Spherical Admixture Models

Joseph Reisinger*, Austin Waters, Bryan Silverthorn, and
Raymond J. Mooney

June 22nd, 2010

* Supported by an NSF Graduate Research Fellowship and a Google Research Award.

- "I want to use LDA..."
- Want to use some set of feature weights capturing semantic content (tf-idf, pmi, etc)
- Empirical benefits to cosine distance in classical IR tasks.

Dhillon and Modha (2001), Strehl et al. (2000), Salton and McGill (1983)

Clustering, soft clustering, and topic models

Latent Dirichlet Allocation

$$
\begin{array}{rcllll}
\boldsymbol{\theta}_d | \boldsymbol{\alpha} & \sim & \mathrm{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t | \boldsymbol{\beta} & \sim & \mathrm{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id} | \boldsymbol{\theta}_d & \sim & \mathrm{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id} | \boldsymbol{\phi}_{z_{id}} & \sim & \mathrm{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

$\boldsymbol{\phi}_1$	$\boldsymbol{\phi}_2$	$\boldsymbol{\phi}_3$
government	wrote	finance
minister	said	economists
state	responding	spending
federal	editor	budget

$\sim \mathrm{Dir}(\beta)$

$\boldsymbol{\phi} =$ 

$d_1 =$ Responding to finance minister Ruth Richardson's May 1991 budget which cut government spending, 15 academic economists from the University of Auckland wrote a letter to the editor of the New Zealand Herald on 6 June 1991. It read: "We wish to state in the strongest possible terms our view that in the present

$\mathbf{d} =$ 

Clustering, soft clustering, and topic models

Latent Dirichlet Allocation

$$
\begin{array}{llllll}
\boldsymbol{\theta}_d | \boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t | \boldsymbol{\beta} & \sim & \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id} | \boldsymbol{\theta}_d & \sim & \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id} | \boldsymbol{\phi}_{z_{id}} & \sim & \text{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

$\boldsymbol{\phi}_1$	$\boldsymbol{\phi}_2$	$\boldsymbol{\phi}_3$
government	wrote	finance
minister	said	economists
state	responding	spending
federal	editor	budget

$\sim \text{Dir}(\beta)$

$\boldsymbol{\phi} =$ 

$d_1 =$ Responding to finance minister Ruth Richardson's May 1991 budget which cut government spending, 15 academic economists from the University of Auckland wrote a letter to the editor of the New Zealand Herald on 6 June 1991. It read: "We wish to state in the strongest possible terms our view that in the present

$\mathbf{d} =$ 

Topic modeling, dimensionality reduction

Latent Dirichlet Allocation

$$
\begin{array}{llllll}
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t|\boldsymbol{\beta} & \sim & \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id}|\boldsymbol{\theta}_d & \sim & \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id}|\boldsymbol{\phi}_{z_{id}} & \sim & \text{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

- Topic modeling is basically the same story as dimensionality reduction, e.g. SVD, PCA, NMF, ...
- Differences:
  - Bayesian
  - More emphasis on interpreting topics
  - Generative models offer more flexibility

Documents as weighted average over topics

Latent Dirichlet Allocation

$$\begin{array}{llllll}
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t|\boldsymbol{\beta} & \sim & \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id}|\boldsymbol{\theta}_d & \sim & \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id}|\boldsymbol{\phi}_{z_{id}} & \sim & \text{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}$$

- We can explicitly represent the multinomial distribution that a document is drawn from integrating out z instead of theta:

$$w_{id} \sim \text{Mult}(\theta_d^\top \Phi)$$

- i.e. a weighted average over the topics.

(Blei et al. 2003)

Spherical mixture modeling intuition

spherical mixture model

$$
\begin{array}{llll}
\phi_k \sim \mathrm{vMF}(\mathbf{m}_0) & k \in K & & \text{(clusters)} \\
z_i \sim H & i \in D & & \text{(assignments)} \\
\boldsymbol{d}_i \sim \mathrm{vMF}(\boldsymbol{\phi}_{z_i}) & i \in D & & \text{(documents)}
\end{array}
$$

von Mises-Fisher Distribution

$$
f(\mathbf{x}; \boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{x}\right)
$$

$$
||\boldsymbol{\mu}|| = 1, \, \kappa \geq 0
$$

- Generalization of spherical k-means / cosine distance
- Embed documents in the unit-hypersphere (L2 norm)
- Cosine distance has been quite successful in IR / document modeling (less sensitive to any one single feature)

(Banerjee et al. 2006)

Latent Dirichlet Allocation

$$
\begin{array}{rcllll}
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t|\boldsymbol{\beta} & \sim & \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{id}|\boldsymbol{\theta}_d & \sim & \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{id}|\boldsymbol{\phi}_{z_{id}} & \sim & \text{Mult}(\boldsymbol{\phi}_{z_{id}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

+

Spherical mixture model

$$
\begin{array}{rclll}
\boldsymbol{\phi}_k & \sim & \text{vMF}(\boldsymbol{\mu}, \xi) & k \in K & \text{(clusters)} \\
z_i & \sim & H & i \in D & \text{(assignments)} \\
\boldsymbol{d}_i & \sim & \text{vMF}(\boldsymbol{\phi}_{z_i}) & i \in D & \text{(documents)}
\end{array}
$$

=

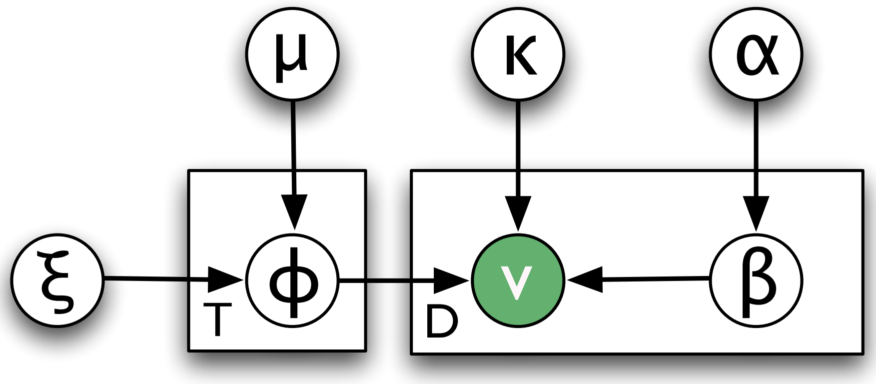
Spherical Admixture Model

$$
\begin{array}{rcllll}
\boldsymbol{\mu}|\kappa_0 & \sim & \text{vMF}(\mathbf{m}, \kappa_0), & & \text{(corpus mean)} \\
\boldsymbol{\phi}_t|\boldsymbol{\mu}, \xi & \sim & \text{vMF}(\boldsymbol{\mu}, \xi), & t \in T, & \text{(topics)} \\
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\bar{\boldsymbol{\phi}}_d|\boldsymbol{\phi}, \boldsymbol{\theta}_d & = & \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\theta}_d), & d \in D, & \text{(spherical average)} \\
\mathbf{v}_d|\bar{\boldsymbol{\phi}}_d, \kappa & \sim & \text{vMF}\left(\bar{\boldsymbol{\phi}}_d, \kappa\right), & d \in D, & \text{(documents)}
\end{array}
$$

Spherical average

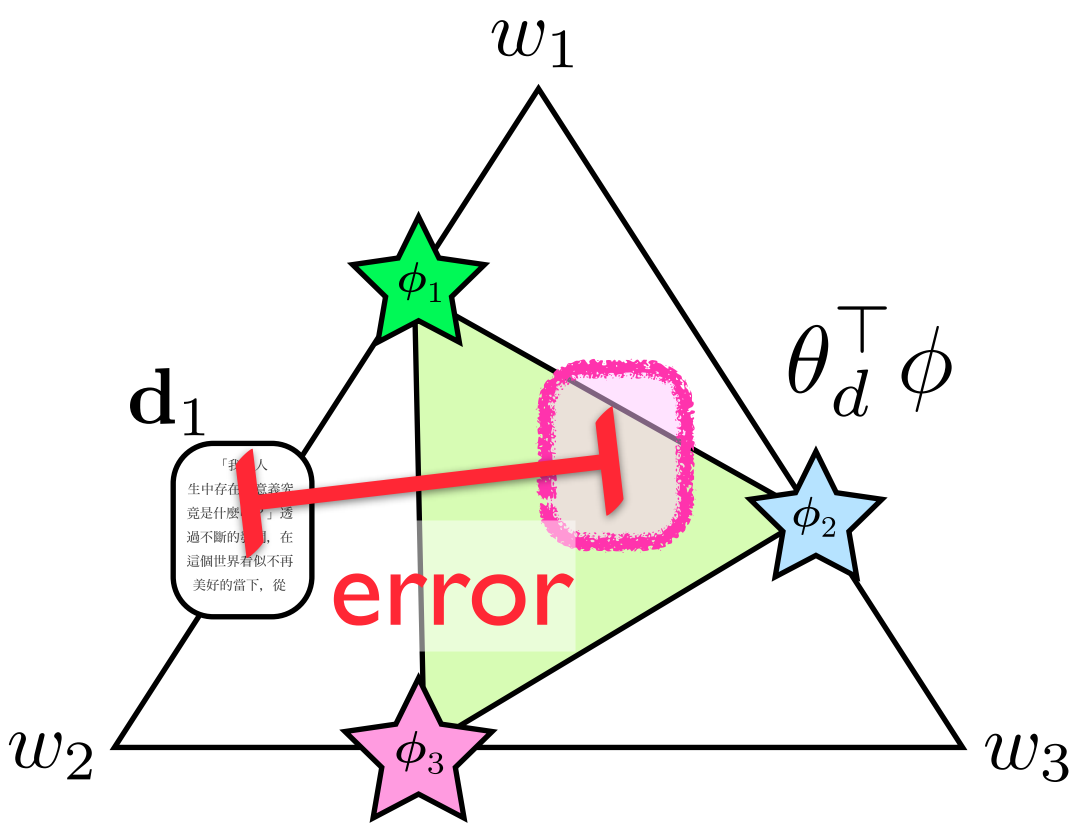
Spherical Admixture Model

$$
\begin{array}{llllr}
\boldsymbol{\mu}|\kappa_0 & \sim & \text{vMF}(\mathbf{m}, \kappa_0), & & \text{(corpus mean)} \\
\boldsymbol{\phi}_t|\boldsymbol{\mu}, \xi & \sim & \text{vMF}(\boldsymbol{\mu}, \xi), & t \in T, & \text{(topics)} \\
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\bar{\boldsymbol{\phi}}_d|\boldsymbol{\phi}, \boldsymbol{\theta}_d & = & \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\theta}_d), & d \in D, & \text{(spherical average)} \\
\mathbf{v}_d|\bar{\boldsymbol{\phi}}_d, \kappa & \sim & \text{vMF}\left(\bar{\boldsymbol{\phi}}_d, \kappa\right), & d \in D, & \text{(documents)}
\end{array}
$$

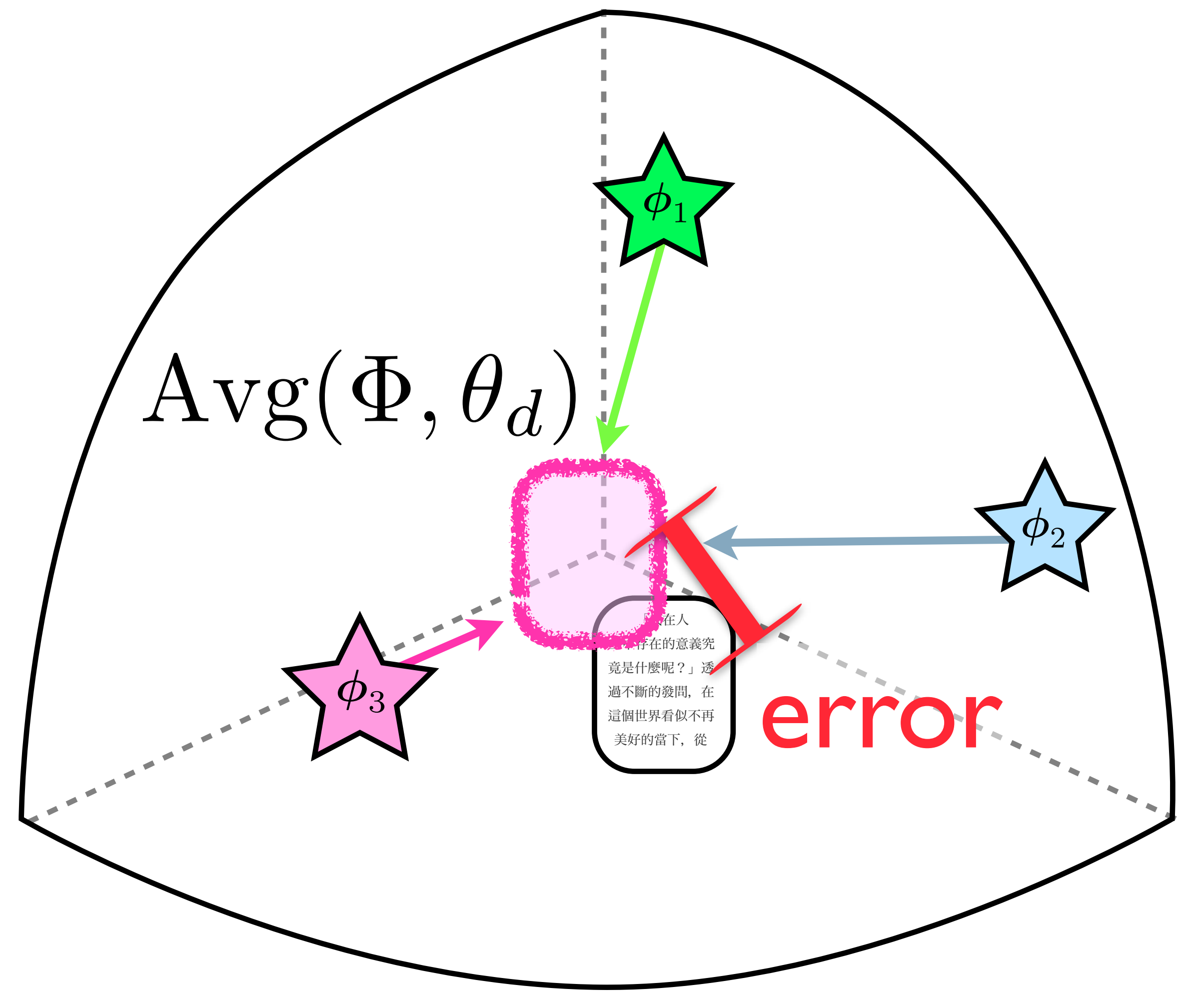


- LDA does an implicit weighted averaging step.
- This is easy with the bag-of-words assumption, slightly harder when we're drawing documents all at once (b/c of L2 norm)
- So we compute the weighted average explicitly.

Drawing documents



Latent Dirichlet Allocation

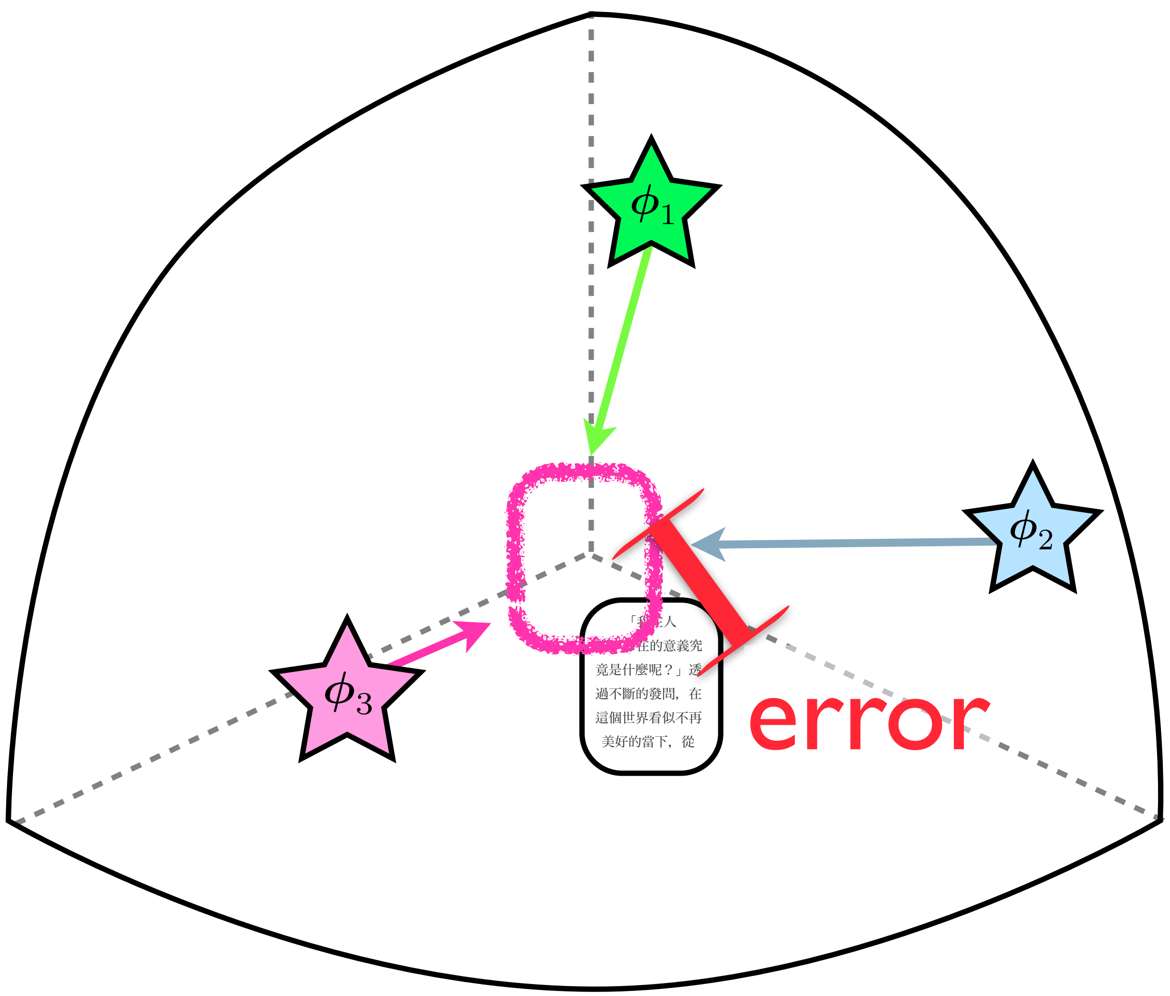


Spherical Admixture Model

Drawing documents

Spherical Admixture Model

$$
\begin{array}{llllr}
\boldsymbol{\mu}|\kappa_0 & \sim & \mathrm{vMF}(\mathbf{m}, \kappa_0), & & \text{(corpus mean)} \\
\boldsymbol{\phi}_t|\boldsymbol{\mu}, \xi & \sim & \mathrm{vMF}(\boldsymbol{\mu}, \xi), & t \in T, & \text{(topics)} \\
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim & \mathrm{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\bar{\boldsymbol{\phi}}_d|\boldsymbol{\phi}, \boldsymbol{\theta}_d & = & \mathrm{Avg}(\boldsymbol{\phi}, \boldsymbol{\theta}_d), & d \in D, & \text{(spherical average)} \\
\mathbf{v}_d|\bar{\boldsymbol{\phi}}_d, \kappa & \sim & \mathrm{vMF}\left(\bar{\boldsymbol{\phi}}_d, \kappa\right), & d \in D, & \text{(documents)}
\end{array}
$$



- Variational EM for inference
- Tractable: ~10k docs in O(hours)

http://www.cs.utexas.edu/~austin

Topic interpretability

NIPS			
(+)	(−)	(+)	(−)
svm kernel margin machines support	network experts units target clusters	genetic fitness crossover population search	mlp tree matrix discriminant lemma

Wikipedia					
(+)	(−)	(+)	(−)	(+)	(−)
navy ships naval submarines aircraft	airport airlines flights bus satellites	album label singles chart song	opera actor films players conservatory	india temple dynasty indian khan	germany borough england france parish

- Observing a term with negative weight is evidence *against* that topic
- Negative weight terms are often semantically similar, near-neighbor topics

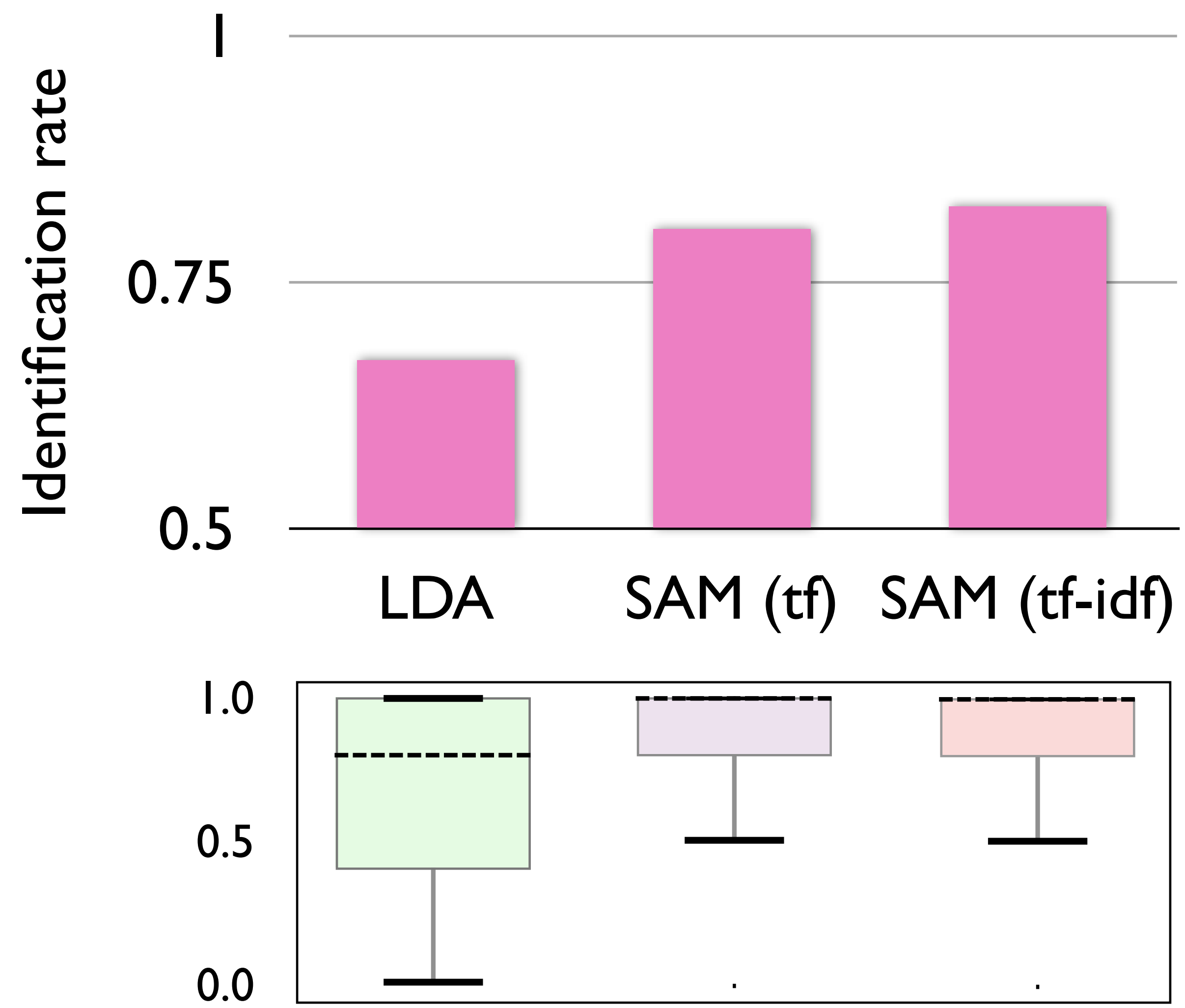
Human studies: topic coherence

LDA	male, mammals, empire, plants, species, birds
	court, crimes, police, law, security, jazz
SAM	vishnu, tamil, kerala, singh, meteorologist, nadu
	oxidation, footballers, protein, potassium, hydrogen, symptoms

- Measure semantic coherence of the highest weighted terms in each topic via a "word intrusion" task
- Human raters were recruited using Mechanical Turk
- Quality control: (1) manually constructed tasks, (2) screening for low LOO inter-annotator agreement

(Chang et al. 2009)

Human studies: topic coherence



- 8 raters per question (632 unique), 50 questions per model
- LDA: 52%, SAM tf 80%, SAM tf-idf 82% identification rate

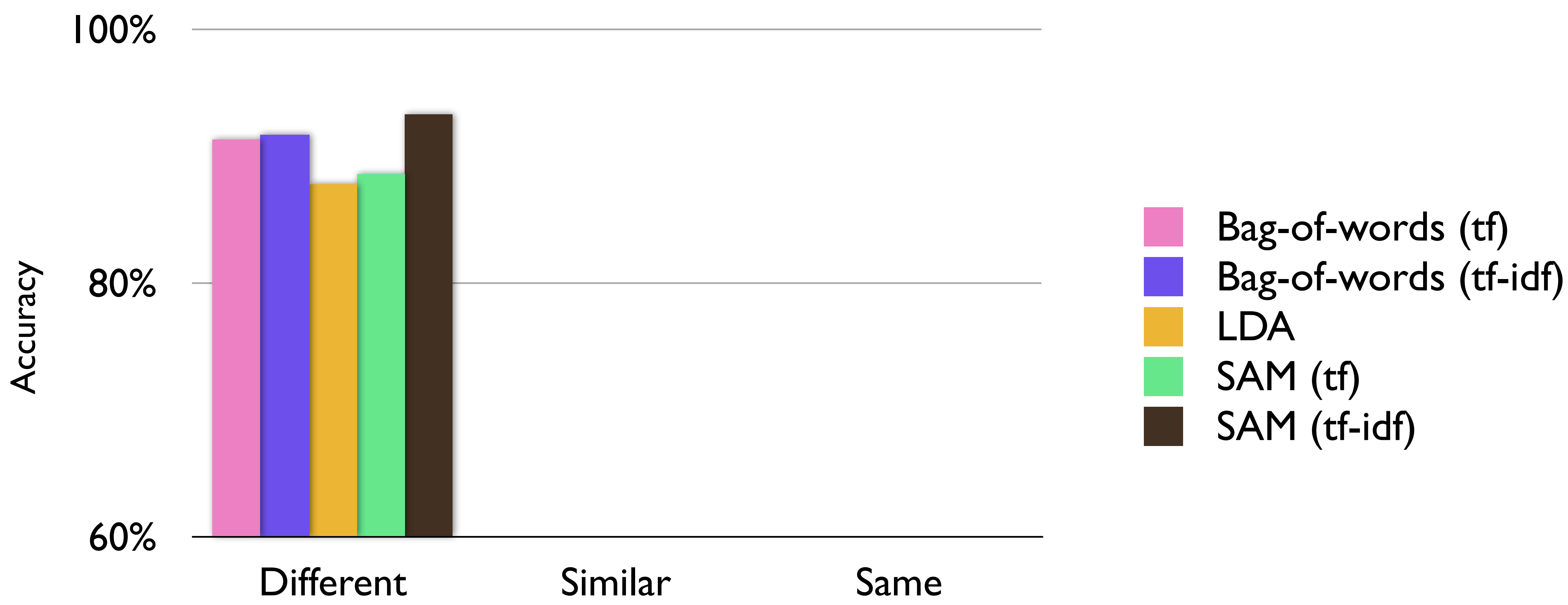
(Chang et al. 2009)

Human studies: topic relevance

- Forced choice: “which set of words best describes the main theme of the article?”
- Discarded 47 articles with low kappa; SAM results preferred 62%

(Chang et al. 2009)

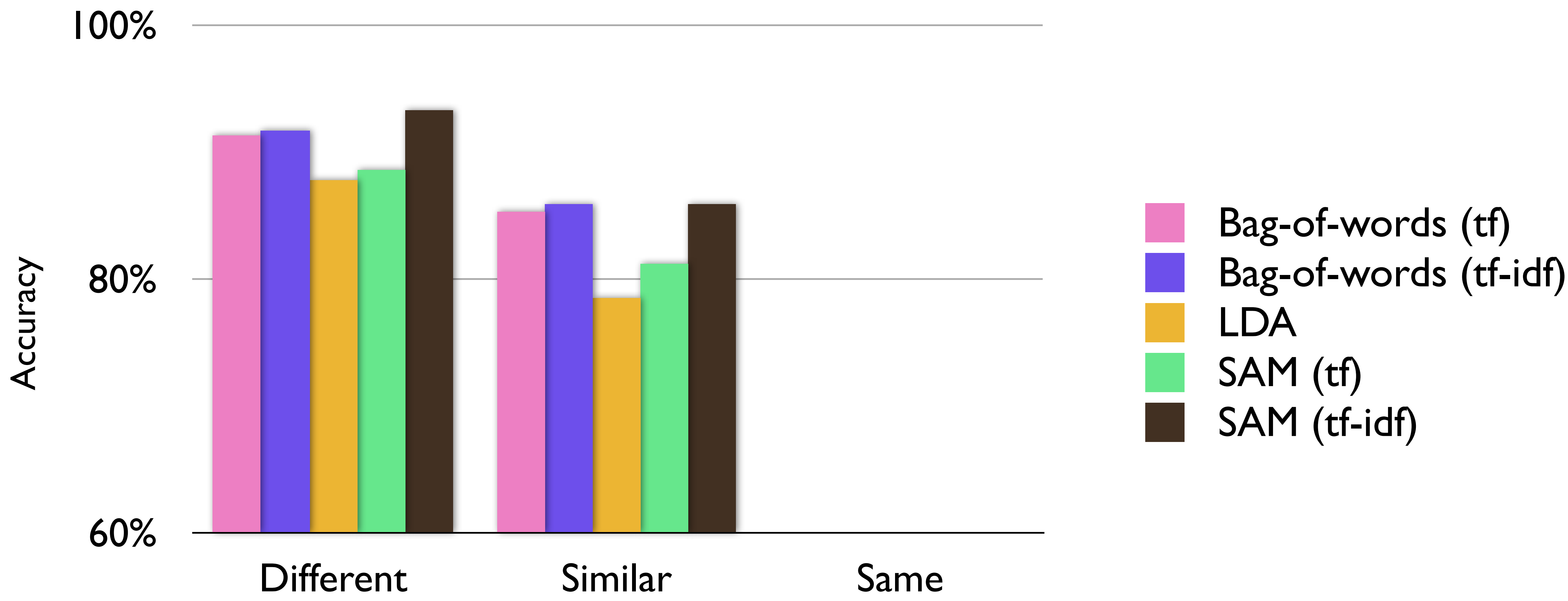
Results: 20 newsgroups



- Three classification tasks:
  - **Different:** `rec.sport.baseball, sci.space, alt.atheism`
  - **Similar:** `rec.sport.baseball, talk.politics.guns, talk.politics.misc`
  - **Same:** `comp.os.ms-windows.misc, comp.windows.x comp.graphics`

(Banerjee and Basu 2007)

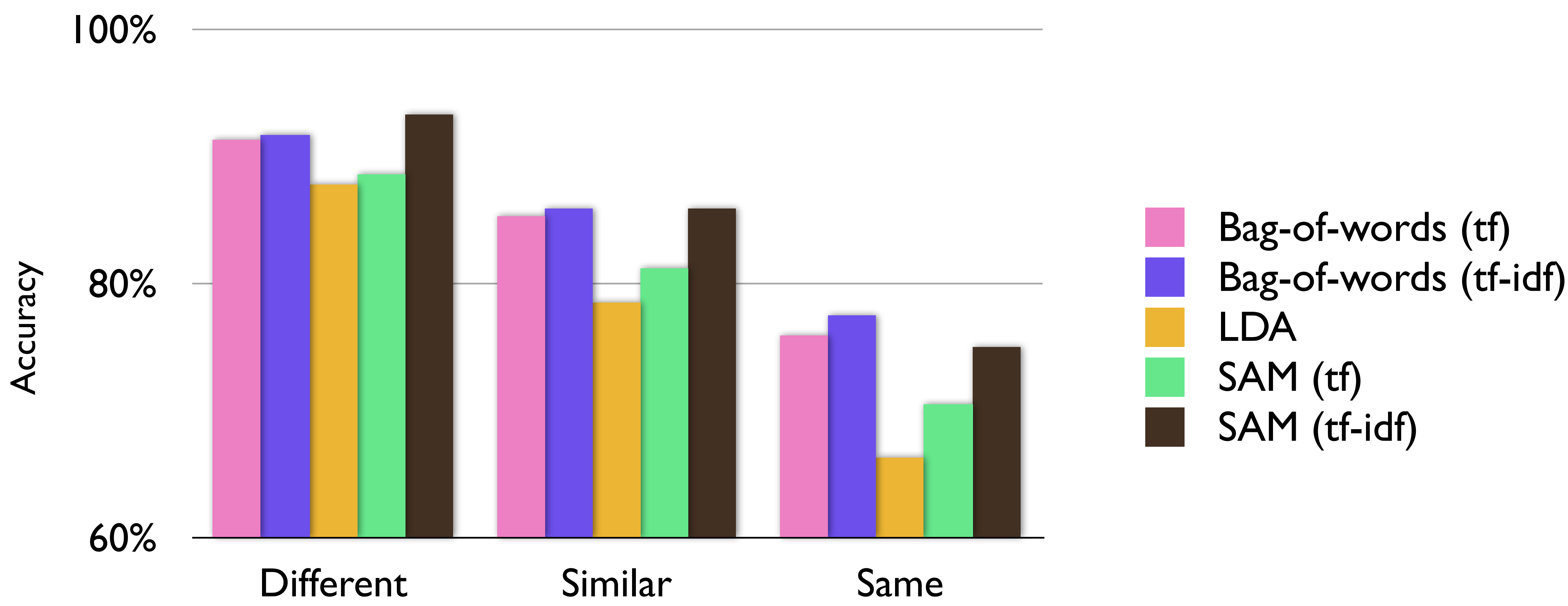
Results: 20 newsgroups



- Three classification tasks:
  - Different: `rec.sport.baseball, sci.space, alt.atheism`
  - Similar: `rec.sport.baseball, talk.politics.guns, talk.politics.misc`
  - Same: `comp.os.ms-windows.misc, comp.windows.x comp.graphics`

(Banerjee and Basu 2007)

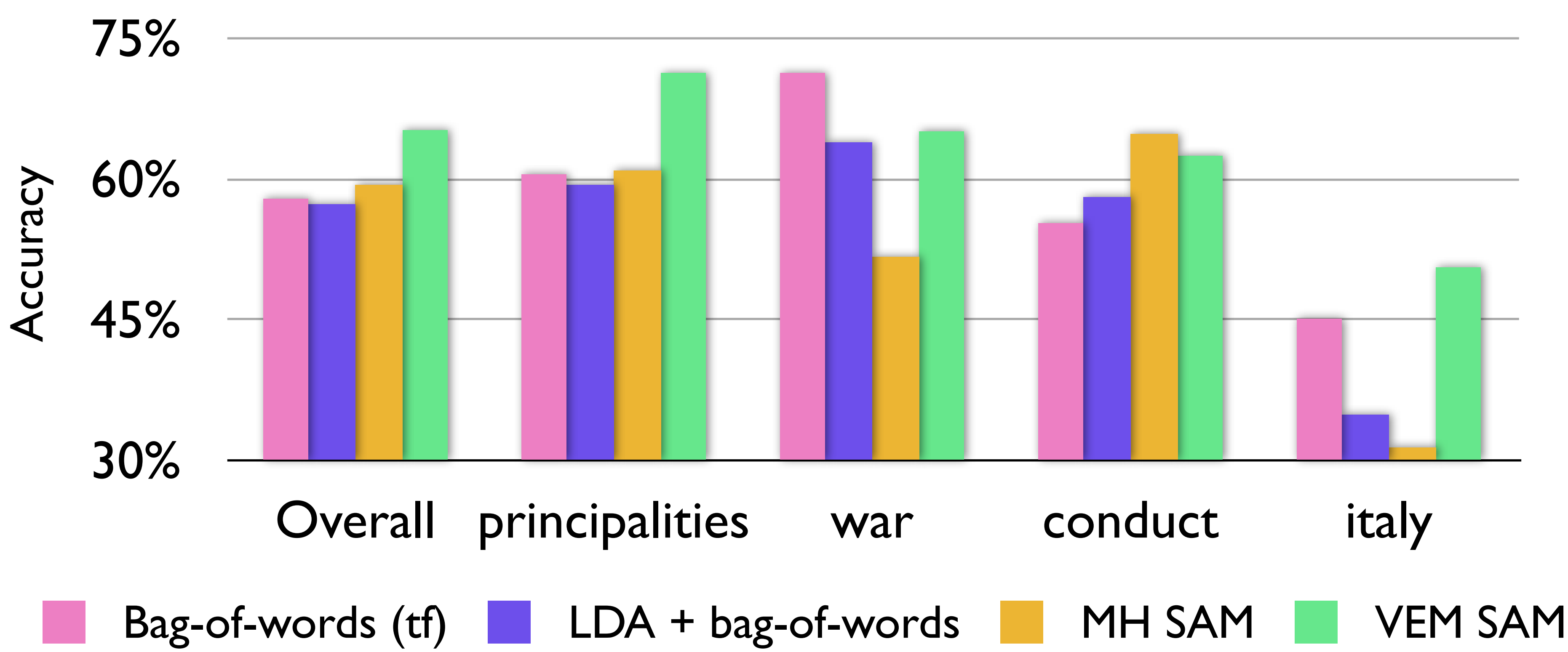
Results: 20 newsgroups

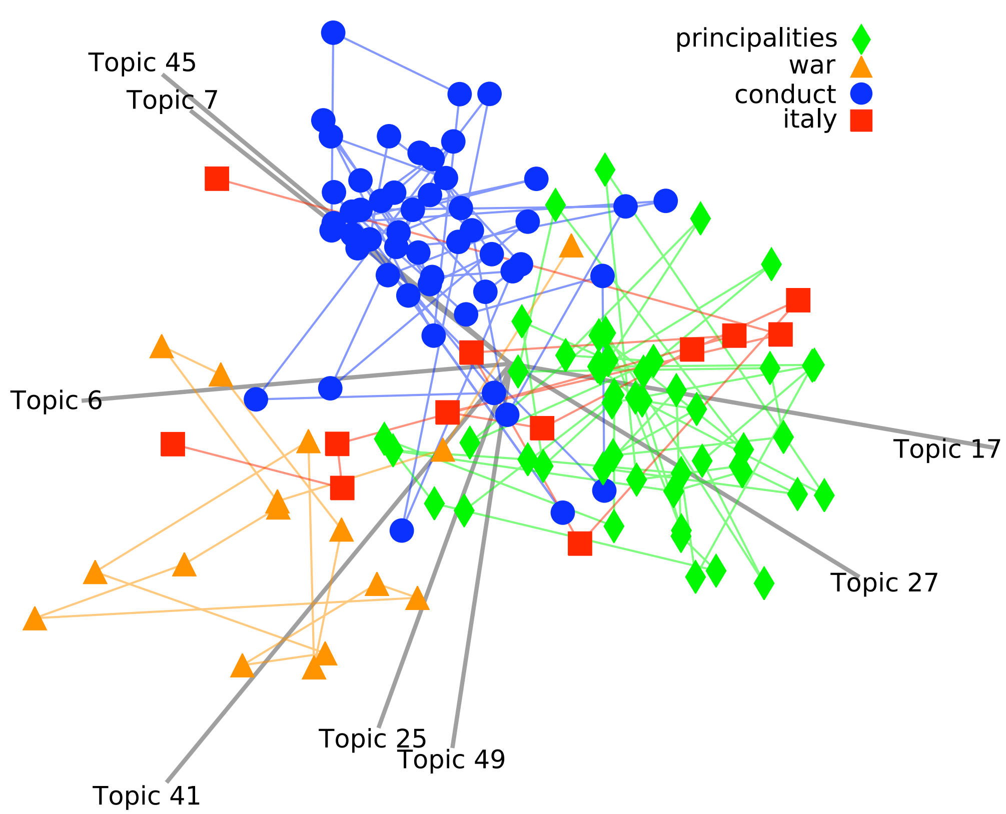


- Three classification tasks:
  - Different: `rec.sport.baseball, sci.space, alt.atheism`
  - Similar: `rec.sport.baseball, talk.politics.guns, talk.politics.misc`
  - Same: `comp.os.ms-windows.misc, comp.windows.x comp.graphics`

(Banerjee and Basu 2007)

Results: *il principe*





- Short, singly-authored, thematically tight
- 4 main themes corresponding to 4 sections:
  - Types of Principalities, Ch 1-11
  - Types of Armies, Ch 12-14
  - The Conduct of Princes, Ch 15-23
  - Political Situation in Italy, Ch 24-26

Why does it work?

- Feature weighting helps dimensionality reduction (less for interpretability).
- Dense topic vectors can account for missing terms.
- Cosine distance may better measure document / topic similarity.

Conclusions

- Replacing multinomial likelihood of LDA with vMF (spherical); inference is tractable
- Cosine distance; dense topic vectors
- Better results as a dimensionality reduction method
- Top weighted terms are more semantically coherent (human raters)
- Benefits are less pronounced for denser data sets (e.g. vision)
- Negative weight terms capture some useful structure.

Thanks!

Spherical Topic Models

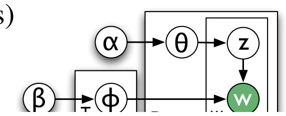
Joseph Reisinger, Austin Waters, Bryan Silverthorn, Raymond Mooney
The University of Texas at Austin

Abstract

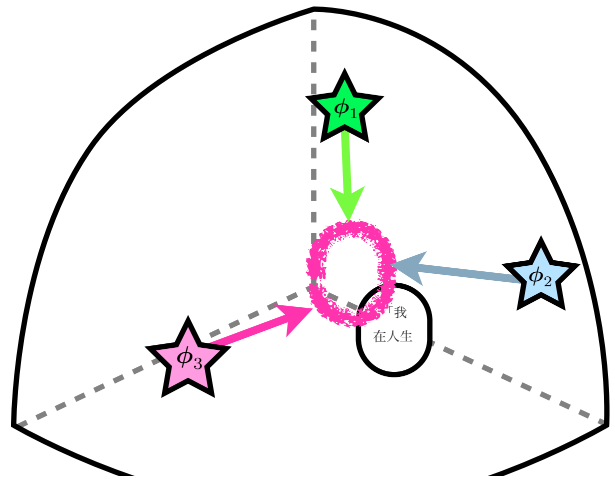
We introduce the Spherical Admixture Model (SAM), an efficient Bayesian topic model for arbitrary L2 normalized data. SAM maintains the same hierarchical structure as Latent Dirichlet Allocation (LDA), but models documents as points on a high-dimensional spherical manifold, allowing a natural likelihood parameterization in terms of cosine distance. Furthermore, SAM is capable of representing negative topic features and word presence/absence, unlike previous models. Performance is evaluated empirically across several disparate classification tasks, from natural language processing and computer vision.

Implicit Averaging in LDA

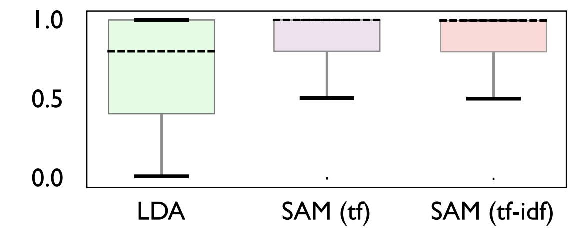
$\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}$	$\sim$	Dirichlet($\boldsymbol{\alpha}$),	$d \in D$,	(topic proportions)
$\boldsymbol{\phi}_t \mid \boldsymbol{\beta}$	$\sim$	Dirichlet($\boldsymbol{\beta}$),	$t \in T$,	(topics)
$z_{id} \mid \boldsymbol{\theta}_d$	$\sim$	Mult($\boldsymbol{\theta}_d$),	$i \in \lvert \mathbf{w}_d \rvert$,	(topic indicators)



Spherical Average



Human Topic Quality Judgements

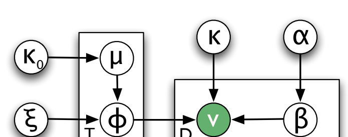


SAM (easy)	1: vishnu, tamil, kerala, singh, nadu, **meteorologist** 2: oxidation, protein, potassium, **footballers**, hydrogen, symptoms
SAM (hard)	1: **saloon**, huron, burlington, county, mississippi, wl 2: tang, hong, **howe**, wu, kong, leone
LDA (easy)	1: male, mammals, **empire**, plants, species, birds 2: court, crimes, police, law, security, **jazz**

Come visit our poster
6pm Oren

Spherical Admixture Model

$\boldsymbol{\mu}_t \mid \kappa_0$	$\sim$	vMF($\mathbf{m}, \kappa_0$),	$t \in T$,	(topic means)
$\boldsymbol{\phi}_t \mid \boldsymbol{\mu}_t, \xi$	$\sim$	vMF($\boldsymbol{\mu}_t, \xi$),	$t \in T$,	(topics)
$\boldsymbol{\beta}_d \mid \boldsymbol{\alpha}$	$\sim$	Dirichlet($\boldsymbol{\alpha}$),	$d \in D$,	(topic proportions)
$\bar{\boldsymbol{\phi}}_d \mid \boldsymbol{\phi}, \boldsymbol{\beta}_d$	$=$	Avg($\boldsymbol{\phi}, \boldsymbol{\beta}_d$),	$d \in D$,	(spherical average)
$\mathbf{v}_d \mid \bar{\boldsymbol{\phi}}_d, \kappa$	$\sim$	vMF$\left(\bar{\boldsymbol{\phi}}_d, \kappa\right)$,	$d \in D$,	(documents)



1. Draw a set of T topics $\boldsymbol{\phi}$ on the unit hypersphere;
2. For each document d, draw topic weights $\boldsymbol{\beta}_d$ from a Dirichlet with hyperparameter $\boldsymbol{\alpha}$;
3. Draw a document vector $\mathbf{v}_d$ from a vMF with mean $\bar{\boldsymbol{\phi}} = \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\beta}_d)$ and concentration κ.

machines	target	population	discriminant
support	clusters	search	lemma

Wikipedia					
(+)	(−)	(+)	(−)	(+)	(−)
navy ships naval submarines aircraft	airport airlines flights bus satellites	album label singles chart song	opera actor films players conservatory	india temple dynasty indian khan	germany borough england france parish

Top positive and negative term weights learned by SAM on the NIPS corpus and Wikipedia. (+) shows the highest weighted words and (−) shows lowest weighted within each topic.

LDA	87.8 ± 0.6	78.5 ± 2.7	66.3 ± 2.6
movMF (tf)	71.4 ± 0.3	64.5 ± 0.6	59.4 ± 0.4
movMF (tf-idf)	71.9 ± 0.3	74.2 ± 0.4	56.0 ± 0.6
SAM (tf)	88.6 ± 0.4	81.2 ± 0.4	70.5 ± 0.5
SAM (tf-idf)	93.3 ± 0.3	85.9 ± 0.3	75.0 ± 0.4
Topic + Bag-of-Words			
LDA	91.8 ± 0.4	85.7 ± 0.7	75.6 ± 0.8
movMF (tf)	91.1 ± 0.3	84.9 ± 0.5	75.8 ± 0.8
movMF (tf-idf)	91.4 ± 0.5	84.9 ± 0.5	75.3 ± 0.6
SAM (tf)	91.9 ± 0.4	86.3 ± 0.5	75.6 ± 0.6
SAM (tf-idf)	94.1 ± 0.3	88.1 ± 0.5	78.1 ± 0.6

Using SAM to generate features for document classification (L1 regularized logistic regression). Three different three-way classification tasks were derived from the 20-news dataset with increasing difficulty (i.e. classes become semantically similar).

Download: http://www.cs.utexas.edu/~austin

Acknowledgements: We would like to thank Arindam Banerjee and Kristen Grauman for helpful suggestions. JR acknowledges the support of an NSF Graduate Research Fellowship and a Google Research Award. Experiments were run on the Mastodon cluster, provded by NSF Grand EIA-0303609.