Man vs. Data: Domain Knowledge + Latent Dirichlet Allocation

Xiaojin Zhu

Department of Computer Sciences
University of Wisconsin-Madison

2011
Acknowledgments

David M. Andrzejewski
did most work; now postdoc at Livermore National Laboratory

Mark Craven  Ben Liblit  Ben Recht
University of Wisconsin-Madison

The Wisconsin Alumni Research Foundation, NSF CAREER Award IIS-0953219,
AFOSR FA9550-09-1-0313, and NIH/NLM R01 LM07050
New Year’s Eve, Times Square
The Wish Corpus

[Goldberg et al., NAACL 2009]

- Peace on earth
- www.accupros.com- Website Optimization & Marketing
- own a brewery
- The safe return of my friends in Iraq
- find a cure for cancer
- To lose weight and get a boyfriend
- I Hope Barack Obama Wins the Presidency
- To win the lottery!
Corpus-wide word frequencies
Some Topics by Latent Dirichlet Allocation (LDA)
[Blei et al., JMLR 2003]

\[ p(\text{word} \mid \text{topic}) \]

“troops”

“election”

“love”
Some Topics by Latent Dirichlet Allocation (LDA)
[Blei et al., JMLR 2003]

\[ p(\text{word} \mid \text{topic}) \]

Major applications: exploratory data analysis
- Research trends [Wang & McCallum, 2006]
- Scientific influence [Gerrish & Blei, 2009]
- Matching papers to reviewers [Mimno & McCallum, 2007]
Quick Statistics Review

Dirichlet

[20, 5, 5]
Quick Statistics Review

Dirichlet
[20, 5, 5]

Multinomial
[0.6, 0.15, 0.25]
Quick Statistics Review

**Dirichlet**

$$[20, 5, 5]$$

- A
- B
- C

**Multinomial**

$$[0.6, 0.15, 0.25]$$

- A
- B
- C

**Observed counts**

$$[3, 1, 2]$$

A, A, B, C, A, C
Quick Statistics Review

**Dirichlet**

\[ [20, 5, 5] \]

**Multinomial**

\[ [0.6, 0.15, 0.25] \]

**Observed counts**

\[ [3, 1, 2] \]

A, A, B, C, A, C

CONJUGACY!
A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\text{argmax}_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w | \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z | w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta | w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\text{argmax}_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)

(WID Optimization 2011) Knowledge + LDA
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:
For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta)$, $\text{argmax}_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for \( p(\phi, \theta, z, w \mid \alpha, \beta) \):

For each topic \( t \)

\( \phi_t \sim \text{Dirichlet}(\beta) \)

For each document \( d \)

\( \theta \sim \text{Dirichlet}(\alpha) \)

For each word position in \( d \)

\( \text{topic } z \sim \text{Multinomial}(\theta) \)
\( \text{word } w \sim \text{Multinomial}(\phi_z) \)

Inference goals: \( p(z \mid w, \alpha, \beta) \), \( \arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta) \)
(reminder on top)
Latent Dirichlet Allocation (LDA) Review

A generative model for $p(\phi, \theta, z, w \mid \alpha, \beta)$:

For each topic $t$

$$\phi_t \sim \text{Dirichlet}(\beta)$$

For each document $d$

$$\theta \sim \text{Dirichlet}(\alpha)$$

For each word position in $d$

- topic $z \sim \text{Multinomial}(\theta)$
- word $w \sim \text{Multinomial}(\phi_z)$

Inference goals: $p(z \mid w, \alpha, \beta), \arg\max_{\phi, \theta} p(\phi, \theta \mid w, \alpha, \beta)$

(reminder on top)
When LDA Alone is not Enough

- LDA is unsupervised
- Often domain experts have knowledge in addition to data, want better topics $\phi$
When LDA Alone is not Enough

- LDA is unsupervised
- Often domain experts have knowledge in addition to data, want better topics $\phi$
- There has been many specialized LDA variants
- This talk: how to do “Logic + LDA”
Example Application: Statistical Software Debugging

[Andrzejewski et al., ECML 2007], [Zheng et al., ICML 2006]

- Insert predicates into a software:

```c
int x = my_func();
if (x > 5) {
    branch_42_true++
    ...
} else {
    branch_42_false++
    ...
}
```
Example Application: Statistical Software Debugging

[Andrzejewski et al., ECML 2007], [Zheng et al., ICML 2006]

- Insert predicates into a software:

```c
int x = my_func();
if (x > 5) {
    branch_42_true++
    ...
} else {
    branch_42_false++
    ...
}
```

- predicates $\rightarrow$ words $w$
- a software run $\rightarrow$ doc $d$
- we know which runs crashed and which didn’t (extra knowledge)
- the hope: run LDA on crashed runs, some topics $\phi$ will correspond to “buggy behaviors”
Hope Crashed

Normal software usage topics dominate. No “bug” topic.
Hope Crashed

Normal software usage topics dominate. No “bug” topic.

Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.
Hope Crashed

Normal software usage topics dominate. No “bug” topic.

Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.
  
- Synthetic success (left) and crashed (right) runs
Hope Crashed

Normal software usage topics dominate. No “bug” topic.

Toy example:

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.

- Synthetic success (left) and crashed (right) runs

- LDA topics on crashed runs
Model success and crashed runs jointly with $T$ topics:

- fix $t < T$
- for all words in success runs, $z \in \{1 \ldots t\}$ (restricted)
- for all words in crashed runs, $z \in \{1 \ldots T\}$

New hope: $\phi_1 \ldots \phi_t$ usage topics, $\phi_{t+1} \ldots \phi_T$ bug topics
New Hope Succeeds

- Actual usage (left) and bug (right) topics. Each pixel is a predicate.
- Synthetic success (left) and crashed (right) runs
- LDA topics on crashed runs
- ΔLDA topics on success and crashed runs
New Hope Succeeds

- 15 bug1 runs
- 15 bug2 runs

- 56 bug1 runs
- 144 bug2 runs

- 12 bug1 runs
- 174 bug2 runs

- 254 bug1 runs
- 106 bug3 runs
- 147 bug4 runs
- 329 bug5 runs
- 206 bug8 runs
- 186 other runs

exif
grep
gzip

moss
Generalize $\Delta$LDA to Topic-in-Set

[Andrzejewski & Zhu, NAACL’09 WS]

- The domain knowledge:
  - For each word position $i$ in corpus, we are given a set $C_i \subset \{1 \ldots T\}$, such that $z_i \in C_i$.
Generalize ΔLDA to Topic-in-Set

[Andrzejewski & Zhu, NAACL’09 WS]

- The domain knowledge:
  - For each word position \( i \) in corpus, we are given a set \( C_i \subset \{1 \ldots T\} \), such that \( z_i \in C_i \).
- Very easy to implement in collapsed Gibbs sampling:

\[
P(z_i = v|z_{-i}, w, \alpha, \beta) \propto \left( \frac{n_{-i,v}^{(d)} + \alpha}{\sum_{u}^{T} n_{-i,u}^{(d)} + \alpha} \right) \left( \frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^{W} (\beta + n_{-i,v}^{(w_i)})} \delta(v \in C_i) \right)
\]

- \( n_{-i,v}^{(d)} \) is the number of times topic \( v \) is used in document \( d \)
- \( n_{-i,v}^{(w_i)} \) is the number of times word \( w_i \) is generated by topic \( v \)
- both excluding position \( i \)

- Easy to relax the hard constraints
Further Generalize Topic-in-Set to Logic

- Topic-in-set is $Z(i, c_1) \lor Z(i, c_2) \lor \ldots \lor Z(i, c_k)$ where $C_i = \{c_1, \ldots, c_k\}$
Further Generalize Topic-in-Set to Logic

- Topic-in-set is $Z(i, c_1) \lor Z(i, c_2) \lor \ldots \lor Z(i, c_k)$ where $C_i = \{c_1, \ldots, c_k\}$
- Fold.all = First-Order Logic latent Dirichlet ALLocation
  - easy for domain experts to write rules
  - can describe very general domain knowledge
    - can encode many existing LDA variants
  - efficient inference
Further Generalize Topic-in-Set to Logic

- Topic-in-set is $Z(i, c_1) \lor Z(i, c_2) \lor \ldots \lor Z(i, c_k)$ where $C_i = \{c_1, \ldots, c_k\}$
- Fold.all = First-Order Logic latent Dirichlet ALLocation
  - easy for domain experts to write rules
  - can describe very general domain knowledge
    - can encode many existing LDA variants
  - efficient inference
- A hybrid Markov Logic Network (MLN) [Wang & Domingos 2008] [Richardson & Domingos 2006], but with fast stochastic optimization
Domain Knowledge in Logic

- Key hidden predicate: $Z(i, t)$ TRUE if topic $z_i = t$
- Observed predicates (anything goes):
  - $W(i, v)$ TRUE if word $w_i = v$
  - $D(i, j)$ TRUE if word position $i$ is in document $j$
  - HasLabel($j, l$) TRUE if document $j$ has label $l$
  - $S(i, k)$ TRUE if word position $i$ is in document $k$
  - ...
Domain Knowledge in Logic

- Key hidden predicate: $Z(i, t)$ TRUE if topic $z_i = t$
- Observed predicates (anything goes):
  - $W(i, v)$ TRUE if word $w_i = v$
  - $D(i, j)$ TRUE if word position $i$ is in document $j$
  - HasLabel$(j, l)$ TRUE if document $j$ has label $l$
  - $S(i, k)$ TRUE if word position $i$ is in document $k$
  - ...

- Domain knowledge-base $(\lambda_1, \psi_1) \ldots (\lambda_L, \psi_L)$
  - rules $\psi$
  - positive weights $\lambda$ indicate strength of rule

Example:

$\lambda_1 = 1, \psi_1 = \text{“}\forall i : W(i, \text{embryo}) \Rightarrow Z(i, 3)\text{“}$

$\lambda_2 = 100, \psi_2 = \text{“}\forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))\text{“}$
Propositionalization

- Let $G(\psi)$ be all ground formulas of $\psi$.
  - $\psi = “\forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg(Z(i, t) \land Z(j, t))”$
  - One ground formula $g \in G(\psi)$ is $W(123, \text{movie}) \land W(456, \text{film}) \Rightarrow \neg(Z(123, 9) \land Z(456, 9))$
Propositionalization

- Let $G(\psi)$ be all ground formulas of $\psi$.
  - $\psi = \forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg (Z(i, t) \land Z(j, t))$’
  - One ground formula $g \in G(\psi)$ is
    $W(123, \text{movie}) \land W(456, \text{film}) \Rightarrow \neg (Z(123, 9) \land Z(456, 9))$
- $|G(\psi)|$ combinatorial.
- Let
  $$\mathbb{1}_g(z) = \begin{cases} 
  1, & \text{if } g \text{ is TRUE under } z \\
  0, & \text{otherwise.}
  \end{cases}$$
Fold.all = LDA + MLN
[Andrzejewski et al. IJCAI 2011]

\[
p(z, \phi, \theta | w, \alpha, \beta)
\propto \left( \prod_t p(\phi_t | \beta) \right) \left( \prod_j p(\theta_j | \alpha) \right) \left( \prod_i \phi_{z_i}(w_i)\theta_{d_i}(z_i) \right)
\]
Fold.all = LDA + MLN

[Andrzejewski et al. IJCAI 2011]

\[ p(z, \phi, \theta \mid w, \alpha, \beta) \]
\[ \propto \left( \prod_{t}^{T} p(\phi_t \mid \beta) \right) \left( \prod_{j}^{D} p(\theta_j \mid \alpha) \right) \left( \prod_{i}^{N} \phi_{z_i}(w_i) \theta_{d_i}(z_i) \right) \]
\[ \times \exp \left[ \sum_{l}^{L} \sum_{g \in G(\psi_l)} \lambda_l \mathbb{1}_g(z) \right] \]
Fold.all Inference

MAP estimate, non-convex objective

\[ Q(z, \phi, \theta) \equiv \sum_{t}^{T} \log p(\phi_t | \beta) + \sum_{j}^{D} \log p(\theta_j | \alpha) \]

\[ + \sum_{i}^{N} \log \phi_{z_i}(w_i) \theta_{d_i}(z_i) + \sum_{l}^{L} \sum_{g \in G(\psi_l)} \lambda_l 1_g(z) \]

Alternating optimization. Repeat:

- fixing \( z \), let \( (\phi^*, \theta^*) \leftarrow \text{argmax}_{\phi,\theta} Q(z, \phi, \theta) \) (easy)
- fixing \( \phi, \theta \), let \( z^* \leftarrow \text{argmax}_z Q(z, \phi, \theta) \) (integer)
Optimizing $z$ Step 1: Relax $1_g(z)$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$

\[-Z(i, 1) \land Z(j, 2)\]
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$  \hspace{1cm} $\neg Z(i, 1) \land Z(j, 2)$
2. Remove negations $(\neg g)^+$  \hspace{1cm} $(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$
Optimizing $z$ Step 1: Relax $1_g(z)$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$
   $$\neg Z(i, 1) \land Z(j, 2)$$

2. Remove negations $(\neg g)_+$
   $$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$$

3. Numeric $z_{it} \in \{0, 1\}$
   $$(z_{i2} + z_{i3}) z_{j2}$$
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$
2. Remove negations $(\neg g)_+$
3. Numeric $z_{it} \in \{0, 1\}$
4. Polynomial $\mathbb{1}_g(z)$

\[
\neg Z(i, 1) \land Z(j, 2) \\
(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2) \\
(z_{i2} + z_{i3})z_{j2} \\
1 - (z_{i2} + z_{i3})z_{j2}
\]
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$$g = Z(i, 1) \lor \neg Z(j, 2), \text{ and } t \in \{1, 2, 3\}$$

1. Take complement $\neg g$
2. Remove negations $(\neg g)_+$
3. Numeric $z_{it} \in \{0, 1\}$
4. Polynomial $\mathbb{1}_g(z)$
5. Relax discrete $z_{it}$

$\neg Z(i, 1) \land Z(j, 2)$

$$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$$

$$(z_{i2} + z_{i3}) z_{j2}$$

$1 - (z_{i2} + z_{i3}) z_{j2}$

$z_{it} \in \{0, 1\} \rightarrow z_{it} \in [0, 1]$
Optimizing $z$ Step 1: Relax $\mathbb{1}_g(z)$

$g = Z(i, 1) \lor \neg Z(j, 2)$, and $t \in \{1, 2, 3\}$

1. Take complement $\neg g$
2. Remove negations $(\neg g)_+$
3. Numeric $z_{it} \in \{0, 1\}$
4. Polynomial $\mathbb{1}_g(z)$
5. Relax discrete $z_{it}$

$\neg Z(i, 1) \land Z(j, 2)$

$(Z(i, 2) \lor Z(i, 3)) \land Z(j, 2)$

$(z_{i2} + z_{i3})z_{j2}$

$1 - (z_{i2} + z_{i3})z_{j2}$

$z_{it} \in \{0, 1\} \rightarrow z_{it} \in [0, 1]$
Optimizing $z$: Step 1: Relax $1_g(z)$

\[ g = \mathbb{Z}(i, 1) \lor \neg \mathbb{Z}(j, 2), \text{ and } t \in \{1, 2, 3\} \]

1. Take complement $\neg g$
2. Remove negations $(\neg g)_+$
3. Numeric $z_{it} \in \{0, 1\}$
4. Polynomial $1_g(z)$
5. Relax discrete $z_{it}$

\[
1_g(z) = 1 - \prod_{g_i \neq \emptyset} \left( \sum_{\mathbb{Z}(i, t) \in (\neg g)_+} z_{it} \right)
\]

\[ (\neg \mathbb{Z}(i, 1) \land \mathbb{Z}(j, 2)) \land (\mathbb{Z}(i, 2) \lor \mathbb{Z}(i, 3)) \land \mathbb{Z}(j, 2)\]

\[ 1 - (z_{i2} + z_{i3})z_{j2} \]

\[ z_{it} \in \{0, 1\} \rightarrow z_{it} \in [0, 1] \]
Optimizing $z$ Step 2: Stochastic Optimization

- $\sum_{l}^{L} |G(\psi_l)| + NT$ terms in $Q$ related to $z$:

  $\max_{z} \sum_{l}^{L} \sum_{g \in G(\psi_l)} \lambda_{l} \mathbb{1}_{g}(z) + \sum_{i}^{N} \sum_{t}^{T} z_{it} \log \phi_{t}(w_{i}) \theta_{d_{i}}(t)$

  s.t. $z_{it} \geq 0, \sum_{t}^{T} z_{it} = 1$.

- Entropic Mirror Descent [Beck & Teboulle, 2003]. Repeat:
  - select a term $f$ at random
  - descent with decreasing step size $\eta$

  $$z_{it} \leftarrow \frac{z_{it} \exp(\eta \nabla_{z_{it}} f)}{\sum_{t'} z_{it'} \exp(\eta \nabla_{z_{it'}} f)}$$
Example: Movie Reviews

\[ \forall i, j, t : W(i, \text{movie}) \land W(j, \text{film}) \Rightarrow \neg (Z(i, t) \land Z(j, t)) \]
Interactive Topic Modeling
[Andrzejewski et al. ICML 2009]
LDA Topics on Wish Corpus:

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|--------------------------------------------------------|
| 0     | love i you me and will forever that with hope           |
| 1     | and health for happiness family good my friends       |
| 2     | year new happy a this have and everyone years         |
| 3     | that is it you we be t are as not s will can           |
| 4     | my to get job a for school husband s that into        |
| 5     | to more of be and no money stop live people           |
| 6     | to our the home for of from end safe all come         |
| 7     | to my be i find want with love life meet man           |
| 8     | a and healthy my for happy to be have baby            |
| 9     | a 2008 in for better be to great job president        |
| 10    | i wish that would for could will my lose can           |
| 11    | peace and for love all on world earth happiness       |
| 12    | may god in all your the you s of bless 2008           |
| 13    | the in to of world best win 2008 go lottery           |
| 14    | me a com this please at you call 4 if 2 www           |
Interactive Topic Modeling

[Andrzejewski et al. ICML 2009]

LDA Topics on Wish Corpus:

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|---------------------------------------------------------|
| 0     | love i you me and will forever that with hope           |
| 1     | and health for happiness family good my friends        |
| 2     | year new happy a this have and everyone years          |
| 3     | that is it you we be t are as not s will can            |
| 4     | my to get job a for school husband s that into         |
| 5     | to more of be and no money stop live people            |
| 6     | to our the home for of from end safe all come          |
| 7     | to my be i find want with love life meet man           |
| 8     | a and healthy my for happy to be have baby             |
| 9     | a 2008 in for better be to great job president        |
| 10    | i wish that would for could will my lose can           |
| 11    | peace and for love all on world earth happiness        |
| 12    | may god in all your the you s of bless 2008            |
| 13    | the in to of world best win 2008 go lottery            |
| 14    | me a com this please at you call 4 if 2 www           |
isolate(50 stopwords from existing topics)

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|----------------------------------------------------------|
| 0     | love forever marry happy together mom back             |
| 1     | health happiness good family friends prosperity        |
| 2     | life best live happy long great time ever wonderful    |
| 3     | out not up do as so what work don was like              |
| 4     | go school cancer into well free cure college           |
| 5     | no people stop less day every each take children       |
| 6     | home safe end troops iraq bring war husband house      |
| 7     | love peace true happiness hope joy everyone dreams     |
| 8     | happy healthy family baby safe prosperous everyone     |
| 9     | better job hope president paul great ron than person   |
| 10    | make money lose weight meet finally by lots hope married|
|       | Isolate and to for a the year in new all my 2008       |
| 12    | god bless jesus loved know everyone love who loves     |
| 13    | peace world earth win lottery around save              |
| 14    | com call if 4 2 www u visit 1 3 email yahoo            |

(WID Optimization 2011) Knowledge + LDA
isolate(50 stopwords from existing topics)

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|---------------------------------------------------------|
| 0     | love forever marry happy together mom back             |
| 1     | health happiness good family friends prosperity        |
| 2     | life best live happy long great time ever wonderful    |
| 3     | out not up do as so what work don was like             |
| MIXED | go school cancer into well free cure college           |
| 5     | no people stop less day every each take children       |
| 6     | home safe end troops iraq bring war husband house      |
| 7     | love peace true happiness hope joy everyone dreams     |
| 8     | happy healthy family baby safe prosperous everyone     |
| 9     | better job hope president paul great ron than person   |
| 10    | make money lose weight meet finally by lots hope married|
| Isolate | and to for a the year in new all my 2008             |
| 12    | god bless jesus loved know everyone love who loves     |
| 13    | peace world earth win lottery around save              |
| 14    | com call if 4 2 www u visit 1 3 email yahoo            |
| Isolate | i to wish my for and a be that the in                 |
\[
\text{split}([\text{cancer free cure well}], [\text{go school into college}])
\]

<table>
<thead>
<tr>
<th>0</th>
<th>love forever happy together marry fall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>health happiness family good friends</td>
</tr>
<tr>
<td>2</td>
<td>life happy best live love long time</td>
</tr>
<tr>
<td>3</td>
<td>as not do so what like much don was</td>
</tr>
<tr>
<td>4</td>
<td>out make money house up work grow able</td>
</tr>
<tr>
<td>5</td>
<td>people no stop less day every each take</td>
</tr>
<tr>
<td>6</td>
<td>home safe end troops iraq bring war husband</td>
</tr>
<tr>
<td>7</td>
<td>love peace happiness true everyone joy</td>
</tr>
<tr>
<td>8</td>
<td>happy healthy family baby safe prosperous</td>
</tr>
<tr>
<td>9</td>
<td>better president hope paul ron than person</td>
</tr>
<tr>
<td>10</td>
<td>lose meet man hope boyfriend weight finally</td>
</tr>
</tbody>
</table>

**Isolate**
- and to for a the year in new all my 2008
- god bless jesus loved everyone know loves
- peace world earth win lottery around save
- com call if 4 www 2 u visit 1 email yahoo 3

**Split**
- job go school great into good college
- mom husband cancer hope free son well
\textbf{split([cancer free cure well],[go school into college])}

<table>
<thead>
<tr>
<th>LOVE</th>
<th>love forever happy together marry fall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>health happiness family good friends</td>
</tr>
<tr>
<td>2</td>
<td>life happy best live love long time</td>
</tr>
<tr>
<td>3</td>
<td>as not do so what like much don was</td>
</tr>
<tr>
<td>4</td>
<td>out make money house up work grow able</td>
</tr>
<tr>
<td>5</td>
<td>people no stop less day every each take</td>
</tr>
<tr>
<td>6</td>
<td>home safe end troops iraq bring war husband</td>
</tr>
<tr>
<td>7</td>
<td>love peace happiness true everyone joy</td>
</tr>
<tr>
<td>8</td>
<td>happy healthy family baby safe prosperous</td>
</tr>
<tr>
<td>9</td>
<td>better president hope paul ron than person</td>
</tr>
<tr>
<td>LOVE</td>
<td>lose meet man hope boyfriend weight finally</td>
</tr>
<tr>
<td>Isolate</td>
<td>and to for a the year in new all my 2008</td>
</tr>
<tr>
<td>12</td>
<td>god bless jesus loved everyone know loves</td>
</tr>
<tr>
<td>13</td>
<td>peace world earth win lottery around save</td>
</tr>
<tr>
<td>14</td>
<td>com call if 4 www 2 u visit 1 email yahoo 3</td>
</tr>
<tr>
<td>Isolate</td>
<td>i to wish my for and a be that the in me get</td>
</tr>
<tr>
<td>Split</td>
<td>job go school great into good college</td>
</tr>
<tr>
<td>Split</td>
<td>mom husband cancer hope free son well</td>
</tr>
</tbody>
</table>
merge([love ... marry...],[meet ... married...])
(10 words total)

| Topic | Top words sorted by $\phi = p(\text{word}|\text{topic})$ |
|-------|--------------------------------------------------|
| Merge | love lose weight together forever marry meet     |
| success | health happiness family good friends prosperity |
| life | life happy best live time long wishes ever years |
| - | as do not what someone so like don much he |
| money | out make money up house work able pay own lots |
| people | no people stop less day every each other another |
| iraq | home safe end troops iraq bring war return |
| joy | love true peace happiness dreams joy everyone |
| family | happy healthy family baby safe prosperous |
| vote | better hope president paul ron than person bush |
| Isolate | and to for a the year in new all my |
| god | god bless jesus everyone loved know heart christ |
| peace | peace world earth win lottery around save |
| spam | com call if u 4 www 2 3 visit 1 |
| Isolate | i to wish my for and a be that the |
| Split | job go great school into good college hope move |
| Split | mom hope cancer free husband son well dad cure |
Generalization and Scalability

Cross-validation

- Training: do Fold.all MAP inference to estimate $\hat{\phi}$
- Testing: use trainset $\hat{\phi}$ to infer testset $\hat{z}$ (no logic rules)
- Evaluation: testset objective $Q$
- “-”: runs more than 24 hours

| Data  | Fold.all | LDA | Alchemy | $\sum_l |\mathcal{G}(\psi_l)|$ |
|-------|----------|-----|---------|-------------------------------|
| Synth | 9.86     | -2.18 | -1.73  | $10^5$                        |
| Comp  | 2.40     | 1.19 | -       | $10^4$                        |
| Con   | 2.51     | 1.09 | -       | $10^3$                        |
| Pol   | 5.67     | 5.67 | -       | $10^9$                        |
| HDG   | 10.66    | 3.59 | -       | $10^8$                        |
Summary

- “Knowledge + data” for latent Dirichlet allocation
- Logic easy for non-CS users, general
- Scalable inference