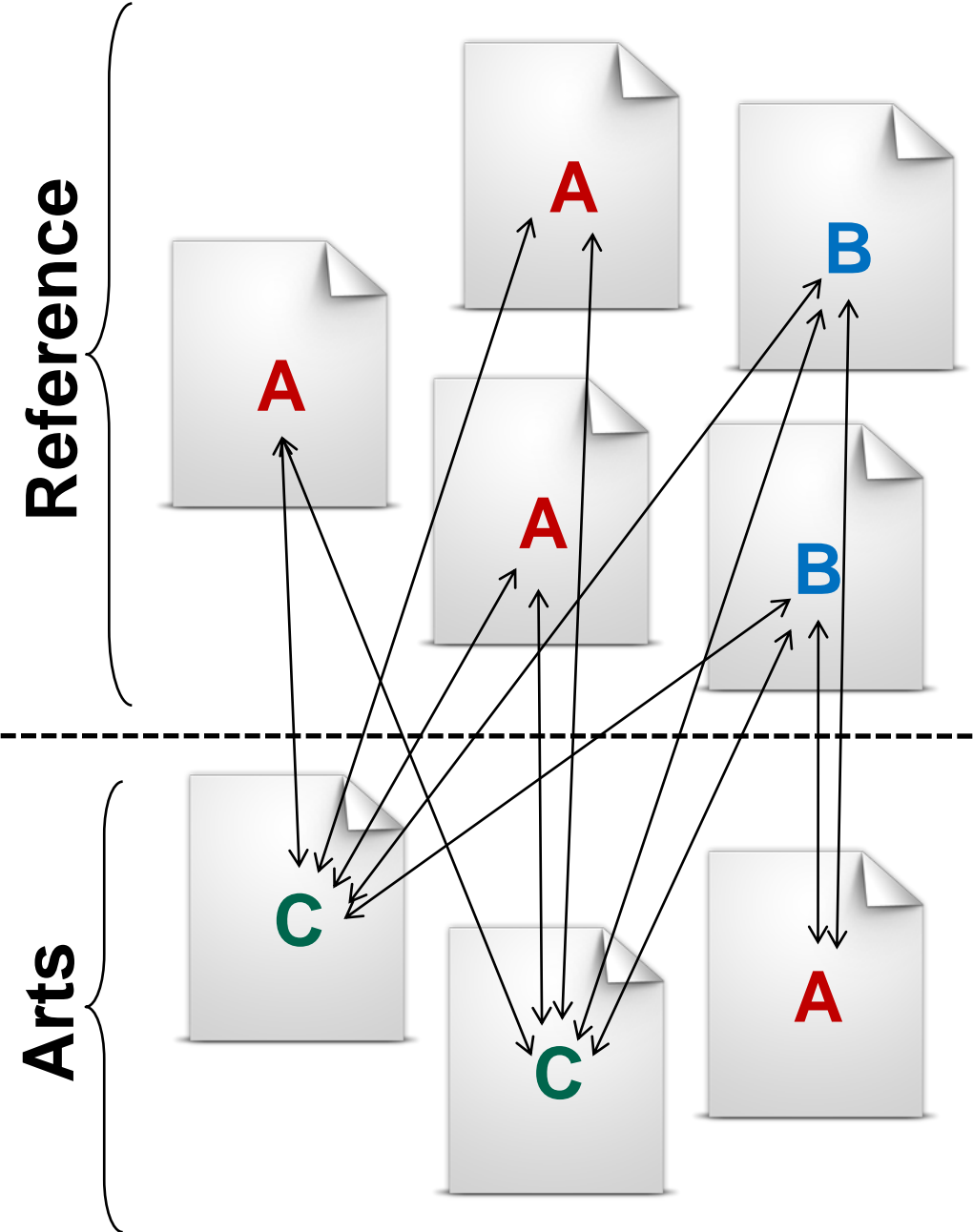
≈ 120 thousand ($\approx 10^5$) posts/day
(versus $\approx 10^6$ blog posts/day)

60–150 million posts

12–75 million ($\approx 10^7$ – 10^8) unique URLs
(versus $\approx 10^9$ – 10^{11} total URLs)



Evaluation: Cluster F1



	Same Label	Different Label
Same Cluster	5	3
Different Cluster	8	12

Cluster Precision: 5/8

Cluster Recall: 5/13

Goal: clustering for information retrieval

- Better user interfaces
 - e.g. Clusty, Vivisimo, Scatter/Gather, and friends
- Collection clustering
 - e.g. Columbia Newsblaster, Google News
- Improved language models for better retrieval
 - e.g. Liu and Croft 2004; Wei and Croft 2006
- Better cluster based-retrieval
 - e.g. Salton 1971

Stanford tag crawl / ODP intersection

ODP Name	#Docs	Top Tags by PMI
Adult	36	blog illustration art erotica sex
Arts	1446	lost recipes knitting music art
Business	908	accounting business lockpicking agency
Computers	5361	web css tools software programming
Games	291	un rpg fallout game games
Health	434	parenting medicine healthcare medical
Home	654	recipes blog cooking coffee food
Kids	669	illusions anatomy kids illusion copyright
News	373	system-unfiled daily cnn media news
Recreation	411	humor vacation hotels reviews travel
Reference	1325	education reference time research dictionary
Science	1574	space dreams psychology astronomy science
Shopping	310	custom ecommerce shop t-shirts shopping
Society	1852	buddhism christian politics religion bible
Sports	146	sport cycling nfl football sports
World	756	speed bandwidth google speedtest maps

K-means feature vectors

Strategy

Feature Space Size

Words



Tags



Tags as Weighted Words



Tags as New Words



Tags+Words



Experiments

		Features		
		Words	Tags	Anchors
Models	Vector Space Model: K-means			
	Generative Model: MM-LDA	2. Extending LDA for multiple feature types		

Result: Sometimes, tags tell you more about cluster membership than words do

	Features	K-means	(MM-)LDA
All	Words	.139	.260
	Tags	.219	.270
	Words+Tags	.225	.307
Programming Languages	Words	.189	.288
	Tags	.567	.463
	Words+Tags	.556	.297
• Social Sciences	Words	.196	.300
	Tags	.307	.310
	Words+Tags	.308	.302

but

“software” applies to only 21% of Computer pages

- K-means wins when the feature space is cleaner