Seminar: “Social Networks”

Clustering the tagged web

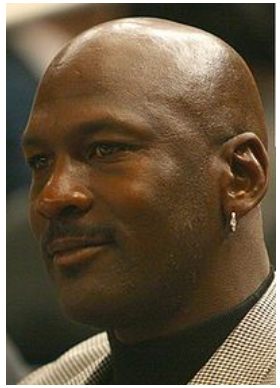
Daniel Ramage
Paul Heymann
Christopher D. Manning
Hector Garcia-Molina

Thomas von Bomhard

Problem: Ambiguity of user queries

- "Barcelona" (City? Football team? Movie?)
- "Michael Jordan"

Michael *I.* Jordan



Michael *J.* Jordan

Google shows only one Michael Jordan

Google michael jordan Search Adv

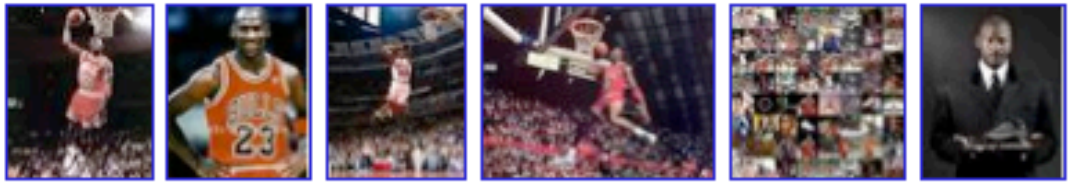
Web Show options... R

Michael Jordan - Wikipedia, the free encyclopedia
Michael Jeffrey **Jordan** (born February 17, 1963) is a retired American professional basketball player and active businessman. His biography on the National ...
Early years - Professional sports career - Olympic career
en.wikipedia.org/wiki/**Michael_Jordan** - Cached - Similar

NBA.com: **Michael Jordan** Bio
Michael Jordan | 23. Season statistics & Notes · Season splits · Game-by-game stats · Bio · Printable player file. 2002-03. Statistics. PPG, 20.0. RPG, 6.10 ...
www.nba.com/playerfile/**michael_jordan**.html - Cached - Similar

NBA.com: **Michael Jordan** Summary
Michael Jordan By acclamation, **Michael Jordan** is the greatest basketball player of all time. Although, a summary of his basketball career and influence on ...
www.nba.com/history/players/**jordan**_summary.html - Cached - Similar

Image results for **michael jordan** - Report images



Video results for **michael jordan**

Better: More diversity in search results

Google michael jordan Search Advanc

Web Show options... Res

Results include your SearchWiki notes for **michael jordan**. Share these notes

Michael Jordan - Wikipedia, the free encyclopedia
Michael Jeffrey **Jordan** (born February 17, 1963) is a retired American professional basketball player and active businessman. His biography on the National ...
Early years - Professional sports career - Olympic career
en.wikipedia.org/wiki/**Michael_Jordan** - Cached - Similar -
30 5 - Picked by 29 other people.

Michael I. Jordan's Home Page
18 Aug 2004 ... Graphical models, variational methods, machine learning, reasoning under uncertainty.
www.eecs.berkeley.edu/~**jordan**/ - Cached -
4 0 - Picked by 3 other people.

NBA.com: **Michael Jordan** Bio
Michael Jordan | 23. Season statistics & Notes · Season splits · Game-by-game stats · Bio · Printable player file. 2002-03. Statistics. PPG, 20.0. RPG, 6.10 ...
www.nba.com/playerfile/**michael_jordan**.html - Cached - Similar -

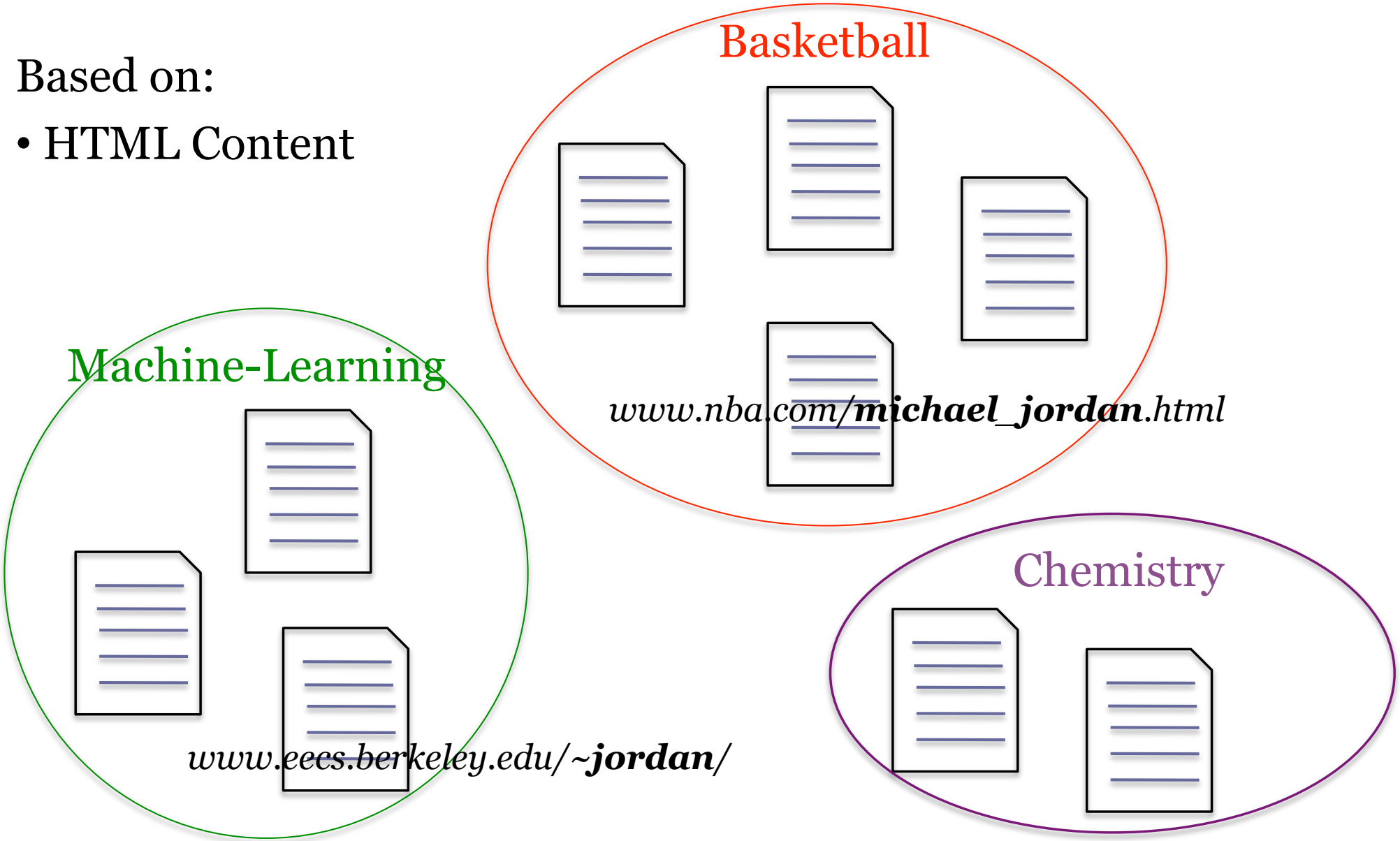
NBA.com: **Michael Jordan** Summary
Michael Jordan By acclamation, **Michael Jordan** is the greatest basketball player of all time. Although, a summary of his basketball career and influence on ...
www.nba.com/history/players/**jordan**_summary.html - Cached - Similar -

Image results for **michael jordan** - Report images

Clustering the Web

Based on:

- HTML Content



Clustering the tagged Web

Based on:

- HTML Content
- User-generated tags from a social bookmarking website like delicious.com

Everyone's Bookmarks for:

Michael I. Jordan's Home Page

www.cs.berkeley.edu/~jordan/

History | Notes

Saved 73 times, first saved by Markus Fix on 13 May 04. **View Chart**

24 DEC 09	replore	ml, researcher
09 NOV 09	akastrin	people
16 OCT 09	angie.h.hsieh	graphical+models, machine+learning, statistics, berkeley, research
14 OCT 09	caiyizhi	people, research, berkeley, machine_learning
01 OCT 09	Dlador	research, AI, vision, graph, learning, people, machine_learning, machinelearning, machine-learning, statistics, ucberkeley, berkeley, researcher
09 SEP 09	大牛 Guibin	people, research, berkeley, machinelearning, ai, ucberkeley, statistics
21 AUG 09	zybler	ai, homepage
08 JUN 09	i_stevenson	people, MachineLearning, bayesian, vision, motorcontrol, u_berkeley, gen1
06 MAY 09	menocinque	[UniEd]-MLPR
05 MAY 09	The statistician. Roger Bilisoly	web_pages
	thuhien7110	berkeley, jordan, machine.learning, researcher, machine

Save this bookmark
Look up another URL

Tags

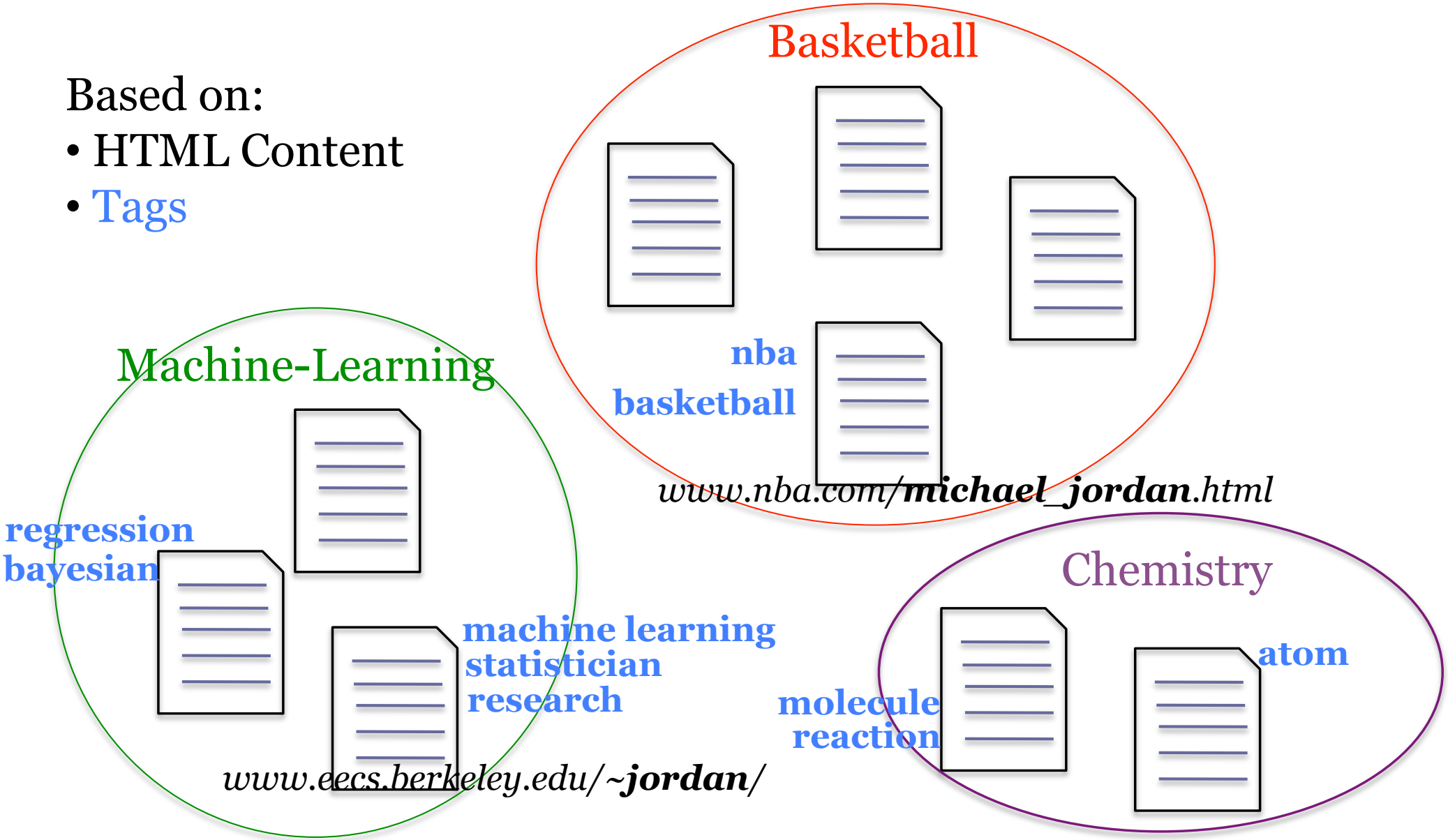
Top Tags

people	25
research	18
statistics	15
berkeley	14
machinelearning	10
machine-learning	10
machine_learning	9
ai	9
jordan	8
researcher	6
ucberkeley	5
learning	5
machine.learning	4
machine	4
vision	3
michael	3
graph	3
homepage	3
bayesian	2
researchers	2
academic	2
bayes	1

Clustering the tagged Web

Based on:

- HTML Content
- Tags



Questions

- Does tagging data improve the performance of clustering methods ?
  - How do we model words and tags of a document ?
  - How do we modify clustering methods in order to include tagging data ?
  - How can we evaluate the clustering results ?

Outlook

- Document Models
- Clustering Methods
  - K-Means
  - (Multi Multinomial) Latent Dirichlet Allocation
- Evaluation Method
- Experiments & Results

Document models for a vector space

Word vocabulary: W

Tag vocabulary: T

Bag of words of a document: B_w

Bag of tags of a document: B_t

- Words Only: $V_w = \langle w_1, w_2, \ldots, w_{|W|} \rangle$

  w_i is tf (or tf-idf) of word i in B_w

  Normalization: $\|V_w\|_2 = 1$

- Tags Only: $V_t = \langle w_1, w_2, \ldots, w_{|T|} \rangle$

Document models for a vector space

- Tags as Words

  Vocabulary: $W' = W \cup T$

  Bag of Words: $B_{w'} = B_t \cup B_w$

$$V_{w'} = \langle w_1, w_2, \ldots, w_{|W'|} \rangle$$

- Tags as New Words:

$$V_{w,t} = \langle w_1, w_2, \ldots, w_{|W|}, w_{|W|+1}, w_{|W|+2}, \ldots, w_{|W|+|T|} \rangle$$

Document models for a vector space

- Words+Tags:

$$V_{w+t} = \left\langle \sqrt{\frac{1}{2}} V_w , \sqrt{\frac{1}{2}} V_t \right\rangle$$

Count and weight words and tags independently !

K-Means Clustering Problem

Given the data: $(x_1, \ldots, x_N)$ $\quad x_i \in \mathbb{R}^d$

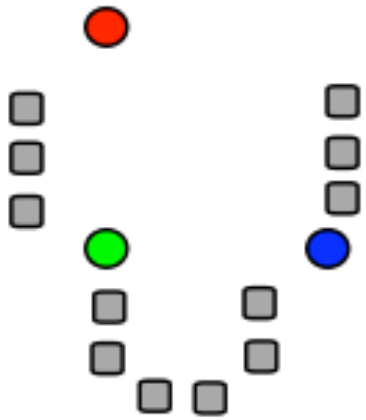
K-Means aims for the clusters: $P = \{C_1, \ldots, C_k\}$
such that:

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \quad \text{is minimal}$$

where μ_i is the mean of cluster C_i

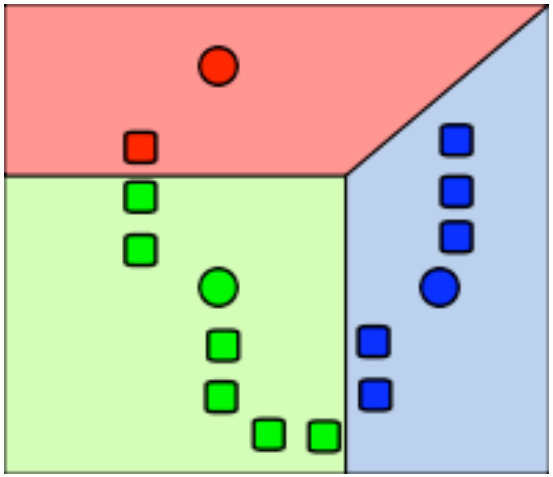
Standard K-Means Clustering Algorithm

Step 1: Choose randomly k datapoints as initial means

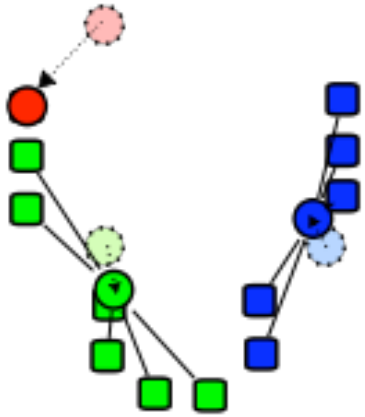


Standard K-Means Clustering Algorithm

Step 2: Assign each datapoint to the cluster with the closest mean.

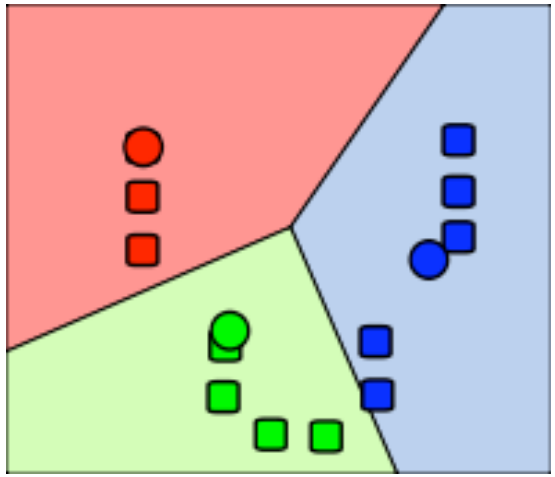


Step 3: Compute centroids of the *k* clusters.
They become the new means.



Standard K-Means Clustering Algorithm

Repeat steps 2 and 3 until convergence has been reached.



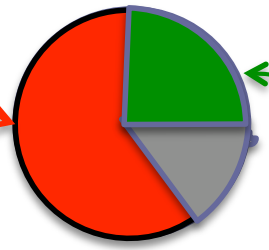
Outlook

- Document models
- Clustering Methods
  - K-Means
  - Multi-Multinomial Latent Dirichlet Allocation
    - Topic Models
    - Latent Dirichlet Allocation
    - Multi-Multinomial Latent Dirichlet Allocation
- Evaluation
- Results

Topic Models

Document 21:

catalog, pricing, logic, Jordan, kitchen, work, math, chemistry, basketball, sport



Topic 3:

Catalog	0.3
Shopping	0.2
Internet	0.1
Buy	0.1
Cart	0.1

Topic 17:

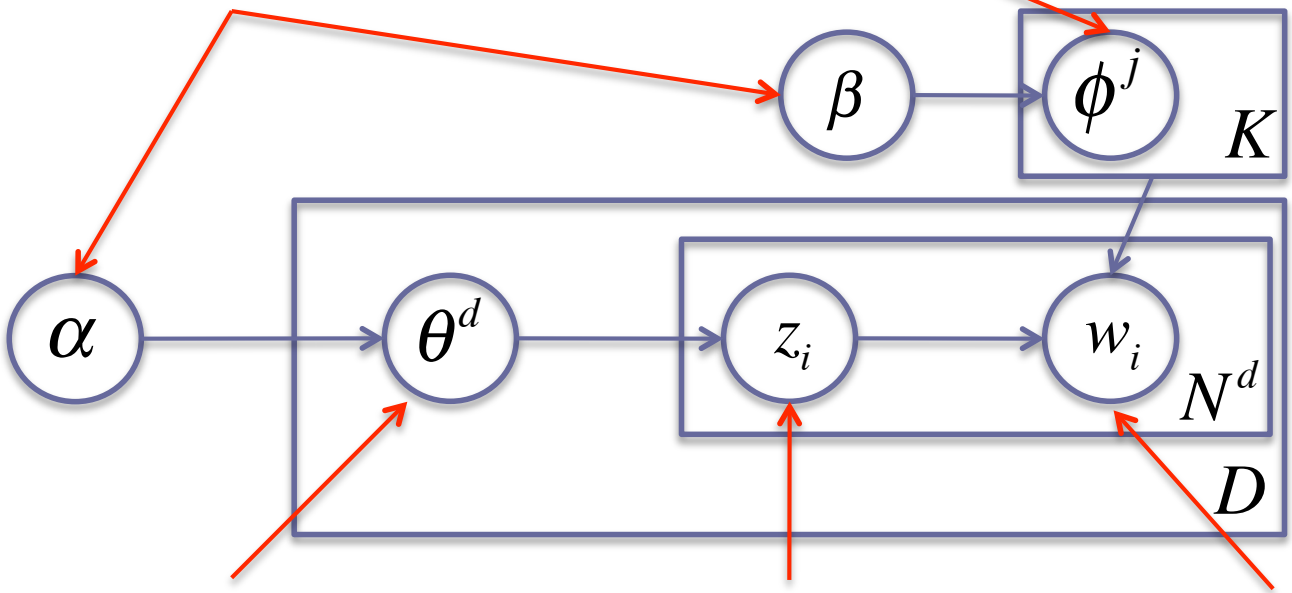
Scientific	0.5
Research	0.1
Knowledge	0.1
Work	0.1
Math	0.1

Latent Dirichlet Allocation

distribution over words for each topic

$\phi^j \sim Dirichlet(\beta)$

Dirichlet priors



distribution over topics for each document

$\theta^d \sim Dirichlet(\alpha)$

topic assignment for each word

$z_i \sim Discrete(\theta^d)$

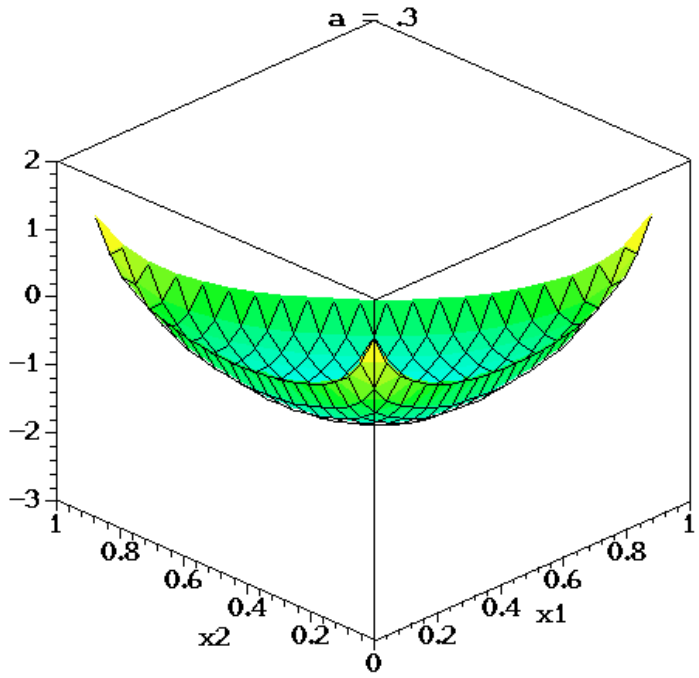
word generated from assigned topic

$w_i \sim Discrete(\phi^{z_i})$

Prior: Dirichlet Distribution

$$p(x_1,\ldots,x_K) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{i=1}^{K}(x_i)^{\alpha-1}$$

- Hyperparameter α determines the form of the Dirichlet D.
- The form determines which kinds of multinomial distributions are more likely or less likely.



K=3 α changes from 0.3 to 2.0

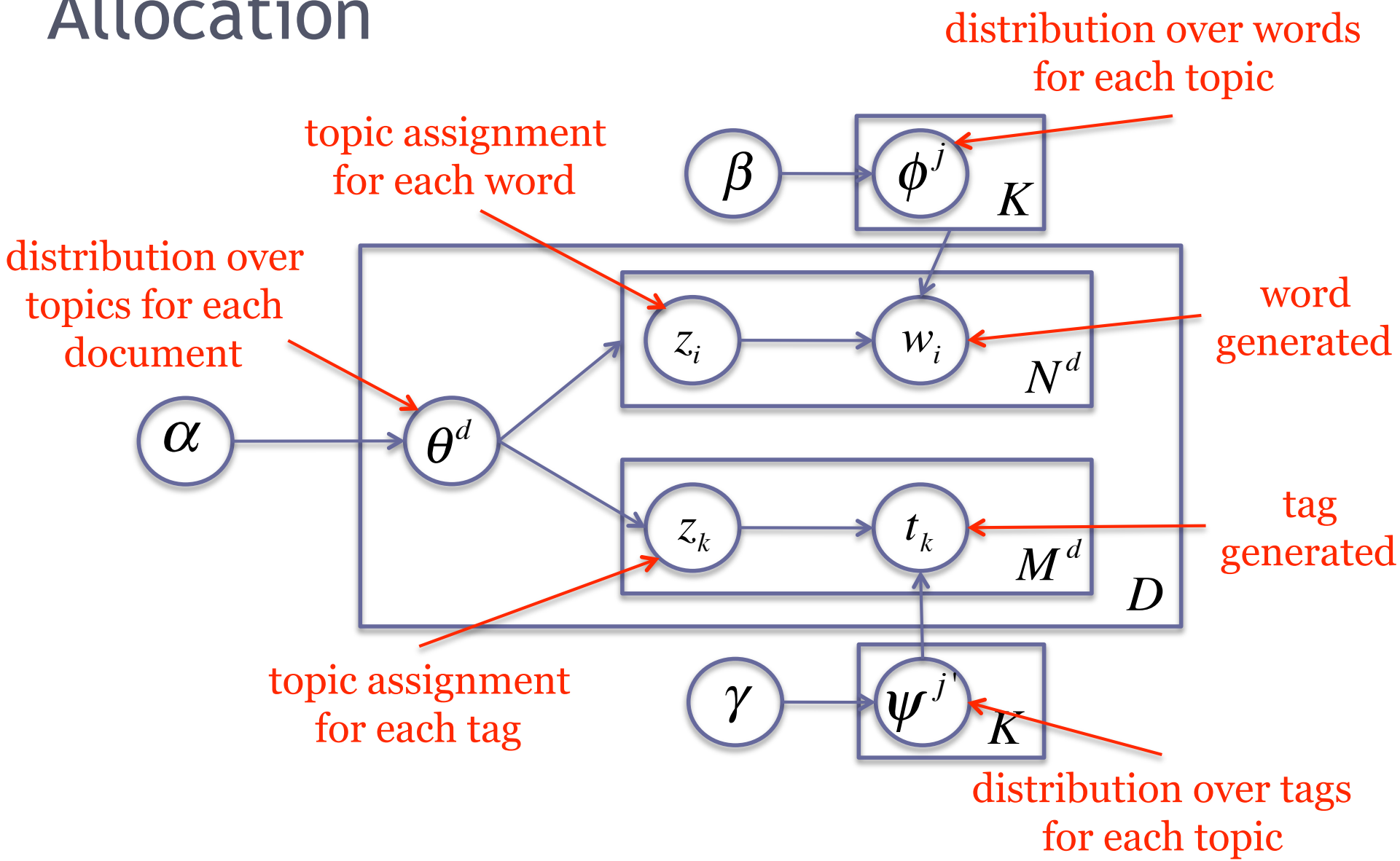
Inverting the generative model

- Maximum likelihood estimation
  - EM-Algorithm: Hofmann (1999)
- Deterministic approximate algorithms
  - Variational EM: Blei, Ng, Jordan (2003)
- Markov Chain Monte Carlo
  - Gibbs Sampler: Griffiths & Steyr (2004)
  - Gibbs Sampler: Wei and Croft (2006)

Document models for (MM)-LDA

- Words only: LDA
- Tags only: LDA
- Tags as Words Times n: Add tags as words with multiplicity of n and use LDA
- Tags as new Words: Add tags as special words (e.g. tag#Basketball) and use LDA
- Words+Tags: Use MM-LDA

Multi Multinomial Latent Dirichlet Allocation



Outlook

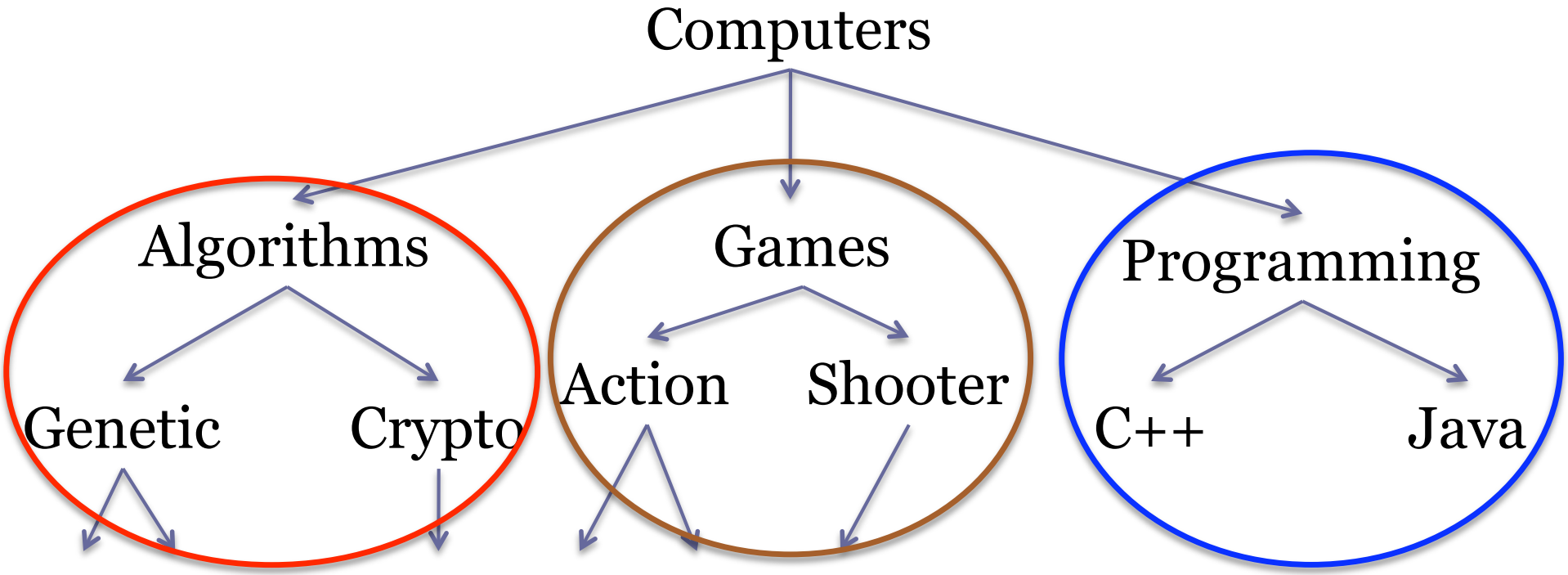
- Document models
- Clustering Methods
  - K-Means
  - (Multi-Multinomial) Latent Dirichlet Allocation
- Evaluation Method
  - Gold Standard Clustering
  - Cluster Evaluation Score
  - Dataset
- Experiments & Results

Gold Standard Clustering

- We create a “gold standard” clustering using the Open Directory Project

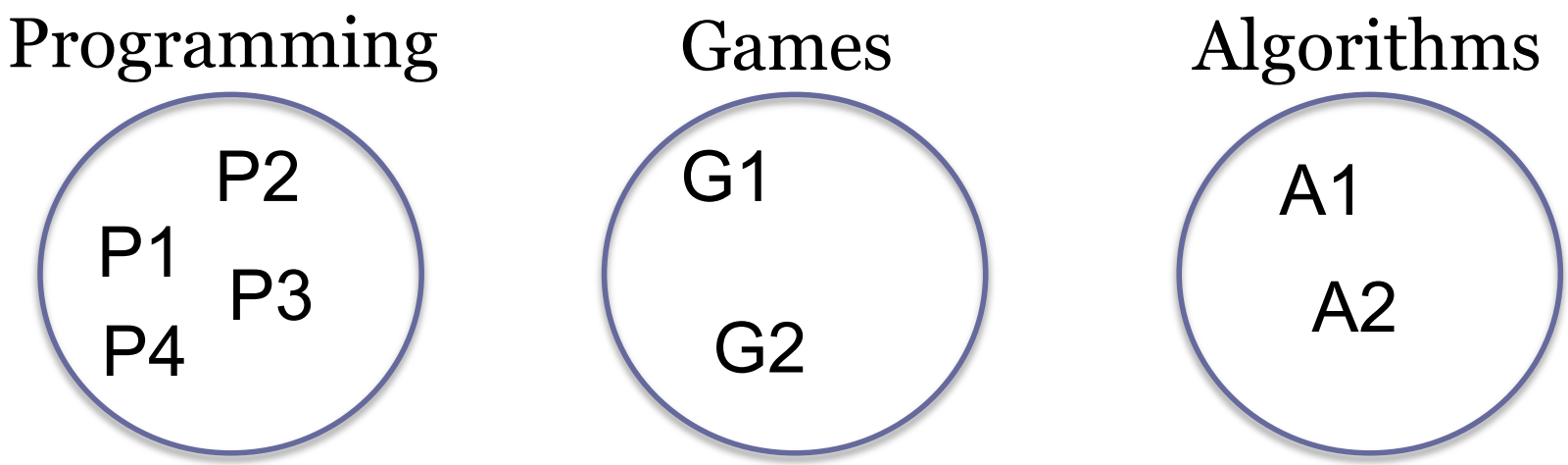
Gold Standard Clustering

- A node in the ODP hierarchy is chosen as root
- Each child (+ its descendants) is treated as one cluster.



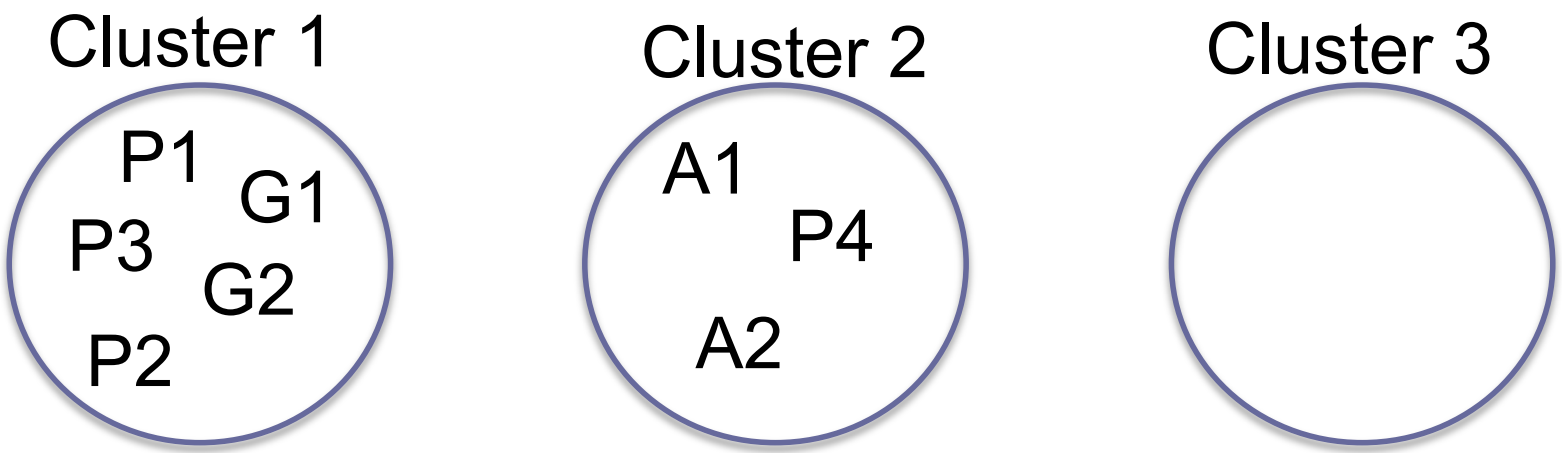
Cluster Evaluation Metric

Gold Standard (GS) says:



Cluster Evaluation Metric

Clustering Algorithm (CA) returns:



Consider a pair of documents:

If the CA placed the two documents in different clusters, but the GS has them in different clusters

→ False Negative (FN)

F1 Cluster Evaluation Score

- The F1 score is the harmonic mean of precision and recall

- Precision : $TP / (TP + FP) = 5/13$
- Recall: $TP / (TP + FN) = 5/8$
- F1: $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall}) \approx 0.476$

Dataset

- ODP Dataset
- Stanford Tag Crawl Dataset: One contiguous month of del.icio.us feeds.

- Consider only documents which are
  - present both in ODP and the Tag Crawl Dataset
  - are in English
  - and their page text is crawled
- Total number: 15,230

Outlook

- Document models
- Clustering Methods
  - K-Means
  - Multi-Multinomial Latent Dirichlet Allocation
- Evaluation Method
- Experiments & Results
  - K-Means (Document models)
  - MM-LDA (Document models)
  - Comparison

Experiment: K-Means on different document models

Averaged F1 – Scores of 10 runs of K-Means applied on 13230 documents using tf-weighting

	K-Means
Words	.139
Tags as Words x 1	.158
Tags as Words x 2	.176
Tags as New Words	.154
Words+Tags	.225

Experiment: (MM)-LDA on different document models

F1 – Scores of LDA and MM-LDA applied on 13230 documents

	(MM-)LDA
Words	.260
Tags as Words x 1	.213
Tags as Words x 2	.198
Tags as New Words	.216
Words+Tags	.307

Comparison: K-Means and MM-LDA

Tag-Augmented K-means

	tags	*words*
1	linux security php opensource vpn unix	linux ircd php beware kernel exe
2	games go game sports firefox gaming	dmg munsey ballparks suppes racer game
3	music research finance audio mp3 lyrics	music research redirect nottingham meta laboratory
4	news business newspaper politics media magazine	v business leadership d news j
5	politics activism travel movies law government	aquaculture terrapass geothermal anarchist wwoof cpsc
6	science physics biology astronomy space chemistry	science wildman foraging collembola physics biology
7	css python javascript programming xml webdesign	squeakland sql coq css python flash
8	food recipes cooking shopping tea recipe	recipes food cooking recipe stylist tea
9	blog blogs fashion design art politics	flf blog comments posted my beuys
10	education art college university school teaching	learning gsapp students education school cutecircuit
11	health medical healthcare medicine solar psychology	health napkin cafepress.com medical care folding
12	java programming development compiler c opensource	java c programming goto code language
13	software windows opensource mac freeware osx	software windows mac download os thinkfree
14	dictionary reference language bible writing english	dictionary english words syw dictionaries spanish
15	internet dns search seo google web	internet shutdown sportsbook epra kbs npower
16	history library books literature libraries philosophy	library tarot peopling ursula guin bowdoin

Multi-Multinomial LDA (MM-LDA)

	tags	*words*
1	web2.0 tools online editor photo office	icons uml powerpoint lucid dreams dreaming
2	guitar scanner chemistry military earthquake groupware	grub outlook bittorrent rendering recovery boot
3	health medical medicine healthcare process gardening	exe health openpkg okino dll polytrans
4	bible christian space astronomy religion christianity	gaelic bible nt bone scottish english
5	politics activism environment copyright law government	war shall power prisoners their article
6	social community web2.0 humor fun funny	press f prompt messages ignoring each
7	reference science education research art books	science research information university search site
8	java database programming development mysql sql	java sql mysql schizophrenia testing test
9	dictionary language english reference translation thesaurus	english writing dictionary spanish words bppv
10	travel search maps google reference map	search deadline call flf conference paper
11	time clock timezones world train md5	quantum thu pfb am pm mf
12	food recipes cooking business shopping finance	my food tea wine me recipes
13	news blog music blogs technology system/unfiled	comments blog he posted news pm
14	programming software webdesign web css design	you can if or not use
15	photography photo compression zip photos photoblog	flash camera eos light e-ttl units
16	mac apple osx games unicode game	dmg u x mac b v

Comparison: K-Means and MM-LDA

	(MM-)LDA	**K-means**
Words	.260	.139
Tags	.270	.219
Words+Tags	.307	.225

Clustering the tagged Web

- Motivation (Ambiguity of queries, Web Categorization, Tags)
- Document models (Words, Tags, Words+Tags,…)
- Clustering Methods
  - K-Means (Problem, Algorithm)
  - (MM)-LDA (Topic Models, LDA, MM-LDA)
- Evaluation Method
  - Gold Standard Clustering
  - F1 Cluster Evaluation Score
  - Dataset
- Experiments & Results
  - K-Means (Document models)
  - (MM)-LDA (Document models)
  - Comparison