Interactive, Multilingual Topic Models

Jordan Boyd-Graber, 2020

The Challenge of Big Data

Every second . . .

- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

The Challenge of Big Data

Every second . . .

- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

Unstructured

No XML, no semantic web, no annotation. Often just raw text.

The Challenge of Big Data

Every second . . .

- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

Unstructured

No XML, no semantic web, no annotation. Often just raw text.

Common task: what's going on in this dataset.

- Intelligence analysts
- Brand monitoring
- Journalists
- Humanists

The Challenge of Big Data

Every second . . .

- 600 new blog posts appear
- 34,000 tweets are tweeted
- 30 GB of data uploaded to Facebook

Unstructured

No XML, no semantic web, no annotation. Often just raw text.

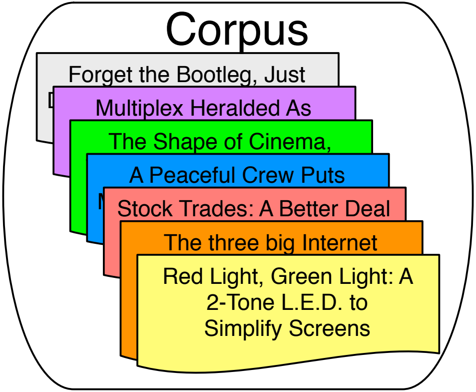
Common task: what's going on in this dataset.

- Intelligence analysts
- Brand monitoring
- Journalists
- Humanists

Common solution: topic models

Topic Models as a Black Box

From an **input corpus** and number of topics K $\rightarrow$ words to topics



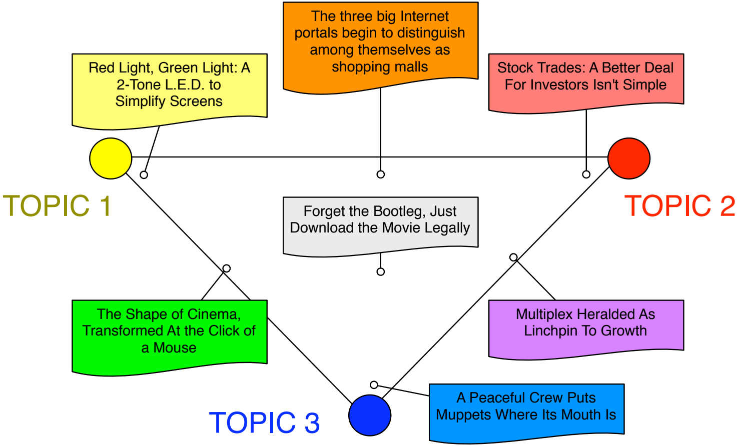
Topic Models as a Black Box

From an input corpus and number of topics K $\rightarrow$ **words to topics**

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Topic Models as a Black Box

From an input corpus and number of topics $K \rightarrow$ words to topics



Word Intrusion

1. Take the highest probability words from a topic

Original Topic
dog, cat, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

dog, cat, apple, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

dog, cat, apple, horse, pig, cow

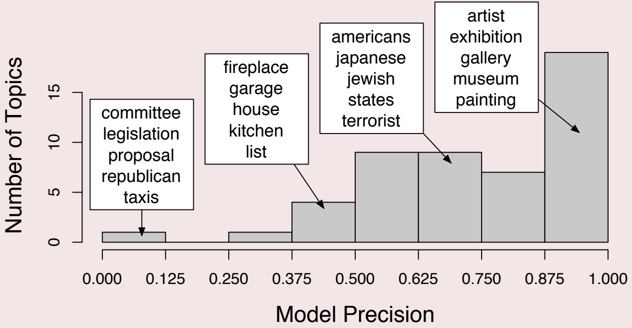
3. We ask users to find the word that doesn't belong

Hypothesis

If the topics are interpretable, users will consistently choose true intruder

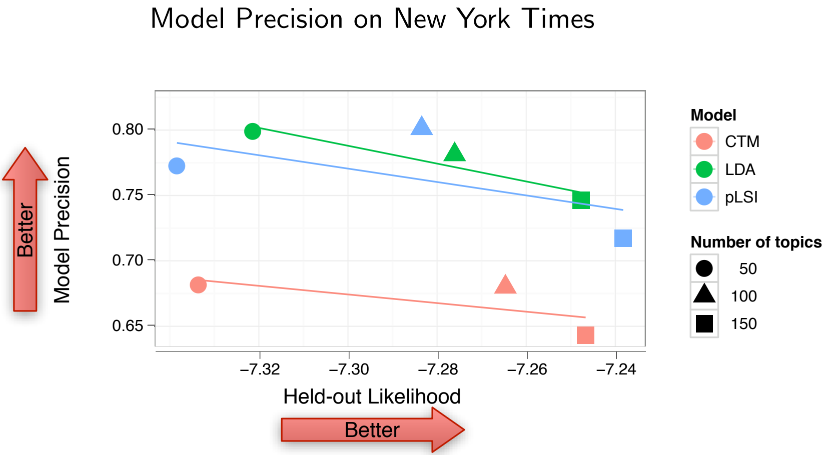
Word Intrusion: Which Topics are Interpretable?

New York Times, 50 Topics



Model Precision: percentage of correct intruders found

Interpretability and Likelihood



within a model, higher likelihood $\neq$ higher interpretability

Interactive Topic Modeling

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive Topic Modeling. Machine Learning, 2014.

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Suggestion

boris, communist, gorbachev, mikhail, russia, russian, soviet, union, yeltsin

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	After
1	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
	⋮
20	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	After
1	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
	⋮
20	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	After
1	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
	⋮
20	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Example: Negative Constraint

Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Example: Negative Constraint

Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial, injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Negative Constraint

spinal_cord, bladder

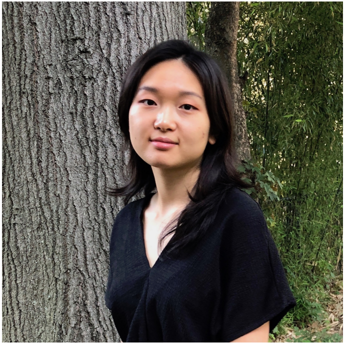
Example: Negative Constraint

Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Topic	Words
318	sci, spinal_cord, spinal_cord_injury, spinal, injury, recovery, motor, reflex, urothelial, injured, functional_recovery, plasticity, locomotor, cervical, locomotion

Negative Constraint

spinal_cord, bladder



Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages

Michelle Yuan, Benjamin Van Durme, and Jordan Boyd-Graber. Neural Information Processing Systems, 2018.


(Source: National Geographic)

- Large text collections often require topic triage quickly in low-resource settings (e.g. natural disaster, political instability).
- Analysts need to examine multilingual text collections, but are scarce in one or more languages.

Generative Approaches

- Polylingual Topic Model [*Mimno et al.* 2009]
- JointLDA [*Jagarlamudi and Daumé* 2010]
- Polylingual Tree-based Topic model [*Hu et al.* 2014b]
- MCTA [*Shi et al.* 2016]

Generative Approaches

- Polylingual Topic Model [*Mimno et al.* 2009]
- JointLDA [*Jagarlamudi and Daumé* 2010]
- Polylingual Tree-based Topic model [*Hu et al.* 2014b]
- MCTA [*Shi et al.* 2016]

These methods are slow, assume extensive knowledge about languages, and preclude human refinement.

farming, livestock, crop, corn, wheat, tractor, cows, 農業 (nóngyè), 牲畜 (shēngchù), 米 (mǐ), 收成 (shōuchéng)

environment, earth, energy, recycling, trash, 碳足跡 (tàn zújì), 太陽能 (tàiyángnéng) 污染 (wūrǎn), 空氣 (kōngqì)

economy, cash, industry, income, services, demand, 經濟 (jīngjì), 收入 (shōurù) 就業率 (jiùyè lǜ), 銀行 (yínháng)

Coral reefs have been damaged by sources of pollution, such as coastal development, deforestation, and agriculture. Destruction of coral reefs could impact food supply, protection, and income …

全球土地總計有三分之一用於生產肉製品與動物製品。如果大豆不需用來餵飼牛群，森林砍伐與土地退化的現象將得以緩解。如果美國將養牛的土地該種大豆，研究人員發現，這一舉措將節約42%的耕地 ……

farming, livestock,
crop, corn, wheat,
tractor, cows,
農業 (nóngyè),
牲畜 (shēngchù),
米 (mǐ),
收成 (shōuchéng)

environment,
earth, energy,
recycling, trash,
碳足跡 (tàn zújì),
太陽能 (tàiyángnéng)
污染 (wūrǎn),
空氣 (kōngqì)

economy, cash,
industry, income,
services, demand,
經濟 (jīngjì),
收入 (shōurù)
就業率 (jiùyè lü),
銀行 (yínháng)

Coral reefs have been damaged by sources of pollution, such as coastal development, deforestation, and agriculture. Destruction of coral reefs could impact food supply, protection, and income …

全球土地總計有三分之一用於生產肉製品與動物製品。如果大豆不需用來餵飼牛群，森林砍伐與土地退化的現象將得以緩解。如果美國將養牛的土地該種大豆，研究人員發現，這一舉措將節約42%的耕地 ……

farming, livestock,
crop, corn, wheat,
tractor, cows,
農業 (nóngyè),
牲畜 (shēngchù),
米 (mǐ),
收成 (shōuchéng)

environment,
earth, energy,
recycling, trash,
碳足跡 (tàn zújì),
太陽能 (tàiyángnéng)
污染 (wūrǎn),
空氣 (kōngqì)

economy, cash,
industry, income,
services, demand,
經濟 (jīngjì),
收入 (shōurù)
就業率 (jiùyè lü),
銀行 (yínháng)

Coral reefs have been damaged by sources of pollution, such as coastal development, deforestation, and agriculture. Destruction of coral reefs could impact food supply, protection, and income …

全球土地總計有三分之一用於生產肉製品與動物製品。如果大豆不需用來餵飼牛群，森林砍伐與土地退化的現象將得以緩解。如果美國將養牛的土地該種大豆，研究人員發現，這一舉措將節約42%的耕地 ……

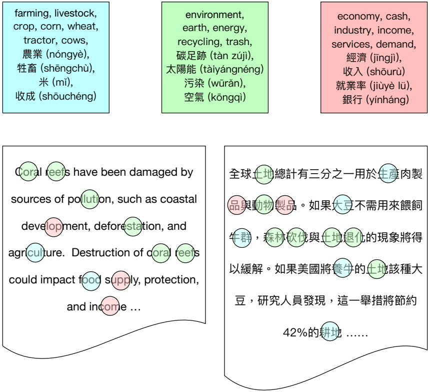
farming, livestock, crop, corn, wheat, tractor, cows, 農業 (nóngyè), 牲畜 (shēngchù), 米 (mǐ), 收成 (shōuchéng)

environment, earth, energy, recycling, trash, 碳足跡 (tàn zújì), 太陽能 (tàiyángnéng) 污染 (wūrǎn), 空氣 (kōngqì)

economy, cash, industry, income, services, demand, 經濟 (jīngjì), 收入 (shōurù) 就業率 (jiùyè lü), 銀行 (yínháng)

Coral reefs have been damaged by sources of pollution, such as coastal development, deforestation, and agriculture. Destruction of coral reefs could impact food supply, protection, and income …

全球土地總計有三分之一用於生產肉製品與動物製品。如果大豆不需用來餵飼牛群，森林砍伐與土地退化的現象將得以緩解。如果美國將養牛的土地該種大豆，研究人員發現，這一舉措將節約42%的耕地 ……



farming, livestock, crop, corn, wheat, tractor, cows, 農業 (nóngyè), 牲畜 (shēngchù), 米 (mǐ), 收成 (shōuchéng)
environment, earth, energy, recycling, trash, 碳足跡 (tàn zújì), 太陽能 (tàiyángnéng) 污染 (wūrǎn), 空氣 (kōngqì)
economy, cash, industry, income, services, demand, 經濟 (jīngjì), 收入 (shōurù) 就業率 (jiùyè lü), 銀行 (yínháng)
Coral reefs have been damaged by sources of pollution, such as coastal development, deforestation, and agriculture. Destruction of coral reefs could impact food supply, protection, and income …
全球土地總計有三分之一用於生產肉製品與動物製品。如果大豆不需用來餵飼牛群，森林砍伐與土地退化的現象將得以緩解。如果美國將養牛的土地該種大豆，研究人員發現，這一舉措將節約42%的耕地 ……

Anchor words

Definition

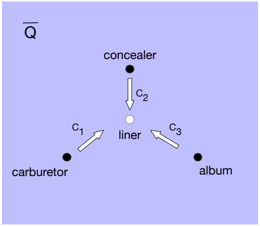
An **anchor word** is a word that appears with *high* probability in one topic but with *low* probability in all other topics.

From Co-occurrence to Topics

- Normally, we want to find $p(\text{word} \mid \text{topic})$ [*Blei et al.* 2003b].
- Instead, what if we can easily find $p(\text{word} \mid \text{topic})$ through using anchor words and conditional word co-occurrence $p(\text{word 2} \mid \text{word 1})$?

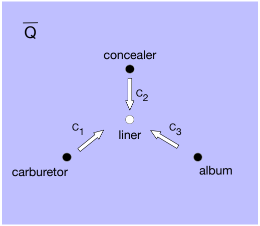
From Co-occurrence to Topics

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i)$$



From Co-occurrence to Topics

$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i)$$



$$\bar{Q}_{\text{liner}} \approx C_1 \bar{Q}_{\text{carburetor}} + C_2 \bar{Q}_{\text{concealer}} + C_3 \bar{Q}_{\text{album}}$$

$$= 0.4 * \begin{bmatrix} 0.3 \\ \cdots \\ 0.1 \end{bmatrix} + 0.2 * \begin{bmatrix} 0.1 \\ \cdots \\ 0.2 \end{bmatrix} + 0.4 * \begin{bmatrix} 0.1 \\ \cdots \\ 0.4 \end{bmatrix}$$

Anchoring

- If an anchor word appears in a document, then its corresponding topic is among the set of topics used to generate document [*Arora et al.* 2012].
- Anchoring algorithm uses word co-occurrence to find anchors and gradient-based inference to recover topic-word distribution [*Arora et al.* 2013].
- Runtime is **fast** because algorithm scales with number of unique word types, rather than number of documents or tokens.

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:
$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:
$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$
2. Given anchor words $s_1, ..., s_K$, approximate co-occurrence distributions:
$$\bar{Q}_i \approx \sum_{k=1}^{K} C_{i,k} \bar{Q}_{s_k} \text{ subject to } \sum_{k=1}^{K} C_{i,k} = 1 \text{ and } C_{i,k} \geq 0.$$

Anchoring

1. Construct co-occurrence matrix from documents with vocabulary of size V:
$$\bar{Q}_{i,j} = p(w_2 = j \mid w_1 = i).$$

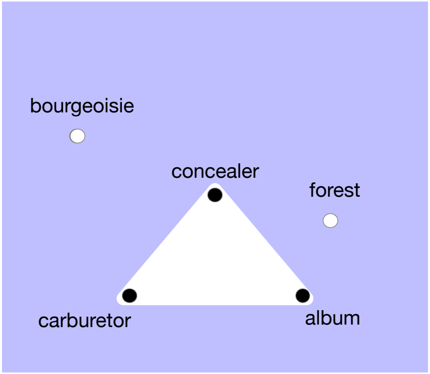
2. Given anchor words $s_1, \ldots, s_K$, approximate co-occurrence distributions:
$$\bar{Q}_i \approx \sum_{k=1}^{K} C_{i,k} \bar{Q}_{s_k} \text{ subject to } \sum_{k=1}^{K} C_{i,k} = 1 \text{ and } C_{i,k} \geq 0.$$

3. Find topic-word matrix:
$$\begin{aligned} A_{i,k} &= p(w = i \mid z = k) \propto p(z = k \mid w = i) p(w = i) \\ &= C_{i,k} \sum_{j=1}^{V} \bar{Q}_{i,j}. \end{aligned}$$

Finding Anchor Words

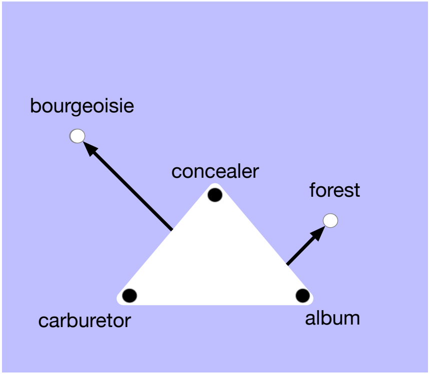
- So far, we assume that anchor words are given.
- How do we find anchor words from documents?

Finding Anchor Words



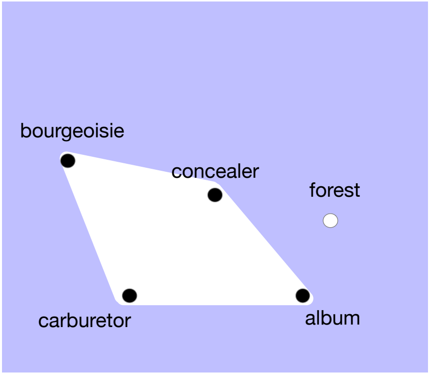
Anchor words are the vertices of the co-occurrence convex hull.

Finding Anchor Words



Anchor words are the vertices of the co-occurrence convex hull.

Finding Anchor Words



Anchor words are the vertices of the co-occurrence convex hull.

Issues with Topic Models

Topics
music concert singer voice chorus songs album
singer pop songs music album chorale jazz
cosmetics makeup eyeliner lipstick foundation primer eyeshadow

Issues with Topic Models

Topics
music concert singer voice chorus songs album
singer pop songs music album chorale jazz
cosmetics makeup eyeliner lipstick foundation primer eyeshadow

Duplicate topics.

Issues with Topic Models

Topics
music band art history literature books earth
bts taehyung idol kpop jin jungkook jimin

Issues with Topic Models

Topics
music band art history literature books earth
bts taehyung idol kpop jin jungkook jimin

Ambiguous topics.
Overly-specific topics.

Interactive Anchoring

- Incorporating interactivity in topic modeling has shown to improve quality of model [*Hu et al.* 2014a].
- Anchoring algorithm offers speed for interactive work, but single anchors are unintuitive to users.
- **Ankura** is an interactive topic modeling system that allows users to choose multiple anchors for each topic [*Lund et al.* 2017].
- After receiving human feedback, **Ankura** only takes a few seconds to update topic model.

Interactive Anchoring

- Incorporating interactivity in topic modeling has shown to improve quality of model [*Hu et al.* 2014a].
- Anchoring algorithm offers speed for interactive work, but single anchors are unintuitive to users.
- **Ankura** is an interactive topic modeling system that allows users to choose multiple anchors for each topic [*Lund et al.* 2017].
- After receiving human feedback, **Ankura** only takes a few seconds to update topic model.

These methods only work for monolingual document collections.

Linking Words

Definition

Language $\mathcal{L}$ is a set of word types w.

Linking Words

Definition

Language $\mathcal{L}$ is a set of word types w.

Definition

Bilingual dictionary $\mathcal{B}$ is a subset of the Cartesian product $\mathcal{L}^{(1)} \times \mathcal{L}^{(2)}$, where $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ are two, different languages.

Linking Words

Definition

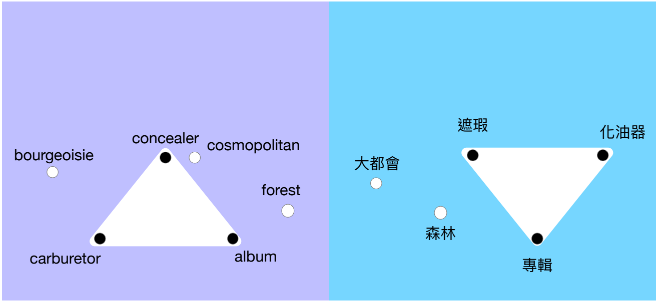
Language $\mathcal{L}$ is a set of word types w.

Definition

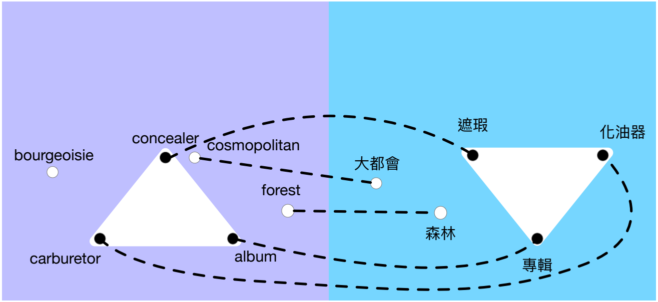
Bilingual dictionary $\mathcal{B}$ is a subset of the Cartesian product $\mathcal{L}^{(1)} \times \mathcal{L}^{(2)}$, where $\mathcal{L}^{(1)}, \mathcal{L}^{(2)}$ are two, different languages.

Idea: If dictionary $\mathcal{B}$ contains entry (w, v), create a link between w and v.

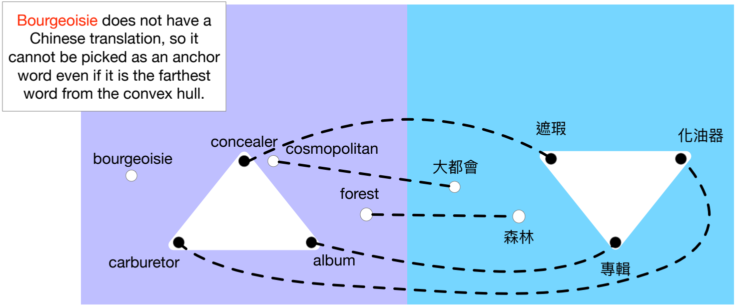
Finding Multilingual Anchors



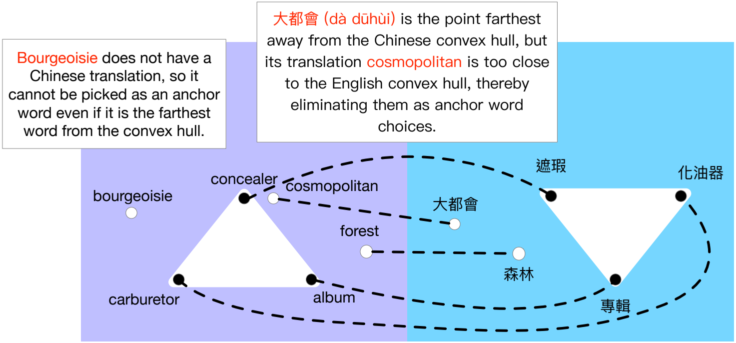
Finding Multilingual Anchors



Finding Multilingual Anchors



Finding Multilingual Anchors

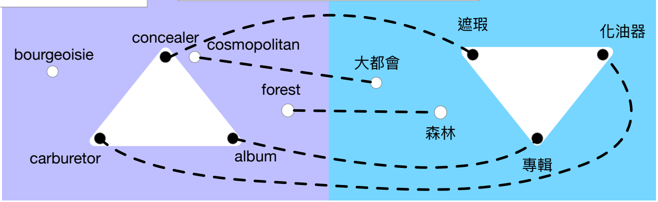


Finding Multilingual Anchors

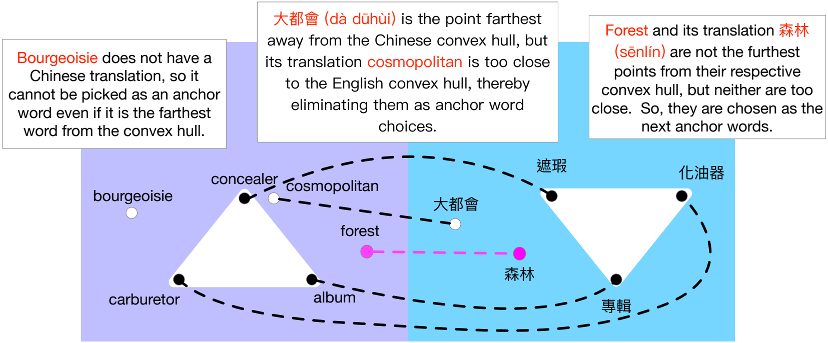
Bourgeoisie does not have a Chinese translation, so it cannot be picked as an anchor word even if it is the farthest word from the convex hull.

大都會 (dà dūhùi) is the point farthest away from the Chinese convex hull, but its translation cosmopolitan is too close to the English convex hull, thereby eliminating them as anchor word choices.

Forest and its translation 森林 (sēnlín) are not the furthest points from their respective convex hull, but neither are too close. So, they are chosen as the next anchor words.



Finding Multilingual Anchors



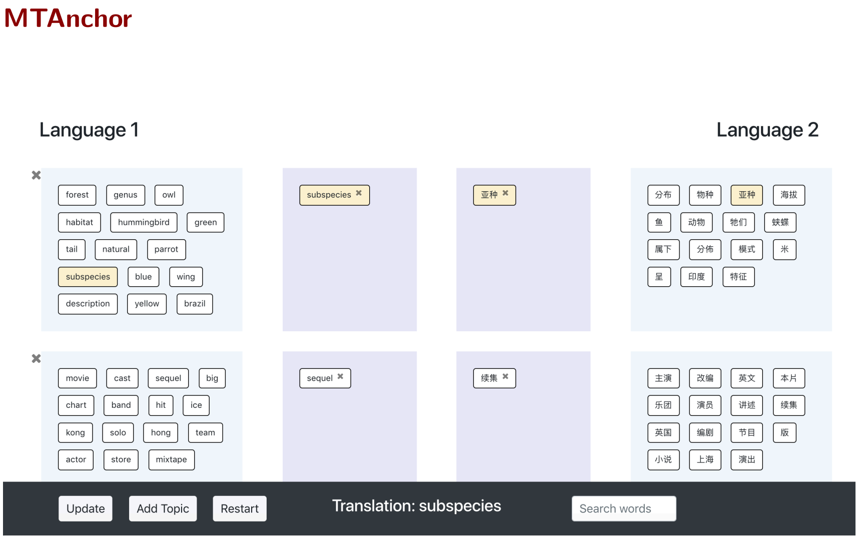
Multilingual Anchoring

1. Given a dictionary, create links between words that are translations of each other.
2. Select an anchor word for each language such that the words are linked and span of anchor words is maximized.
3. Once anchor words are found, separately find topic-word distributions for each language.

- What if dictionary entries are scarce or inaccurate?
- What if topics aren't aligned properly across languages?

- What if dictionary entries are scarce or inaccurate?
- What if topics aren't aligned properly across languages?

Incorporate human-in-the-loop topic modeling tools.


MTAnchor
Language 1
Language 2
forest
genus
owl
habitat
hummingbird
green
tail
natural
parrot
subspecies
blue
wing
description
yellow
brazil
subspecies
亚种
分布
物种
亚种
海拔
鱼
动物
牠们
蛱蝶
属下
分佈
模式
米
呈
印度
特征
movie
cast
sequel
big
chart
band
hit
ice
kong
solo
hong
team
actor
store
mixtape
sequel
续集
主演
改编
英文
本片
乐团
演员
讲述
续集
英国
编剧
节目
版
小说
上海
演出
Update
Add Topic
Restart
Translation: subspecies
Search words

Experiments

Datasets:

1. Wikipedia articles (EN, ZH)
2. Amazon reviews (EN, ZH)
3. LORELEI documents (EN, SI)

Experiments

Metrics:

1. Classification accuracy
   - Intra-lingual: train topic model on documents in one language and test on other documents in the *same* languages
   - Cross-lingual: train topic model on documents in one language and test on other documents in a *different* language.
2. Topic coherence [*Lau et al.* 2014].
   - Intrinsic: use the trained documents as the reference corpus to measure local interpretability.
   - Extrinsic: use a large dataset (i.e. entire Wikipedia) as the reference corpus to measure global interpretability.

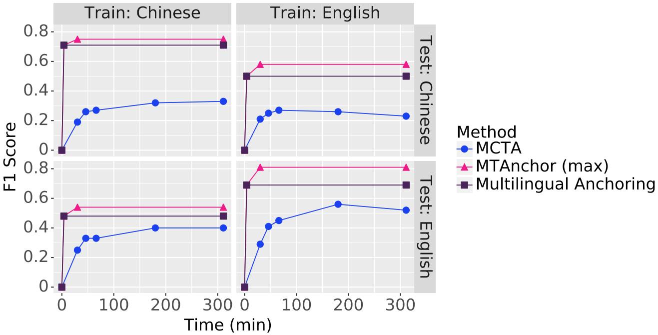
Comparing Models

		Classification accuracy			
Dataset	Method	EN-I	ZH-I SI-I	EN-C	ZH-C SI-C
Wikipedia	Multilingual anchoring	69.5%	71.2%	50.4%	47.8%
	MTAnchor (maximum)	**80.7%**	**75.3%**	**57.6%**	**54.5%**
	MTAnchor (median)	69.5%	71.4%	50.3%	47.2%
	MCTA	51.6%	33.4%	23.2%	39.8%
Amazon	Multilingual anchoring	**59.8%**	**61.1%**	**51.7%**	**53.2%**
	MCTA	49.5%	50.6%	50.3%	49.5%
LORELEI	Multilingual anchoring	**20.8%**	**32.7%**	**24.5%**	**24.7%**
	MCTA	13.0%	26.5%	4.1%	15.6%

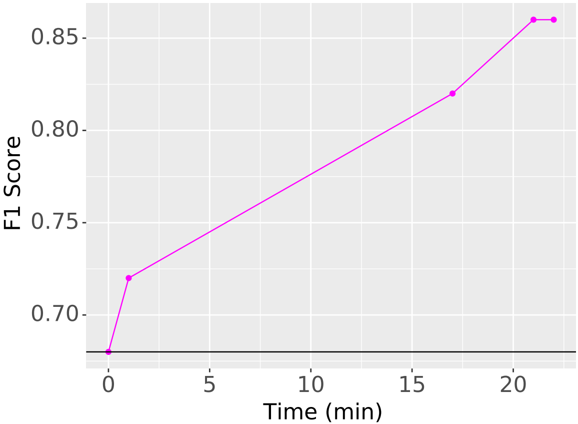
Comparing Models

Dataset	Method	Topic coherence EN-I	ZH-I SI-I	EN-E	ZH-E SI-E
Wikipedia	Multilingual anchoring	0.14	0.18	0.08	0.13
	MTAnchor (maximum)	**0.20**	**0.20**	**0.10**	**0.15**
	MTAnchor (median)	0.14	0.18	0.08	0.13
	MCTA	0.13	0.09	0.00	0.04
Amazon	Multilingual anchoring	**0.07**	**0.06**	**0.03**	**0.05**
	MCTA	-0.03	0.02	0.02	0.01
LORELEI	Multilingual anchoring	0.08	0.00	0.03	n/a
	MCTA	**0.13**	0.00	**0.04**	n/a

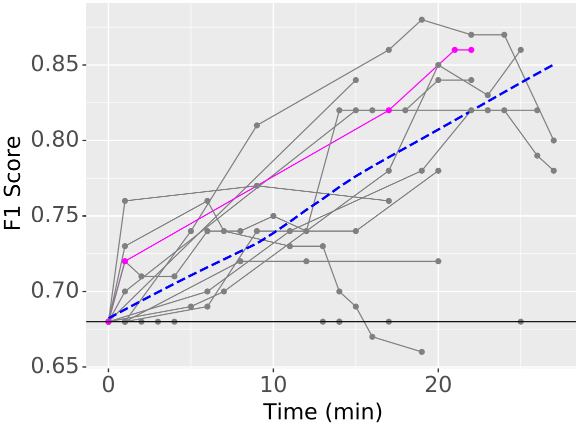
Multilingual Anchoring Is Much Faster



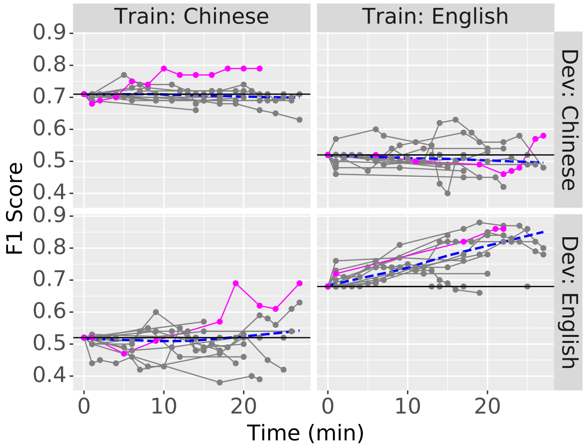
Improving Topics Through Interactivity



Improving Topics Through Interactivity



Improving Topics Through Interactivity



Comparing Topics

Dataset	Method	Topic
Wikipedia	MCTA	dog san movie mexican fighter novel california 主演 改編 本 小説 拍攝 角色 戰士
	Multilingual anchoring	**adventure** daughter bob kong hong robert movie 主演 改編 本片 飾演 **冒險** 講述 編劇
	MTAnchor	**kong hong movie** office martial box reception 主演 改編 飾演 本片 **演員 編劇** 講述
Amazon	MCTA	woman food eat person baby god chapter 來貨 頂頂 水 耳機 貨物 張傑 傑 同樣
	Multilingual anchoring	eat diet food recipe **healthy** lose weight **健康** 幫 吃 身體 全面 同事 中醫
LORELEI	MCTA	help need floodrelief please families needed victim
	Multilingual anchoring	aranayake warning landslide site missing nbro areas

Why Not Use Deep Learning?

- Neural networks are data-hungry and unsuitable for low-resource languages
- Deep learning models take long amounts of time to train
- Pathologies of neural models make interpretation difficult [*Feng et al.* 2018]

Summary

- Anchoring algorithm can be applied in multilingual settings.
- People can provide helpful linguistic or cultural knowledge to construct better multilingual topic models.

Future Work

- Apply human-in-the-loop algorithms to other tasks in NLP.
- Better understand the effect of human feedback on cross-lingual representation learning.

ALTO: Active Learning with Topic Overviews for Speeding Label Induction and Document Labeling

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. Association for Computational Linguistics, 2016.

Many Documents

Sort into Categories

Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best

Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best

Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples
  - Compare classifier labels to expert judgements

Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples (don't tell them how many labels)
  - Compare classifier labels to expert judgements

Evaluation

- User study
- 40 minutes
- Sort documents into categories
- What information / interface helps best
  - Train a classifier on human examples (don't tell them how many labels)
  - Compare classifier labels to expert judgements (purity)

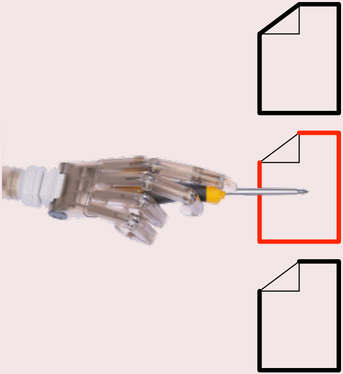
$$\text{purity}(\mathbf{U}, \mathbf{G}) = \frac{1}{N} \sum_{l} \max_{j} |U_l \cap G_j|, \tag{1}$$

Which is more Useful?

Who should drive?

Which is more Useful?

Active Learning



Topic Models

TOPIC 1

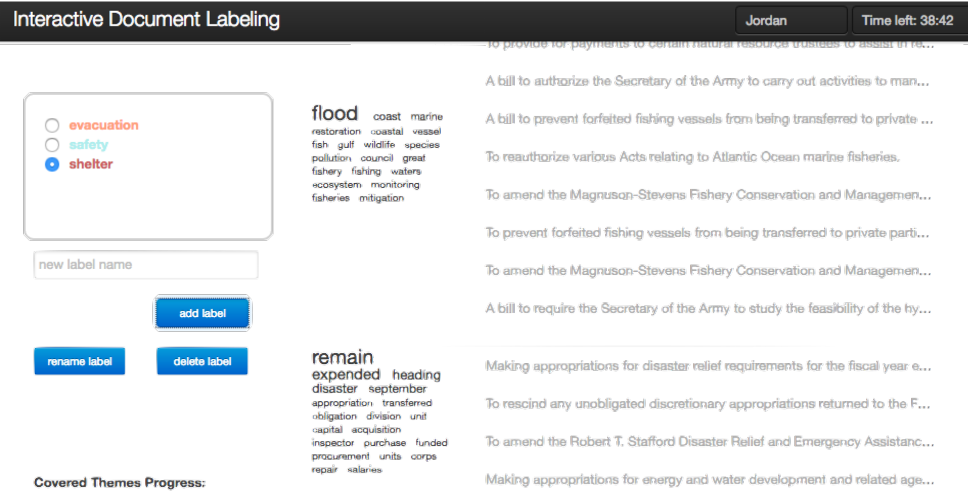
computer, technology, system, service, site, phone, internet, machine

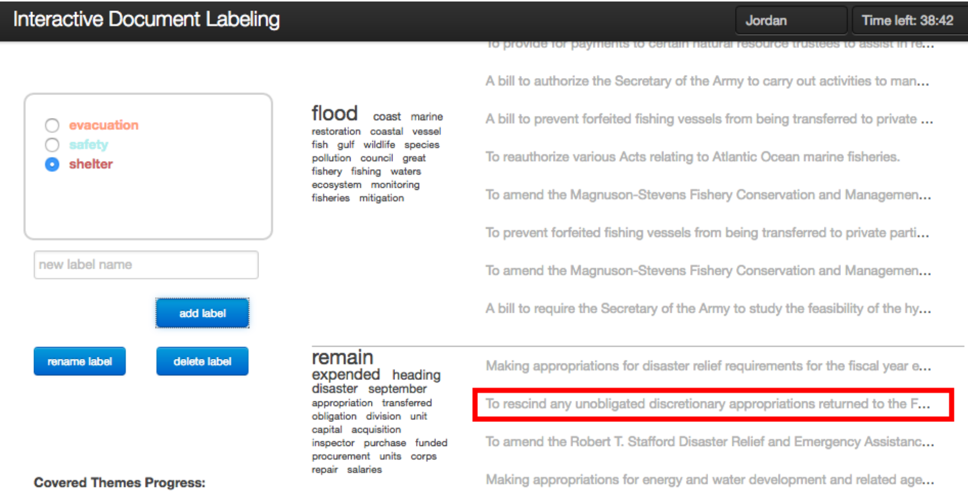
TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

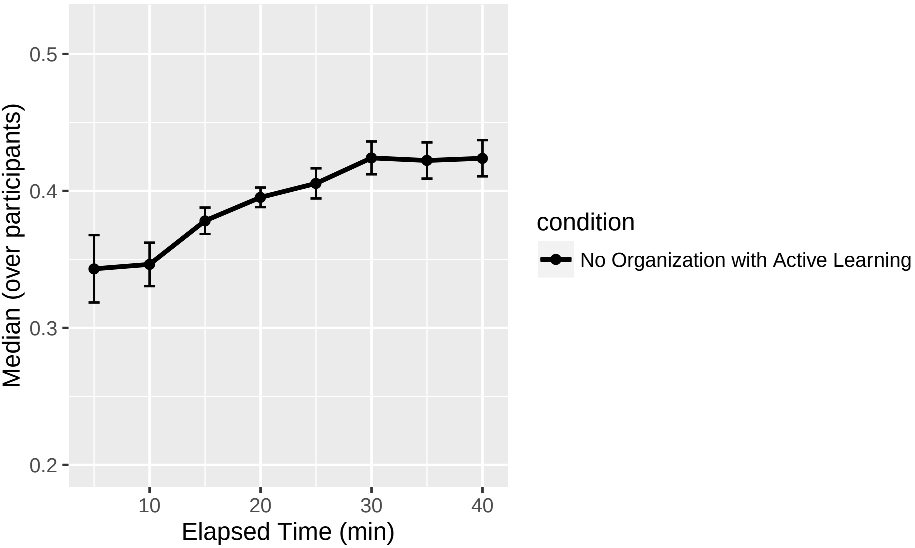
TOPIC 3

play, film, movie, theater, production, star, director, stage

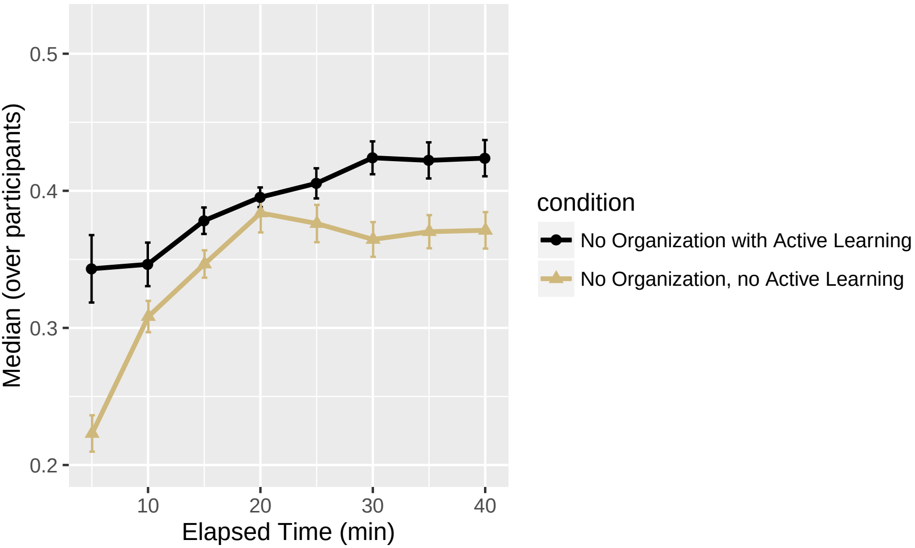

Interactive Document Labeling
Jordan
Time left: 38:42
evacuation
safety
shelter
new label name
add label
rename label
delete label
Covered Themes Progress:
flood coast marine restoration coastal vessel fish gulf wildlife species pollution council great fishery fishing waters ecosystem monitoring fisheries mitigation
A bill to authorize the Secretary of the Army to carry out activities to man...
A bill to prevent forfeited fishing vessels from being transferred to private ...
To reauthorize various Acts relating to Atlantic Ocean marine fisheries.
To amend the Magnuson-Stevens Fishery Conservation and Managemen...
To prevent forfeited fishing vessels from being transferred to private parti...
To amend the Magnuson-Stevens Fishery Conservation and Managemen...
A bill to require the Secretary of the Army to study the feasibility of the hy...
remain expended heading disaster september appropriation transferred obligation division unit capital acquisition inspector purchase funded procurement units corps repair salaries
Making appropriations for disaster relief requirements for the fiscal year e...
To rescind any unobligated discretionary appropriations returned to the F...
To amend the Robert T. Stafford Disaster Relief and Emergency Assistanc...
Making appropriations for energy and water development and related age...



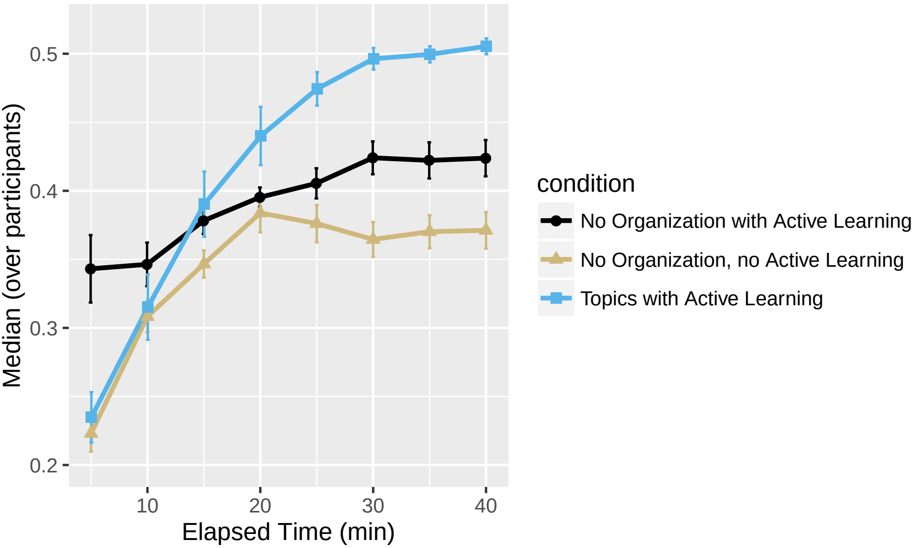
Direct users to document



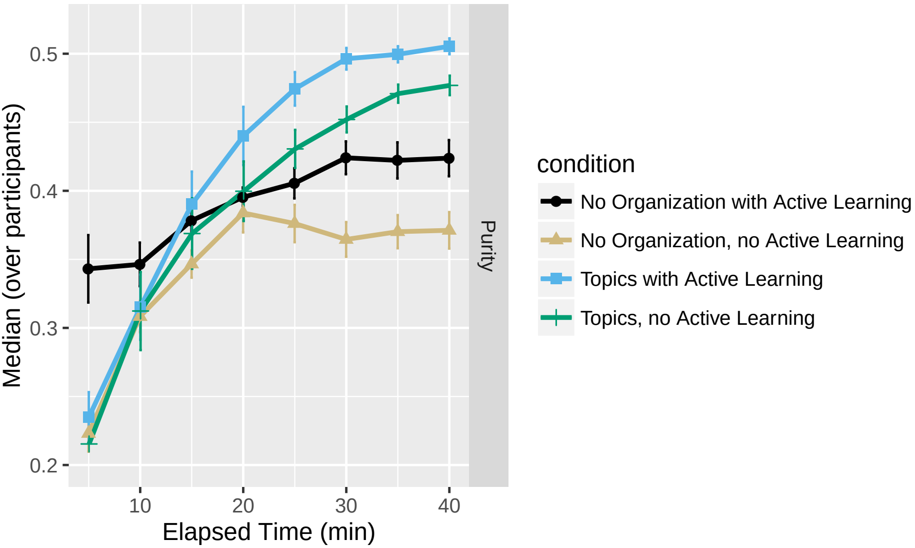
Active learning if time is short



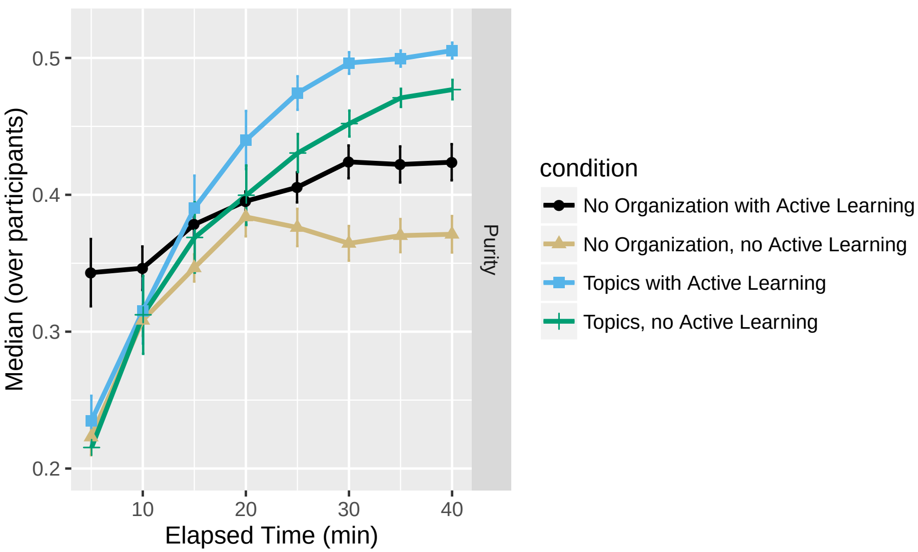
Better than status quo



Active learning can help topic models



Topic models help users understand the collection



Moral: machines and humans together (if you let them)

Ongoing and Future Work

- Embedding interactivity in applications
- Visualizations to measure machine learning explainability
- Using morphology in infinite representations
- Multilingual analysis

Ongoing and Future Work

- Embedding interactivity in applications
- Visualizations to measure machine learning explainability
- Using morphology in infinite representations
- Multilingual analysis

Thanks

Collaborators

Yuening Hu (UMD), Ke Zhai (UMD), Viet-An Nguyen (UMD), Dave Blei (Princeton), Jonathan Chang (Facebook), Philip Resnik (UMD), Christiane Fellbaum (Princeton), Jerry Zhu (Wisconsin), Sean Gerrish (Sift), Chong Wang (CMU), Dan Osherson (Princeton), Sinead Williamson (CMU)

Funders

References

Alvin Grissom II, He He, Jordan Boyd-Graber, and John Morgan.
2014.
Don't until the final verb wait: Reinforcement learning for simultaneous machine translation.
In *Empirical Methods in Natural Language Processing*.

Sanjeev Arora, Rong Ge, and Ankur Moitra.
2012.
Learning topic models–going beyond SVD.
In *Foundations of Computer Science (FOCS)*.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu.
2013.
A practical algorithm for topic modeling with provable guarantees.
In *Proceedings of the International Conference of Machine Learning*.

David M. Blei, Andrew Ng, and Michael Jordan.
2003a.
Latent Dirichlet allocation.
Machine Learning Journal, 3.

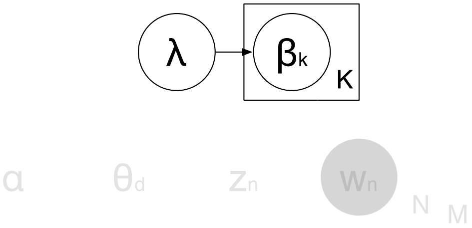
David M. Blei, Andrew Y. Ng, and Michael I. Jordan.
2003b.
Latent Dirichlet allocation.
Machine Learning Journal.

Jordan L. Boyd-Graber, Sonya S. Nikolova, Karyn A. Moffatt, Kenrick C. Kin, Joshua Y. Lee, Lester W. Mackey, Marilyn M. Tremaine, and Maria M. Klawe.
2006.
Participatory design with proxies: Developing a desktop-PDA system to support people with aphasia.
In *International Conference on Human Factors in Computing Systems*.

Latent Dirichlet Allocation: A Generative Model

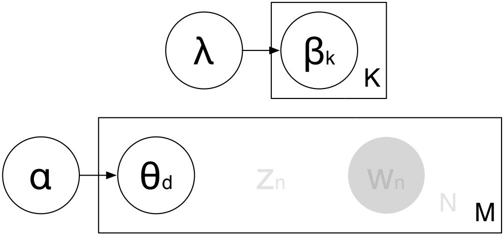
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA) [*Blei et al.* 2003a], a fully Bayesian version of pLSI [*Hofmann* 1999], probabilistic version of LSA [*Landauer and Dumais* 1997]

Latent Dirichlet Allocation: A Generative Model



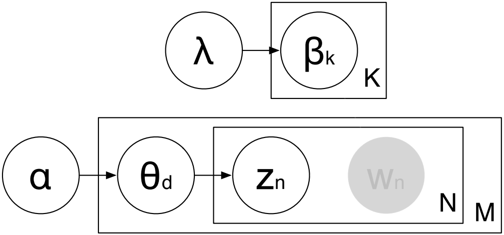
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Latent Dirichlet Allocation: A Generative Model



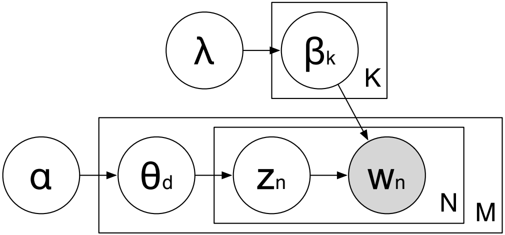
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Latent Dirichlet Allocation: A Generative Model



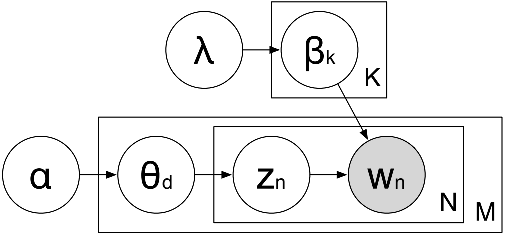
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

Latent Dirichlet Allocation: A Generative Model



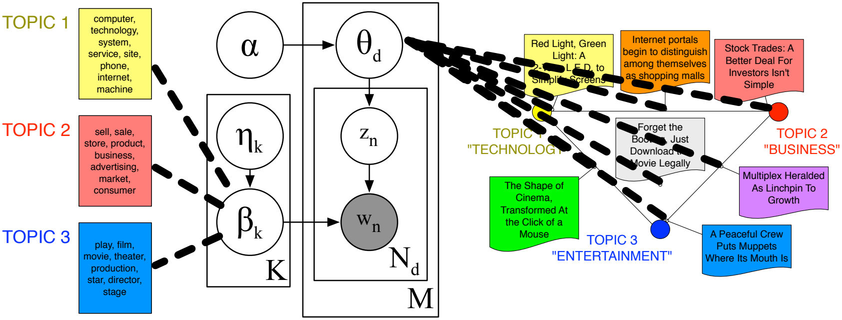
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Latent Dirichlet Allocation: A Generative Model



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

We use statistical inference to uncover the most likely unobserved


TOPIC 1
computer, technology, system, service, site, phone, internet, machine
TOPIC 2
sell, sale, store, product, business, advertising, market, consumer
TOPIC 3
play, film, movie, theater, production, star, director, stage
α
θd
ηk
zn
βk
wn
K
Nd
M
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Internet portals begin to distinguish among themselves as shopping malls
Stock Trades: A Better Deal For Investors Isn't Simple
TOPIC 1 "TECHNOLOGY"
Forget the Bootleg, Just Download the Movie Legally
TOPIC 2 "BUSINESS"
Multiplex Heralded As Linchpin To Growth
The Shape of Cinema, Transformed At the Click of a Mouse
TOPIC 3 "ENTERTAINMENT"
A Peaceful Crew Puts Muppets Where Its Mouth Is

Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \tag{2}$$

Inference

- We are interested in posterior distribution

$$p(Z|X,\Theta) \tag{2}$$

- Here, latent variables are topic assignments z and topics θ. X is the words (divided into documents), and Θ are hyperparameters to Dirichlet distributions: α for topic proportion, λ for topics.

$$p(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{w},\alpha,\lambda) \tag{3}$$

Inference

- We are interested in posterior distribution

$$p(Z|X,\Theta) \tag{2}$$

- Here, latent variables are topic assignments z and topics θ. X is the words (divided into documents), and Θ are hyperparameters to Dirichlet distributions: α for topic proportion, λ for topics.

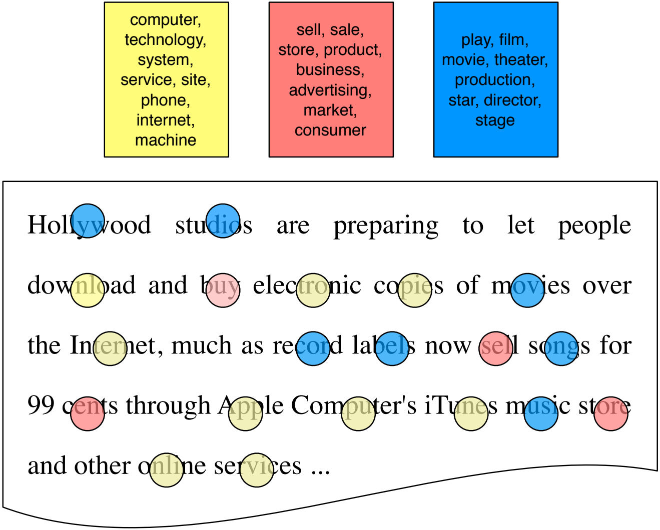
$$p(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{w},\alpha,\lambda) \tag{3}$$

$$p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\lambda) = \prod_k p(\beta_k|\lambda)\prod_d p(\theta_d|\alpha)\prod_n p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{z_{d,n}})$$

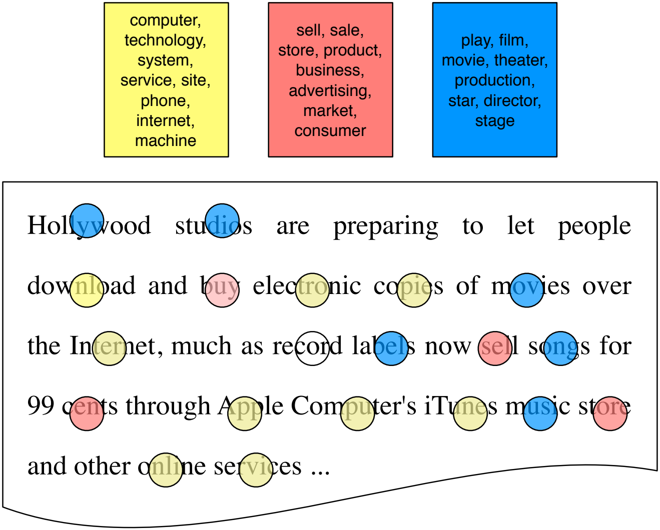
Gibbs Sampling

- A form of Markov Chain Monte Carlo
- Chain is a sequence of random variable states
- Given a state $\{z_1, \ldots z_N\}$ given certain technical conditions, drawing $z_k \sim p(z_1, \ldots z_{k-1}, z_{k+1}, \ldots z_N | X, \Theta)$ for all k (repeatedly) results in a Markov Chain whose stationary distribution *is* the posterior.
- For notational convenience, call $\boldsymbol{z}$ with $z_{d,n}$ removed $\boldsymbol{z}_{-d,n}$

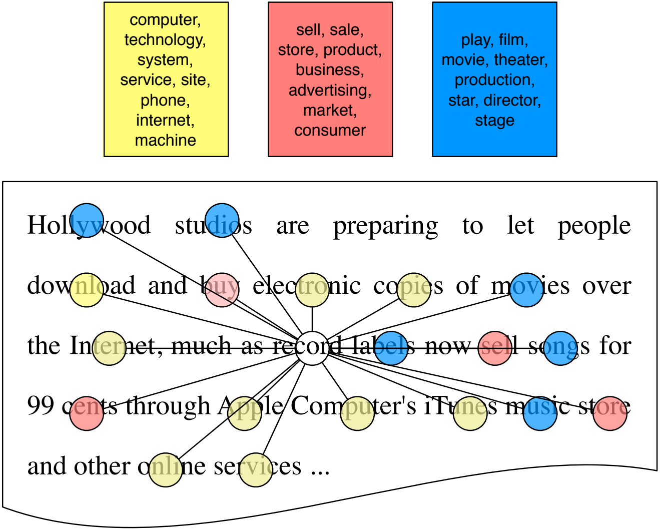
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

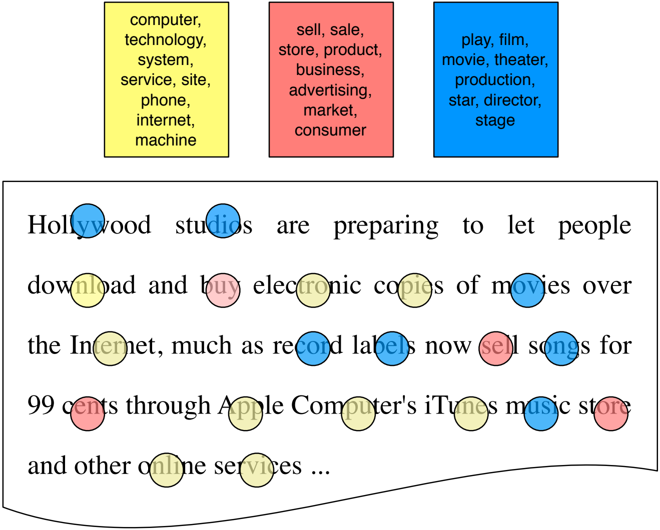
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

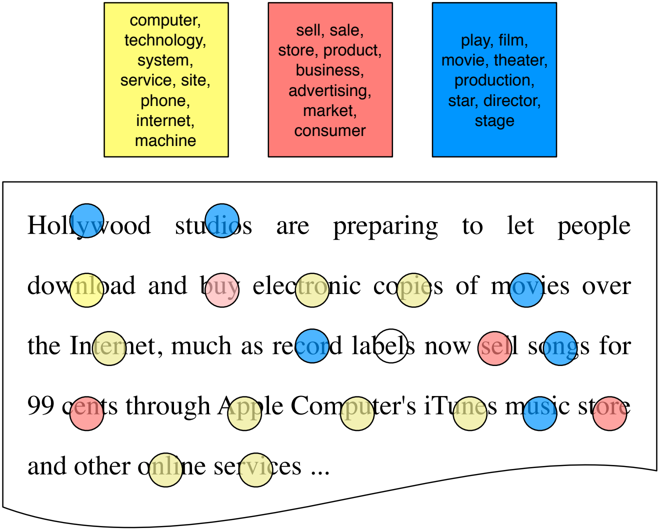
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

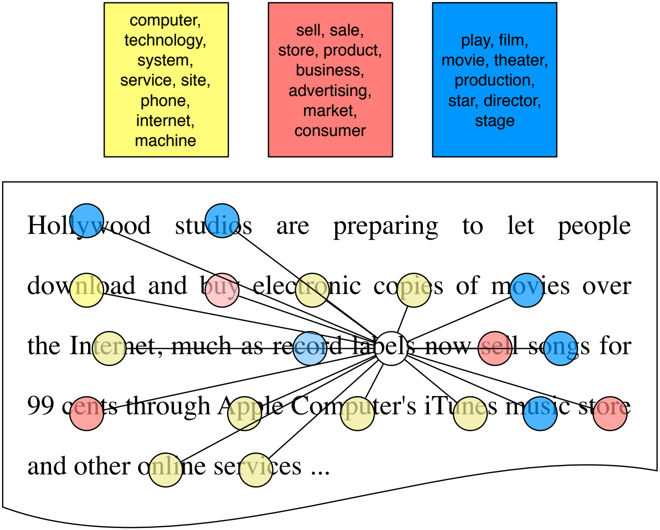
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

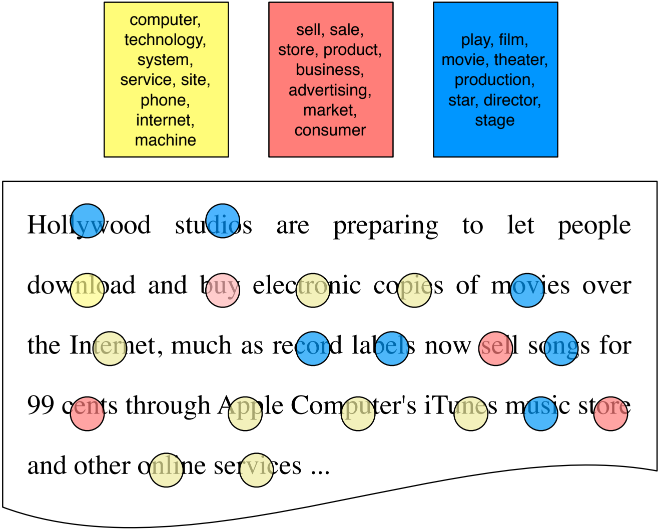
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}{p(\boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}$$

Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}{p(\boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}$$

- The topics and per-document topic proportions are integrated out / marginalized / collapsed
- Let $n_{d,i}$ be the number of words taking topic i in document d. Let $v_{k,w}$ be the number of times word w is used in topic k.

$$= \frac{\int_{\theta_d} \left(\prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i} - 1}\right) \theta_d^{\alpha_k + n_{d,i}} d\theta_d \int_{\beta_k} \left(\prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) \beta_{k,w_{d,n}}^{\lambda_i + v_{k,i}} d\beta_k}{\int_{\theta_d} \left(\prod_i \theta_d^{\alpha_i + n_{d,i} - 1}\right) d\theta_d \int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) d\beta_k}$$

Gibbs Sampling

- For LDA, we will sample the topic assignments
- The topics and per-document topic proportions are integrated out / marginalized / Rao-Blackwellized
- Thus, we want:

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \alpha, \lambda) = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k = \frac{\prod_i^V \mid \beta_i + v_{k,i}}{\mid \sum_i^V \beta_i + v_{k,i}}$$

Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i+v_{k,i}-1}\right) d\beta_k = \frac{\prod_i^V \mid \beta_i + v_{k,i}}{\mid \sum_i^V \beta_i + v_{k,i}}$$

- So we can simplify

$$\frac{\int_{\theta_d} \left(\prod_{i\neq k} \theta_d^{\alpha_i+n_{d,i}-1}\right) \theta_d^{\alpha_k+n_{d,i}} d\theta_d \int_{\beta_k} \left(\prod_{i\neq w_{d,n}} \beta_{k,i}^{\lambda_i+v_{k,i}-1}\right) \beta_{k,w_{d,n}}^{\lambda_i+v_{k,i}} d\beta_k}{\int_{\theta_d} \left(\prod_i \theta_d^{\alpha_i+n_{d,i}-1}\right) d\theta_d \int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i+v_{k,i}-1}\right) d\beta_k} =$$

$$\frac{\frac{\mid \alpha_k+n_{d,k}+1}{\mid \sum_i^K \alpha_i+n_{d,i}+1} \prod_{i\neq k}^K \mid \alpha_k + n_{d,k}}{\frac{\prod_i^K \mid \alpha_i+n_{d,i}}{\mid \sum_i^K \alpha_i+n_{d,i}}} \frac{\frac{\mid \lambda_{w_{d,n}}+v_{k,w_{d,n}}+1}{\mid \sum_i^V \lambda_i+v_{k,i}+1} \prod_{i\neq w_{d,n}}^V \mid \lambda_k + v_{k,w_{d,n}}}{\frac{\prod_i^V \mid \lambda_i+v_{k,i}}{\mid \sum_i^V \lambda_i+v_{k,i}}}$$

Gamma Function Identity

$$z = \frac{\Gamma(z+1)}{\Gamma(z)} \tag{4}$$

$$\frac{\frac{| \, \alpha_k + n_{d,k} + 1}{| \sum_i^K \alpha_i + n_{d,i} + 1} \prod_{i \neq k}^K | \, \alpha_k + n_{d,k}}{\frac{\prod_i^K | \, \alpha_i + n_{d,i}}{| \sum_i^K \alpha_i + n_{d,i}}} \frac{\frac{| \, \lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1}{| \sum_i^V \lambda_i + v_{k,i} + 1} \prod_{i \neq w_{d,n}}^V | \, \lambda_k + v_{k,w_{d,n}}}{\frac{\prod_i^V | \, \lambda_i + v_{k,i}}{| \sum_i^V \lambda_i + v_{k,i}}}$$

$$= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{5}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

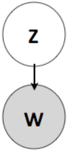
Sample Document

Etruscan	trade	price	temple	market

Sample Document

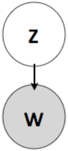
Etruscan	trade	price	temple	market

Randomly Assign Topics

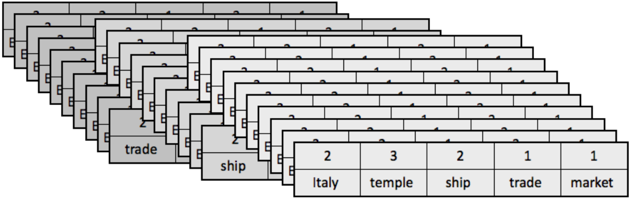


3	2	1	3	1
Etruscan	trade	price	temple	market

Randomly Assign Topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	50	0	1

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	50	0	1

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

We want to sample this word . . .

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

We want to sample this word . . .

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**8**	1
...			

Decrement its count

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**7**	1
...			

What is the conditional distribution for this topic?

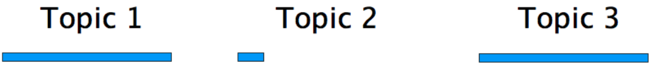
3	?	1	3	1
Etruscan	trade	price	temple	market

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market



Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
trade	10	7	1

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

Sampling Equation

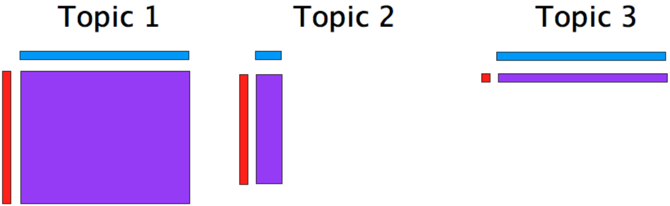
$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market

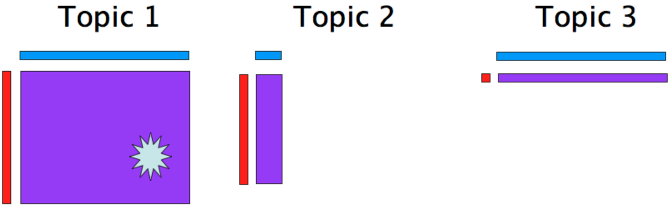
Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market



Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market



Update counts

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10**	7	1
...			

Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market

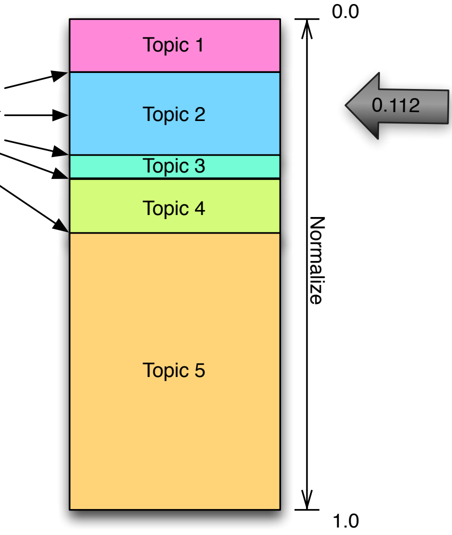
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
...			

Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market

Details: how to sample from a distribution

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$



Algorithm

1. For each iteration i:
   1.1 For each document d and word n currently assigned to z_{old}:
      1.1.1 Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
      1.1.2 Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
      1.1.3 Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

Naïve Implementation

Algorithm

1. For each iteration i:
   1.1 For each document d and word n currently assigned to z_{old}:
      1.1.1 Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
      1.1.2 Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
      1.1.3 Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

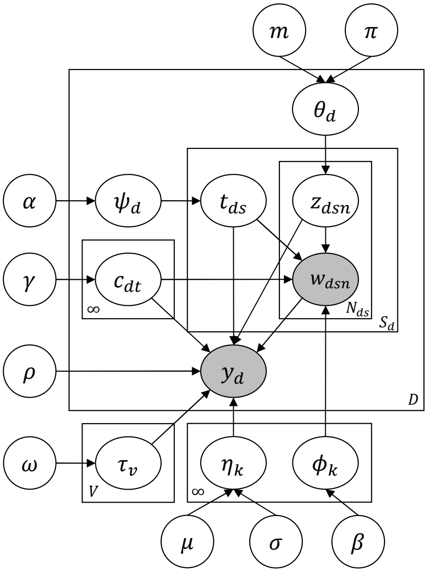
Desiderata

- Hyperparameters: Sample them too (slice sampling)
- Initialization: Random
- Sampling: Until likelihood converges
- Lag / burn-in: Difference of opinion on this
- Number of chains: Should do more than one

Available implementations

- Mallet (http://mallet.cs.umass.edu)
- LDAC (http://www.cs.princeton.edu/ blei/lda-c)
- Topicmod (http://code.google.com/p/topicmod)

SHLDA Model


m
π
θ_d
α
ψ_d
t_{ds}
z_{dsn}
γ
c_{dt}
w_{dsn}
∞
N_{ds}
S_d
ρ
y_d
D
ω
τ_v
V
η_k
ϕ_k
∞
μ
σ
β

Infvoc Classification Accuracy

$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$			
		fixvoc	vb-dict	45.514

Table: Classification accuracy based on 50 topic features extracted from *20 newsgroups* data.

Infvoc Classification Accuracy

$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$	*fixvoc*	vb-dict	45.514
		fixvoc-hash	vb-dict	52.525

Table: Classification accuracy based on 50 topic features extracted from *20 newsgroups* data.

Topics learned with *hashing* are no longer interpretable, they can only be used as features.

Infvoc Classification Accuracy

$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$	*infvoc*	$\alpha^\beta = 3k$ $T = 40k$ $U = 10$	52.683
		fixvoc	vb-dict	45.514
		fixvoc-hash	vb-dict	52.525

Table: Classification accuracy based on 50 topic features extracted from *20 newsgroups* data.

Topics learned with *hashing* are no longer interpretable, they can only be used as features.

Infvoc Classification Accuracy

$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$	*infvoc*	$\alpha^\beta = 3k$ $T = 40k$ $U = 10$	52.683
		fixvoc	vb-dict	45.514
		fixvoc	vb-null	49.390
		fixvoc	hybrid-dict	46.720
		fixvoc	hybrid-null	50.474
		fixvoc-hash	vb-dict	52.525
		fixvoc-hash	vb-full $T = 30k$	51.653
		fixvoc-hash	hybrid-dict	50.948
		fixvoc-hash	hybrid-full $T = 30k$	50.948

Table: Classification accuracy based on 50 topic features extracted from *20 newsgroups* data.

Topics learned with *hashing* are no longer interpretable, they can only be used as features.

Infvoc Classification Accuracy

$S = 155$	$\tau_0 = 64$ $\kappa = 0.6$	*infvoc*	$\alpha^\beta = 3k$ $T = 40k$ $U = 10$	52.683
		fixvoc	vb-dict	45.514
		fixvoc	vb-null	49.390
		fixvoc	hybrid-dict	46.720
		fixvoc	hybrid-null	50.474
		fixvoc-hash	vb-dict	52.525
		fixvoc-hash	vb-full $T = 30k$	51.653
		fixvoc-hash	hybrid-dict	50.948
		fixvoc-hash	hybrid-full $T = 30k$	50.948
	dtm-dict $tcv = 0.001$			62.845

Table: Classification accuracy based on 50 topic features extracted from *20 newsgroups* data.

Topics learned with *hashing* are no longer interpretable, they can only be used as features.

Unassign $(d, n, w_{d,n}, z_{d,n} = k)$

1: $T: \ T_{d,k} \leftarrow T_{d,k} - 1$
2: If $w_{d,n} \notin \Omega^{old}$,
  $P: \ P_{k,w_{d,n}} \leftarrow P_{k,w_{d,n}} - 1$
3: Else: suppose $w_{d,n} \in \Omega^{old}_m$,
  $P: \ P_{k,m} \leftarrow P_{k,m} - 1$
  $W: \ W_{k,m,w_{d,n}} \leftarrow W_{k,m,w_{d,n}} - 1$

SparseLDA

$$
\begin{aligned}
p(z = k | Z_-, w) &\propto (\alpha_k + n_{k|d}) \frac{\beta + n_{w|k}}{\beta V + n_{\cdot|k}} \\
&\propto \underbrace{\frac{\alpha_k \beta}{\beta V + n_{\cdot|k}}}_{s_{\mathrm{LDA}}} + \underbrace{\frac{n_{k|d} \beta}{\beta V + n_{\cdot|k}}}_{r_{\mathrm{LDA}}} + \underbrace{\frac{(\alpha_k + n_{k|d}) n_{w|k}}{\beta V + n_{\cdot|k}}}_{q_{\mathrm{LDA}}}
\end{aligned}
\tag{6}
$$

Tree-based sampling

$$
\begin{aligned}
& p(z_{d,n} = k, l_{d,n} = \lambda | Z_-, L_-, w_{d,n}) \\
& \quad \propto (\alpha_k + n_{k|d}) \prod_{(i \to j) \in \lambda} \frac{\beta_{i \to j} + n_{i \to j|k}}{\sum_{j'} (\beta_{i \to j'} + n_{i \to j'|k})}
\end{aligned} \tag{7}
$$

Factorizing Tree-Based Prior

$$p(z = k|Z_{-}, w) \propto (\alpha_k + n_{k|d})\frac{\beta + n_{w|k}}{\beta V + n_{\cdot|k}} \tag{8}$$

$$\propto \underbrace{\frac{\alpha_k \beta}{\beta V + n_{\cdot|k}}}_{s_{\mathrm{LDA}}} + \underbrace{\frac{n_{k|d}\beta}{\beta V + n_{\cdot|k}}}_{r_{\mathrm{LDA}}} + \underbrace{\frac{(\alpha_k + n_{k|d})n_{w|k}}{\beta V + n_{\cdot|k}}}_{q_{\mathrm{LDA}}}$$

Factorizing Tree-Based Prior

$$
\begin{aligned}
p(z = k | Z_-, w) &\propto (\alpha_k + n_{k|d}) \frac{\beta + n_{w|k}}{\beta V + n_{\cdot|k}} \\
&\propto \underbrace{\frac{\alpha_k \beta}{\beta V + n_{\cdot|k}}}_{s_{\mathrm{LDA}}} + \underbrace{\frac{n_{k|d}\beta}{\beta V + n_{\cdot|k}}}_{r_{\mathrm{LDA}}} + \underbrace{\frac{(\alpha_k + n_{k|d}) n_{w|k}}{\beta V + n_{\cdot|k}}}_{q_{\mathrm{LDA}}}
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
s &= \sum_{k,\lambda} \frac{\alpha_k \prod_{(i\to j)\in\lambda} \beta_{i\to j}}{\prod_{(i\to j)\in\lambda} \sum_{j'} (\beta_{i\to j'} + n_{i\to j'|k})} \\
&\leq \sum_{k,\lambda} \frac{\alpha_k \prod_{(i\to j)\in\lambda} \beta_{i\to j}}{\prod_{(i\to j)\in\lambda} \sum_{j'} \beta_{i\to j'}} = s'.
\end{aligned}
\tag{9}
$$

```
for word w in this document do
    sample = rand() ∗(s′ + r + q)
    if sample < s′ then
        compute s
        sample = sample ∗(s + r + q)/(s′ + r + q)
        if sample < s then
            return topic k and path λ sampled from s
        end if
        sample − = s
    else
        sample − = s′
    end if
    if sample < r then
        return topic k and path λ sampled from r
    end if
    sample − = r
    return topic k and path λ sampled from q
end for
```

Number of Topics				
	T50	T100	T200	T500
Naive	5.700	12.655	29.200	71.223
Fast	4.935	9.222	17.559	40.691
Fast-RB	2.937	4.037	5.880	8.551
Fast-RB-sD	2.675	3.795	5.400	8.363
Fast-RB-sW	2.449	3.363	4.894	7.404
Fast-RB-sDW	2.225	3.241	4.672	7.424
Number of Correlations				
	C50	C100	C200	C500
Naïve	11.166	12.586	13.000	15.377
Fast	8.889	9.165	9.177	8.079
Fast-RB	3.995	4.078	3.858	3.156
Fast-RB-sD	3.660	3.795	3.593	3.065
Fast-RB-sW	3.272	3.363	3.308	2.787
Fast-RB-sDW	3.026	3.241	3.091	2.627